



ENCYCLOPEDIA OF  
**Biodiversity**

VOLUME 4

---

ACADEMIC PRESS



# MARINE AND AQUATIC COMMUNITIES, STRESS FROM EUTROPHICATION

Jonathan H. Sharp  
*University of Delaware*

---

- I. Introduction
  - II. Extrapolating from Lakes to the Sea
  - III. Nearshore Ocean Nutrient Response
  - IV. Stoichiometry
  - V. Contaminants and Stress
- 

depletion of dissolved oxygen concentration, leading to very low oxygen (hypoxia) or no oxygen (anoxia).

---

## GLOSSARY

**eutrophication** Overenrichment of aquatic systems with nutrients, often leading to harmful algal blooms and subsurface oxygen depletion.

**harmful algal blooms** Development of sufficient numbers of cells of algal species to cause “harmful” effect to ecosystem.

**microbial response** Expected responses of microscopic algae to nutrient enrichment is excess production beyond what can be consumed by grazers and species shift to noxious species; this expected response is not necessarily what happens.

**nutrient enrichment** Usually excess of nitrogen and phosphorus nutrients to aquatic systems.

**redfield ratio** From large-area and time averaging, ratio of carbon, nitrogen, phosphorus, and oxygen for normal ocean plankton and deep-ocean nitrate and phosphate pools.

**subsurface oxygen depletion** Due to isolation of waters below the surface, metabolic breakdown of organic matter from the surface waters can cause

**MOST LAKES, ESTUARIES, AND COASTAL OCEAN REGIONS** in the proximity of large human populations have experienced significant nutrient enrichment when contrasted to more “pristine” conditions. This stress of nutrient enrichment is viewed as the cause of eutrophication, with classical symptoms of subsurface water oxygen depletion from excess algal production and proliferation of noxious algal species. The typical eutrophication phenomenon has been defined in lakes as a direct cause-and-effect relationship to a single macronutrient, usually phosphate. For both ecological research and resource management, this lake concept has been extended to estuarine and coastal waters, with some incorrect conclusions. Estuaries and coastal waters, as well as lakes, have complex microbial communities of phytoplankton algae plus heterotrophic bacteria and protozoa that together act as the beginning of the food web. The response of different members of these communities to different stimuli as well as differences of grazer pressure on different ecosystems must be taken into consideration. In estuaries and coastal waters, more complex biogeochemical reactions and tides and currents further complicate understanding the impact of nutrients. Ratios of the macronutrients nitrogen and phosphorus to each other and to silicon vary in both time and space,

exerting varying influences on the microbial response. In addition, trace nutrients, which are often considered unimportant in lakes, probably have a major influence on estuarine and coastal primary production. Trace metals and organic compounds also have potential contaminant influences on the overall microbial response. Eutrophication as a stress in estuarine and coastal marine waters is not a simple cause-and-effect phenomenon.

## I. INTRODUCTION

Perhaps the single anthropogenic stress on aquatic and marine environments that is considered to be most ubiquitous is described with the broad term eutrophication. Human activities have mobilized nitrogen and phosphorus through agriculture, urban and suburban sewage, and atmospheric emissions sufficiently to greatly increase fluxes to aquatic environments, especially to lakes, estuaries, and coastal ocean waters. The overly simplistic view of eutrophication is that loading of nitrogen and phosphorus causes increased production and biomass of planktonic algae, with decreased species diversity. Thus, nutrient loading creates a stress to the aquatic community causing adverse community impact. Characteristic symptoms of eutrophication are depletion of oxygen in subsurface waters from excess algal biomass and development of blooms of noxious algal species (Richardson, 1997). Limnological studies of eutrophication have given concepts and observations about the phenomenon which, when applied to estuarine and coastal waters, often lead to incorrect conclusions. Contrary to simple models, nutrient loading in nearshore marine waters often does not support the level of primary production and phytoplankton biomass that would otherwise be expected from the high nutrient concentrations. Depletion of oxygen in subsurface waters is probably more a function of the physics of the specific aquatic system and lack of grazer consumption than it is of nutrient-stimulated excess algal growth. A bloom of a noxious algal species is probably more a function of the response by the entire aquatic community to the overall chemical milieu than it is to nutrient-stimulated growth of the algal species.

This lack of a simple cause-and-effect response to nutrient loading is especially important to evaluate in light of two essentially opposite resource management actions. In nearshore waters, the traditional approach to eutrophication is removal of nutrient inputs to prevent algal production and growth. This approach has been successful in some cases. However, its success may be

overestimated and probably, nutrient removal will have little of the intended impact in many estuarine and coastal waters. The second and opposite action with nutrient enrichment comes from recent proposals for large-scale engineering projects to fertilize waters of the open ocean to increase primary production (Cullen and Chisholm, 1999). In the one case, there is the simple goal of decreasing algal production (which is considered bad) by reducing nutrient inputs and in the other case a goal of increasing algal production (which is considered good) by adding nutrients; it's not that simple. A principal reason that we often misinterpret marine eutrophication is that insufficient consideration is given to the hydrodynamic and biogeochemical complexity of the environment as well as the biodiversity of the microbial community (Paerl, 1998).

In discussing biological diversity, Norse (1993) considers hierarchical levels that range from genetic to species to ecosystem. Stress appears to decrease community diversity, often also decreasing the number of species within an individual function, such as primary production. It is critical to consider that increased diversity at one trophic level may decrease diversity at another. An interesting example comes from a recent examination of thermal stress on a planktonic community. Microcosms were studied with examination of several trophic levels (Petchy *et al.*, 1999). It was found that environmental warming caused losses of top predators and herbivores, with increasing dominance by autotrophs and bacterivores. The warming increased extinction of predators with little effect on primary producers and bacterivores. Primary producer and bacterivore biomass increased, bacterial biomass did not change, and there were idiosyncratic impacts on total biomass. Warming directly increased primary production through temperature-dependent physiology and indirectly through changes in trophic structure. Warming also increased decomposition, probably both through physiology and indirectly through food web structure. The results of this quantitative study are consistent with the general descriptive view of eutrophication where primary production and bacterial decomposition appear to increase while much of the rest of the trophic structure decreases (Vollenweider *et al.*, 1992).

Many anthropogenic inputs to the aquatic environment that may be considered potential stressors have a dual impact, both stimulating growth at low concentrations and decreasing growth at higher concentrations. At a functional level, like primary producers, it is well known that temperature has a stimulating influence on individual species until a threshold is reached, above which it can inhibit. There has been

considerable interest recently about the influence of trace metals on primary producers showing this dual stimulating–inhibiting influence (Bruland *et al.*, 1991). The impact of nutrients is clearly similar if one looks at the full ecosystem. Overcoming limiting levels increases primary production and can increase entire ecosystem production, but only to an extent. Then too much nutrient loading may become a negative stress on the ecosystem. It is necessary to discriminate whether the effect is from nutrients or other inputs.

In the next section, the extension to the marine environment of the limnological example of eutrophication is examined. Following that is a section examining relatively low growth in high-nutrient environments and a section on ecosystem stoichiometry. The last section addresses the need for more thorough understanding of multiple stressor effects on the entire community, including grazer and trophic transfer.

## II. EXTRAPOLATING FROM LAKES TO THE SEA

### A. Lake Eutrophication

The classic example of reducing input of one nutrient to ameliorate eutrophication can be seen with the relatively successful cleanup effort that has taken place with lakes (e.g., Edmondson, 1991). Vollenweider's models of the 1960s have led to the single-element, phosphorus, approach to lake eutrophication. These were successfully applied to the experimental lakes program, where focus on P proved successful (Schlinder, 1981). In these studies, if carbon or nitrogen were in shortage, it could be brought in from atmospheric or sedimentary sources. Schlinder also concluded that micronutrients were not important except in rare cases. The accepted overview for lakes is that phytoplankton production is controlled by annual P-loading and that P-loading is directly related to P concentration and P concentration and chlorophyll are well correlated. Generalizations of this nature may be more valid for lakes than for marine and estuarine waters due to more thorough experimental information and the hierarchical approach used to study lake eutrophication (Hecky and Kilham, 1988). In addition, physical properties in estuarine and marine waters may make these systems more complex. Indeed, Schlinder (1981) has stated, "The control of eutrophication in estuaries has appeared to be much more complex than it is in lakes."

Estuarine and coastal ocean waters have considerable transport circulation driven by both tidal energy and

riverine discharge and this circulation is variable on predictable and unpredictable periodicities. These environments also have complicating factors from variations in suspended sediments through specific input sites, circulatory resuspension of bottom sediments, and coagulation along estuarine salinity gradients. Additionally, the salinity gradients of estuaries cause significant physical stress such that populations of organisms will change and moreover the change in ionic strength will influence chemical speciation and solute–solvent interactions. The more complicated and varying physics gives rise to less predictability of populations on a seasonal basis. This pertains not only to primary producers but also to food web structure. There is a large literature on the subject of coastal ocean eutrophication, with recognition of difficulties in applying eutrophication concepts to estuarine and coastal waters, where physics and biogeochemistry complicate the picture (Vollenweider *et al.*, 1992).

### B. Food Web Complexity

In addition to the complicated nature of coastal waters, it is well recognized that food web structure can bring significant variability to all aquatic ecosystems, lakes as well as coastal marine waters. Schlinder *et al.* (1997) demonstrated that the primary production enhancement from nutrient enrichment was less in a piscivore-dominated lake than in a planktivore-dominated lake. The reason for this difference was the suppression of phytoplankton by large zooplankton in the piscivore lake. The study focus included evaluation of drawdown of atmospheric CO<sub>2</sub> from nutrient-enhanced production and confirmed that changes in aquatic CO<sub>2</sub> fugacity could be successfully manipulated in lakes and open-ocean ecosystems.

The recognition of complexity in oceanic microbial ecosystems led to the concept of the "microbial loop," where organic matter from phytoplankton is rapidly consumed by bacteria, which are in turn consumed by protozoans that are consumed by small zooplankton that otherwise would be herbivores exclusively (Azam *et al.*, 1989). This shunt from the phytoplankton–herbivore direct trophic transfer means lower efficiency in the path to higher metazoans and led to debates of bacteria as "source or sink" in oceanic as well as estuarine environments (Sherr *et al.*, 1987). Subsequent work with the microbial loop has led to considering the combined phytoplankton and microbial heterotrophs (bacteria and protozoans) as the primary producer community supporting the metazoans (Sherr and Sherr, 1991). With the primary producer function from a multicom-

partment ecosystem, it is not surprising to find phytoplankton and heterotrophic bacteria being influenced by different stimuli. For example, Pace (1993) showed bacteria being controlled by phosphorus while phytoplankton were controlled by nitrogen and phosphorus in lake nutrient enrichment experiments. In most cases with any aquatic environment, nutrient addition will bring variable influences on phytoplankton and heterotrophic bacteria and may also influence heterotrophic protozoans and metazoan grazers.

It has been well recognized that phytoplankton production, or the entire primary producer community, is influenced by removal (top-down control) as well as by resource limitation (bottom-up control). In a recent review of eutrophication in planktonic ecosystems, Glibert (1998) pointed out that grazing and nitrogen recycling are intricately connected in controlling planktonic nitrogen availability. Another important recognition is that top-down control has a major impact on export from the pelagic system (Wassman, 1998). Wassman warned that to view only bottom-up controls (nutrient influence) will not successfully guide biogeochemical studies of marine systems. Thus, it seems obvious that a nutrient influence on phytoplankton should not be considered in the absence of the rest of the beginnings of the ecosystem.

### C. Response of Nearshore Waters to Nutrient Enrichment

The idea that a single nutrient controls primary production comes from classical ecological theory. With a single nutrient, a single phytoplankton species should have an advantage over others and dominate by outcompeting. The fact that multiple species can coexist within an apparent niche was considered a paradox (Hutchinson, 1961) and was the subject of a massive amount of excellent aquatic research. Recognizing the necessity to consider guilds rather than a single species makes understanding of primary production more complicated. Add to this the more recent recognition that multiple compartments of phytoplankton plus bacterial and protozoan heterotrophs may be considered as a primary producer community, and the need for whole-ecosystem experiments is obvious.

There is an oversimplified view that nutrient concentrations (or loading) above those of some "pristine" condition directly cause phytoplankton response, with negative impact in estuarine and coastal waters. The overall impression is that increased nutrients cause increased algal growth with the consequence of excess algal production causing oxygen depletion or the conse-

quence of a bloom of noxious algae. Oxygen depletion from nutrient-enriched phytoplankton growth occurs in environments where summer stratification isolates bottom waters, for example, Chesapeake Bay and mid-Atlantic coastal waters. Oxygen depletion is often quite variable on an innerannual basis and moderated by meteorological forcing. Thus, the occurrence and extent of oxygen depletion are complicated and not simply predictable as a function of nutrient loading.

There is concern that unusual and noxious algal blooms are increasing globally in both geographic extent and intensity, although there is debate on the quantitative significance of such claims (Anderson, 1997). It is important to be careful in defining harmful algal blooms (HABs) as Smayda (1997) has indicated. In most cases, the sign of the "bloom" is the appearance of numbers of cells of a species of a harmful alga sufficient to have a negative environmental impact. Analysis of individual HABs shows that generally the HAB taxa have no unique ecophysiology, including higher affinity for nutrients, and often that the HAB taxa have growth rates lower than those of phytoplankton in general (Smayda, 1997). Many HAB taxa have allelochemically enhanced competition with other algal species and have allelopathic defense against predators as well as against a broad group of other microbial taxa. It appears that noxious algae bloom from their ability to dominate rather than their ability to outcompete other species for nutrients or to grow fast. What actually stimulates these taxa to express their domination is an area in need of more research. Full-ecosystem studies are needed to better understand noxious algal proliferations.

The experimental lakes program mentioned above has provided great empirical evidence to combine with theory in limnology. It is more difficult to manipulate whole parts of estuaries and coastal oceans than it is to do so with small lakes. Controlled mesocosms are a good compromise. The Marine Ecosystems Research Laboratory (MERL) at the University of Rhode Island has been one of the largest and most successful versions of realistic estuarine ecosystems (Oviatt *et al.*, 1986). In these, sufficient volumes of water have been used to overcome many problems of confinement and attempts have been made to simulate estuarine physical and biogeochemical influences. Some excellent research has been done and much learned about the complexity of estuarine responses to nutrients and other stressors.

A much used picture that was developed with information from the MERL experiments and from comparison of phytoplankton biomass and production in various estuaries and coastal waters has been shown by Nixon. The picture indicates phytoplankton increasing

proportionately with increasing nutrient concentrations or loading. Figure 1A shows a generic version of this with phytoplankton production versus nitrogen loading. Note that the nitrogen loading is portrayed on a logarithmic scale (as was done in Nixon and Pilson, 1983), giving the appearance of regularly increasing production as a function of increasing nitrogen over a broad range of nutrient loadings. With transformation to a linear axis for N-loading, it is obvious that production becomes asymptotic after an initial linear increase. There have been several articles in which this conceptual picture has been shown and expanded upon; most

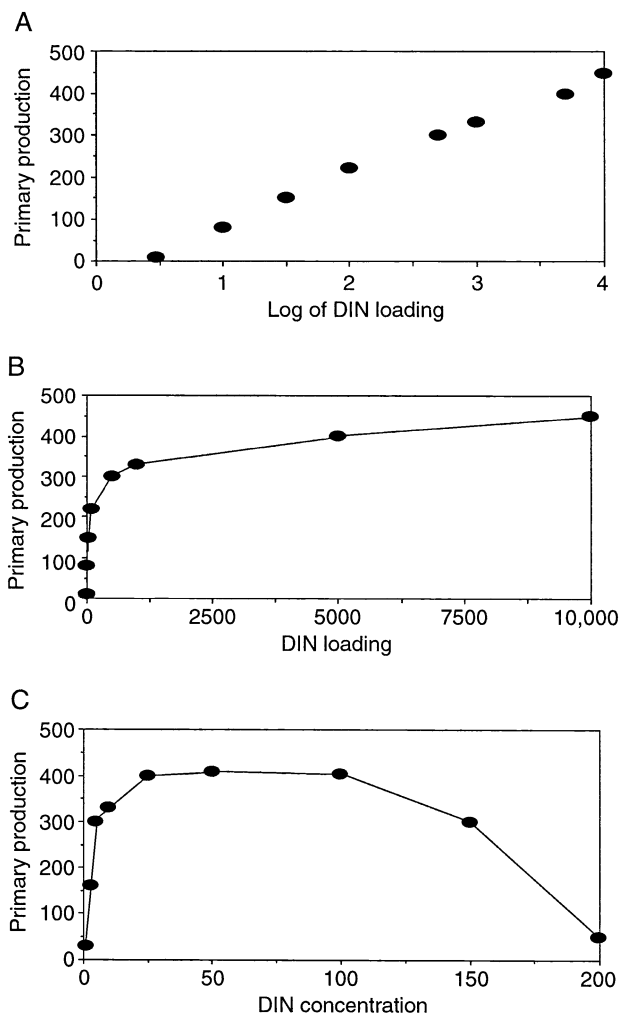


FIGURE 1 Theoretical relationship of dissolved inorganic nitrogen loading to primary production. Frame A shows the often-depicted version (from Nixon and Pilson, 1983) with N-loading on a logarithmic scale, while transformation to a linear scale for N-loading is shown in frame B. Frame C shows inhibition at high N concentration (note concentration rather than loading).

of them are extensive evaluations of data from many published works, with the authors indicating that the relationship is complex (e.g., Nixon *et al.*, 1986). As has been cautioned by Nixon and others using it, the relationship is intended to cover a large range of nutrient conditions and to compare a number of different environments. However, a simplistic extension has been made suggesting that there is a simple linear relationship between nutrient loading and adverse phytoplankton production.

The behavior shown in Fig. 1B is probably more correct to indicate that phytoplankton response to increased nutrients is not linear along a very long loading or concentration scale. Relatively small increases in nutrient concentrations and loadings will cause a large increase in primary production, but continued increases do not. In fact, it is likely that with very high nutrient concentrations a decreased phytoplankton response will be encountered as is shown with the theoretical curve in Fig. 1C. Thus, it is not necessarily the case that nutrient enrichment leads to excess algal production. Perhaps, we should be addressing the question of why there is not greater phytoplankton production in estuarine and coastal waters from nutrient enrichment.

### III. NEARSHORE OCEAN NUTRIENT RESPONSE

#### A. The Oceanic HNLC Concept

Although it is well known that enhancement of primary production is not simply and directly proportional to enrichment by a single nutrient, there is a tendency to oversimplify this relationship in both nearshore and oceanic waters. I would like to use the oceanic HNLC (high nutrient, low chlorophyll) concept to examine enrichment of nearshore waters. The HNLC concept came from interest in iron limitation in the ocean that has a history going back to at least the 1920s (Martin *et al.*, 1990) but is best recognized in relation to Martin's proposed Antarctic and equatorial Pacific experiments (Coale *et al.*, 1996). Underlying the HNLC is the observation that some areas of the open ocean have relatively high concentrations of nitrogen and phosphorus nutrients but do not support proportionately large phytoplankton biomass; see Table I. In open-ocean experiments, iron as a trace element added to overcome limitation has been shown to increase phytoplankton production with concomitant drawdown of atmospheric  $\text{CO}_2$  in both the equatorial Pacific and the Antarctic oceans. The very nature of considering a trace

TABLE I

---

Characteristics of oceanic HNLC (high nutrient, low chlorophyll) environments

1. Relatively high concentrations of macronutrients
2. Low standing stock of phytoplankton
3. Moderate phytoplankton growth rates

Characteristics of estuarine HNLG (high nutrient, low growth) environments

1. High concentrations and fluxes of macronutrients
  2. Moderate to high phytoplankton standing stock
  3. Comparatively low phytoplankton growth rates
  4. Sometimes, domination of flora by "undesirable" species
- 

constituent as critical to oceanic production indicates that the ecosystem is more complex than a simple cause-and-effect relationship with a major nutrient. In addition, the details of responses and trophic complexities in these oceanic experiments need more study before simple conclusions should be reached. However, this does not stop engineering plans for commercial fertilization of the ocean that are based on very simple cause-and-effect relationships. The same oversimplification in nearshore waters leads to proposals to solve eutrophication based upon very simple cause-and-effect relationships, i.e., reducing input of a single major nutrient.

An explanation for the observed HNLC conditions is the simultaneous control by grazing and micronutrients, such as iron. Cullen *et al.* (1992) have demonstrated with modeling for the equatorial Pacific upwelling region that grazing is the main control of standing stock but that a trace nutrient (e.g., Fe) might ultimately regulate overall productivity by influencing species composition and food web structure. Frost and Franzen (1992), with a chemostat model using a multiple-step food chain, demonstrate that simultaneous grazing control and trace nutrient limitation (e.g., Fe) could account for the observed conditions. Armstrong (1994) has shown with his multiple-species model that it is possible for each phytoplankton size class to be controlled by herbivores, while at the same time micronutrient limitation (e.g., Fe) may limit the number of size classes that can exist in a community and thus the total phytoplankton biomass that can be supported. A model has been presented (Armstrong, 1999) that shows HNLC conditions controlled by a combination of iron limitation of algal growth rates, ammonium inhibition of nitrate uptake leading to reduced uptake, and dependence of both processes on cell size. This dependence on cell size affects phytoplankton community structure

and community uptake of nitrate. Recognition of the combined effect of a bottom-up influence (Fe limitation) and top-down influence (grazer control) would suggest that primary production in any aquatic environment has complex multiple controls.

## B. High Nutrients and Low Growth in Estuaries

I suggest that we use a concept called HNLG (high nutrient, low growth) in estuarine waters that is similar to the oceanic HNLC, but with a different twist; see Table I. A characteristic of the estuarine phenomenon is that phytoplankton show a comparatively low growth rate and relatively high biomass. Probably a number of factors, individually or in combination, lead to the HNLG phenomenon. Proportions of macronutrients are often inappropriate for sustained high growth and limitation by micronutrients is also likely. Partial light limitation from algal biomass and from nonbiological suspended sediments also is involved in limiting growth. In addition, there is probably a negative influence on phytoplankton by contaminants and also inhibition of grazers by contaminants.

While lakes are generally considered P-limited, the traditional view is that marine waters are N-limited (Ryther and Dunstan, 1971). More recent evaluations have suggested that estuarine waters have alternating controls by nitrogen, phosphorus, and light. It is important to consider that nutrient enrichment is not necessarily an extension of nutrient limitation and that nutrient enrichment will not necessarily cause a direct proportional increase in phytoplankton. In some cases, fairly direct increases can be shown such as the indication that chlorophyll levels in the Chesapeake Bay have increased appreciably, with a doubling in dissolved inorganic nitrogen (DIN) over several decades. In contrast to this are examples in estuaries that do not show predicted increases. For example, Balls *et al.* (1996) show no change in chlorophyll in the Ythan River estuary between 1960 and 1990 when there was a fourfold increase in DIN. Alpine and Cloern (1992) showed a decline in primary production with increasing nutrient enrichment over time in the San Francisco Bay estuary. Reviewing conditions in a number of shallow coastal environments, Cloern (1999) showed a nonlinear response between N-loading and phytoplankton production and suggested that the simple eutrophication model in lakes does not have a current analog in coastal eutrophication.

Examining data on DIN concentration versus measured primary production for summer estuarine tran-

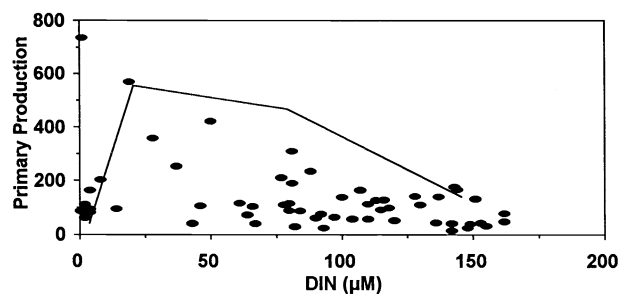


FIGURE 2 Primary production (integrated areal values as millimoles of C/m<sup>2</sup>/day) versus total dissolved inorganic nitrogen (DIN) concentration for samples along the full gradient of the Delaware Estuary. Composite data from summer samplings over 3 years (unpublished data of J. H. Sharp *et al.*, 1986–1988). Highest DIN concentrations are found in the urban freshwater region of the estuary, with slightly lower values immediately upstream in the tidal freshwater region and downstream going into the salinity gradient. Lowest DIN concentrations are found in the lower estuary near the mouth of the Delaware Bay.

sects along the Delaware Estuary, we can show a negative relationship (Fig. 2). While I will not suggest that the high nitrogen concentrations directly cause a decrease in phytoplankton production, it is clear that there is not a simple continual increase in production proportional to high nitrogen content. Note the similarity of Fig. 2 to the theoretical Fig. 1C, where a small increase in DIN causes a large increase in production at concentrations near limiting, reaching an asymptote followed by a decline at very high DIN concentrations. To normalize measured primary production, the ratio of productivity to chlorophyll biomass (P/B) is often used. With data for the 1980s and 1990s from the Delaware Estuary, average summer P/B from the high-nutrient upper estuary is 61 (grams of C fixed per gram of chlorophyll) contrasted to the lower estuary production maximum region P/B value of 225 (J. H. Sharp *et al.*, unpublished data).

In most estuaries, nutrient increases have not been uniform. Often increases in DIN are accompanied by smaller increases or decreases in dissolved inorganic phosphorus (PO<sub>4</sub>) such that the N/P ratio has increased. For example, this is the case with the nitrogen increases in the Chesapeake Bay and the Delaware Estuary (Sharp, 1988).

## IV. STOICHIOMETRY

### A. Redfield Ratios

From extensive averages of planktonic CNP composition and deep-ocean NO<sub>3</sub> and PO<sub>4</sub> nutrient concentra-

tions, it was noted that a regular and predictable N/P ratio was found (Redfield *et al.*, 1963). This Redfield ratio concept was extended to carbon and oxygen for ecosystem utilization of elements. The concept has been extensively applied to aquatic systems for over a half century and is a valuable guide in understanding biogeochemical fluxes. Analysis of particulate CNP in a number of fresh-water lakes indicates that a variety of conditions exist, ranging from N and P deficiency to sufficiency (Hecky *et al.*, 1993). They concluded that ocean plankton is not as N and P deficient as is lake plankton. From a different viewpoint, Flynn (1990) concluded that only in the presence of excess NH<sub>4</sub> is the cellular response to N-stress fully suppressed. Thus, plankton throughout the ocean show some symptoms of N-stress. He suggested that there are three forms of N status: N-replete (no stress), N-sufficiency (enough stress to depress NO<sub>3</sub> transport and assimilation), and N-deplete (maximum stress, no growth). With this classification, most estuarine, coastal, and open-ocean waters are N-sufficient. To evaluate whether or not there is nutrient stress from too little or too much, it is probably necessary to look over sufficient time to be in essentially steady-state conditions. For instance, it has been suggested that in Southern Ocean waters, Redfield CNP ratios for plankton use were only obtained when averaging over the full vegetative season of the Austral summer (Hoppema and Goeyens, 1999). They suggested that the Redfield ratio was reached only because of nutrient-replete conditions.

### B. Adding Silicon

The concept of the Redfield ratio has been extended to silicon for many studies of aquatic ecology. From laboratory culture studies, the average Si/N ratios for small and large diatom species is close to 1/1 (Brezinski, 1985). This would give an overall Redfield ratio for CNPSiO of 106/16/1/16/–276 for balanced diatom growth. For a long time, it has been assumed that with deficiency of Si, phytoplankton populations will shift from diatom domination to that of other groups of algae. However, recent research would suggest a more absolute limitation of “healthy” ecosystem production by Si. In oceanic HNLC environments, it has been suggested that “new” production (that which is supported by upwelled NO<sub>3</sub>) is reduced by Si limitation and thus export from pelagic primary production is controlled by Si availability. In coastal upwelling regions, it has been demonstrated that iron limitation will cause diatoms to increase the Si/N uptake ratio, depleting the water of Si, leading to secondary Si limitation. Clearly,



Si is very important and can influence N response of the primary producers.

Comparative Si availability may be a major feature in the apparent eutrophication response seen in nearshore waters. In a recent 20-year comparison in the Bay of Brest, there was a large decrease in the Si/N ratio, “but, contrary to what has been observed in other coastal ecosystems, phytoplankton stocks have not increased” (LePape *et al.*, 1996). In light of the discussion in Section III.B, maybe this is less of an exception than LePape *et al.* interpreted. In some cases, an increase in phytoplankton biomass is seen, but not always with a shift from diatoms, and rarely is there an increase in higher trophic level consumption of the primary production. In an extensive study of a very long term record for the Mississippi River outflow into the Gulf of Mexico, Rabalais *et al.* (1996) have shown a large decrease in the Si/N ratio accompanied with an increase in primary production but also an increase in the deposition of biogenic silica in the sediments underlying an increasingly large hypoxia region. The explanation is that with relative Si scarcity, diatoms that are in the plankton are not grazed efficiently, and they fall to subsurface waters and contribute to hypoxia.

### C. Changing Nutrient Ratios

Probably most estuarine waters with impact from human activities show greatly changed N/Si as well as N/P ratios. In Table II, average values for total dissolved inorganic nitrogen (NO<sub>3</sub> plus NH<sub>4</sub>), PO<sub>4</sub>, and Si are shown for several nutrient-enriched estuaries. All of the examples show large increases in DIN and most have smaller proportional increases in P so that the N/P is usually considerably higher than would be the

case without the anthropogenic influence. Since Si is not usually a byproduct of human activity, the Si concentration has not changed much; there is probably a large natural variation depending upon the nature of the land drained for the estuary. A few systems probably have had significant decreases in Si due to decreased natural land erosion (dams, diked river banks); this is definitely the case with the Mississippi (Rabalais *et al.*, 1996). As a result, the N/Si ratio is much different from that prior to human impacts. Inverting this as Si/N, the pristine condition is about 10/1 and most of our nutrient-enriched systems show values of 1/1 or lower. This very likely has a serious negative impact on the primary production community. The importance of Si in relation to eutrophication has been recognized in the past, but usually only in relation to shift from diatom to flagellate flora (e.g., Officer and Ryther, 1980). With more recent information on interactive influences of Si, N, and Fe and on the fate of Si-limited diatom production, it is timely to reinvestigate the role of Si on eutrophication. While species responsible for HABs do not necessarily show greater affinity for nutrients in general, giving them ability to outcompete more “normal” phytoplankton like diatoms, it is probable that changing ratios of N and P to Si do favor some of the HAB flagellates (Smayda, 1990).

The large changes in N/P ratios are often not documented because of lack of complete nutrient records from long-range monitoring. In a data set from the Delaware Estuary, dissolved inorganic nitrogen has been measured regularly along the full axis of the estuary for over 35 years, but parameters for phosphorus measurements have varied over that period. Total P, a composite that includes dissolved organic and particulate phosphorus as well as PO<sub>4</sub>, has been measured

TABLE II  
Approximate Nutrient Concentrations at the Beginning of the Salinity Gradient for Several Nutrient-Enriched Estuaries<sup>a</sup>

Estuary	DIN	PO <sub>4</sub>	Si	N/Si/P	Reference
Scheldt	550	15	250	37/17/1	Zwolsman, 1994, 1999
Delaware	250	5	125	40/20/1	Sharp, unpublished data
Mississippi	114	7.7	108	15/14/1	Rabalais <i>et al.</i> , 1996
Chesapeake	75	1	50	57/50/1	Malone <i>et al.</i> , 1996
Northern San Francisco Bay	40	2	200	20/100/1	Peterson <i>et al.</i> , 1985
Pristine	10	0.5	100	20/200/1	Fanning and Maynard, 1978; Meybeck, 1982

<sup>a</sup> Average concentrations of nutrients (in  $\mu\text{M}$  element) approximated from publications listed. Averages for total dissolved inorganic nitrogen (DIN), dissolved phosphorous (PO<sub>4</sub>), and silicate (Si) and ratios normalized to P are listed. Values for pristine estuaries approximated from data for the Zaire and Magdalena River outflow systems.

consistently. A comparison of the N/P ratio change, based on the total P, over a 30-year period is shown in Fig. 3. Recognizing that the majority of the P in the estuary today is  $\text{PO}_4$  and that in the past  $\text{PO}_4$  was probably a larger portion of the total, it is possible to view the N/P ratio as indicative of available P. This dramatic N/P ratio change is probably largely due to reduced input of detergent phosphorus and the same change has occurred in many U.S. estuaries (N. A. Jaworski, unpublished data, 1998). In the earlier situation, almost the entire estuary would appear to be replete in relation to P since the N/P was considerably below the Redfield ratio; in the more modern situation, N/P ratios are in the 30–60 range. However, it must be recognized that transport and availability of phosphorus in estuaries is a complex function that also involves geochemical influences. In the past 20 years, the P geochemical reactivity in the Delaware Estuary has changed due to increased pH and dissolved oxygen. As a result, the N/P ratio based on  $\text{PO}_4$  only for average concentrations of the entire salinity gradient of the estuary has decreased in the past 20 years from about 90/1 to 40/1. More thorough analysis of many estuaries may show this dual trend of long-term decrease of N/P loading in an upper estuary but of more available dissolved  $\text{PO}_4$  being delivered to the lower estuary. It is important to understand the full biogeochemical picture of estuarine phosphorus before accurate conclusions of nutrient impacts can be made.

In addition to long-term changes in nutrient ratios, there are large spatial changes in estuaries at any single time. Figure 4 shows nutrient ratios along the full length of the Delaware Estuary from sampling in the spring. In the spring bloom condition in the estuary,  $\text{NH}_4$ ,  $\text{PO}_4$ , and Si are exhausted from the mouth of the estuary moving up toward a strong-light-limiting turbidity

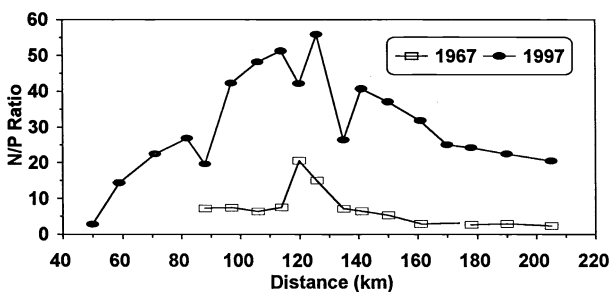


FIGURE 3 N/P ratios for the Delaware Estuary using total dissolved inorganic nitrogen (DIN) versus total P from summer transects of the entire estuary in summers of 1967 and 1997 data from Delaware River Basin Commission routine monitoring (unpublished data of J. H. Sharp and E. Santoro).

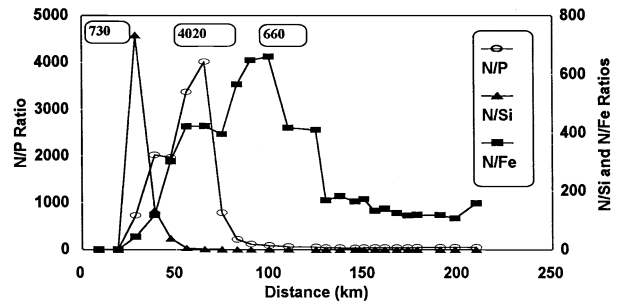


FIGURE 4 Ratios from sampling along the distance axis of the Delaware Estuary in the spring from measurements of total dissolved inorganic nitrogen (N), P, Si, and Fe (unpublished data of J. H. Sharp *et al.*, 1987). The N/P ratio is generally in the 30–100 range over much of the estuary during much of the year and  $<16$  only near the mouth of the bay. The N/Si ratio is generally between 1 and 2 during most of the year. The N/Fe ratio is in the 100–600 range except near the bay mouth, where it can become  $<100$ .

maximum (Pennock and Sharp, 1994). The very high ratios of P and Si to N are due to the large excess  $\text{NO}_3$  concentration of the river water as it is advected downstream. It is interesting to see that different regions of the estuary appear to have large differences in the nutrient that could be most limiting. Also, from this picture, it appears that Fe could be more limiting in the upper estuary than near the mouth of the bay. This greater proportional availability of Fe is a year-round occurrence. A noxious algal group that has caused considerable international concern recently is responsible for brown tides. A recent analysis of brown tide occurrence suggests the macronutrient levels are not implicated but that Fe is.

In the Delaware Estuary, the area of the greatest phytoplankton production throughout the year is the lower bay (Pennock and Sharp, 1994). Figure 5 shows

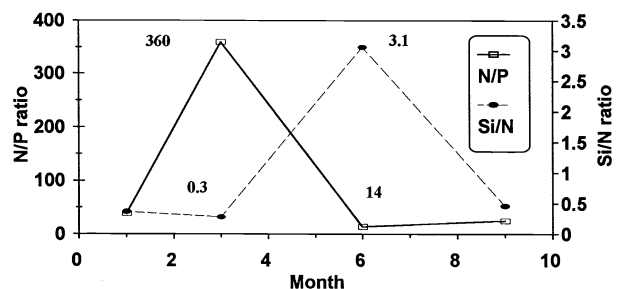


FIGURE 5 Average ratios of N/P and N/Si for stations in the lower Delaware Bay for winter, spring, summer, and fall (unpublished data of J. H. Sharp *et al.*, 1986–1988). During the spring (month 3), the maximum biomass for the entire estuary for the year is found in the lower bay. The maximum estuarine primary production is found in the summer in the lower bay.

macronutrient ratios for this region averaged on a seasonal basis. The maximum biomass is achieved in the spring bloom, where N/P and Si/N ratios indicate exhaustion of P and Si with residual  $\text{NO}_3$ ; at this time, nutrient regeneration is minimal and there is also little herbivore grazing. Usually, the highest seasonal primary production is found in the summer. At this season, it would appear that grazing controls chlorophyll (Pennock and Sharp, 1994) and grazing is in balance with nutrient regeneration, which is sufficient to allow measurable  $\text{NH}_4$ ,  $\text{PO}_4$ , and Si levels). At this season, the N/P ratio is close to Redfield and the Si/N ratio is sufficient to support healthy diatom growth although the flora is dominated by small flagellates. Acknowledging the caution of Hecky and Kilham (1988) that nutrient concentration does not equal nutrient utilization, we have demonstrated limitation in several ways (Pennock and Sharp, 1994). Looking at the entire estuary with highly nutrient enriched tidal freshwater region, a light-limited turbidity maximum in the oligohaline region, and clearer and nutrient-diluted lower bay, it would appear that the only time and place that near to Redfield ratio of nutrients is found is in the lower bay in the summer. We have performed preliminary simple mesocosm experiments and find that the CNPO fluxes approach Redfield stoichiometry only in the lower estuary in the summer (Sharp *et al.*, unpublished data). In the lower estuary in the spring bloom, close to Redfield ratios of NPSiO occur but there appears to be an accumulation of C with the biomass accumulation. In the nutrient-rich upper estuary, nothing close to expected stoichiometry is seen. Further research on this anomalous stoichiometry is currently underway in my laboratory.

## V. CONTAMINANTS AND STRESS

The lack of high growth rates of estuarine and coastal phytoplankton in the presence of high nutrient concentrations leads to the suggestion that anthropogenic contaminants, other than nutrients, may have an influence on ecosystem response. The MERL research facility mentioned earlier has extensive and expensive controls of the large tanks for the studies. A number of excellent experiments have been carried out also in less sophisticated smaller mesocosms in other estuarine areas. One interesting study done in Chesapeake Bay mesocosms over several seasons showed that nutrient enrichment by N and P caused growth of "beneficial" diatom species over flagellates (Sanders *et al.*, 1987). A conclusion of that study was that many other factors probably control

community dynamics. As discussed earlier, the lack of predation is often a factor in noxious algal species proliferation. Also discussed earlier is the suggestion that ratios of nutrients may be more important than quantities in causing eutrophication responses. It is possible also that contaminants cause stress to the "desirable" algal species and to grazers that would otherwise consume the primary production.

Most estuaries with large anthropogenic influences have had chronic exposure to many chemicals in addition to nutrients. Chronic exposure to arsenic appeared to cause reductions in phytoplankton cell size with less trophic transfer while chronic exposure to silver resulted in essentially the opposite. Thus, arsenic would appear to be partially responsible for eutrophication response. Many estuaries also have frequent or continuous inputs of chlorination byproducts which must have a selected influence. Sanders (1984) showed that one diatom and a chrysophyte would not grow in aged chlorinated water (with levels below detection for total residual chlorine) where a more resistant flagellate would grow. This result would also favor species other than "normal" with potential to decrease trophic transfer.

Paerl (1998) has illustrated positive and negative interactions and feedback from nutrient loading, emphasizing that negative influences on grazing can increase the impact of primary production. In lakes, it has been shown that combinations of nutrient additions and zooplankton size can have major influence on phytoplankton sizes and thus on trophic transfer. The variability in nutrient and trace metal impacts on phytoplankton, bacterioplankton, heterotrophic protozoa, copepods, fish, and benthos caused variable trophodynamic responses in estuarine mesocosm experiments (Breitburg *et al.*, 1999). These authors concluded that trace elements may mask the response of high nutrient loadings in eutrophic systems. In lake studies, N, P, and C had different controlling effects on zooplankton, phytoplankton, and bacterioplankton, with variable responses in different seasons.

An overall conclusion is that eutrophication as a stress in estuarine and coastal marine environments is not a simple cause-and-effect phenomenon. Nutrient enrichment elicits complex and variable responses from the phytoplankton, bacterioplankton, and protozoa that make up the primary producer community. Increased nutrient concentrations and loadings may cause an overall increase in phytoplankton biomass, but not invariably from higher growth rates. Shifts in dominant species and shunts through the microbial loop may decrease the trophic transfer to higher metazoan levels.

The biodiversity of the aquatic community may allow some resilience to the system (Patrick, 1988), but ultimately changes in the diversity are probably more important than a direct response to increased levels of nutrients.

## See Also the Following Articles

COASTAL BEACH ECOSYSTEMS • ESTUARINE ECOSYSTEMS • LAKE AND POND ECOSYSTEMS • PLANKTON, STATUS AND ROLE OF • RIVER ECOSYSTEMS

## Bibliography

- Anderson, D. M. (1997). Turning back the harmful red tide. *Nature* 388, 513–514.
- Armstrong, R. A. (1999). An optimization-based model of iron–light–ammonium colimitation of nitrate uptake and phytoplankton growth. *Limnol. Oceanogr.* 44, 1436–1446.
- Azam, F., Fenchel, T., Field, J. G., Ray, J. S., Meyer-Reil, L. A., and Thingstad, F. (1989). The ecological role of water-column microbes in the sea. *Mar. Ecol. Prog. Ser.* 10, 257–263.
- Brezinski, M. A. (1985). The Si:C:N ratio of marine diatoms: Interspecific variability and the effect of some environmental variables. *J. Phycol.* 21, 347–357.
- Bruland, K. W., Donat, J. R., and Hutchins, D. A. (1991). Interactive influences of biactive trace metals on biological production in oceanic waters. *Limnol. Oceanogr.* 36, 1555–1577.
- Cloern, J. E. (1999). The relative importance of light and nutrient limitation of phytoplankton growth: A simple index of ecosystem sensitivity to nutrient enrichment. *Aquat. Ecol.* 33, 3–16.
- Coale, K. H., et al. (1996). A massive phytoplankton bloom induced by an ecosystem-scale iron fertilization experiment in the equatorial Pacific Ocean. *Nature* 383, 495–501.
- Cullen, J. J., and Chisholm, S. W. (1999). Commercial ocean fertilization: We know enough to know better. *EOS* 80, OS66–OS67.
- Hecky, R. E., and Kilham, P. (1988). Nutrient limitation of phytoplankton in freshwater and marine environments: A review of recent evidence on effects of enrichment. *Limnol. Oceanogr.* 33, 796–822.
- Hutchinson, G. E. (1961). The paradox of the plankton. *Am. Nat.* 95, 137–145.
- Martin, J. H., Gordon, R. M., and Fitzwater, S. E. (1990). Iron in Antarctic waters. *Nature* 345, 156–158.
- Meybeck, M. (1982). Carbon, nitrogen, and phosphorus transport by world rivers. *Am. J. Sci.* 282, 401–450.
- Nixon, S. W., and Pilson, M. E. Q. (1983). Nitrogen in estuarine and coastal marine ecosystems. *Nitrogen in the Marine Environment*, pp. 565–648. Academic Press, New York.
- Nixon, S. W., Oviatt, C. A., Frithsen, J., and Sullivan, B. (1986). Nutrients of the productivity of estuarine and coastal marine ecosystems. *Limnol. Soc. Southern Africa* 12, 43–71.
- Norse, E. A. (1993). *Global Marine Biological Diversity*. Island Press.
- Officer, C. B., and Ryther, J. H. (1980). The possible importance of silicon in marine eutrophication. *Mar. Ecol. Prog. Ser.* 3, 83–91.
- Oviatt, C. A., Keller, A. A., Sampou, P. A., and Beatty, L. L. (1986). Patterns of productivity during eutrophication: A mesocosm experiment. *Mar. Ecol. Prog. Ser.* 28, 69–80.
- Pace, M. L. (1993). Heterotrophic microbial processes. In *The Trophic Cascade in Lakes*, pp. 252–277. Cambridge Univ. Press, Cambridge, UK.
- Paerl, H. W. (1998). Structure and function of anthropogenically altered microbial communities in coastal waters. *Curr. Opin. Microbiol.* 1, 296–302.
- Patrick, R. (1988). Importance of diversity in the functioning and structure of riverine communities. *Limnol. Oceanogr.* 33, 1304–1308.
- Pennock, J. R., and Sharp, J. H. (1994). Temporal alternation between light- and nutrient-limitation of phytoplankton production in a coastal plain estuary. *Mar. Ecol. Prog. Ser.* 111, 275–288.
- Petchey, O. L., McPhearson, P. T., Casey, T. M., and Morin, P. J. (1999). Environmental warming alters food-web structure and ecosystem function. *Nature* 402, 69–72.
- Rabalais, N. N., Turner, R. E., Justic, D., Dortch, Q., Wiseman, W. J., and Sen Gupta, B. (1996). Nutrient changes in the Mississippi River and system responses on the adjacent continental shelf. *Estuaries* 19, 386–407.
- Redfield, A. C., Ketchum, B. H., and Richards, F. A. (1963). The influence of organisms on the composition of sea water. In *The Sea*, Vol. 2, pp. 26–77.
- Richardson, K. (1997). Harmful or exceptional phytoplankton blooms in the marine ecosystem. *Adv. Mar. Biol.* 31, 301–385.
- Ryther, J. H., and Dunstan, W. M. (1971). Nitrogen, phosphorus, and eutrophication in the coastal ocean environment. *Science* 171, 1008–1013.
- Sanders, J. G., Cibik, S. J., D'Elia, D. F., and Boynton, W. R. (1987). Nutrient enrichment studies in a coastal plain estuary: Changes in phytoplankton species composition. *Can. J. Fish. Aquat. Sci.* 44, 83–90.
- Schlinder, D. W. (1981). Studies of eutrophication in lakes and their relevance to the estuarine environment. In *Estuaries and Nutrients*, pp. 71–82. Humana Press, Clifton, NJ.
- Schlinder, D. E., Carpenter, S. R., Cole, J. J., Kitchell, J. F., and Pace, M. L. (1997). Influence of food web structure and carbon exchange between lakes and the atmosphere. *Science* 277, 248–251.
- Sharp, J. H. (1988). Trends in nutrient concentrations in the Delaware Estuary. In *Ecology and Restoration of the Delaware River Basin* (S. K. Majumdar, E. W. Miller, and L. E. Sage, Eds.), pp. 78–192. Pennsylvania Academy of Sciences.
- Sherr, B. F., Sherr, E. B., and Albright, L. J. (1987). Bacteria: Link or sink? *Science* 235, 88.
- Sherr, E. B., and Sherr, B. F. (1991). Planktonic microbes: Tiny cells at the base of the ocean's food web. *Trends Ecol. Evol.* 6, 50–54.
- Smayda, T. J. (1990). Novel and nuisance phytoplankton blooms in the sea: Evidence for a global epidemic. In *Toxic Marine Phytoplankton* (E. Graneli, B. Sundstrom, L. Edler, and D. M. Anderson, Eds.), pp. 29–40. Elsevier, Amsterdam.
- Smayda, T. J. (1997). What is a bloom? A commentary. *Limnol. Oceanogr.* 42, 1132–1136.
- Vollenweider, R. A., Marchetti, R., and Viviani, R. (1992). *Marine Coastal Eutrophication*. Elsevier, Amsterdam.





# MARINE ECOSYSTEMS

J. Frederick Grassle

*Institute of Marine and Coastal Sciences, Rutgers University*

---

- I. Marine Ecosystems
  - II. Biodiversity of Marine Ecosystems
  - III. Ecosystem Function
  - IV. Ecosystem Diversity
  - V. Potential Consequences for Anthropogenic Change
- 

## GLOSSARY

**benthic** Pertaining to the bottom of the sea or other aquatic environment.

**benthos** Organisms living on, in, or near the seabed or at the bottom of some other aquatic environment.

**coastal** Estuaries, semi-enclosed seas, and shallower regions of the ocean, including areas influenced by rivers and runoff from land.

**community** A group of species co-occurring in an area and interacting through trophic and spatial relationships.

**coral reef** Benthic environments characterized by reef-building corals with symbiotic dinoflagellates.

**deep sea** Volumes of water or areas of ocean bottom at depths greater than 200 m.

**ecosystem** A community of organisms and their physical environment interacting as an ecological unit.

**habitat** The locality or three-dimensional space occupied by an organism.

**mangrove** Environments characterized by mangrove trees.

**nekton** Actively swimming pelagic organisms.

**pelagic** Pertaining to the water column in aquatic environments.

**plankton** Organisms that float freely in the water column and do not maintain their position independent of water movements. Phytoplankton (literally plant plankton) is plankton with photosynthetic pigments and zooplankton is animals of the plankton.

---

**MARINE ECOSYSTEMS** may be defined as major units of ecological function in the marine environment. Ecosystems are communities of organisms and their physical, chemical, and geological environment—distinct assemblages of species coevolved with a particular environment over long periods of evolutionary history. As units of function, ecosystems have measurable imports and exports of material and energy. In comparison to ecosystems on land, ocean ecosystems have less clearly defined boundaries, a greater variety of major taxonomic divisions of organisms, and a long evolutionary history that preceded colonization of land. As the diversity of life in the oceans is explored, the importance of previously unrecognized aspects of ocean circulation, flux of energy and materials, and bottom characteristics to marine ecosystems are becoming better understood.

## I. MARINE ECOSYSTEMS

### A. Ecosystem Units

On land, ecosystems are separated into two-dimensional biomes, land areas defined by characteristic primary producing plants such as trees, grasses, and shrubs. Most shallow lakes and streams are similarly two-dimensional; however a few freshwater deep, ancient lakes, such as Lake Baikal in Siberia, and large rivers such as the Amazon have spatial complexity comparable to many coastal marine ecosystems. The ocean biosphere has an average depth of 4 km and comprises 99.5% of the biosphere. The dense seawater medium allows at least part of the life cycle of almost all marine organisms to be transported and dispersed by ocean currents. One ocean phylum is entirely pelagic, and about a third of the ocean phyla have representatives that spend their entire life cycle in near-surface waters as plankton. The boundaries that define ocean habitats and communities may involve a variety of overlapping criteria such as depth, distance from land, separation by landmasses, ocean currents, water masses of characteristic salinity and temperature, depth, and sea bottom characteristics such as sediment texture, composition, and surface topography. In addition, interactions with land and rivers and patterns of ocean circulation, light, nutrients, hydrology, and physical energy of water movements can strongly influence the distribution of species.

Descriptions of species boundaries are few and biogeographical classification depends heavily on the groups of organisms considered and how well they have been sampled. The ocean generally lacks the obvious barriers to dispersal characteristic of terrestrial environments. There may be multiple criteria for defining biogeographical provinces or marine ecosystems.

Major estuaries, where fresh water from rivers mixes with ocean water, are among the smallest individual ecosystem units in area. The largest units are regions defined by major boundary currents features such as the Gulf Stream, Kuroshio, and Brazil currents, and the north and south subtropical ocean gyres (the Sargasso Sea and South Atlantic Gyre in the Atlantic and the North Pacific Subtropical and South Pacific Subtropical Gyres in the Pacific). In the far north, the Arctic Ocean ecosystem is a distinct ocean basin covered by ice and the southern ocean around Antarctica is separated from the circulation of the Atlantic, Indian, and Pacific Oceans by the cyclonic circulation of the Antarctic Circumpolar Current.

As with terrestrial environments, marine ecosystems

may be classified by their characteristic primary producers (i.e., single-celled phytoplankton that float in the surface layers of the ocean, marsh grasses, sea grasses, mangrove trees, seaweeds such as those forming kelp beds, the single-celled plants called zooxanthellae that live symbiotically with corals, and the chemosynthetic bacteria living in water, sediments, or symbiotically with other organisms at hydrothermal vents or other seep environments rich in chemically reduced compounds such as sulfide or methane).

Using combinations of coastline, coastal bathymetry, ocean current systems, surface winds, and biota, the near-surface pelagic layer of the ocean where primary productivity occurs has been classified into 51 provinces (Fig. 1) by Longhurst (1998). Similar criteria have been used to classify coastal areas (Briggs, 1974). Marine sediments cover almost the entire surface of the ocean floor, yet a consistent global biogeographic classification of these benthic ecosystems has yet to be developed (Snelgrove *et al.*, 1997).

### B. Comparison of Marine Environments with Land

The ocean occupies 71% of the surface area of the globe and the deep sea at depths below 200 m occupies 63.5% of the earth's surface. Seawater is 830 times more dense than air and supports most of the biomass in the ocean. The volume of seawater in the ocean provides 99.5% of the livable volume of the earth (Cohen, 1994).

Concentrations of near-surface chlorophyll in the ocean are measured according to wavelengths of light reflected from the surface of the ocean, which are sensed by earth-orbiting satellites. Extensive studies of the relationship between near-surface chlorophyll and primary production allow satellite-derived information on chlorophyll to be converted to maps of primary productivity. Until very recently, overall primary production was thought to be approximately half that on land. Using distribution of chlorophyll in satellite photographs and models, primary productivity of the oceans has been shown to be about the same as that on land ( $\sim 45\text{--}50$  Pg C per annum in the ocean and  $\sim 55$  Pg C per annum on land; Falkowski *et al.*, 1998). For regions without ice cover, average net primary productivity (NPP) per area in the ocean is a third of that on land (ocean:  $140$  g C  $\text{m}^{-2}$   $\text{year}^{-1}$ , and land:  $426$  g C  $\text{m}^{-2}$   $\text{year}^{-1}$ ). Only about 1.7% of the ocean surface area has NPP greater than  $500$  g C  $\text{m}^{-2}$   $\text{year}^{-1}$  compared to 25% for land. Most productivity in the marine environment is from phytoplankton. Attached, multicellular algae contribute only about 2%. The highest productivity occurs in estu-

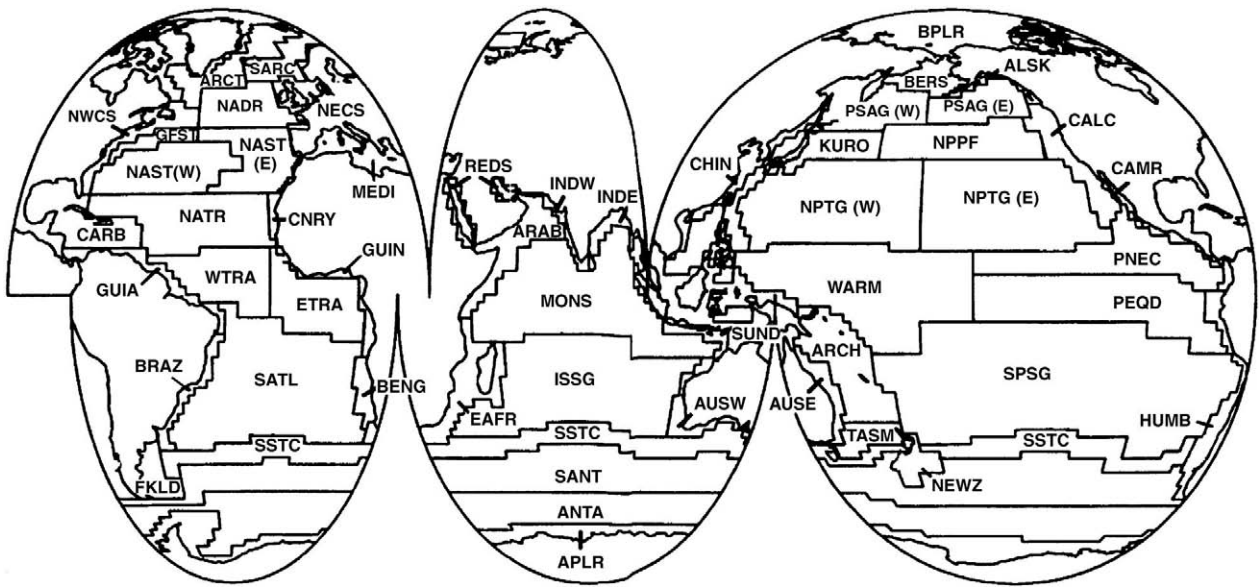


FIGURE 1 Pelagic biomes (Longhurst, 1998).

aries and upwelling areas—these highly productive areas contribute approximately 18% to net ocean primary productivity. In the open ocean, the greatest primary productivity is near the equator and at midtemperate latitudes in the Northern Hemisphere where there are regional maxima in terrestrial productivity. A smaller peak in productivity occurs in the Southern Subtropical Convergence where physical processes supply high concentrations of nutrients to surface waters (Falkowski *et al.*, 1998; Field *et al.*, 1998).

Marine primary producers are small and mobile whereas terrestrial primary producers are mostly large and rooted in the ground—trees account for approximately 80% of the primary production in terrestrial systems. By contrast, in central ocean gyres, photoautotrophic bacteria less than  $2\ \mu$  in diameter and short generation times account for most of the primary production. Oceanic biomass is extremely dilute and filtering of organic particles is an important mode of feeding in marine environments.

Oceanic food webs have an average food chain length of nearly six trophic links as opposed to four trophic links in terrestrial systems (Cohen, 1994). The number of species of smallest marine organisms, such as the various groups of one-celled marine organisms, are extremely poorly known. The relationship between the spectrum of individual body size and the spectrum of rates of population growth differs in marine and terrestrial systems (Fig. 2). In open ocean food webs, the hierarchy of size is not apparent at the lower trophic

levels because of the broad overlap in size of consumers and primary producers (Fig. 3, Karl, 1999).

The pattern of temporal variability of the physical environment differs between oceans and land. Marine ecosystems are characterized by about the same environmental variation over weeks and years as over days—variability is constant at frequencies ranging from days to decades. In terrestrial environments the variance of environmental parameters (e.g., temperature) increases steadily from frequencies of hours to millennia. Beyond 50 years the variance increases with increasing frequency as it does over the entire time spectrum on land (Steele, 1985).

On land, individual organisms have a high probability of surviving the relatively predictable patterns of environmental variation that occur over time periods up to decades. For example, individual trees and many vertebrate animals resist adverse effects of variation at all frequencies up to several decades because of their large size and long generation time. In the open ocean, time series measurements at a single station show that primary production varies significantly on periods from days to decades (Karl, 1999). Both seasonal and daily differences in cloud cover may result in three-fold variation in light at the surface. Vertical displacements of phytoplankton by internal waves further increase the amount of light absorbed by seawater before it reaches the photosynthetic organisms, creating a further source of variability. Small bacterial and flagellate primary producers have reduced the adverse effect of this variation



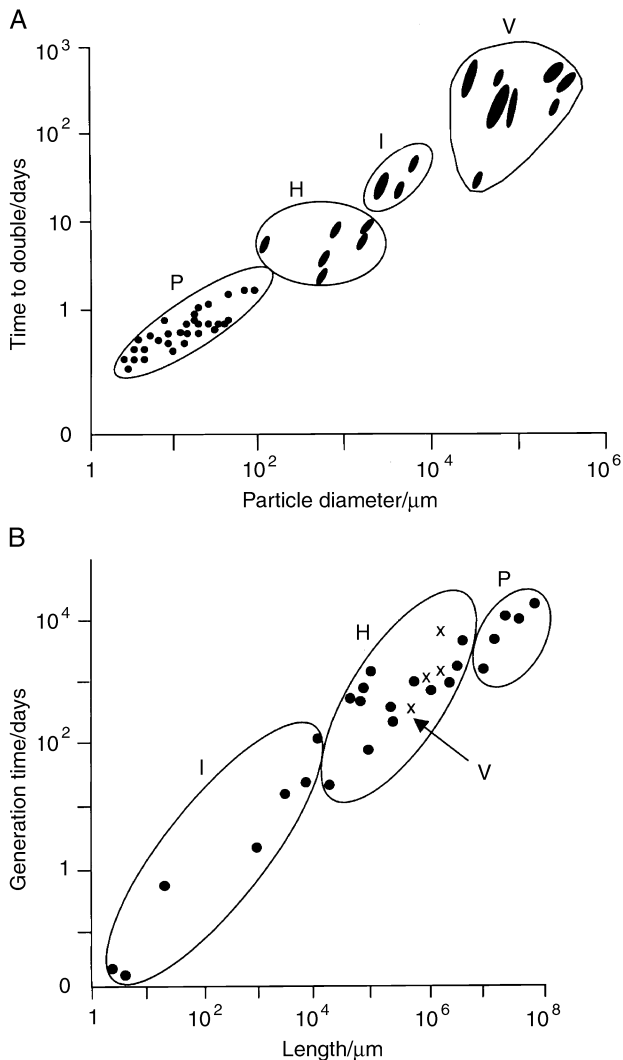


FIGURE 2 Relation of size to growth for plants (P), herbivores (H), other invertebrates (I), and vertebrates (V). (a) From Sheldon *et al.* (1972) for pelagic marine ecosystems. (b) From Bonner (1965, reprinted by permission of Princeton University Press) using only the terrestrial species. Derived from Cohen (1994, p. 60).

in light by supplementing their diet from the pool of dissolved organic matter excreted by other organisms.

Other distinctive features of marine populations are outlined by Cohen (1994) and in a U.S. National Academy of Sciences book on marine biological diversity (National Academy of Sciences, 1995). Plant and animal populations in marine ecosystems generally spend part of their life cycle as floating or swimming stages in the plankton. Unlike most terrestrial systems, the connections between benthic and planktonic life-history stages assume great significance and there is an unusually broad range of dispersal abilities, reproductive rates,

and generation times. Almost all species have the ability to disperse in the water column as larval stages produced by some form of sexual reproduction. As a consequence, marine ecosystems are largely open and distant marine habitats can be linked by dispersing larvae. Terrestrial systems are more localized functionally and localized extinction of species occurs more frequently. Invertebrate predators and grazers generally have very high reproductive output, which makes population fluctuations more likely. Fluctuations at the highest trophic levels affect interactions among species at successively lower trophic levels. This cascading effect often has unpredictable consequences, and even the lowest trophic level of primary producers may be controlled from the top down. Bottom-up control of food webs is exerted through the effects of nutrients and physical processes on primary productivity.

## II. BIODIVERSITY OF MARINE ECOSYSTEMS

### A. Higher Taxa

The three main biological lineages are the Bacteria, Archaea, and Eukarya (includes plants, fungi, protists, and animals). Recent advances in molecular-biological techniques permit the first measurements of highly diverse oceanic assemblages of bacteria and archaea that cannot presently be cultured in the laboratory. Bacteria are more abundant in the photic zone and archaea are more abundant in deeper water.

The Eukarya (all taxa except the Bacteria and Archaea) are divided into 71 well-defined monophyletic groups with no apparent taxonomic affinity with one another on the basis of cell organization (Patterson, 1999). Each of these groups includes taxa formerly assigned to the protists. By this classification animals and their relatives the choanoflagellates, and fungi and their relatives the chytrids, are defined as a single group. Plants are in another group altogether with 11 categories (~7000 species) of green algae.

Important groups of primary producers have affinities with several other monophyletic groups. The red algae are a distinct group with about 4000 known species; the ~1000 species of dinoflagellates are related to the ciliates. The ~10,000 species of diatoms are in a highly diverse lineage that includes kelps and other brown algae. The conspicuous red, green, and brown seaweeds of rocky shores are divided among three separate lineages. The two most important primary producers in the open ocean were formerly called blue-green

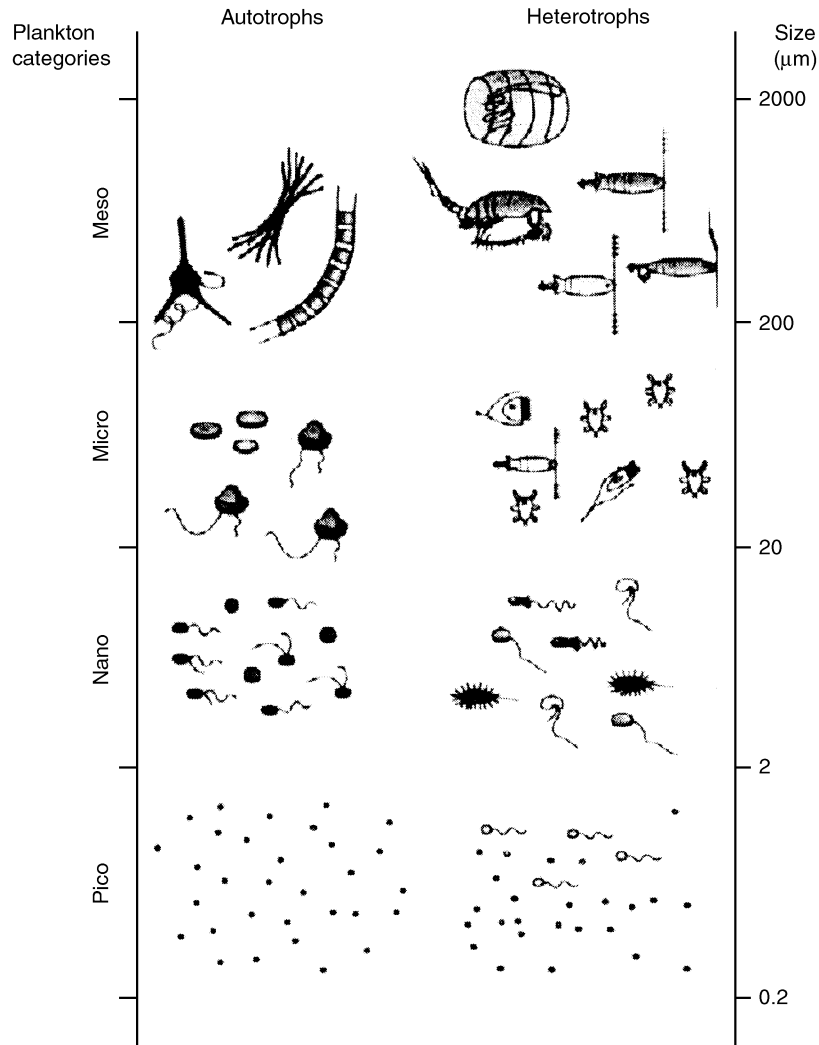


FIGURE 3 Representative classification of planktonic organisms by size showing the diversity of various autotrophic and heterotrophic groups. Size, per se, cannot be used to separate autotrophs from heterotrophs in NPSG plankton assemblages. Courtesy of Albert Calbet in Karl (1999).

algae. They are actually prokaryotic bacteria in two groups: the *Synechococcus* with three lineages, and the *Prochlorococcus* group with two lineages. These organisms account for most of the phototrophic standing stock and primary production in the open ocean (Andersen *et al.*, 1996).

Among the many nonphotosynthetic unicellular marine organisms, the ubiquitous Foraminifera are common both on the bottom at all depths and as pelagic organisms. Two abundant, poorly described benthic groups, the Komokiacea and the Xenophyophora (~40,000 known species), are big enough to be seen on the surface of deep-sea sediments. A leaflike form

of Xenophyophora may be as large as 25 cm in diameter. These groups are separate lineages with no obvious relatives.

In the classification of marine, free-living, multicellular animals there are 29 phyla. Figure 4 (modified from May, 1994) compares the described diversity and abundance among marine benthic, marine pelagic, freshwater, and terrestrial environments. Of the 29 known Phyla, all are known to have lived in the ocean and 14, or about half, are known only from the ocean. Living representatives of the Phylum Onychophora are presently found only on land in the Southern Hemisphere, but are also known from fossil organisms that

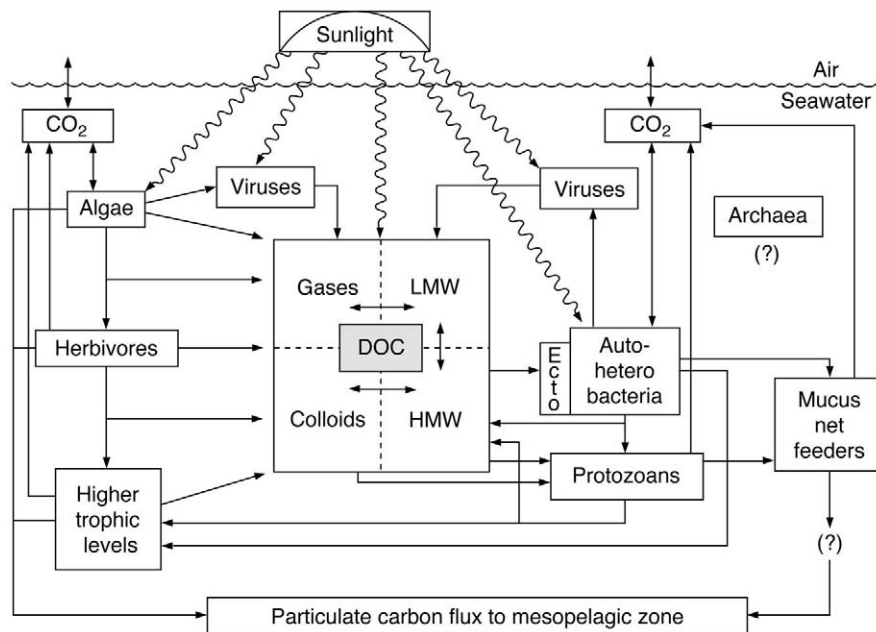


FIGURE 4 Schematic representation of the oceanic food web showing, on left, the classic pathway of carbon and energy flow through the photosynthetic eukarya to herbivores and on to higher trophic levels. Depicted on the right is the microbial food web, which uses energy stored in the nonliving, detrital carbon pool to produce microbial biomass that can reenter the classic pathway of carbon and energy flow. Cell-associated ectoenzymes (ECTO) enable bacteria to use high molecular weight (HMW) DOC in addition to the more traditional low molecular weight (LMW) and gaseous carbon substrates. Also shown in the microbial food web are viral particles and *archaea*. At the present time, there is only rudimentary knowledge of the role of *archaea* in the oceanic food web. Shown at the bottom of this diagram is the downward flux of particulate carbon (and energy), which is now thought to fuel most subeuphotic zone processes. The classic algae-herbivore grazer pathway is most important in this regard. From Karl (1999).

lived in the ocean more than 300 million years ago. Species diversity on land is dominated by insects and trees, groups that play a significant role only at the margins of the marine environment. Only about 15% of described species are found in the marine environment, but this may reflect the much greater cumulative effort devoted to species descriptions on land, rather than an actual difference in the number of species (May, 1994).

## B. Species

Species are the basic units of evolution and represent the biological variability for future generations of life. For whole collections, species diversity is measured as the number of species and their relative abundances within and between habitats, regions, or other ecological or geographical units. Species richness is measured by collecting enough samples to represent very large numbers of individuals over very large areas. Ideally,

communities should be sampled until the rate at which new species are found declines, and a plot of species versus area approaches a constant number of species. This level of sampling effort is achieved for groups with few rare species (e.g., larger animals including most vertebrates, planktonic organisms, and macrophytic plants). For species-rich taxa of bottom-dwelling invertebrates from coral-reef or deep-sea habitats, this level of sampling has not been attained. Where habitats are patchy and the vast majority of species are rare, it is seldom possible to collect and process enough samples to estimate species richness accurately.

For individual samples, indices based on the absolute number of species and the relative abundance of species are used to study species diversity. The most commonly used index is the Shannon-Wiener information function,  $H'$ , which equals the frequency of each species  $p_i = s_i/\sum s_i$  multiplied by  $\log_2 p_i$  summed over the number of species ( $n$ ) collected (e.g.,  $\sum p_i \log_2 p_i$ ). Another measure, Hurlburt rarefaction, calculates a species versus

individuals curve for each sample based on the expected number of species in successively smaller subsamples drawn from an actual sample. These species diversity curves are especially useful in comparing samples of unequal size.

There are approximately 200,000 described species of animals in the marine environment (Table I). The

TABLE I  
Free-Living Animal Phyla and Their Relative Numbers of Described Species (4 = >10<sup>4</sup>, 3 = >10<sup>3</sup>, 2 = >10<sup>2</sup>, 1 = present)

Phylum	Marine		Freshwater	Terrestrial
	Benthic	Pelagic		
Annelida	4	1	2	3
Arthropoda	4	3	3	5
Brachiopoda	2	0	0	0
Bryozoa	3	0	1	0
Chaetognatha	1	1	0	0
Chordata	3	3	3	3
Cnidaria	3	2	1	0
Ctenophora	0	1	0	0
Dicyemida	0	0	0	0
Echinodermata	3	1	0	0
Echiura	2	0	0	0
Gastrotricha	2	0	2	0
Gnathostomulida	2	0	0	0
Hemichorodata	1	0	0	0
Kamptozoa	1	0	1	0
Kinorhyncha	2	0	0	0
Loricifera	1	0	0	0
Mollusca	4	2	3	4
Nematoda	3	0	3	3
Nematomorpha	0	0	0	0
Nemertea	2	1	1	1
Onychophora	0	0	0	1
Phoronida	1	0	0	0
Placozoa	1	0	0	0
Plathyhelminthes	3	1	3	2
Pogonophora	2	0	0	0
Porifera	3	0	1	0
Priapula	1	0	0	0
Rotifera	1	1	2	1
Sipuncula	2	0	0	0
Tardigrada	1	0	2	1
Total (30)	27	11	14	10
endemic	13		0	1

most species-rich and least well known areas are coral reefs and the sediments of the deep-sea floor. There are no precise estimates for these environments but estimates for coral reefs alone exceed 600,000 species (Reaka-Kudla, 1997). Based on quantitative analysis of 233 box core samples from the Atlantic Ocean continental slope and rise off the east coast of North America, Grassle and Maciolek (1992) estimated 1 to 10 million macrofaunal species in the deep sea (Gage and Tyler, 1991). May (1994) estimated 0.5 million based on the portion of species previously undescribed in the Grassle and Maciolek study. Poore and Wilson (1993) analyzed samples from the Southern Pacific Ocean off Australia and, on the same basis, estimated that there are 5 million species in deep-sea sediments. Multicellular animals small enough to pass through a 1 mm sieve (meiofauna), such as nematode worms, are even less well known and Lamshead has argued that there may be 100 million species if nematodes are included (Lamshead 1993). Reasons for high diversity of species in the ocean include the long evolutionary history of the ocean, the vast area of deep-sea floor ( $3 \times 10^8$  km<sup>2</sup>) with relatively few barriers to dispersal, and the episodic nature of patch formation within and between habitats on a variety of spatial and temporal scales.

### C. Genes

Genetic diversity is the heritable variation among individuals measured as allelic diversity at a broad sampling of genetic loci or as genetic sequence information at the molecular level within populations. Genetic variation occurs among subpopulations as well as within populations. Differentiation among subpopulations results from natural selection for genetic variants adapted to local patterns of environmental variation or random loss of genetic variants in small isolated subpopulations. Species with relatively high rates of dispersal are less likely to form subpopulations and species with very poor dispersal ability are more likely to diverge from parent populations as a result of random processes. In coastal areas, genetic divergence is related to the length of life of dispersal stages and barriers to current flow from one place to another along a coastline. For some shallow-water species, genetic isolation of island populations is related to distances among islands. The archipelagos in the central Indopacific in the vicinity of Indonesia and Papua New Guinea have a high richness of species, which then declines eastward to relatively isolated peripheral island archipelagos (Planes and Galzin, 1997; Stehli, 1965). In the same region, in a study of population differentiation in four species of sea urchins,

Palumbi (1997) found high genetic diversity (mitochondrial DNA sequence diversity) in the central area (1.6% variation among individuals) and much lower genetic diversity (0.5% variation among individuals) in peripheral island localities to the east. For these species, genetic diversity and species diversity covary across gradients suggesting a similarity in the processes maintaining gradients in diversity despite different mechanisms for the origin of the variation. Fluctuations in population size in relatively isolated populations could result in both loss of genetic variants and reductions in number of species (Palumbi, 1997).

In the deep ocean, hydrothermal vents are analogous to islands in the sense that these fluid flows support widely separated biological communities, linearly aligned along the Mid-Ocean Ridge. The patterns of deep-sea ocean currents that transport dispersal stages of species restricted to hydrothermal vents are poorly understood, but it is possible to make estimates of gene flow from the extent of genetic differentiation among populations of individual species. The flow of hydrothermal fluids, containing energy-rich reduced compounds such as hydrogen sulfide, supports chemosynthetic primary productivity. At East Pacific Rise vents, the flow of hydrothermal fluid may last only a decade or two at any one site and all populations are maintained by dispersal over considerable distances. Species can be divided into three categories: those that show no geographic pattern of genetic differentiation, those that are isolated by distance, and species without a free-living larval dispersal stage, which apparently have good dispersal to sites along a single ridge segment but poor dispersal between separated ridge segments (Vrijenhoek, 1997). The latest methods for measuring genetic diversity have been applied to very few marine species and rapid advances in this area of research can be expected.

### III. ECOSYSTEM FUNCTION

It is useful to classify members of species assemblages according to their feeding relationships with other species in the ecosystem. A trophic unit includes all species that eat the same kinds of foods or are consumed by the same kinds of consumers. Within a food chain, there is a hierarchy of consumers from primary producers to primary consumers followed by a further sequence of consumers. Each step in a food chain results in a reduction in biomass, and simple food chains are often described as a pyramid with plants at the base and apex predators at the top. In the water column, unicellular

phytoplankton form the first trophic level of marine food chains. The second level is formed by herbivores and detritivores and subsequent levels are formed by successive levels of predators. Species at the highest trophic levels can affect the food web relationships among species at lower levels. For example, removal of a top predator can have cascading effects on herbivores and ultimately on primary producers.

Because of the dilute seawater medium, a great many marine species have developed both active and passive means for filtering or trapping food particles from the dilute seawater medium. Copepods, the most common animals in the water column, have filtering appendages and gelatinous zooplankton cast mucous nets to feed on phytoplankton. Baleen whales filter zooplankton (krill) from the water column. On the sea bottom, clams and sea cucumbers pump water past internal filters and many animals in sediments pump water through burrows in order to feed. Other bottom animals have appendages protruding above the sediments that trap or filter food particles. In many marine organisms, the distinction between producers and consumers is blurred. Reef-building corals use their tentacles to trap zooplankton yet may take most of their sustenance from photosynthetic dinoflagellates living symbiotically in their tissues. Other animal-plant relationships of this sort are found in tropical clams and one-celled radiolarians and foraminifera.

Some marine species play another important functional role by providing habitat for other species, either on a large spatial scale—as with coral or coralline algae reefs, polychaete worm reefs, seagrasses, kelps, marsh grasses, and mangrove trees. On a smaller scale, biogenic sediment structures (tubes, burrows, mounds, fecal aggregations) and more persistent structures made by tube builders, sponges, or shell-bearing animals may serve as habitat for other species.

Some species significantly affect the ecosystem by regenerating nutrients that limit primary production. Burrowing animals release nitrogen into the water column and stimulate phytoplankton growth. In chemically reduced sediments, animals pump water into sediments for respiration or feeding and supply oxygen to chemosynthetic primary producers living in the burrow. The role of single species is often not obvious, and several different criteria may be used to assign species to functional groups within an ecosystem. In general, redundancy of ecosystem function within a functional group has the potential to stabilize ecosystem processes despite fluctuations in the environment. Loss of functional groups implies drastic changes in ecosystem function.

## IV. ECOSYSTEM DIVERSITY

### A. The Edge of the Ocean

#### 1. Intertidal Beaches

Beaches can be classified according to topography, organic content of sediments, and wave action. Reflective beaches are dominated by low wave energy, low organic content, and coarse sand. Reflective beaches have waves 1 m high or less and are generally found on open coasts with deep embayments, tropical coasts, and coasts of polar seas. Surging wave action filters and drains large volumes of water through the interstices of the sediments, resulting in well-flushed and highly oxygenated coarse sand deposits (Alongi, 1998). Dissipative beaches, at the other extreme of a continuum, are produced by a combination of high waves ( $>2.5$  m) and fine sand deposits with higher amounts of organic matter. These are common on the west coasts of Australia and Southern Africa and seasonally on the west coast of North America where high wave swells and fine sands are abundant. Intertidal sand and mudflats are common on dissipative beaches.

Many beaches have adjacent seagrass beds, kelp beds, or other sources of macrodetritus, which are deposited as thick layers of wrack on the beach. These accumulations support communities that include both marine and terrestrial invertebrates (e.g., beach hoppers, beetles, and kelp-fly larvae). Other beaches are more dependent on growth of diatoms in the sediments and input of small, filterable organic particles. Many animals live in the sediments, and in some high energy situations animals such as mole crabs and small bivalves move up and down the beach with the tides filtering particles from the waves. Large areas of sand flats, such as the Wadden Sea in the Netherlands, may be especially productive and support high standing stocks of grazing invertebrates.

#### 2. Kelp Beds

Kelps attach to the bottom and form a surface canopy at depths up to  $\sim 20$  m. Under the most favorable conditions these large marine plants form subtidal forests and attain rates of primary production in excess of  $1000 \text{ g C m}^{-2} \text{ d}^{-1}$ . These forests provide protection and food for a rich community of fish and invertebrates. The biomass and abundance of kelps may be regulated by sea urchin consumers. Sea otters play an important role in maintaining kelp forests by controlling the abundance of sea urchins. In the absence of sea otters, kelp forests are reduced by urchins to a pavement of encrusting algae and sea urchins. Kelp forests are impor-

tant nursery areas for many species of fish and their detrital production enhances the abundance of benthic populations (United Nations Environmental Programme, 1995). Kelp populations are influenced over large scales by oceanographic climate. Nutrient-rich conditions during La Nina years result in increased growth and reproduction of the competitively dominant, canopy kelp species, *Macrocystis pyrifera*. Interdecadal-scale shifts in community composition result from fluctuations in kelp density (Dayton *et al.*, 1999).

#### 3. Rocky Shores

Rocky coasts exposed to the open ocean are characterized by wave action resulting in communities of attached seaweeds and filter-feeding bivalve mollusks, such as mussels that provide physical structure for other species. Wave energy enhances the productivity of these ecosystems by continually renewing nutrients and food. The shore face and the organisms that reside on the shore can be divided into zones according to tidal height and length of exposure to air and the interactions of the dominant species with herbivores such as snails (gastropod mollusks) and predators (particularly snails, starfish, and birds). The large-scale pattern of rocky-shore communities depends on the distribution of rocky outcrops and sporadic changes in climate resulting in unusually heavy waves, ice cover, or sedimentation from rivers. The interaction of physical change and biological relationships among species at a variety of spatial scales (from local to regional) and temporal scales (from annual storm events to interdecadal climatic change) are most clearly worked out for rocky intertidal ecosystems.

#### 4. Coral Reefs

Coral reef ecosystems occur where conditions are favorable for growth of reef-forming corals with dinoflagellate primary producers living symbiotically in their tissues. Growth of corals over many generations in geologic time results in major limestone structures such as coral atolls or the Great Barrier Reef off Australia. Dense growths of coral can sometimes occur in the deep sea, but these species lack photosynthetic symbionts, grow relatively slowly, and do not form major reef structures.

Reefs grow in strong light and clear water at temperatures from  $18^{\circ}\text{C}$  to  $30^{\circ}\text{C}$  at latitudes between  $30^{\circ}\text{N}$  and  $30^{\circ}\text{S}$ . Coral reefs are adversely affected by high nutrient concentrations, runoff of sediments from land, direct removal, and overfishing. The midrange of primary production of corals in combination with their symbiotic dinoflagellates is about  $25 \text{ g C m}^{-2} \text{ d}^{-1}$  and

varies greatly from species to species. Over large areas, net primary productivity of the most actively growing reef crests and slopes ranges from 1 to 5 g C m<sup>-2</sup> d<sup>-1</sup>.

Reefs support an enormous species richness and complexity of interactions among species. Conspicuous large animals include enormous coral heads and large fish such as groupers, stingrays, and manta rays. Many of the colorful reef fish do not move far and develop complex behavioral relationships both within and between species. Some live symbiotically with other species, for example, individual anemone fish live in close association with patches of anemones. Cleaner fish set up cleaning stations where they feed on the ectoparasites attached to the gills of other fish. Some species mimic the cleaner fish and take bites out of the fish expecting to be cleaned of parasites.

## B. Continental Shelves

Continental shelf coastal areas, on the order of 10,000 km<sup>2</sup> or more, have been called “large marine ecosystems” (Sherman, 1993). These are separated from other areas of the ocean by continental shelf depth and ocean currents, and the shapes of coastlines form major seas, bays, or gulfs. Examples include the Baltic, North, Mediterranean, Black, Caspian, Red, Arabian, Barents, Bering, Okhotsk, Japan, Yellow, East China, Sulu, Celebes, and Caribbean Seas; Bay of Bengal and Walvis Bay; and Gulfs of Alaska, California, and Mexico. Primary productivity in these systems ranges from below 35 g C m<sup>-2</sup> yr<sup>-1</sup> in the low latitude, warm waters of the Red Sea and high latitude, cold waters of the Beaufort Sea (10–20 g C m<sup>-2</sup> yr<sup>-1</sup>) to the very high primary productivity of Eastern Boundary Current upwelling areas in the Southern Hemisphere (1000–2000 g C m<sup>-2</sup> yr<sup>-1</sup>) of the Peru Current and Walvis Bay (Walsh, 1988). Most of the world’s major fisheries are on continental shelves in midlatitudes.

## C. The Open Ocean and Deep Sea

### 1. Pelagic

The largest ecosystems in the ocean are the central gyres of the Atlantic, Pacific, and Indian Oceans. Ecosystem processes in the North Pacific Subtropical Gyre (NPSG) have been summarized by Karl (1999). This ecosystem is the largest circulation feature on the planet (2 × 10<sup>7</sup> km<sup>2</sup>) and one of the most persistent, its boundaries having remained approximately the same for the past 10<sup>7</sup> years. The NPSG has a clockwise circulation of less than 4 cm s<sup>-1</sup> and forms a circumscribed, stable, and

relatively homogenous habitat. The surface mixed layer varies from 40 m to 100 m depth and is characterized by surface temperatures are 24°C or higher low nitrate concentrations but relatively high dissolved organic nitrogen, and low standing stocks of organisms. The zone of primary productivity can be divided into two layers: an upper layer where chlorophyll increases in the winter and decreases in the summer and lower layer (100–175 m) where chlorophyll increases in the spring and declines in the fall. Recharge of nutrients is from deeper water below as a result of vertical eddy diffusion and episodic mixing events leading to considerable spatial variability in mixing processes and nutrient concentrations varying by as much as three orders of magnitude. Phytoplankton primary production was once thought to be mostly by Eukaryotes (diatoms and flagellates), but is now known to be more than 90% from the small bacterial taxa *Synechococcus* and *Prochlorococcus*. The standing stock of these autotrophic bacteria groups comprise 80% of chlorophyll a and feed a microbial loop that internally regenerates nutrients and maintains a pool of dissolved organic matter, which supports them (Fig. 4). The abundance of these auto-heterotrophs is controlled by light, nutrients, and predation by bacteria and a mixed assemblage of protists. Viral infection may also be an important source of mortality for these organisms. Archaea are abundant but it is not clear whether these are significant chemosynthetic primary producers because little is presently known about these organisms.

Very little organic matter escapes remineralization and the microbial loop provides negligible subsidy to the rest of the food web. The classic food chain pathway of eucaryote phytoplankton to copepod herbivores and on to higher trophic-level fish is ephemeral and occurs more frequently in surface waters during the summer. Organic matter produced by the eucaryotic phytoplankton food chain produces most of the exportable carbon during aperiodic, pulsed events.

Falkowski *et al.* (1998) provide a summary of biogeochemical processes controlling primary production in the open ocean. The central ocean gyres in the Atlantic, Pacific, and Indian Oceans have been considered analogous to deserts on land with low primary productivity and contain only ~0.2 mg m<sup>-3</sup> of chlorophyll. Coastal upwelling regions, seasonally mixed regions of temperate and boreal seas, divergent subpolar gyres, and meso-scale features with eddy-induced pumping have sufficient vertical flux of nutrients to support 5 mg m<sup>-3</sup> of chlorophyll. Throughout most of the coastal and open ocean, primary production is limited by the availability of inorganic fixed nitrogen. In some instances, the cyanobacteria that fix nitrogen in the open ocean are limited

by iron and an important source of iron to the ocean is dust carried from land by winds. Limitation of primary production by lack of iron is especially notable in the South Pacific (Falkowski *et al.*, 1998).

## 2. Benthic

The deep-sea floor is divided into major ocean basins by continents and the Mid-Ocean Ridge. Communities within ocean basins may be further divided according to depth, sediment type, and level of energy of deep-sea currents. The deep ocean floor is the least-known part of the planet but, through use of manned and unmanned submersibles, distinct ecosystem processes at hydrothermal vents, continental margin seeps, seamounts, ocean trenches, and areas of strong bottom currents are being explored and described.

The largest ocean basins and deep ocean trenches each have some species that live only in that basin and nowhere else. Hydrothermal processes along the Mid-Ocean Ridge mix seawater through porous rock at high temperatures yielding an energy-rich fluid containing reduced compounds. These compounds support chemosynthetic microorganisms that provide primary production for a discrete ecosystem clustered around each hydrothermal vent. Flow of subsurface fluid seeps out of sediments deposited along some ocean margins providing similarly energy-rich fluid to chemosynthetic organisms.

The food supply for the deep sea comes from the productivity of surface waters. When diatoms bloom, or gelatinous animals such as salps multiply rapidly, they die and sink, so that organic material accumulates in low areas of the uneven surface of the sea floor and in burrows and depressions left by the larger inhabitants. Even in the central ocean gyres where export production is low, the dead remains of fish, marine mammals, or terrestrial plant material carried seaward sink and form widely separated organic patches on the sea floor. Species respond to these patches at different rates and the probability that two species reach the same patch at the same time is low. This reduces the likelihood of species competing and of one species eliminating another. Most deep-sea species are small and many species, including most fish species, are relatively slow growing, long lived, and late in maturation. Attempts to sustain deep-water fisheries have proven unsuccessful because low rates of population growth cannot keep up with rates of removal.

Species that grow relatively fast characteristically respond to patchy but concentrated sources of food from the ocean surface, such as wood from rivers, or the bodies of pelagic animals that settle to the bottom. For

example, wood-boring bivalves rapidly colonize pieces of wood, grow to maturity in a few months feeding on their wood habitat, and produce thousands of eggs and larvae to colonize the next piece of wood that settles to the sea floor. Other species of bivalves grow very slowly in relatively homogeneous sediments, take several decades to reach maturity, and may produce only one egg at a time—in contrast to the rapid maturation and production of millions of eggs produced by most shallow-water bivalves.

Submarine canyons form conduits for sediment from continental shelves into the deep ocean. Unpredictable events of sediment erosion or scouring by intense currents result in relatively few species in the soft sediments at the bottom and sides of canyons. Seamounts are undersea mountains formed by the same processes at the hot spots on the ocean floor that form volcanic islands. Seamounts often support large populations of fish, and more than 70 species of commercially important fish have been reported. Interactions of currents with the steep topography of seamounts results in areas of enhanced primary productivity and concentrations of zooplankton that provide food for fish and dense concentrations of bottom animals (Rogers, 1994).

## D. Mid-Ocean Ridges and Hydrothermal Vents

The 40,000 nautical mile Mid-Ocean Ridge system is the largest feature on the deep-sea floor. In 1977 a unique ecosystem was discovered at sites where a plume of high-temperature fluid rich in reduced compounds pours out into the water column. It is now known that sulfur oxidizers are among the most numerous bacteria and form a major base of the food chain. Other energy sources include reduced iron, manganese, and hydrogen. In the Pacific, large, red-plumed worms up to 2 m long and large clams and mussels dominate the vents. These animals feed on organic compounds produced by symbiotic sulfur bacteria living in their tissues. Vents in the Atlantic have some of the same kinds of animals, but the most conspicuous are shrimp, which swarm over the surface of vent chimneys. Vents usually have a restricted distribution on any given ridge segment and persist for about 10 to 20 years, until there is local extinction of the vent community. Animals colonize new vents rapidly, grow fast, and produce enough offspring to colonize the next vent. In comparison with the rest of the deep sea, few species have adapted to the extreme thermal (4°C up to temperatures in excess of 150°C), chemical (high concentrations of cadmium, lead, cobalt, and arsenic) conditions at hydrothermal



vents (Grassle, 1986). Most species found at hydrothermal vents live exclusively in this environment. Of the 443 species found at hydrothermal vents, 15 have been found in other sulfide-rich environments and only 30 species are known from elsewhere in the deep sea (Tunnicliffe *et al.*, 1998).

## V. POTENTIAL CONSEQUENCES OF ANTHROPOGENIC CHANGE

### A. Eutrophication

Eutrophication is the increase in the rate of supply of organic matter to an ecosystem. Increases in global inputs of nitrogenous fertilizers and the mining of phosphate rock have generated increased concern about the effects of eutrophication on enclosed marine ecosystems (Nixon, 1995). Eutrophic ecosystems have algal production in excess of  $300 \text{ g C m}^{-2} \text{ y}^{-1}$ , which results in areas of anoxia and loss of habitat for fish and other organisms. Relatively high rates of denitrification on continental shelves remove excess nitrogen originating from land sources and, in concert with dilution, help prevent adverse eutrophication effects in open coastal areas.

### B. Overfishing

Globally, about 30% of commercial fish stocks are overfished and another 44% are being fished at or near the maximum potential long-term catch rate. Atlantic halibut, cod, orange roughy, and many species of salmon are now severely depleted. Significant changes in community structure as a result of overfishing have occurred in ecosystem structure in the Bering, Barents, and Baltic Seas (National Academy of Sciences, Committee on Ecosystem Management for Sustainable Marine Fisheries, 1999). Bottom-fishing has been shown to result in physical destruction of some bottom habitats. Fishing gear, when dragged over the bottom, levels structures such as worm tubes, burrows, and shell hash necessary for the survival of many species.

Overfishing has resulted in major changes in coral reef ecosystems. Normally, herbivorous fish heavily graze the attached algae, ensuring enough open reef surface for corals to settle and grow. This is especially true following major storms when wave action reduces coral coverage and circumstances are favorable for rapid algal growth. In the Caribbean, under normal circumstances, sea urchin grazing may compensate for reductions in fish grazing. A combination of overfishing and

the decimation of sea urchin grazers by disease favored algal growth following a hurricane, which has resulted in reefs dominated by algae (National Academy of Sciences, 1995).

### C. Invasive Species

Unwanted, exotic species are sometimes introduced to new geographic regions both deliberately to start new fisheries and accidentally through release from aquaria or ballast water carried by ships, sometimes with disastrous consequences. The Asian clam became established in the San Francisco Bay in 1986 and quickly displaced other species from large areas of the seabed and altered the water chemistry of the bay (National Academy of Sciences, 1995). The introduction of predatory green crabs to coastal environments on the east coast resulted in major reductions in shellfish beds. In short, invasive species have become a significant problem in many marine coastal environments and considerable effort is needed to curb this severe problem.

In summary, the oceans encompass a broad array of habitats that differ in their diversity, function, and vulnerability. Much of the vast area of the oceans is poorly described, but we have some understanding of a variety of globally essential ecosystem processes, and species loss may threaten not only the organisms themselves but also the many ecological processes that serve the rest of the planet and its human populations.

### See Also the Following Articles

COASTAL BEACH ECOSYSTEMS • ENDANGERED MARINE INVERTEBRATES • ESTUARINE ECOSYSTEMS • INTERTIDAL ECOSYSTEMS • INVERTEBRATES, MARINE, OVERVIEW • MANGROVE ECOSYSTEMS • MARINE ECOSYSTEMS, HUMAN IMPACT ON • PELAGIC ECOSYSTEMS • REEF ECOSYSTEMS • VENTS

### Bibliography

- Alongi, D. M. (1998). *Coastal Ecosystem Process*. CRC Press, Boca Raton.
- Briggs, J. C. (1974). *Marine Zoogeography*. McGraw Hill, New York.
- Cohen, J. E. (1994). Marine and continental food webs: Three paradoxes? *Phil Trans. R. Soc. Lond. B* 343, 57–69.
- Dayton, P. K., Tegner, M. J., Edwards, P. B., and Riser, K. L. (1999). Temporal and spatial scales of kelp demography: The role of oceanographic climate. *Ecol. Monogr.* 69, 219–250.
- Falkowski, P. G., Barber, R. T., and Smetacek, V. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science* 281, 200–206.
- Gage, J., and Tyler, P. A. (1991). *Deep-Sea Biology: A Natural History*

- of *Organisms at the Deep-Sea Floor*. Cambridge University Press, Cambridge.
- Grassle, J. F. (1986). The ecology of deep-sea hydrothermal vent communities. *Advances in Marine Biology* 23, 443–452.
- Gray, J. S. (1997). Marine biodiversity: Patterns, threats and conservation needs. *Biodiversity and Conservation* 6, 153–175.
- Karl, D. M. (1999). A sea of change: Biogeochemical variability in the North Pacific Subtropical Gyre. *Ecosystems* 2, 181–214.
- Longhurst, A. (1998). *Ecological Geography of the Sea*. Academic Press, San Diego, CA.
- May, R. M. (1994). Biological diversity: Differences between land and sea. *Phil. Trans. R. Soc. Lond. B* 343, 105–111.
- National Academy of Sciences, Committee on Biological Diversity in Marine Systems. (1995). *Understanding Marine Biodiversity*. National Academy Press, Washington, D.C.
- National Academy of Sciences, Committee on Ecosystem Management for Sustainable Marine Fisheries. (1999). *Sustaining Marine Fisheries*. National Academy Press, Washington, D.C.
- Palumbi, S. R. (1997). Molecular biogeography of the Pacific. *Coral Reefs* 16, Suppl.: S47–S52.
- Patterson, D. J. (1999). The diversity of Eukaryotes. *American Naturalist* 154, Suppl., S96–S124.
- Reaka-Kudla, M. L. (1997). The global biodiversity of coral reefs: a comparison with rain forests. In *Biodiversity II* (M. L. Reaka-Kudla, D. E. Wilson, and E. O. Wilson, Eds.). Joseph Henry Press, Washington, D.C.
- Rogers, A. D. (1994). The biology of seamounts. *Advances in Marine Biology* 30, 305–350.
- Sheldon, R. W., Prakash, A., and Sutcliffe, W. H., Jr., (1972). The size distribution of particles in the ocean. *Limnol. Oceanogr.* 17, 327–340.
- Sherman, K. (1993). Large marine ecosystems as global units for marine resources management—An ecological perspective. In *Large Marine Ecosystems* (K. Sherman, L. M. Alexander, and B. D. Gold, Eds.). AAAS Press, Washington, D.C.
- Snelgrove, P. V. R., Blackburn, T. H., Hutchings, P. A., Alongi, D. M., Grassle, J. F., Hummel, H., King, G., Koike, I., Lamshead, P. J. D., Ramsing, N. B., and Solis-Weiss, V. (1997). The importance of marine sediment biodiversity in ecosystem processes. *Ambio* 26, 578–583.
- Stehli, F. G. (1965). Taxonomic diversity gradients in pole location: The recent model. In *Evolution and Environment* (E. T. Drake, Ed.). Yale University Press, New Haven, CT.
- Tunncliffe, V., McArthur, A. G., and McHugh, D. (1998). A biogeographical perspective of the deep-sea hydrothermal vent fauna. *Advances in Marine Biology* 34, 353–442.
- United Nations Environmental Programme. (1995). In *Global Biodiversity Assessment* (V. H. Heywood, Exec. Ed., R. T. Watson, Chair), pp. 370–399. Cambridge University Press, Cambridge.
- Vrijenhoek, R. C. (1997). Gene flow and genetic diversity in naturally fragmented metapopulations of deep-sea hydrothermal vent animals. *Journal of Heredity* 88, 285–293.





# MARINE ECOSYSTEMS, HUMAN IMPACTS ON

Juan Carlos Castilla  
*Pontificia Universidad Católica de Chile*

---

- I. Introduction
  - II. Human Impacts on Marine Communities and the Effects on Species Diversity and Functioning
  - III. Nonanthropogenic Environmental Changes and Variability
  - IV. Conclusions
- 

## GLOSSARY

**alien: introduced, exotic, nonindigenous, nonnative, invasive species** A species that has been transported by human activity (i.e., mariculture), intentionally or accidentally, to a site at which it does not naturally occur.

**ballast water** Water carried by a vessel to improve stability.

**benthic organism** An organism pertaining to the seabed; bottom-dwelling.

**biodiversity** The variability among living organisms from all sources and the ecological systems of which they are a part.

**disturbance** Any relatively discrete event in time that disrupts ecosystem, community, or population structure and changes resources, substrate availability, or the physical environment.

**ecosystem** A complex nonlinear community of organisms in their physical environment.

**ecosystem engineer species** Species that directly or indirectly modulate the availability of resources (other than themselves to other species) by causing physical state changes in biotic or abiotic material

and in so doing modify, maintain, and/or create habitats.

**eutrophication** Enrichment of a body of water with nutrients causing excessive growth of phytoplankton, seaweed, or vascular plants and often accompanied by a depletion of oxygen.

**food web, trophic web** A network of interconnected trophic chains in a community. A network of consumer–resource interactions among a group of organisms, populations, or aggregate trophic units.

**guild** A group of species having similar functional roles in the community (i.e., herbivores).

**keystone species** A group of species whose effects on the structure, dynamics, and functioning of the community is disproportionately large relative to its abundance.

**pelagic organism** A free-swimming (nekton) or floating (plankton) organism that lives exclusively in the water column.

**resilience** The resistance to a disturbance of a system and the speed of return to an equilibrium point, or the disturbance that can be absorbed before the system changes in structure by the change of variables and processes that control system behavior.

**species diversity** The number of species in a given community (= species richness) and the way the species' abundances (i.e., number, biomass, and cover) are distributed among species (= species evenness).

**trophic level** Feeding level in a food chain or pyramid (e.g., carnivores).

---

MARINE ECOSYSTEMS represent the greater part of the earth's total biological system. At the present time these marine communities are threatened by human effects, both direct and indirect, such as resource extraction (e.g., fishing), introduction of alien species, pollution, and water temperature modification. These effects demonstrate the unique ability of humans to profoundly influence the status of ecosystems.

## I. INTRODUCTION

The main threats to marine ecosystems are the human alteration of habitats, the excessive extraction of resources, pollution (Castilla, 1996), invasive species (i.e., introduction through mariculture and ballast water; Cohen and Carlton, 1998), eutrophication, and nonanthropogenic environmental changes [National Research Council (NRC), 1999; Castilla and Camus, 1992]. Furthermore, multiple and compounded perturbations related to physically and biological based disturbances are resulting in communities entering new domains or "ecological surprises" (Paine *et al.*, 1998), with important modifications in their structure (i.e., species composition) and dynamics (i.e., alternative states).

Single, multiple, or compounded impacts on ecosystems may directly or indirectly affect their structure, including species diversity and functioning. Ecosystems are complexly linked nonlinear systems and their dynamics may be sensitive to past conditions and subjected to shifts when exposed to anthropogenic and nonanthropogenic environmental stress (NRC, 1999).

The concept of biological diversity (biodiversity; Heywood, 1995) is defined as: the variability among living organisms from all sources and the ecological system to which they are part. The analysis of biodiversity considers four levels: genetic, species, community, and ecosystems. This article focuses on the species diversity (richness, the number of species in a given community; evenness, species abundance), community resilience, and ecosystem functioning. One of the best avenues to integrate species diversity functioning and community resilience (Holling, 1973) is to study their dynamics through long-term manipulations. The article reviews long-term experiments and impacts on marine communities and ecosystems in which humans are one of the key ecological factors (Castilla, 1999).

## II. HUMAN IMPACTS ON MARINE COMMUNITIES AND THE EFFECTS ON SPECIES DIVERSITY AND FUNCTIONING

### A. Rocky Intertidal Communities

Castilla (1999), based on a 16-year intertidal human exclusion experiment in central Chile (Las Cruces fenced Marine Coastal Preserve; ECIM), summarized the ecological roles played by humans as top predators on rocky mid-intertidal marine communities. The functional intertidal food web, without humans (inside the ECIM preserve) and with humans (outside ECIM), differed substantially. On these rocky shores the impact of intertidal food gatherers is significant (Durán *et al.*, 1986). The collectors target mainly the keystone muricid snail *Concholepas concholepas*, locally known as "loco" (Castilla *et al.*, 1998). The high density of locos inside ECIM, following its closure to collectors in 1982, resulted in strong loco predation on the competitive dominant mussel *Perumytilus purpuratus*, which cannot "escape in size" from its predator. Therefore, a few years after the fencing of ECIM, the original dense mid-intertidal mussel beds inside ECIM were almost completely eliminated by the locos (Castilla, 1999). The primary space, so liberated, was readily invaded by two species of barnacles, *Jehlius cirratus* and *Notochthamalus scabrosus*, and several species of algae. Despite the fact that the loco also consumes barnacles, they have persisted for several years since they have a "weed recruitment strategy" (Castilla, 1988): After removal they keep reinvading the shore. This is not the case for *P. purpuratus*, which requires special substratum conditions to reinvade the shore (Navarrete and Castilla, 1990). Following the closure of the rocky shore at ECIM, species richness and evenness of sessile organisms using primary substrata increased inside ECIM. Outside ECIM (control), under reduced loco density due to food gathering, primary space is still dominated almost exclusively by the competitive dominant mussel *P. purpuratus*, and the biological diversity of the sessile primary substrata users is reduced since the mussels are long-term winners and appropriate the rock resource (Fig. 1). Castilla (1999) provided a detailed account of direct and indirect human impacts on these communities and discussed differences in their functioning. For instance, it was noted that the settlement of keyhole limpets, *Fissurella* spp., was indirectly negatively impacted inside ECIM since their recruitment substratum, the beds

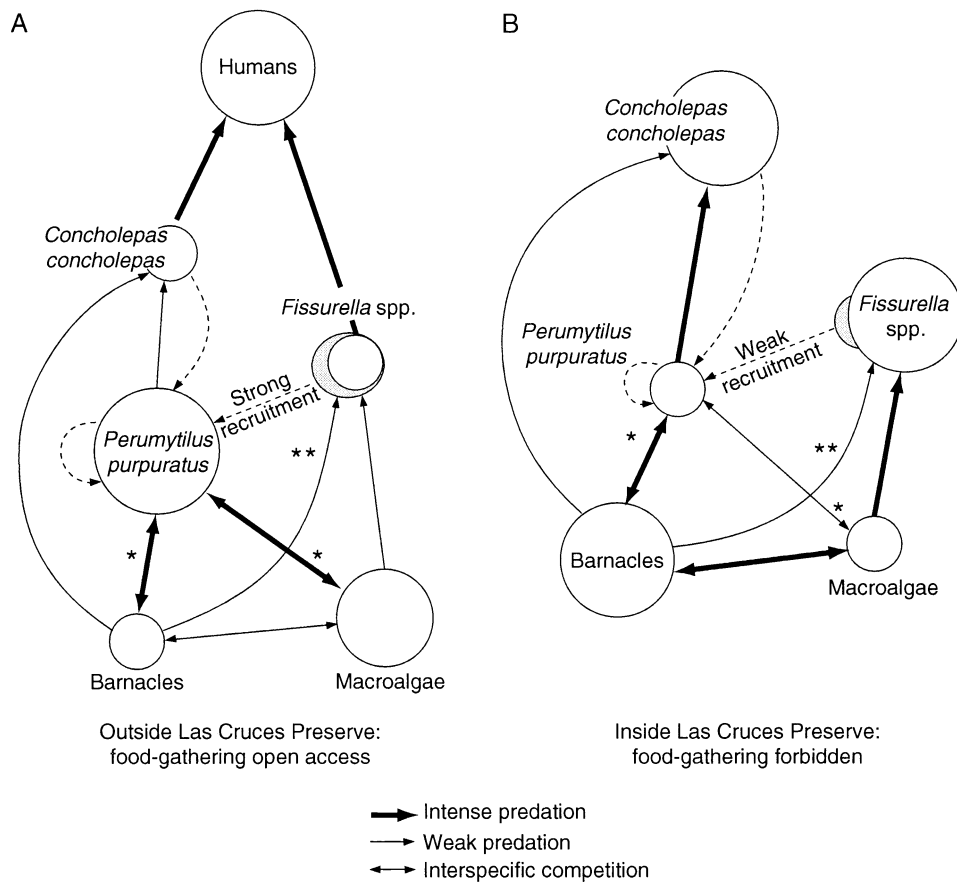


FIGURE 1 Diagrammatic representation of rocky intertidal food webs and human impacts outside (A) and inside (B) the Las Cruces (ECIM) Marine Preserve, central Chile. The size of the circles represents the approximate density of populations. An arrow with a single asterisk indicates predation. The point of the arrow shows the flow of energy and the width indicates strong (wide) or weak (narrow) interactions. A double-asterisk arrow represents interspecific competition and the point width indicates the long-term competitive dominant (wider) or subordinate (narrower) species (intraspecific interactions are not considered). Settlement is shown by dashed lines and the arrows on these lines show the settler facilitator. One asterisk indicates that barnacles and macroalgae, apart from their ability to settle directly on rock, settle on top of mussel shells. A double asterisk indicates keyhole limpet browsing on young barnacles. *Concholepas concholepas* is a carnivore muricid. *Fissurella* spp. are herbivore gastropods (reprinted from Castilla, Rocky intertidal food webs and human impacts © 1999, p. 281, with permission of Elsevier Science).

of the mussels *P. purpuratus*, were absent due to loco's direct predatory impacts (Fig. 1).

Nevertheless, in the papers previously noted, no mention was made that rocky intertidal species diversity should be viewed in a more comprehensive way so as to include the secondary substrata generated by *P. purpuratus*, an ecosystem engineer species (Jones *et al.*, 1994). Mussel matrices allow for the establishment of a rich macroinvertebrate and algal community composed of dozens of species (Paredes and Tarazona, 1980; Lohse, 1993) which live inside the matrices and on

mussel shells. Although in central Chile this effect has not been evaluated, the *P. purpuratus* matrices enhance species richness (for southern Chile, see López and Osorio, 1977) in sites impacted by humans (outside ECIM) compared to those not impacted (inside ECIM, J. Castilla, unpublished results).

Similar ecological direct and indirect human impacts and drastic modification in rocky intertidal species evenness and intertidal community functioning (Fig. 2) have been reported at Mehuín's southern Chile coastal preserve (Moreno *et al.*, 1984). Lindberg *et al.*

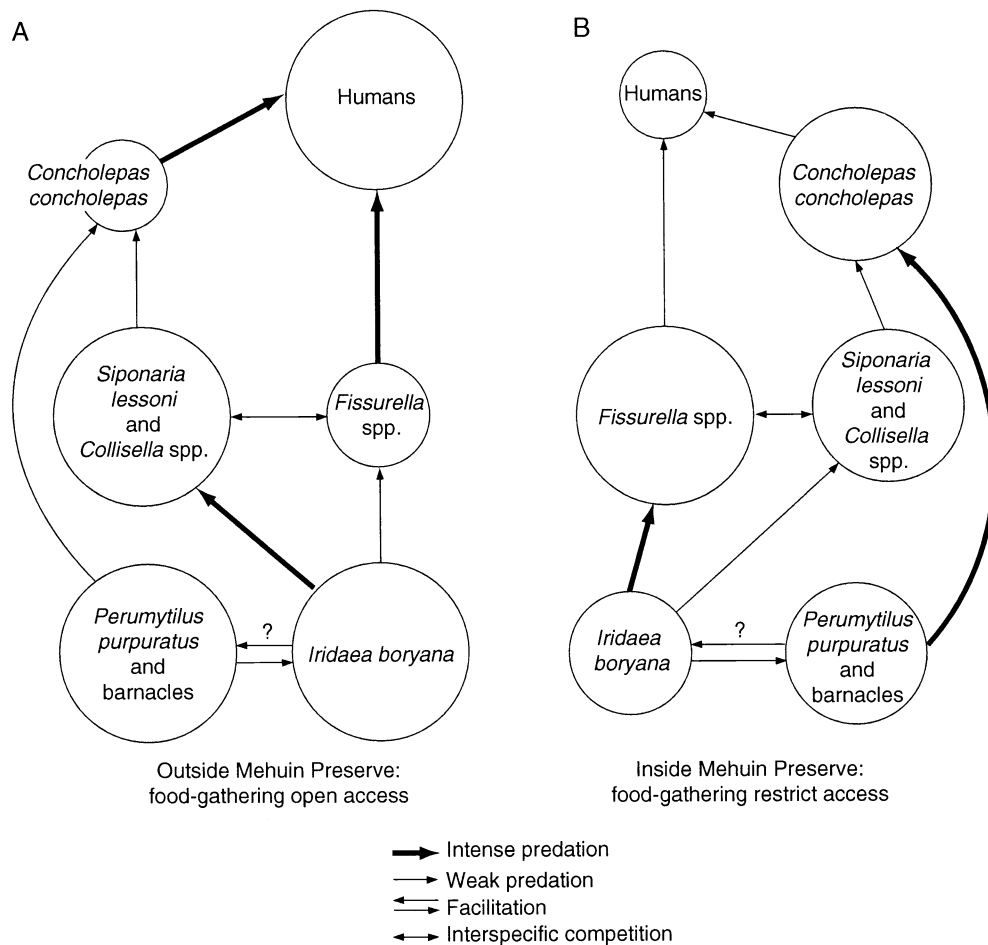


FIGURE 2 Diagrammatic representation of rocky intertidal food webs and human impacts outside (A) and inside (B) the Mehuin's Marine Preserve, southern Chile. Symbols are as described in the legend to Fig. 1 (reproduced with permission from Moreno, 1986).

(1998), through manipulative and "natural" experiments, described a three-trophic-level interaction among the American black oystercatcher (*Haematopus bachmani*), limpets (*Lottia spp.*), and erect fleshy algae in rocky intertidal bench communities of central and southern California. Human disturbances, such as the selective collection of large-size limpets and the reduction of shorebirds (in shores frequented by humans), drive the communities to a state dominated by small limpets and high cover of fleshy algae. Intertidal benches in relatively isolated islands (e.g., San Nicolas in central California) with large densities of oystercatchers and an absence of limpet human collection present communities in a different alternative state, which is characterized by large-size limpet populations and comparatively reduced fleshy algal cover.

## B. Rocky Subtidal Communities

The Cape rock lobster *Jasus lalandii*, commercially the most important lobster species in South Africa, causes profound direct and indirect effects on subtidal competitive dominant mussel species, such as *Choromytilus meridionalis* and *Aulacomya ater* (Griffiths and Seiderer, 1980), severely modifying species diversity and community functioning. Barkai and Branch (1988a, b) compared the nearshore benthic communities of two adjacent islands on the west coast of South Africa: Malgas and Marcus Islands (33°S, 18°E), which are approximately 4 km apart. The biotas of both islands have been protected from human exploitation since 1929. In the 1960s both islands supported populations of rock lobsters, but later, due to overfishing, a management plan

was established which included a catch quota. Currently, Malgas still supports an unusually dense population of *J. lalandii* (probably partly due to the management plan) with densities of up to 10 individuals per square meter, whereas Marcus has a very reduced adult population of lobster. The benthic communities of both islands have only 34% of species in common. The biota of Malgas is dominated by numerous species of algae, whereas that of Marcus consists of thick beds of the black mussel *C. meridionalis*, an autogenic ecosystem engineer species that has a rich and diverse associated fauna (Barkai and Branch, 1988a). At Malgas, the predatory lobsters have eliminated a large proportion of spatial competitors, including mussels and barnacles, and sea urchins are absent. As a consequence, macroalgae proliferated. At Marcus, due to the absence of lobsters, the competitive dominant *C. meridionalis* formed dense beds, outcompeting other species of mussels, such as *A. ater* and algae; sea urchins are common (Castilla *et al.*, 1994). Barkai and Branch (1988a, b) discussed this ecological situation and argued for the

existence of alternative stable states on the contrasting islands. Figure 3 provides a summary of the main species involved, relative biomass, and direct, indirect, positive, and negative interactions between organisms on both islands.

The ecological impact of the Cape rock lobster at Malgas was experimentally demonstrated by Barkai and McQuaid (1988). The experiments showed that the drastic community differences between the islands were due to the dense population of lobster at Malgas and its absence at Marcus. In fact, the introduction of 1000 lobsters at Marcus ended amazingly: The lobsters were attacked by thousand of snails, *Burnupena* sp., which exist at Marcus in densities of up to 250 per square meter, and the lobsters perished within 30 min. This may explain their absence at Marcus, supporting the existence of an alternative ecological state.

In South African waters, it is unknown to what extent the commercial exploitation of rock lobsters or conservation measures (i.e., coastal closures) have impacted the nearshore rocky subtidal communities or in how

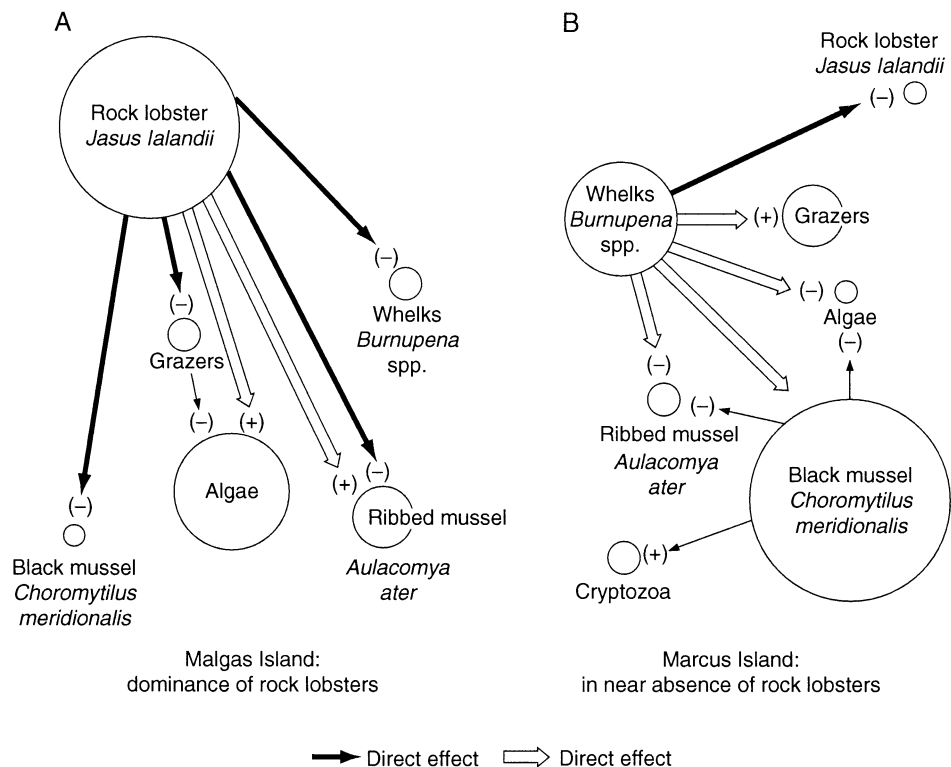


FIGURE 3 Rock lobster direct (+) and indirect (-) effects on mussel, welk, and grazer (sea urchin) preys in two South African islands. (A) Malgas, with a high density of adult lobsters. (B) Marcus, with a virtual absence of lobsters. The circles indicate relative biomasses (reproduced with permission from Castilla *et al.*, 1994).



many cases (other than Marcus and Malgas Islands) alternative stable states have been reached. This is a classical example in which both extreme attitudes—overexploitation and total conservation (no-take areas)—can result in drastically different species diversity and community functioning, mediated by the role of a high-trophic-level predator.

### C. Humans and Linkages between Coastal and Oceanic Waters

*Enhydra lutris*, the northern sea otter, is found in near-shore environments ranging across the Pacific rim from Hokkaido (Japan) to Baja California (Mexico). The exploitation of their pelts led to the near extinction of otter populations in approximately 1911, when unregulated hunting was ended. Since then, the recovery of otter populations has occurred, particularly in the Aleutian Island chain, where by the 1970s the populations reached near maximum densities in some areas, were growing rapidly in others, and remained absent from others. Otters as keystone species (Power *et al.*, 1996) control the local biomass and the abundance of sea urchins, which regulate benthic algae biomass and productivity. Aleutian interisland comparisons (Estes *et al.*, 1998) have shown that kelp deforestation occurred in islands with low sea otter densities due to the increased density of sea urchins, whereas islands with high sea otter densities showed high kelp biomass. Estes *et al.* reported the complete transformation of a subtidal kelp forest in islands of the Aleutian Archipelago from three to four trophic-level systems and the release of sea urchin populations from the limiting influence of their predator, *E. lutris*. In the original circumstances, in the absence of sea otters, sea urchin populations increased rapidly and overgrazed the kelp forest, setting in motion a suite of different ecological impacts which drastically transformed the coastal ecosystems. These transformations had implications in the functioning of the communities and affected species diversity. Humans are highly involved in Estes *et al.*'s findings. In recent years in western Alaska, declines of *E. lutris* populations have been observed. The authors have argued that this is probably due to the recent increased predation on sea otters by killer whales, *Orcinus orca*. *Orcinus* may have initiated predatory influences that cascaded down successively lower trophic levels, first through the reduction of densities of sea otters, which triggered the increase of sea urchin populations, and ultimately the depletion of kelp biomass due to overgrazing. Estes *et*

*al.*'s paper includes documented information on declines of sea otter populations and increases in the density and intensity of grazing of sea urchins on the kelp beds. Sea otters and killer whales have coinhabited the Aleutian Archipelago for millennia and Estes *et al.* attributed the sudden change of behavior of killer whales to a shift in their prey resource base. This has probably resulted from the collapse of pinniped populations, such as the Stellar sea lion and harbor seals, which were among the killer whale's main prey items. It has been suggested that the pinniped populations may have collapsed due to the Northwest Pacific midwater-trawl overfishing of walleye pollock (*Theragra chalcogramma*) (Alverson *et al.*, 1994) and/or increases in the ocean temperature. The authors recognized that some of their arguments contained speculations and that the critical one refers to the direct/indirect impacts of humans on marine ecosystems. In fact, sea otters, pinnipeds, and whales are under national and international protection in the Aleutians through different treaties and agreements signed dozens of years ago, but it also has to be recognized that their food resources have been depleted independently through overfishing. For instance, there is evidence that in the case of the pinnipeds a reduction (population collapses in some cases) has occurred mostly due to overfishing of pinnipeds, or of their fish resources, and also to climate changes. Overfishing is directly linked to human activities, and in Estes *et al.*'s scenario, humans and not killer whales may be considered as the apex predator. Humans have redirected the functioning of oceanic and coastal marine ecosystems in these localities and modified trophic linkages.

These examples indicate that there are at least two aspects of human ecological influences on marine communities that are difficult to evaluate and hence demonstrate an indisputable cause-effect situation. First, in many cases, the functioning of the marine communities is affected indirectly by anthropogenic activities—for example, human overfishing of pinniped's fish resources, collapse of pinniped populations, a shift in the prey item of killer whales, predation on the sea otter, population explosion of sea urchins, and overgrazing of kelp beds. The cascading down to successively lower trophic levels is complex and requires long-term observation and experiments to be understood. Furthermore, nonanthropogenic impacts also need to be considered. Second, limited knowledge exists on the resilience properties of marine communities and ecological conclusions on linkages between marine ecosystems are based on preliminary data.

## D. Humans and Ecosystem Engineer and Invasive Species

Ecosystem engineer species are species that directly or indirectly modulate the availability of resources (other than themselves) to other species by causing physical state changes in biotic or abiotic materials, and in so doing they modify, maintain, and/or create habitats (Jones *et al.*, 1994). Jones *et al.* distinguished (i) autogenic engineers, when the changes in the environment occurred via their own physical structure, living or dead tissues (e.g., coral reefs), and (ii) allogenic engineers, when they produced changes in the environment through the transformation of living or nonliving materials from one physical state to another via mechanical means (e.g., rabbits and burrows). In marine coastal communities, there are numerous autogenetic engineer species playing roles in the functioning of the community and ecosystem and creating the physical conditions for other species to exist (e.g., mussels; Lohse, 1993). In the Southern Hemisphere, rocky littoral zone tunicates of the genus *Pyura* play such a role (see Fielding *et al.*, 1994, for *P. stolonifera* in S. Africa). These tunicates are also important as species extracted for food and/or bait by recreational fishers, divers, and intertidal food gatherers (for *Pyura praeputialis* in Australia, see Fairweather, 1991; for *P. praeputialis* in Antofagasta, northern Chile, see Castilla, 1998). The tunicates form dense intertidal and subtidal belt monocultures and attain collective cemented beds, creating microhabitats used by several dozen macroinvertebrates and algae. Fielding *et al.* identified 83 taxa of macroinvertebrates in intertidal and subtidal *Pyura stolonifera* beds around Durban, South Africa, whereas more than 100 taxa of macroinvertebrates and algae have been found in intertidal *P. praeputialis* beds in Antofagasta.

The *P. praeputialis* beds in Chile present a very restricted geographical distribution of only 60–70 km around Antofagasta Bay (Clarke *et al.*, 1999). According to a working hypothesis (J. Castilla, work in progress), the species might have been introduced recently to Antofagasta by ships or arrived on floating objects from Australia. In Antofagasta, a contrasting situation concerning species richness is found in sites with *P. praeputialis*, with more than 100 taxa in the *Pyura* beds, as opposed to sites without the tunicate, which have about one-third to one-fourth of the species. It is unknown how much ecological damage, if any, human extraction causes on the dynamics of *Pyura* populations or on species diversity at a local scale. However, preliminary information at Antofagasta indicates that following

*Pyura* removals by waves, predators, or humans, the species reinvades intertidal sites (the center of its distribution) within 1 year (J. Castilla, work in progress). A higher rate of anthropogenic or nonanthropogenic removal of engineer species than the rate of recovery may be key to local species diversity.

Invasive species are displacing native species throughout the world. They are altering the physical nature of habitats (e.g., the effects of the Asian clam *Potamocorbula amurensis* in the San Francisco Bay) and causing changes in food webs of economically important species (NRC, 1999). The best reported case is that of the Bay of San Francisco, in which ship activities (i.e., the elimination of ballast waters) have increased drastically the number of exotic species in the bay's benthic communities (Carlton, 1996). At the pelagic level the introduction in the bay of the zooplanktonic mysid *Acanthomysis* sp., which displaced another species of mysid, *Neomysis mercedis*, a major food item of the striped bass *Morone saxatilis*, is partly responsible for a severe decline in the bay's bass population (NRC, 1999). Furthermore, there are recent reports showing that the predator green crab *Carcinus maenas* has invaded the San Francisco Bay and is spreading through the coastal waters of California (Cohen and Carlton, 1998).

## E. Mariculture

The intensive and extensive marine farming of fish, shellfish, and algae has a long history and is a controversial issue. For instance, mariculture production expectations have not been achieved (NRC, 1999) and adverse environmental effects, such as contamination of surface waters by fish wastes, eutrophication, spread of diseases, introduction of unwanted species, and deterioration of coastal habitats (e.g., mangroves in connection with shrimp farming in Asia and Latin America), have occurred (Chamberlain, 1997; Anderson, 1997). The introduction of exotic cultured species may be a serious and irreversible event to native ecosystems which merits careful consideration. For instance, oysters have been transported by man from country to country and there are several cases of the concomitant spread of pests (unwanted species) and diseases, even under strict import controls. The introduction of the American oyster *Cassostrea virginica* into English waters (late 1800s and 1939) brought in several exotic species, the worst being the American oyster drill *Urosalpinx cinerea* and the gastropod competitor *Crepidula fornicata* (Edwards, 1990). Critical epizootic disease events in the Gulf of

St. Lawrence that caused serious oyster stock depletions were ascribed to the transplant of oysters in 1914 from New England to Canada (Edwards, 1990). No comprehensive ecological reports on the ecological effects of these species introductions and diseases on local species diversity or community functioning have been published.

The intensive farm-raising of high-value species, such as shrimp and salmon, is far from trouble-free. There are concerns about the increase in the deposition of particulates and accumulation of organic matter under salmon cages in intensive mariculture installations due to unwanted effects, such as anoxic conditions and the production of toxic gases (Beveridge, 1996). Coastal ecosystem destruction, nutrient loading, antibiotics wastes, accidental release of alien or genetically altered organisms, and disease spreading to native species are some of the threats to community and ecosystem functioning.

### F. Human Overfishing, Diseases, and Trophic Cascades

Hughes (1994) and Jackson (1997) reported major ecological effects on coral reef communities as a consequence of the overexploitation of herbivorous fishes and a disease killing sea urchins. In Caribbean coral reefs, a chain of effects, starting with the overfishing of herbivorous fishes, appeared following category 5 hurricane Allen in 1980. Allen severely damaged coral reefs in Jamaica, but by 1983 there was evidence of their recuperation. Nevertheless, at that time a disease devastated the herbivorous populations of the sea urchin *Diadema antillarum*. The elimination of the herbivore guild caused dramatic food cascading effects, resulting in reefs overgrown by algae and the detention of their recuperation. Species diversity and community functioning severely changed: The coral cover was reduced from approximately 52% in 1977 to 3% in the early 1990s, and cover of macroalgae increased from approximately 3 to 92% (Hughes, 1994).

### G. Pollution and Artificial Reefs

The cases exemplified are among the best known ecological situations in which human impacts and the function of communities or ecosystems, combined with changes in species diversity, have been observed or studied. However, there are additional examples show-

ing anthropogenic negative, as well as positive, impacts on marine communities and ecosystems. Among negative impacts on marine species diversity and community functioning, the most conspicuous (not discussed here) is pollution (Castilla, 1996). Among positive impacts is the building of marine reefs for fishing enhancement and recreational purposes. Artificial habitats may locally enhance species diversity and resources and drive community structure toward alternative states (Buckley, 1982).

## III. NONANTHROPOGENIC ENVIRONMENTAL CHANGES AND VARIABILITY

Nonanthropogenic environmental changes and impacts on marine populations and communities have been well documented. For instance, Soutar and Isaacs (1974) reported large fluctuations in the density of scales of hake, anchovy, and sardines in sediment cores during the past 2000 years, well before fishing was a factor. Large-scale ocean climate changes, such as El Niño Southern Oscillation (ENSO) events, have dramatic negative (Arntz and Fahrbach, 1996) or positive (Castilla and Camus, 1992) impacts on fish, shellfish, and algae populations in the Southeastern Pacific. ENSO also causes multiple positive and negative oceanic, freshwater, and terrestrial impacts throughout the world.

Barry *et al.* (1995) reported changes between 1931 and 1933 and between 1994 and 1995 in species richness and evenness of intertidal invertebrates at a rocky intertidal transect in the Hopkins Marine Station, Monterey, California. They reported species' latitudinal range shifting northward, suggesting a consistency with predictions associated with anthropogenic-linked climate warming (but see alternative explanation by Denny and Paine, 1998). Nevertheless, it is debatable whether the current global warming trend, due partly to the build-up of several greenhouse gases, is part of a long-term climatic trend. In any case, marine species with different geographical origins would have different responses to water temperature alterations (Castilla and Camus, 1992). Moreover, in the case of the oceans, water temperature modifications would be just one of the potential factors affecting the distribution of species. For instance, temperature effects on the turbulence of the ocean waters, and their association with wind stress, may

have major implications for plankton dispersal. Also, the predicted north–south interhemispheric asymmetry, due to the thermal inertia in the south, must be considered before drawing firm conclusions on marine species latitudinal shifts (Bernal, 1994). Furthermore, since the ocean is affected simultaneously by several climate forces (including anthropogenic greenhouse effects), it is difficult to determine the real cause of any observed change, such as that in surface seawater temperature. Shifts in marine populations, community structure, and their functioning represent the integrated response of species assemblages to nonanthropogenic long-term climate changes superimposed on the effects of numerous short-term factors, including anthropogenic forcing.

#### IV. CONCLUSIONS

This article discussed several marine examples in which direct anthropogenic and nonanthropogenic impacts (or combinations), such as species extraction and oceanic water temperature modifications, caused drastic ecological shifts on marine benthic intertidal, subtidal, and coastal–oceanic communities, and thereby modified species diversity and the functioning of associated communities. Interestingly, extreme conservation measures (e.g., the establishment of no-take areas) to protect species, habitat, community, or ecosystem may also cause drastic modifications in the functioning of marine communities and drive communities into alternative ecological states (Castilla *et al.*, 1994; Estes *et al.*, 1998; Castilla, 1999). This article highlighted that anthropogenic activities (e.g., mariculture) and impacts (e.g., overfishing) on different ecological categories of species (predator, keystone, engineer, invasive, and competitive dominant) translate into differential responses and functioning at the species diversity and community level. The unique ecological role played by humans and their apex keystone position in trophic webs were discussed.

#### Acknowledgments

I acknowledge financial support received through the Cátedra Presidencial en Ciencias, Chile (awarded in 1997) and the Pew Charitable Fund as a Pew Fellow in Marine Conservation (1997). Minera Escondida Ltd. supported fieldwork in Antofagasta, Chile. Comments and suggestions on a preliminary manuscript made by a referee are acknowledged.

#### See Also the Following Articles

AQUACULTURE • INTERTIDAL ECOSYSTEMS • MARINE AND AQUATIC COMMUNITIES, STRESS FROM EUTROPHICATION • MARINE ECOSYSTEMS • RESOURCE EXPLOITATION, FISHERIES

#### Bibliography

- Alverson, D. L., Freeberg, M. H., Murawski, S. A., and Pope, J. G. (1994). A global assessment of fisheries bycatch and discards, FAO Fisheries Technical Paper No. 339. Food and Agriculture Organization, Rome.
- Anderson, J. L. (1997). The growth of salmon aquaculture and the emerging new world order of the salmon industry. In *Global Trends in Fisheries Management* (E. Pikitch, D. D. Huppert, and M. Sissenwine, Eds.), American Fisheries Society Symposium No. 20, pp. 175–184. American Fisheries Society, Bethesda, MD.
- Arntz, W. E., and Fahrback, E. (1996). *El Niño: Experimento Climático de la Naturaleza*. Fondo de Cultura Económica, México D. F., México.
- Barkai, A., and Branch, G. M. (1988a). Contrast between the benthic communities of subtidal hard substrata at Marcus and Malgas Islands: A case of alternative stable states? *S. Afr. J. Mar. Sci.* 7, 117–137.
- Barkai, A., and Branch, G. M. (1988b). The influence of predation and substrata complexity on recruitment to settlement plates: A test of the theory of alternative states. *J. Exp. Mar. Biol. Ecol.* 124, 215–237.
- Barkai, A., and McQuaid, C. D. (1988). Predator–prey role reversal in a marine benthic ecosystem. *Science* 242, 62–64.
- Barry, J. P., Baxter, R. D., Sagarin, R. D., and Gilman, S. E. (1995). Climate-related long-term faunal changes in a California rocky intertidal community. *Science* 267, 672–675.
- Bernal, P. A. (1994). Global climate change in the ocean: A review. In *Earth System Response to Global Change: Contrast between North and South America* (H. A. Mooney, E. R. Fuentes, and B. Kronberg, Eds.), pp. 1–15. Academic Press, San Diego.
- Beveridge, M. C. M. (1996). *Cage Aquaculture*. Fishing News Book, Oxford.
- Buckley, R. M. (1982). Marine habitat enhancement and urban recreational fishing in Washington. *Mar. Fish. Rev.* 44, 28–37.
- Carlton, J. T. (1996). Biological invasions and cryptogenic species. *Ecology* 77, 1653–1655.
- Castilla, J. C. (1988). Earthquake-caused coastal uplift and its effects on rocky intertidal kelp communities. *Science* 242, 440–443.
- Castilla, J. C. (1996). Copper mine tailing disposal in northern Chile rocky shores: *Enteromorpha compressa* (Chlorophyta) as a sentinel species. *Environ. Monit. Assess.* 40, 171–184.
- Castilla, J. C. (1998). Las comunidades intermareales de la Bahía San Jorge: Estudios de Línea Base y el Programa Ambiental de Minera Escondida Ltda. en Punta Coloso. In *Minería del Cobre, Ecología y Ambiente Costero* (D. Arcos, Ed.), pp. 221–224. Editora Anibal Pinto S. A., Talcahuano, Chile.
- Castilla, J. C. (1999). Coastal marine communities: Trends and perspectives from human-exclusion experiments. *TREE* 14, 28–83.
- Castilla, J. C., and Camus, P. A. (1992). The Humboldt-El Niño scenario: Coastal benthic resources and anthropogenic influences, with particular reference to the 1982/82 ENSO. *S. Afr. J. Mar. Sci.* 12, 703–712.
- Castilla, J. C., Branch, G. M., and Barkai, A. (1994). Exploitation of

- two critical predators: The gastropod *Concholepas concholepas* and the rock lobster *Jasus lalandii*. In *Rocky Shores: Exploitation in Chile and South Africa* (W. R. Siegfried, Ed.), Ecological Studies Vol. 103, pp. 101–130. Springer-Verlag, Berlin.
- Castilla, J. C., Manríquez, P., Alvarado, J., Rosson, A., Pino, C., Espoz, C., Soto, R., Oliva, D., and Defeo, O. (1998). Artisanal "Caletas" as units of production and co-managers of benthic invertebrates in Chile. In *Proceedings of the North Pacific Symposium on Invertebrates Stock Assessment and Management* (G. S. Jameison and A. Campbell, Eds.), Can. Spec. Publ. Fish. Aquat. Sci., pp. 407–413.
- Chamberlain, G. (1997). Sustainability of world shrimp farming. In *Global Trends in Fishery Management* (E. Pikitch, D. D. Huppert, and M. Sissenwine, Eds.), American Fisheries Society Symposium No. 20, pp. 195–209. American Fisheries Society, Bethesda, MD.
- Clarke, M., Ortiz, V., and Castilla, J. C. (1999). Does early development of the Chilean tunicate *Pyura praeputialis* (Heller, 1878) explain the restricted distribution on the species? *Mar. Sci. Bull.* 65, 745–754.
- Cohen, A., and Carlton, J. T. (1998). Accelerating invasions rate in a highly invaded estuary. *Science* 279, 555–558.
- Denny, M. W., and Paine, R. T. (1998). Celestial mechanics, sea level changes, and intertidal ecology. *Biol. Bull.* 194, 108–115.
- Durán, L. R., Castilla, J. C., and Oliva, D. (1986). Intensity of human predation on rocky shores at Las Cruces, central Chile. *Environ. Conserv.* 14, 143–149.
- Edwards, E. (1990). Be careful with introductions. *Fish Farming Int.* 17.
- Estes, J. A., Tinker, M. T., Williams, T. M., and Doak, D. F. (1998). Killer whale predation on sea otters linking oceanic and nearshore ecosystems. *Science* 282, 473–476.
- Fairweather, P. G. (1991). A conceptual framework for ecological studies of coastal resources: An example of a tunicate collected for bait on Australian seashores. *Ocean Shoreline Manag.* 15, 5–142.
- Fielding, P. J., Weeters, K. A., and Forbes, A. T. (1994). Macroinvertebrate communities associated with intertidal and subtidal beds of *Pyura stolonifera* (Heller) (Tunicata: Ascidiacea) on the Natal coast. *S.-Afr. Tydskr. Dierk.* 29, 46–51.
- Griffiths, C. L., and Seiderer, J. L. (1980). Rock lobsters and mussels limitations and preferences in a predator–prey interaction. *J. Exp. Mar. Biol. Ecol.* 44, 95–109.
- Heywood, V. H. (1995). *Global Biodiversity Assessment*, United Nations Environmental Programme. Cambridge Univ. Press, Cambridge, UK.
- Holling, C. S. (1973). Resilience and stability of ecological systems. *Annu. Rev. Ecol. Syst.* 4, 1–24.
- Hughes, T. P. (1994). Catastrophes, phase shifts, and large-scale degradation of a Caribbean coral reef. *Science* 265, 1547–1551.
- Jackson, J. J. (1997). Reefs since Columbus. *Coral Reefs* 16(Suppl.), S23–S32.
- Jones, C. G., Lawton, J. H., and Schachak, M. (1994). Organisms as ecosystems engineers. *Oikos* 69, 373–386.
- Lindberg, D. R., Estes, J. A., and Warheit, K. I. (1998). Human influences on trophic cascades along rocky shores. *Ecol. Appl.* 8, 880–890.
- Lohse, D. P. (1993). The importance of secondary substratum in a rocky intertidal community. *J. Exp. Mar. Biol. Ecol.* 166, 1–17.
- López, M. T., and Osorio, C. (1977). Diversidad biológica en una comunidad intermareal de Putemún, Chiloé. *Bol. Soc. Biol. Concepción* 51, 123–127.
- Moreno, C. A. (1986). Un resumen de las consecuencias ecológicas de la exclusión del hombre en la zona intermareal de Mehuín-Chile. *Estud. Oceanol.* 5, 59–66.
- Moreno, C. A., Sutherland, J. P., and Jara, F. H. (1984). Man as a predator in the intertidal zone of southern Chile. *Oikos* 42, 155–160.
- National Research Council (1999). *Sustaining Marine Fisheries*. National Academy Press, Washington, D.C.
- Navarrete, S. A., and Castilla, J. C. (1990). Barnacle walls as mediators of intertidal mussel recruitment: effects of patch size on the utilization of space. *Mar. Ecol. Prog. Ser.* 68, 113–119.
- Paine, R. T., Tegner, M. J., and Johnson, E. A. (1998). Compounded perturbations yield ecological surprises. *Ecosystems* 1, 535–545.
- Paredes, C., and Tarazona, J. (1980). Las comunidades de mitilidos del mediolitoral rocoso del Departamento de Lima. *Rev. Peruana Biol.* 2, 59–73.
- Power, M. E., Tilman, D., Estes, J. A., Menge, B., Bond, W. J., Scott-Mills, L., Daily, G., Castilla, J. C., Lubchenco, J., and Paine, R. T. (1996). Challenges in the quest for keystones. *BioScience* 46, 609–620.
- Soutar, A., and Isaacs, J. D. (1974). Abundance of pelagic fish during the 19th and 20th centuries as recorded in anaerobic sediments off California. *Fish. Bull. U.S. Fish Wildlife Service* 72, 257–275.
- White, P. S., and Pickett, S. T. A. (1985). Natural disturbance and patch dynamics: An introduction. In *The Ecology of Natural Disturbance and Patch Dynamics* (S. T. A. Pickett and P. S. White, Eds.), pp. 3–13. Academic Press, New York.



# MARINE MAMMALS, EXTINCTIONS OF

Glenn R. VanBlaricom,\* Leah R. Gerber,† and  
Robert L. Brownell, Jr.‡

\*U.S. Geological Survey and University of Washington, †University of California,  
Santa Barbara, and ‡National Marine Fisheries Service

- I. Introduction
  - II. Patterns and Case Studies of Extinction in  
Marine Mammals
  - III. Discussion
- 

## I. INTRODUCTION

### A. Taxonomic Definition of “Marine Mammals”

The marine mammals include one extinct order and three major extant taxa that were or are fully aquatic, in most cases occurring entirely in the marine habitats of the major ocean basins and associated coastal seas and estuaries. In addition, a few species of largely terrestrial taxa are currently regarded as marine mammals. We consider 127 recent mammal species in total to be marine mammals for purposes of this review. We acknowledge that species numbers within any taxon are subject to revision as new systematic methods and philosophies emerge. Our primary bases for defining our list of marine mammal species are the protocols of the U.S. federal government, determined largely by the U.S. Marine Mammal Protection Act (MMPA) of 1972 [16 U.S.C. §§1361-62, 1371-84, and 1401-07 (Supp. IV 1974)] as amended (MMPA) and managed by two U.S. federal agencies, the National Marine Fisheries Service (NMFS) and the Fish and Wildlife Service (FWS). Our choice of defining criteria is arbitrary. Our

principal source for taxonomic nomenclature, including common names, is the recent review of Rice (1998).

The order Cetacea includes whales, dolphins, and porpoises (Table I). The “pinnipedia” is a group of species in three families in the mammalian order Carnivora (Table I). The pinnipeds include the seals, fur seals, sea lions, and walrus. The term pinnipedia is no longer recognized formally by marine mammal taxonomists, but it continues to appear in the systematic vernacular as a matter of tradition and convenience. The order Sirenia includes the extant manatees and dugong and the extinct Steller’s sea cow (Table I). The order Desmostylia is the only recognized order of marine mammals to become entirely extinct.

Two largely terrestrial families of the order Carnivora also include species recognized as marine mammals (Table I). Sea otters and chungungos (family Mustelidae) live entirely or primarily in marine habitats. Polar bears (family Ursidae) also spend a significant proportion of time at sea.

Many other species of mammal utilize aquatic or marine habitats, including monotremes, ursids, mustelids, canids, primates, rodents, bats, and ungulates. Ultimately, the distinction among aquatic, marine, and terrestrial taxa is arbitrary. Thus, our reliance on definitions and protocols of MMPA, NMFS, and FWS is subjective, although it is consistent with common practice at least in the United States.

We use the term “marine” to refer to large, contiguous aqueous habitats with significant dissolved salt content in ambient waters. Thus, we apply the term marine

TABLE I

Major Taxa and Species Numbers of Marine Mammals<sup>a</sup>

Taxon	No. of species
Cetacea: Whales	83
Mysticeti: Baleen whales	12
Balaenidae: Right whales	2
Neobalaenidae: Pygmy right whale	1
Eschrichtidae: Gray whale	1
Balaenopteridae: Rorquals	8
Odontoceti: Toothed whales	71
Physeteridae: Sperm whales	1
Kogiidae: Pygmy sperm whales	2
Ziphiidae: Beaked whales	20
Platanistidae: Indian river dolphin	1
Iniidae: Amazon river dolphin	1
Lipotidae: Chinese river dolphin	1
Pontoporiidae: La Plata dolphin	1
Monodontidae: Beluga and narwhal	2
Delphinidae: Dolphins	36
Phocoenidae: Porpoises	6
Carnivora, "Pinnipedia"	36
Otariidae: Sea lions and fur seals	16
Odobenidae: Walrus	1
Phocidae: Seals	19
Carnivora, other marine taxa	3
Mustelidae: Marine otters	2
Ursidae: Polar bear	1
Sirenia: Manatees, dugongs, and sea cows	5
Trichechidae: Manatees	3
Dugongidae: Dugong and Steller's sea cow	2
Total species	127

<sup>a</sup> Following the conventions of Rice (1998).

to the world's oceans, seas, and estuaries. We apply the term "aquatic" to aqueous habitats without significant measurable dissolved salt concentrations in ambient waters, such as lakes and rivers above the elevation of significant mixing with marine waters, and to inland saline lakes that lack outlet streams connecting to marine habitats. "Terrestrial" habitats are those lacking standing water in normal circumstances. As indicated in Table I, our list of "marine mammals" includes marine and aquatic species.

### 1. General Features and Habitat Boundaries

Compared to terrestrial mammals, marine mammals are characterized by many striking modifications in anatomy, physiology, and ecology (Table II). In some cases, the modifications are sufficiently extreme that phylogenetic linkages to terrestrial ancestry are obscured and difficult to resolve. The degree of modification is corre-

TABLE II

Distinguishing Characteristics of the Major Marine Mammal Taxa

Characteristic	Cetacea	Sirenia	Pinnipedia
Body streamlined	x	x	x
Limbs modified	x	x	x
Rear limbs modified as flippers			x
Rear limbs and pelvic girdle absent	x	x	
Propulsion by caudal spine and flukes	x	x	
Loss of pelage	x	x	
Subcutaneous blubber layer	x	x	x
Simplification of dentition	x	x	x
Expansion of anterior skull	x		
Development of acoustic capability for communication and echolocation	x <sup>d</sup>		x <sup>d</sup>
Amphibious capability			x

<sup>d</sup> Echolocation capability is known only for the odontocete cetaceans.

lated approximately with the duration of the evolutionary history of the major marine mammal taxa.

Although marine mammals are largely defined by marked departures from the terrestrial mammalian model, it is instructive to consider major features of terrestrial mammals retained in marine mammals. In the context of extinction processes in general, and anthropogenic extinctions in particular, two retained features are of particular importance. First, although most marine mammals spend most of their lives immersed at sea, they retain largely terrestrial respiratory architecture and must surface and breathe in order to exchange respiratory gasses. Second, marine mammals are homeothermic, with core body temperatures typically near 38°C, like their terrestrial relatives. The need to breathe at the surface and the need for major anatomical adjustment to minimize rates of heat loss are constraints that foster vulnerability to unsustainable rates of exploitation and to certain types of pollution. The significance of these constraints is developed in the case studies we present later.

The diving capabilities of marine mammals define the three-dimensional nature of their habitats at sea. Nearly all extant marine mammals dive to forage, although the ranges of diving capability and pattern are broad. Most marine mammals also spend signifi-

cant time submerged while traveling, socializing, or breeding.

Among cetaceans, sperm whales and beaked whales likely dive the deepest and longest compared to other species. Sperm whales can dive to 1500 m, remaining submerged for 20 min or more. The diving behavior of beaked whales is poorly known, but there is emerging evidence that beaked whales may also routinely make repetitive dives of long duration to great depth. Baleen whales may make long deep dives during breeding season. Foraging dives of baleen whales normally are relatively shallow and brief. Many of the smaller cetaceans commonly dive for less than 10 min at a time to depths no greater than a few hundred meters.

Among pinnipeds, elephant seals (*Phocidae*) have maximum diving capabilities comparable to the sperm whales, and they are known to make remarkably long sequences of repetitive deep (to 1500 m), long (20 min or more) dives with surface intervals of only 2 or 3 min. These sequences may be maintained day and night for tens of days at a time. Many other phocid seals are thought to have similar capabilities. The sea lions and fur seals (*Otariidae*), in contrast, usually dive for only a few minutes at a time, and usually to maximum depths of a few hundred meters, although many otariids are known to be capable of continuous sequences of repetitive shallow dives of 10–12 hr or more. Walrus are known to dive as deep as 80 m, with maximum durations of 10 min.

In contrast to cetaceans and pinnipeds, sirenians are weak divers, normally remaining in shallow water (<20 m) and diving for only 2 or 3 min when active. Deeper dives (to 70 m) may occur on occasion, and dive duration can be quite long (up to 24 min) when animals are resting at the bottom. Sea otters are capable of diving to 100-m depth and remaining submerged for a maximum of approximately 5 min, although most dives are to 30 m or less and last only for 1 or 2 min. To our knowledge, there are no data available on the diving capabilities of the chungungo.

Few field observations of Steller's sea cow were made prior to extinction, but morphological analysis suggests that sea cows were unable to dive below the sea surface, surviving instead by foraging on macroalgae floating on the surface. Polar bears are able to make shallow dives but do not typically engage in the extended repetitive dive sequences typical of many marine mammals. They apparently do not forage while diving. Polar bears instead use stealth, quickness, and great strength to capture seals, their primary prey, at seal breathing holes on the ice surface. When confined in ice-bound seas

with small breathing holes, beluga whales and narwhals are also taken as food by polar bears. Thus, the extent to which the at-sea habitat of marine mammals is truly three-dimensional varies widely among the major taxa and the individual species. Within species, there is also marked ontogenetic variation in diving capability and pattern.

The marine mammals are geographically ubiquitous in the world's oceans, seas, and estuaries. Cetaceans occur in marine environments at all latitudes. For example, killer whales and minke whales may have the largest natural geographic ranges of the earth's mammals. Most of the mysticetes and some of the larger odontocetes have global ranges or are distributed antitropically. Smaller cetaceans are widely dispersed as well, although individual populations typically concentrate in regions of predictably high local biological productivity. Several species of small cetacean, including two delphinids, a phocoenid, and the four monotypic families of river dolphins, are found in major river systems in South America and Asia. Beluga whales also spend significant time in river habitats. Pinnipeds occur in all the world's major marine habitats, but most species are concentrated in middle or high latitudes, in close association with regions of high productivity. In addition, there are several pinniped species or populations confined to isolated large lakes in Europe, Asia, and North America. Most sirenians are limited to tropical or subtropical latitudes, in shallow seas that provide adequate macrophytic food and refuge from predation, and are thermally tolerable. The sea otter is confined to the coastal North Pacific Rim and the chungungo to the temperate coastal southeastern Pacific. Polar bears occur only in the Arctic and subarctic, rarely traveling south of 60°N latitude except in the relatively frigid northwestern Atlantic and Hudson Bay.

## 2. Cetacea

The distinguishing anatomical and functional features of the cetaceans are summarized in Table II. The two major taxonomic subdivisions of cetaceans are the suborders Odontoceti, or "toothed whales," and Mysticeti, or "baleen whales." The odontocetes are the most diverse of the major marine mammal taxa, with 10 families and 71 species (Table I). The best known families are the Delphinidae and Phocoenidae. In ecological terms, the family Ziphiidae is among the most sparsely studied groups of mammals on Earth. The mysticetes (Table I) include species that, by a wide margin, are the largest animals in the earth's history. The blue



whale, largest of all, reaches 30 m in length and 200 tons in mass.

Most cetaceans dwell in the open sea or in the seas and estuaries of the continental margins. The exceptions are two delphinids, a phocoenid, and four odontocete species known as river dolphins (Table I). Of the river dolphins, one species, the Franciscana or La Plata dolphin, is found mainly in coastal marine waters of Brazil, Uruguay, and Argentina. The other three river dolphins are exclusively aquatic. One of these, the boto, is known to leave river channels and travel within the adjacent flooded forests of the Amazon basin during the wet season. The Irrawaddy dolphin, a delphinid, occupies coastal marine habitats and major river systems in southeastern Asia, some Indo-Pacific islands, and the northeastern coast of Australia. The tucuxi, another delphinid, occupies the Amazon watershed and coastal marine habitats of tropical Atlantic South America and southern Central America. The finless porpoise, a phocoenid, occurs in the Yangtze River watershed and other large southern Asian rivers and coastal marine habitats from the Persian Gulf to Japan.

Mysticetes often segregate feeding and breeding activities, both in time and space, connecting the two categories of activity with extensive seasonal migration. Feeding is done primarily at high latitude during summer, and breeding and parturition are done primarily at low latitude during winter. Thus, a significant poleward migration is required during spring, and an equatorward return trip is necessary in autumn. The segregation of feeding and breeding, and the associated migratory behavior, is best developed and understood in the largest cetaceans. However, some large mysticetes, such as the bowhead whale of the Arctic region and the Bryde's whale of tropical and subtropical latitudes, undertake only modest seasonal migrations in contrast to species such as the humpback whale or gray whale. The smaller cetaceans, including most of the odontocetes, appear to disperse primarily on the basis of food availability and productivity at sea. Odontocetes typically do not engage in the dramatic seasonal migrations known for some of the mysticetes, although there are exceptions among the larger odontocetes.

Healthy mysticetes never leave the water in favor of land. A few odontocetes occasionally strand intentionally on beaches or river banks for brief periods while in pursuit of prey. Perhaps the most familiar example is the brief intentional stranding of individual killer whales while foraging on pinnipeds in the surf zone of Argentine beaches. In general, however, the cetaceans are not functionally amphibious. Aside from the noted exceptions, strandings of cetaceans can be considered

abnormal, even pathological, behavior typically resulting in death or serious injury.

### 3. Pinnipedia

Although pinnipeds and cetacea often use similar aqueous habitats, the pinnipeds have many obvious differences from the cetacea in form and function. Features shared by most pinnipeds, including the phocids, otariids, and odobenids, are summarized in Table II. The largest pinnipeds are adult males of the highly dimorphic southern elephant seal, exceeding 5 m in length and reaching 3700 kg in mass. Adult male northern elephant seals are only slightly smaller. Walruses are also quite large, reaching 3.5 m and 1500 kg.

Despite many anatomical, physiological, and ecological features obviously associated with life at sea, pinnipeds are best regarded as amphibious. All species utilize solid substrata for breeding or for postbreeding maternal care. Solid substrata are also used as short-term resting sites and for protracted periods in some species during molting of the skin and pelage. Although the proportion of time spent on land ("hauled out") over the long term varies significantly among species and age and sex categories, generally the pinnipeds spend a major portion of their lives on land near shore or on pack or shorefast ice at sea.

Otariids use terrestrial habitats near shore for breeding, postpartum maternal care, molting, and resting. Preferred hauling sites are those near areas of high oceanic productivity and those free of large terrestrial predators. Thus, hauling grounds for otariids typically are islands or mainland locations protected by cliffs or rough terrain from land predators. Often, the hauling grounds are localized at high latitude or the upwelling zones of mid-latitude eastern boundary currents, such as the Humboldt, California, and Benguela currents, where production of preferred food species is high and temporally predictable. Optimal hauling grounds for breeding otariids are few in number and often limited in size. Breeding activities of otariids are highly synchronous, occurring during a narrow time window when food availability for lactating females and newly produced juveniles is seasonally optimal. As a consequence of the various spatial and temporal constraints, pinniped breeding typically involves conditions of extreme crowding on haul-out sites.

Phocids and walruses haul out for the same purposes as otariids. Unlike otariids, phocids and walruses use two very different kinds of substrata. About half of the phocid species use coastal land for hauling grounds, selecting sites for largely the same reasons described for otariids. Thus, timing and location of breeding for

land-breeding phocids and otariids are generally similar. As a consequence, reproductive activities for many land-breeding phocids also occur under conditions of extreme crowding. Known exceptions include some populations of harbor seals with spatially dispersed, largely aqueous breeding systems and the three recent species of monk seal (one now extinct) with temporally asynchronous breeding systems at low latitude. The walrus and the remaining phocids breed or care for young, rest, and molt on ice at high latitudes rather than on land. Ice as a substratum varies widely over time and among locations in stability, vulnerability to predators, and provision of access to the surrounding sea. Thus, among ice-breeding pinnipeds there are significant resultant variations in social and breeding strategies and in the degree of crowding at hauling sites. A major predator of ice-hauling phocids, the polar bear, is present in ice-covered marine habitats of the Arctic region but not the Antarctic. This pattern has many interesting consequences for interhemispheric differences in the ecology of ice-hauling pinnipeds.

Some phocid species or subspecies occur only in aquatic habitats. Two subspecies of ringed seal occur only in Lake Saimaa, Finland, and Lake Ladoga, Russia, respectively. The Caspian seal is found only in the Caspian Sea and the Baikal seal only in Lake Baikal, both in Russia. A population of harbor seals occurs year-round in Lake Iliamna, Alaska, but the degree of exchange, via river connection to populations of harbor seals in nearby Bristol Bay, is unknown.

#### 4. Sirenia

Several of the defining anatomical and functional features of the sirenians are convergent with those of the cetaceans (Table II). However, in contrast to the cetaceans there are few Holocene sirenians (Table I). The sirenians are the only extant herbivorous marine mammals, sharing common ancestry with desmostylians and terrestrial subungulates (e.g., aardvarks, elephants, and hyraxes).

Sirenians are large in body mass and linear dimension compared to most terrestrial mammal species. The Steller's sea cow was the largest of the modern species, reaching 7 m in length. To our knowledge, body mass of the Steller's sea cow was never directly measured, but the estimated maximum is 10,000 kg. Maximum adult lengths and masses of the three manatee species range from 3 to 4 m and from 450 to 1600 kg, respectively. Adult dugongs reach maxima of 3.3 m in length and 400 kg in mass.

All modern sirenians are fully aquatic or marine and are incapable of leaving the water. Sirenians feed on

large plants growing on the bottom, in midwater, at the surface, or closely overhanging the surface. They forage exclusively in shallow habitats. Manatees utilize freshwater, estuarine, and fully marine habitats, often interchangeably. Use of rivers by manatees is influenced by rainfall patterns and river discharge rates. Manatees generally concentrate in habitats that are relatively warm and physically protected from extremes of weather and sea. Dependence on relatively warm-water temperatures may result from the combination of obligate homeothermy and a relatively low basal metabolic rate compared to that of other marine mammals. The limited tolerance of low water temperature likely contributes to seasonal migration and a tendency to concentrate at high density in warm-water refugia during the winter. Manatees may also congregate near sources of fresh water, although fresh water is not a physiological requirement. The smallest of the modern sirenians, the Amazonian manatee, occurs only in the freshwater habitats of the Amazon River watershed of South America. Dugongs are fully marine and forage primarily on benthic seagrasses in shallow coastal tropical marine habitats. Steller's sea cow, known only from isolated parts of the subarctic North Pacific, was the most aberrant of the holocene sirenians. Sea cows likely foraged exclusively on kelps and other large algae along exposed shores.

#### 5. Desmostylia

Desmostylians are the only known extinct order of marine mammals. The small number (<10) of recognizable species in the fossil record are of Oligocene and Miocene age and are confined geographically to the North Pacific region. Desmostylians were quadrupedal amphibians sharing common evolutionary ancestry with the sirenians. Habitats of desmostylians probably were shallow waters supporting productive populations of algae and aquatic vascular plants, their primary food, in latitudes ranging from subtropical to cool temperate. In habits and superficial morphology, desmostylians often are characterized as similar to the modern hippopotamus. Some have argued that at least some of the desmostylians fed on clams and other benthic invertebrate prey, but the consensus is that they were primarily, if not strictly, herbivores.

#### 6. Marine Otters

There are 13 recognized extant species of otters worldwide, comprising the mustelid subfamily Lutrinae. Here, we consider 2 species, the sea otter of the North Pacific Rim and the chungongo of Peru, Chile, and southernmost Argentina. Both are marine species with

amphibious characteristics. Other otter species may utilize marine environments, but they also have obligatory associations with aquatic and terrestrial habitats. The sea otter and chungongo do not appear to utilize freshwater habitats, except occasionally and facultatively.

The sea otter is arguably the most derived of the lutrines. It is the largest of the mustelids, with some adult males reaching 45 kg in mass and 1.6 m in total length, but among the smallest of the marine mammals. Sea otters are relatively weak divers compared to most marine mammals, and they feed almost entirely on large-bodied, sessile or slow-moving benthic invertebrates. Sea otters often haul out on coastal beaches and reefs to rest and conserve heat, especially in the northern portions of their geographic range during periods of harsh weather or reduced sea surface temperature. Sea otters are not known to utilize freshwater habitats for any purpose.

The smallest of the marine mammals, the chungongo, reaches maxima of 6 kg in mass and 1.1 m in length. Chungongos are morphologically similar to the seven congeneric species of otters. The ecological characteristics of chungongos are not well known. They feed primarily on small crustaceans, mollusks, and fish taken during dives in nearshore marine habitats along open coasts. They may also forage in fresh water, taking small crustaceans. They haul out between foraging periods on exposed rocky shores and appear to maintain shoreline dens that are focal areas for social and reproductive behavior.

## 7. Polar Bears

Polar bears are one of seven recognized bear species. Several of the other bear species utilize marine and aquatic habitats for foraging, but polar bears are more dependent on marine habitats for food than are other bears. Although generally similar to other bears morphologically, polar bears have several distinguishing features that reflect their associations with frigid terrestrial and sea ice habitats and with Arctic marine ecosystems. Polar bears are small compared to many marine mammals but large compared to most other bears. Adult males reach 2.5 m in length and 800 kg in mass. Along with chungongos, polar bears are perhaps the least modified morphologically, compared to terrestrial mammals, of the world's recognized marine mammal species. Although polar bears do not dive repetitively in the manner typical of many marine mammals, they are efficient swimmers, able to traverse large expanses of open water. Polar bears also cover large distances at sea by walking or running across sea ice. Although primary prey are pinnipeds taken from the surface on sea ice,

polar bears may have a diverse diet. They are the only recognized species of marine mammal that travels extensively on land away from the shoreline and the only species that consumes both plant and animal species as a regular part of the diet. In addition, polar bears are unique among marine mammals in producing altricial young.

## B. Synopsis of Evolutionary Histories of Major Marine Mammal Taxa

### 1. Cetacea

The oldest recognized cetaceans are Eocene fossils of the cetacean suborder Archaeoceti. Archaeocete fossils are found primarily in rocks of present-day Egypt, Pakistan, and India in strata thought to be associated with the Tethyan Sea of ancient times. Thus, it is presumed that cetaceans originated in the Old World Tethyan environment. Cetaceans share common ancestry with an extinct terrestrial ungulate taxon known as the mesonychia. The earliest recognizable cetacean fossils date to approximately 55 million years ago (Ma). The archaeocetes included many "missing link" fossils, displaying intermediate forms with regard to progressive reduction and loss of the hindlimbs, elongation of the anterior skull, and modification of dentition.

Archaeocetes were largely extinct at the beginning of the Oligocene, approximately 38 Ma. The first precursors to the modern suborders Odontoceti and Mysticeti appear in the fossil record during the Oligocene, but the first fossils linked unambiguously to modern cetacean families appear primarily during the Miocene. For example, the earliest sperm whales appear in early Miocene strata. Beaked whales appeared first in the middle Miocene, and the earliest dolphins and porpoises appeared in the late Miocene approximately 11 Ma. In the mysticetes, the earliest rorquals and right whales also are in Miocene strata. The oldest gray whale fossils are from the Pleistocene. Thus, there are no known fossils providing insight to the early evolution of the modern gray whales. There is general agreement that the cetaceans are monophyletic.

It is apparent from the fossil record that the modern taxa of cetaceans have been preceded by many extinct species, likely outnumbering extant species by a considerable number. For example, an early mysticete group, the family Cetotheriidae, contains about 60 known species dating from the Oligocene. The existence of extinct large taxa implies significant episodes of diversification and subsequent extinction well before the modern families of cetaceans appeared. The record also suggests a

dynamic pattern of biogeographic variation, such as the occurrence of ancestral monodont fossils in Miocene strata of Mexico. The Monodontidae now include only the belugas and narwhals of high northern latitudes. The dynamic evolutionary record almost entirely predates homonid evolution and the emergence of anthropogenic influences on pattern of extinction.

## 2. Pinnipeds

The earliest known pinnipeds are represented by fossils of the late Oligocene or early Miocene, approximately 25 Ma. The two major recognized lineages culminate in the modern Phocidae and the modern superfamily Otarioidea, the latter including the Otariidae and the Odobenidae. The earliest fossils of both lineages are of similar late Oligocene or early Miocene age. In contrast to the cetaceans, there are recognized linkages of good quality between the earliest pinniped fossils and the major modern pinniped taxa.

The traditional view of pinniped evolution is that the phocids and otarioids evolved independently. Phocids are said to have emerged in the North Atlantic region from ancestral forms linked to modern mustelids. Appearance of phocid species in the Pacific likely occurred much more recently, possibly during events related to the extinction of early Pacific otariids and odobenids. Otarioids are thought to have evolved from ancestral ursids in the North Pacific. The diphyletic model is supported by traditional analyses of cranial morphology and by the absence of early fossil otarioids from strata of the North Atlantic region. Recent analyses of postcranial material and of molecular data support an alternative model in which the pinnipeds are a monophyletic lineage sharing common ancestry with the modern mustelids. Current consensus favors the monophyletic model.

Recently published evidence indicates controversy regarding the affinities of the Odobenidae. Some analyses indicate that odobenids are in fact more closely related to phocids than otariids, whereas others favor the more traditional view, with odobenids closely allied to otariids in the Otarioidea. We follow the traditional view here but acknowledge the evidence in support of the alternative scenario.

Ancestral taxa of otariids and odobenids show a high level of diversity compared to modern forms. The recognized modern genera of odobenids and otariids appeared primarily during the late Pliocene or the Pleistocene. The fossil record for phocids is not well developed, especially in some North Atlantic regions thought to be important in understanding early evolution in the group. Miocene fossils from the Southern Hemisphere

include significant numbers of monachine (phocid subfamily Monachinae) seals (ancestral to the modern monk seals), elephant seals, and Antarctic ice seals. Ancestry and relationships of the more derived forms, including the modern phocine (phocid subfamily Phocinae) seals of northern temperate and polar latitudes, have not been resolved definitively. As with the cetaceans, the pinniped fossil record indicates significant episodes of diversification and extinction before the emergence of the modern forms, clearly predating anthropogenic influences.

## 3. Sirenians

The earliest fossil sirenians are from early Eocene strata, approximately 55 Ma. Significant radiation into at least three families of sirenians had occurred by the time the earliest cetacean fossils appeared in the Eocene. Thus, sirenians appear to be the oldest order of living marine mammals on Earth. The fossil record does not provide clear evidence of the ancestral groups that gave rise to the sirenians, although studies of modern forms indicate common ancestry with other subungulates. The oldest sirenian fossils are from Jamaica, but the sirenians are thought to have emerged first in the Old World Tethyan environment. The early radiations of the sirenians appear to share common ancestry with the extinct family Protosirenidae. Protosirenids appear to have given rise to the modern families Trichechidae (manatees) and Dugongidae (dugongs and sea cows). Dugongids probably first appeared in Mediterranean waters during the Eocene, whereas trichechids apparently first evolved along the South American coast during the Miocene. Sea cows first appeared in the southeastern Pacific during the early Miocene, later radiating throughout the Pacific basin. Some sea cows were unusual among sirenians in their great body size and use of relatively exposed cold-water habitats. They occurred along the coasts of California, Japan, and the subarctic North Pacific Rim during the Pliocene and Pleistocene. Although the predominant sirenian family in the Holocene, trichechids have a relatively poor fossil record. The allopatric distribution of the three modern species apparently resulted from temporary geological isolation of the Amazonian watershed and from a chance colonization of West African coastal waters from an ancestral Caribbean population.

## 4. Desmostylians

As noted previously, desmostylians apparently first appeared in the Oligocene, approximately 35 Ma. The oldest known fossils are from the coasts of Washington and Oregon in the northeastern Pacific, and all desmo-

styan fossils come from the North Pacific. Desmostylians share common ancestry with sirenians and terrestrial subungulates, although many details of these relationships are unknown.

### 5. Marine Otters and the Polar Bear

The sea otter, chungungo, and polar bear represent three separate, relatively recent entries into marine environments by largely terrestrial or aquatic taxa. Chungongos still resemble other otters so closely that a meaningful fossil record of their evolution as marine mammals does not exist to our knowledge. There are some Pleistocene fossils of polar bears, but neither fossil nor modern forms represent significant departures from the ancestral ursid morphology. Thus, for both chungongos and polar bears, the history of adaptation to marine life is best inferred from modern biological data.

The sea otter appears to have a somewhat more extensive fossil ancestry in the marine environment than the polar bear or chungungo. There are two extinct genera of sea otters in the fossil record. Species of *Enhydriodon* have been found in Africa and Europe in late Miocene and Pliocene strata. *Enhydritherium lluecai* is known from the late Miocene of Europe, and *Enhydritherium terraenovae* is known from the late Miocene through the middle Pliocene in Florida and California. *Enhydritherium* is thought to be ancestral to *Enhydra*. *Enhydra* is confined to the North Pacific region. The extinct *Enhydra macrodonta* is known only from the late Pleistocene. The single surviving species of sea otter (*Enhydra lutris*), dates from the early Pleistocene.

## C. General Factors Contributing to the Vulnerability of Marine Mammals to Overexploitation and Extinction

### 1. Obligatory Dependence on the Sea Surface for Respiration

Marine mammals must exchange respiratory gasses directly with the atmosphere, in the same manner as their terrestrial relatives. Thus, unlike marine fishes and invertebrates, marine mammals at sea are never entirely free of their association with the sea surface. Marine mammals must return periodically to the surface to breathe. The process of breathing at the surface is often associated with conspicuous activities, such as splashing, exhalations audible over long distances, and the production of clearly visible clouds of condensed water vapor associated with exhalations. Marine mammals are often physiologically obligated to remain at the surface for several minutes, allowing multiple ex-

changes of gas volumes contained in their lungs, in order to set the biochemical stage for successful subsequent dives.

Human travel in watercraft at the sea surface is relatively efficient and advanced. By understanding the respiratory behavior of marine mammals, people can position themselves to facilitate close contact with surfacing animals. The result is high vulnerability of marine mammals at sea to human hunters.

### 2. Large Body Mass and Linear Dimensions

The marine mammals, on average, are large compared to most other animals. Large body size probably evolved in response to certain constraints associated with life at sea, most notably those associated with thermoregulatory and hydrodynamic efficiency, foraging ecology, reproductive ecology and physiology, and habitat preference.

The energetic return to the consumer per unit of hunting effort will increase with the mean size of the prey, all other factors being equal. Among the marine mammals, all the mysticetes and many of the odontocetes, pinnipeds, and sirenians are large enough in body mass to be highly desirable as targets by human hunters. With twentieth-century refinements to the technology of marine mammal hunting at sea, pursuit of even the most mobile and dangerous of the marine mammals produced highly desired rates of economic return as long as stocks were not depleted. Thus, large body size, per se, increases the vulnerability of marine mammal populations to extinction simply by improving the economic return on investment in hunting activity.

### 3. Relatively High Predictability of Spatial and Temporal Distributions in Association with Regions of High Biological Productivity at Sea

With certain exceptions, marine mammals are rarely far removed from locations in which they can forage efficiently. The metabolic generation of heat, fueled by food consumption at high rates, is the only option available to marine mammals at sea for replacing heat lost continuously in a heat-consumptive environment. In species with significant annual migrations to food-poor areas for breeding, high rates of intake during the feeding season are vital for survival of extended travel and for maximizing rates of reproductive fitness. In species whose breeding systems include extensive seasonal fasts, seasonal hyperphagy, requiring proximity to abundant food, may be crucial to reproductive fitness and to long-term survival of both sexes.

Most marine mammals feed on planktonic inverte-

brates and small schooling fishes. Over the long term, seasonal and spatial patterns of production of zooplankton and forage fish are relatively predictable. Successful tracking of resources that vary predictability in space and time is vital to survival and reproductive success. Accumulated ecological data for marine mammals indicate that most populations successfully track food resource productivity most of the time. The result is an array of stereotypical seasonal and spatial movements by marine mammals that, in many cases, are readily identified. Clearly, an understanding of movements of marine mammal stocks over time reduces the investment risk in developing strategies for efficient hunting of marine mammals. Thus, spatial and temporal predictability in marine mammal foraging facilitates efficient hunting by humans and adds to the risk of anthropogenic extinction as a result of direct exploitation of marine mammal populations.

#### 4. Impaired Mobility, Contagious Dispersion, and Temporal and Spatial Predictability When Hauled Out on Land

Pinnipeds, the two marine otters, and polar bears are amphibious marine mammals. Hauling behavior of pinnipeds is particularly synchronous and predictable, producing seasonally dense hauled-out aggregations that can be anticipated readily in space and time. These patterns are most extreme for some of the land-breeding pinnipeds that dwell at middle or subpolar latitudes, but they are prevalent in many other pinniped species as well. Two primary factors contribute to the pattern. First, good hauling sites that are near seasonally predictable foraging locations, and that are free of the disruptive effects of terrestrial predators such as bears or wolves, are typically few in number and small in size. Pinnipeds depending on such hauling sites have evolved high site fidelity and strong navigational capabilities in order to minimize the risk that good hauling sites will be overlooked when needed for breeding, molting, and resting. Second, in the case of pinnipeds at middle and high latitudes, highly productive foraging locations near good hauling sites tend to be strongly seasonal in food availability.

Pinnipeds are awkward on land and can be captured easily by human hunters on hauling sites, even if methods are primitive. Because seasonal hauling patterns are easily recognized in space and time, human hunters can plan highly efficient hunting of hauled pinnipeds with a very low risk of poor return on invested effort. In addition to directed harvest, human activity on preferred hauling grounds can cause significant unin-

tended disruption of breeding activity and social interaction among exploited pinnipeds in addition to directed harvest. The stampeding of panicked animals can lead to premature births, trampling of small pups, disrupted dominance hierarchies, permanent abandonment of haul-outs by adults, and other forms of disturbance. The net added result of hunter-associated disruptions is increased mortality and reduced birth rate in the short term. Repeated disruptions associated with human activity can lead to increased long-term risk of local extinction for populations at particular hauling sites.

Sea otters are awkward when hauled out. Thus, they are also vulnerable to harvest and disturbance by people. However, hauling patterns of sea otters are less predictable in time and space than patterns for pinnipeds. Polar bears are highly mobile on land and are dangerous to human hunters. Nevertheless, time spent on land increases the vulnerability of polar bears to hunting and disturbance by people.

#### 5. Subsistence and Commercial Market Demand for Oil, Blubber, Meat, Baleen, Pelts, Ivory, and Other Body Parts

By virtue of their frequently large size, morphology, physiology, and chemical composition, the harvested carcasses of marine mammals provide many products significant in subsistence and commercial contexts. Harvested marine mammals provide large quantities of meat, organs, and blubber per unit of hunting effort invested. Meat and organs are used directly for human consumption or to feed domestic animals, such as sled dogs, on which human enterprise may depend. Meat may also be marketed commercially for human consumption or as pet food. Blubber is also consumed directly, but traditionally the majority of blubber is rendered and refined into products such as oil or fuel, used both for subsistence and as commercial commodities. In the case of the odontocetes, oil taken from the organs of acoustic transmission may be refinable to high-quality lubricants that, until recent decades, could not be duplicated synthetically.

Skeletal material from hunted marine mammals has been used traditionally by subsistence cultures for tools, boat construction, dwelling construction, and as raw material for handicrafts and objects of ceremonial significance. Hand-crafted articles made from marine mammal skeletons often have significant commercial value as well. Mysticete baleen and the vibrissae of walrus and other pinnipeds have, been marketed commercially as components of clothing or toiletry articles, although such uses are largely in the past. Hides and

pelts may be used for clothing and for construction of boats or dwellings. The pelts of sea otters and fur seals have been sought for centuries as commodities of high commercial value for clothing or as adornments for ceremonial robes and artifacts. Teeth and other body parts from sea otters are also known to have had ceremonial significance to indigenous peoples. Specialized tusks in narwhals and walruses are extremely valuable and have put these species at continued risk during much of the past two centuries. Polar bears and walruses have been hunted for meat, oil, hides, and pelts by subsistence cultures for millennia and were targets of trophy hunters through much of the nineteenth and twentieth centuries.

Much of the demand for marine mammal products centers on tissues involved in thermoregulation. The blubber of cetaceans, pinnipeds, and sirenians and the pelts of fur seals, sea otters, and polar bears are the respective primary organs of heat retention, allowing the preservation of homeothermy in a chilling environment. Thus, the homeothermic physiology of marine mammals underlies their desirability as commodities and is a significant contributor to the vulnerability of marine mammals to anthropogenic extinctions.

#### 6. Low Demographic Potential for Rapid Recovery from Disturbance or Overexploitation

The marine mammals are significantly convergent in many aspects of life history. Mean litter size is one for all species except polar bears and chungongos, and multiple simultaneous births are rare in all species of cetaceans, pinnipeds, sirenians, and in sea otters. The age of first reproduction is often relatively high, especially in sirenians and the larger odontocetes. None of the extant marine mammals has a birth interval of less than 1 year, and for many species the birth interval is at least several years. In all species, parental care is entirely maternal, and the energetic costs of lactation and other forms of care are extensive for the adult female. Reproductive success of newly mature females is often very low in marine mammals, increasing only with experience. Survival rates of weaned offspring may be low during the first few years of independence.

The combined effects of the previously discussed characteristics are low potential rates of growth in marine mammal populations, even when free of the constraints imposed by food limitation, competition, predation, or natural disturbance. Maximum potential annual rates of population growth typically are 2–8% for the cetaceans and sirenians and 10–15% for the pinnipeds

and sea otters. Exceptional cases, both higher and lower than mean rates, are known for both groups. Realized rates of growth, affected by variations in food supply and the effects of disturbance, predation, competition, and possibly other factors, often are much closer to zero and may be constrained to negative values. Given these patterns, recovery from depletion associated with excessive exploitation or disturbance may require many years, and intervening additional harvest or disturbance can increase the risk of extinction to high levels.

#### 7. Morphological, Physiological, and Ecological Predisposition for Bioaccumulation of Lipophilic Anthropogenic Contaminants and Toxins

Most marine mammals utilize a well-developed subcutaneous blubber layer as the primary means of thermoregulation. The blubber layer is also a primary organ for energy storage, as are fat deposits in polar bears. The blubber layer consists of lipids, fatty acids, and connective tissues. Relative proportions vary by species, age, and reproductive status of individuals and by location on the body.

Many species of cetaceans and pinnipeds, the Amazonian manatee, and some polar bears have an extensive seasonal fast each year, relating to migration away from primary feeding areas, an extended haul-out or denning period in association with reproductive activity, or shifts in habitat structure. During such fasts a significant proportion of the blubber layer or fat deposits is metabolically mobilized to meet energy and water demands during the fasting period. Following the fast, animals return to the feeding grounds and forage intensively to reconstitute the reduced blubber layer or fat deposits. In the case of the odontocete cetaceans, the acoustic melon is a second concentration of lipid-based tissue.

Odontocetes, pinnipeds, the two marine otters, and polar bears occupy high trophic levels in marine food webs. Many stable lipophilic contaminants are transmitted through food webs, such that top-level consumers may be exposed to high levels of contaminants. Such patterns are particularly well-known for environmental contaminants such as organochlorines, a group including the polychlorinated biphenyls (PCBs) and the various derivatives of dichlorodiphenyltrichloroethane (DDT). In this context, many marine mammals face “double jeopardy.” First, a high position in their respective food webs confers the risk of high exposure to stable lipophilic contaminants. Second, extensive, met-

abolically active lipid-based tissues are vital to their survival but vulnerable as sites for accumulation of contaminants. In several cases, lipophilic contaminants have been correlated with reduced immune competence, disease outbreaks, and significant mass mortalities in marine mammal populations. There is also evidence that contaminants may cause endocrine disruption and reproductive malfunctions such as premature deliveries of pups. Thus, the combination of lipophilic contaminants in the marine environment, the pattern of foraging at high trophic levels by marine mammals, and lipid-based tissues that are functionally significant and metabolically active in many marine mammals results in increased vulnerability of marine mammals to anthropogenic extinction.

#### 8. Overlap of Diet or Habitat with Commercial or Recreational Fisheries

Marine mammals often feed preferentially on prey species also sought by commercial, recreational, or subsistence fisheries, or they forage in habitats in which significant commercial fishing activity occurs. These patterns create two types of problems that may enhance the vulnerability of marine mammals to anthropogenic extinction. In the first case, marine mammal populations are viewed by commercial, recreational, or subsistence fishers as competitors for a common resource. As a consequence, legal recourse may be sought to actively reduce the range or numbers of marine mammals by killing, translocation, or harassment in order to reduce the intensity of competition in favor of harvesting interests. Illegal activities may also result, including unauthorized killing or harassment also intended to reduce the intensity of competition between marine mammal populations and fisheries. Such circumstances can lead to conflicting management goals by interested parties, particularly if the involved marine mammal populations are small. From the perspective of the involved marine mammal species, competition from harvesting interests may alter the quantity or distribution of food availability and may produce significant consequences at the population level.

The second type of problem is inadvertent or incidental take of marine mammals by entanglement in fishing gear. Such taking may include injury or death of individual animals, again producing possibly significant effects at the population level. Potential responses to such problems include tolerance of taking by management authority, changes to fishing techniques or effort to reduce rates of taking, or displacement of fishing effort to other locations. Such interactions may still

result in population-level effects if reduction in rates of taking is inadequate, or if parties affected by displacement of fishing effort resort to illegal taking of involved marine mammals as a form of retribution. Thus, both types of problems may contribute to increased risks of extinction, especially in cases in which the involved marine mammal population is small at the time that conflicts are recognized.

### D. General Factors Hindering Effective Identification and Monitoring of Marine Mammal Populations Vulnerable to Extinction

#### 1. Availability Bias

Marine mammals at sea spend most of their time below the sea surface. Depending on water clarity, typical depth of dive, angle of observation, and platform of observation (i.e., surface vessel or aircraft), a varying proportion of individual marine mammals in a field of view cannot be seen and thus cannot be enumerated in a population survey. The proportion of animals not visible because of submergence is the availability bias of the survey. Availability bias reduces both accuracy and precision of population estimates, and it reduces the statistical power of a survey effort to detect population trends correctly. Availability bias can be estimated with detailed information on water clarity in the survey area, diving characteristics of target species, and detection characteristics of survey observers. Estimates of bias allow the effects of the bias to be incorporated into calculations of correction factors and coefficients of variation (CVs) for population estimates. Elimination of availability bias generally is not possible for surveys at sea.

Availability bias may also be a problem for surveys directed to hauled animals on shore. For example, topographic irregularities such as overhangs may obscure animals that are present in the defined survey area, and animals at high density may obscure one another. Judicious timing and modified survey angles may sometimes reduce availability bias in surveys of hauled animals to nearly zero.

#### 2. Observer Bias

Observer bias, also known as detection bias, results from the inability of survey observers to correctly enumerate the number of animals visible in a field of view. Unlike availability bias, observer bias can be either a high or low bias. Observer bias has the same implica-



tions for population estimation as summarized previously for availability bias. Observer bias can be a significant source of error in both surveys at sea and surveys of hauled animals on shore. Observer bias can be reduced with increased experience of individual observers, and it can be estimated using double-counting techniques with paired independent observers or by comparing observer counts in the field with counts from aerial photographs taken at the same time and place.

An important form of observer bias involves errors in estimates of group size in marine mammal surveys. Typically, marine mammal surveys involve counting of groups in the survey area. The group count is then multiplied by the mean group size, often based on a separate survey effort, as the first step in population estimation for the survey area. Estimates of group size are subject to observer bias, contributing to an increase in the CV for population estimates. Observer bias in group size estimates can be assessed under good conditions by comparing group size counts by observers with group size counts based on aerial photographs of the same group of animals taken during the surveys.

### 3. Low Statistical Power of Population Survey Data to Detect Trends in Population Size

In many cases, the single most important type of information for assessing the status of a marine mammal population is the trend in population size over time. A trend is simply a time series of counts in which the slope of a fitted line is significantly different from zero. In a statistical context, the ability to detect a trend correctly is influenced by four factors. The first is the strength of the trend. Strong trends are those in which the absolute value of the slope of the fitted line is large. When other factors are constant, strong trends are more likely to be detected correctly than weak trends. The second factor is estimation error. Other factors being equal, trends in survey-based estimates of population size over time are more likely to be detected correctly if the associated CV is small than if it is large. The third factor is the number of replicate surveys available for a given time period, to be used to calculate a single estimate of population size. Increased replication reduces the CV associated with a given estimate. Thus, the probability of correctly detecting a trend increases with the number of replicate surveys used to calculate each point in the population time series, other factors being equal. The fourth factor is the number of years in which surveys are done. The probability of correctly detecting a trend, with other factors being equal, is increased with an increasing number of survey years.

Many marine mammal surveys have characteristics that reduce the probability of correctly detecting trends. Weak trends may portend significant conservation concerns for marine mammals but are inherently difficult to detect if CVs are large and replication is minimal. Large CVs are common in all types of marine mammal population surveys, although CVs are gradually being reduced by the extraordinary efforts of involved researchers and managers. The level of replication generally is directly dependent on levels of funding. Implementation of field surveys in marine mammal science is often compromised by the challenge of limited funding and competing priorities. Well-executed marine mammal population survey programs generally detect strong population trends successfully. However, many surveys lack the statistical power to detect weak trends that may nevertheless be important in the context of avoiding eventual extinction of marine mammal populations. The only solution is to extend survey effort over many years, thus improving the odds of recognizing a trend. Such an approach carries the obvious risk of potentially delaying the recognition of a significant conservation problem for the target population and inevitably increases the overall monetary cost of the monitoring effort. In many cases, there is no alternative to extension of the timescale of the monitoring project. Policies that require decisions about trends over a short time frame are likely to fail in providing appropriate management if trends are weak.

### 4. Inadequate Understanding of Vital Demographic Parameters and Population Structure

Determination of effective measures to eliminate a negative trend in a marine population often depends on a reasonable understanding of the demographic characteristics of the subject population. Such an understanding improves the odds that limited resources for conservation work will be applied where the greatest benefits will accrue. The demographic parameters of marine mammal populations are often known only with poor levels of accuracy or precision. In such cases, conservation effort can be readily misdirected. For example, measures to reduce preweaning mortality in a marine mammal population will be ineffective if the larger problem is a high rate of adult female mortality obscured by imprecise or inaccurate estimates of adult female survival rate. Misdirection of limited resources for conservation has obvious effects on extinction probabilities for populations in jeopardy.

Accurate, precise measurements of population parameters in marine mammals are difficult to obtain.

Because marine mammals are long-lived species, the time line necessary to obtain good parameter estimates is always lengthy, and the best data come from studies that extend beyond a decade in duration. One of the important results of long-term demographic research on marine mammals is evidence, in some cases, of marked interannual variability in demographic parameters. Such patterns further emphasize the critical need for a long timescale in such studies. For many species, good estimates of demographic parameters require tagging of individuals, an invasive process that can impose risks of reduced survival for the tagged individual and risks of disturbance to groups of animals such as pinniped breeding colonies on haul-outs. In some cases, tagging has the potential to bias the demographic data recorded from the tagged individual. Finally, because of the lengthy duration and labor intensity involved, demographic studies may be quite costly and therefore difficult to implement.

Population structure is poorly understood in many marine mammal species. Species with large geographic ranges and high mobility may appear to be genetically panmictic. However, recent data suggest significant within-species genetic structure in several cases. Examples include several delphinids and baleen whales. Uncertainties about population structure impose a risk of inappropriate scale in the application of management policies. For example, bycatch in fishing nets may occur at high rates only in one portion of the geographic range of a delphinid. If the population is panmictic across the entire range, management authorities may deem the bycatch rate acceptable. Such a policy would, however, increase extinction risk if the area of high bycatch rate supports a population genetically distinct from the remainder of the species' range.

#### 5. Inadequate Understanding of Effects of Environmental Uncertainty on Dynamics of Populations

There is limited evidence that apparently stochastic environmental fluctuations may have important effects on the dynamics of marine mammal populations. The best known cases involve the pinnipeds, primarily because pinnipeds breed or care for young on solid substrata and therefore have more readily observed and better known population dynamics than most other marine mammal taxa. For example, during the early 1980s and mid-1990s, food supplies for many temperate pinniped populations in the North Pacific were disrupted by global-scale oceanographic disturbances generally known as "El Niño Southern Oscillation" (ENSO) events. ENSO events involve a suite of changes in global

wind and ocean current patterns, with major large-scale effects on patterns of biological productivity. Although often drastic, local changes in productivity typically are temporary, returning to normal levels over time periods from a few months to a few years. ENSO events have stochastic characteristics regarding both the frequency and the intensity of occurrence and may be variable in their effects on survival rate and population trajectory of marine mammals. Both of the referenced ENSO events caused significant reduction in some pinniped population sizes and in mortality rates in pups of the year. There have been very few opportunities to observe the effects of stochastic disturbance during other time periods or on other marine mammal taxa.

In the context of extinction, the major problem in incorporating naturally occurring stochastic events into conservation planning is the lack of good quality data from an adequate number of events. Thus, it is not possible to generalize about effects of stochastic events, nor is it possible to reasonably consider mitigation measures to minimize increased extinction risks associated with stochastic events.

#### 6. High Cost of Survey Efforts

Despite their obvious conservation value, survey efforts for marine mammal populations often are compromised or eliminated by funding constraints. The central problem is the high cost per unit effort of a good-quality survey for marine mammals at sea. Most surveys at sea use either aircraft or large surface vessels as observation platforms. Both types of platforms are extremely costly to operate at the level of rigor and safety necessary to obtain statistically defensible estimates of target population size. Indeed, the monetary costs of good at-sea surveys may be a significant proportion of the research and management budgets of resource-oriented government agencies, even in wealthy countries such as the United States. In most cases, low-cost alternatives simply do not exist. Thus, only the most serious issues of population trend in marine mammal conservation can reasonably attract a quantitatively rigorous level of survey effort. These funding realities increase the risks of extinction by improving the odds that trends of concern will be overlooked for lack of the necessary survey effort.

#### 7. Lack of Consensus and Consistency in Definition of Objective Criteria for the Determination That Particular Populations Are Vulnerable

There are many forms of institutional or organizational protocols designed to provide protection for depleted

populations of wildlife, including marine mammals. Many such protocols originate from and have the backing of government agencies. However, few protocols for the protection of species in peril include explicit, objective criteria for determining the level of jeopardy or for specifying recovery from jeopardy status. Moreover, those cases in which objective criteria are specified rarely accommodate the quantitative uncertainties in population estimation, demographic parameterization, characteristics of ecological disturbance, and related issues that are universal problems in population data for depleted wildlife. Surveys of criteria for status determination of wildlife in peril reveal little consistency, and often only minimal consideration of fundamental biological patterns, even within particular political jurisdictions.

Resolution of the problem of objective criteria will be difficult. Different taxa of organisms have different population characteristics and may warrant different approaches to the development of objective criteria for jeopardy and recovery. Differences in culture, values, and political traditions among jurisdictions also complicate any effort to establish common objective criteria. However, the separation of jeopardy criteria from arbitrary judgments, often intertwined with political considerations, is crucial to ensure that extinction risk is measured by biological criteria. A failure to achieve such separation may increase the odds of extinction for some populations in peril.

### E. Problems in Distinguishing “Natural” from “Unnatural” Extinction

As indicated previously, marine mammals have been present on Earth since the early Eocene. The majority of the marine mammals that have evolved on Earth are now extinct, and virtually all extinctions occurred before the evolution of the hominid primates, particularly *Homo sapiens*. In the context of conserving the earth's biodiversity, however, we are primarily concerned about anthropogenic loss of species. Thus, our time window is narrowed to the Pleistocene and Holocene and our conceptual focus to extinctions that result from conscious human action rather than from other, natural processes such as changing habitat, disturbance and catastrophe, or displacement of one species by another. We suggest that not all anthropogenic extinctions are necessarily unnatural. However, separating the natural from the more philosophically objectional unnatural is a great challenge, much burdened by the complicating considerations of values, culture, politics, and economics. Thus, the distinction,

however important, is well beyond the scope of this article.

## II. PATTERNS AND CASE STUDIES OF EXTINCTION IN MARINE MAMMALS

### A. Extinction or Near Extinction of Major Taxa over Evolutionary Time

#### 1. Odobenidae

Odobenids have a fossil record of striking diversity, including some species resembling modern walruses and others superficially similar to the modern sea lions and fur seals. The odobenids seem to have diverged from the otariids in the early Miocene. Peak fossil diversity is in strata from the late Miocene and early Pliocene. Diversity declined abruptly during the late Pliocene and Pleistocene. The single surviving species, the modern walrus, appeared in the fossil record during the Pleistocene. Extinct odobenids were also broadly distributed across latitude. The modern walrus is limited to high northern latitudes and is a distributionally aberrant relative to the extinct odobenids.

The fossil record indicates that modern walruses are the single relict of a largely lost taxon. Most extinctions of odobenids seem to be associated with global-scale cooling and related large-scale habitat change during the Pliocene. However, ecological characteristics of the extinct odobenids are in many cases difficult to understand because the surviving contemporary model seems atypical. Thus, it is difficult to develop meaningful functional models of extinction in the odobenids.

#### 2. Sirenia

Sirenian diversity clearly peaked during the Miocene, a period of warm global climate coincident with extensive warm shallow coastal marine habitats. Dugongids, now represented by a single surviving species, were the most diverse family of sirenians through the fossil record. Sirenian diversity declined sharply during the relatively cool Pliocene and Pleistocene, and modern forms are relicts of a largely extinct order. Two of four recognized sirenian families are now fully extinct.

The surviving sirenians seem to provide good ecological models for the extinct forms. Modern sirenians typically consume macrophytes in shallow waters and, except for Steller's sea cow, clearly prefer warm protected waters. Thus, it is likely that Pliocene cooling and the coincident widespread loss of warm shallow seas were major factors in the decline of the sirenia.

Surviving species were those able to retreat to low-latitude refugia or, in the case of Steller's sea cow, a subpolar refuge with abundant food and no significant predators.

### 3. Desmostylia

Desmostylians did not diversify to nearly the extent of the other major marine mammal taxa, and they did not survive beyond the end of the Miocene. Lacking modern ecological models, we prefer not to speculate on specific ecological mechanisms of extinction in the desmostylians. However, loss of the taxon coincided in time with the decline of sirenians, with which desmostylians share common ancestry. Thus, habitat needs and associations may have been similar between late Miocene sirenians and desmostylians, and loss of optimal habitats during Pliocene cooling could have had similar effects on both groups.

## B. Modern Anthropogenic Extinctions of Species, Subspecies, and Major Populations

### 1. Species Level

Following the taxonomic format of Rice (1998), three species of modern marine mammal are known to be extinct. In all cases, the probable causes are anthropogenic. In addition, there is evidence that a recent species of pinniped, undescribed and entirely unknown to science, once occurred in the Chagos Archipelago and Seychelles Islands of the tropical southwestern Indian Ocean (Rice, 1998). If such a species occurred, it is now extinct as a result of unknown factors. There is substantial skepticism that such a species ever existed.

#### a. Caribbean Monk Seal: *Monachus tropicalis* (Gray, 1850)

The Caribbean monk seal is one of three recent species of *Monachus*. All occur in tropical or subtropical latitudes. The preexploitation range of *M. tropicalis* included the islands of the Caribbean Sea, the coastal regions of Venezuela and Caribbean Colombia, southern Florida, the east coast of Mexico south of the Bay of Campeche, and the Caribbean coasts of Belize, Guatemala, Honduras, Nicaragua, Costa Rica, and Panama. Estimates of preexploitation population sizes are not available. Intensive hunting of seals for meat and oil began with the arrival of Columbus in 1492. Seal populations were reported as depleted as early as the seventeenth century, but a few animals survived into the

middle of the twentieth century. The last confirmed sighting of a Caribbean monk seal was at Seranilla Bank, west of Jamaica, in 1952. Directed surveys for seals in the later 1950s and 1960s found no living animals.

#### b. Steller's Sea Cow: *Hydrodamalis gigas* (Zimmerman, 1780)

Steller's sea cow was first observed by a scientist in 1741, at which time it occurred only along the shorelines of the Commander Islands, east of Kamchatka in the Russian Far East. At the time of discovery, sea cows probably numbered no more than a few thousand individuals in total, but population surveys were never done. There are reports that the sea cow had previously occurred along the eastern shore of mainland Kamchatka and in the Near Islands of the Aleutian Archipelago, but the primary evidence is stranded material that could have resulted from the drift of carcasses. Sea cows were subject to intensive hunting, soon after discovery, by crews of sea otter hunters working in the Commander and Aleutian Islands. Harvest of sea cows provided high-quality meat and blubber in great quantity and facilitated extended harvesting expeditions by otter hunters. The last sea cow observation was recorded in the Commander Islands in 1767. The only scientific observations of living Steller's sea cows were made by G. W. Steller while shipwrecked at Bering Island in the Commander Islands during the winter of 1741–1742.

Three factors probably contributed to the extinction of Steller's sea cow. First, directed hunting was intensive and unmanaged. Second, at the time living sea cows were observed by Steller, they occurred only on islands lacking aboriginal human populations. The pattern suggests that aboriginal hunters may have previously reduced the range and numbers of sea cows, predisposing them to extinction when hunting intensified. Third, sea cows foraged on nearshore benthic kelps. The rapid depletion of sea otter populations by otter hunters in the Commander Islands may have allowed sea urchins, the primary prey of otters in the region, to overgraze their preferred food, the same kelps utilized by sea cows. Thus, catastrophic loss of food supply could have facilitated the rapid extinction of sea cows after reduction of population size(s) and range.

#### c. Japanese Sea Lion: *Zalophus japonicus* (Peters, 1866)

The Japanese sea lion is sometimes considered a subspecies (*Z. californianus japonicus*) of the California sea lion. Rice (1998) regards the Japanese sea lion as a separate species. The Japanese sea lion originally ranged

along the shores of Japan and Korea and the southern Pacific shores of Russia. The species was subject to a long history of hunting for meat and oil. No population surveys were done. The sea lion was thought to be extinct at the end of the nineteenth century, but a group of animals was reported from a Japanese island in 1952. There have been no subsequent sightings and the species is now considered extinct. All available data indicate that directed harvest was the primary factor contributing to extinction, but other unreported factors could have facilitated loss of the species.

## 2. Subspecies or Population Level

### a. North Atlantic Gray Whale: *Eschrichtius robustus* (Lilljeborg, 1861)

Whaling records and subfossil remains indicate that a population of gray whales (*Eschrichtius robustus*), apparently not taxonomically distinct from North Pacific populations, once occurred along both coasts of the North Atlantic. Available evidence suggests that gray whales occurred along the Atlantic coast of North America into the seventeenth century but probably not beyond. We are not aware of any available data on population size, foraging or breeding habitats, or migratory corridors. Subfossil specimens have been found in Europe from the central Baltic coast of Sweden to Cornwall in the United Kingdom. In North America, subfossil finds range from New York to South Carolina. In our opinion, the most likely explanation for extinction is prolonged excessive harvest by the whaling industry, but a definitive explanation for extinction does not exist.

## C. Species, Subspecies, and Populations in Imminent Peril of Extinction

### 1. Synopsis

Here, we summarize the status of 11 marine mammal taxa or populations that in our opinion face a substantial probability of extinction during the twenty-first century (Table III). We provide more detailed summaries for five examples of the category, chosen arbitrarily based on our relative familiarity with the cases. In all but one of the eleven cases, available data suggest total abundances of less than 1000 individuals. Most of the taxa or populations considered in this category have several significant past or current anthropogenic sources of mortality that facilitate extinction risk. In nearly all cases, the scope of operational resource investment and human societal adjustment necessary to avoid extinction seems to us very high.

### 2. North Pacific and North Atlantic Populations of the Right Whale, *Balaena glacialis* Müller, 1776

Right whales occur in three major regions: the North Pacific, the North Atlantic, and the Antarctic Ocean, together with adjoining regions of the South Pacific, South Atlantic, and South Indian Oceans. Because of intervening landmasses and the antitropical distribution of the species, rates of migration and genetic exchange among the three regions are very low. The primary conservation problem for the northern populations is very low population sizes, imposed largely by centuries of unregulated commercial whaling.

Right whales were identified by the earliest whalers as targets of choice because of their abundance, large size, high yields of meat and blubber, relatively docile behavior, and relative buoyancy postmortem. Historical records suggest that significant exploitation of the North Atlantic populations began along the European coast early in the second millennium A.D. Changes in hunting effort over time and space followed the stereotypical pattern of overexploitation. As coastal European stocks of whales were depleted, whalers expanded efforts westward to Greenland, Newfoundland, and Labrador beginning in the fifteenth century. Subsequent stock depletion led to further expansion southward in the seventeenth century to the waters of Nova Scotia and the United States. Right whale populations off the northeastern United States were depleted to low numbers after the middle eighteenth century. The historical record of right whale harvest is less lengthy for the North Pacific, but it almost certainly followed the same general pattern. Commercial whaling was most intensive in the nineteenth century. What ever recovery of North Pacific right whales occurred during the twentieth century was damaged by substantial illegal harvest by whalers from the USSR during the 1960s.

Approximately 300 individual right whales are thought to be currently present in two populations in the North Atlantic. The eastern population, off the coast of Europe, probably contains only a few individuals and is in extreme jeopardy of extinction. The western population is found mainly along the coast of Canada and the United States. During the twentieth century, estimated annual growth rates of the western population have never exceeded 2.5%, and the population is now declining. The two populations recognized in the North Pacific are also quite small, although estimates of size and trend have not been made. The western population, ranging from the Sea of Okhotsk to mainland China, is thought to be somewhat larger than the eastern population, which ranges from the eastern Be-

TABLE III  
Species, Subspecies, or Populations Thought to Be in Imminent Peril of Extinction

Taxon or population	Range/habitat	Population identity structure	Estimated population size	Primary risk factors
Eastern population of the North Pacific right whale: <i>Balaena glacialis</i> Müller, 1776	Pelagic, eastern margins of North Pacific at middle and high latitudes	One population	No survey data, perhaps a few hundred animals	Commercial whaling, including illegal Soviet whaling; incidental take in fishing nets
North Atlantic right whale: <i>Balaena glacialis</i> Müller, 1776	Pelagic, eastern and western margins of north Atlantic at middle and high latitude	Two populations—eastern and western	300	Commercial whaling, ship strikes, incidental take in fishing nets
Davis strait, Hudson Bay, Spitsbergen Barents Sea, and Sea of Okhotsk populations of the bowhead whale: <i>Balaena mysticetus</i> Linnaeus, 1758	Coastal and pelagic, in regions as indicated	Four separate populations	Davis Strait and Hudson Bay, about 450 combined; Spitsbergen, less than 50; Okhotsk, a few hundred	Commercial whaling, disturbance from offshore petroleum exploration and development
Western North Pacific gray whale: <i>Eschrichtius robustus</i> (Lilljeborg, 1861)	Coastal, western Pacific, northern Sakhalin Island, Russia, during summer; winter range thought to be off Korea or China	One population	<100	Commercial whaling, offshore petroleum exploration and development in summer range
Vaquita: <i>Phocoena sinus</i> Norris and McFarland, 1958	Coastal, Northern Gulf of Mexico, California	One population	567 (95% CI: 177–1073) in 1997 survey	Incidental take in fishing nets
Baiji: <i>Lipotes vexillifer</i> Miller, 1918	Aquatic in lower and middle Yangtze River, China	Population structure unknown	<100	Incidental take in setline fisheries, directed harvest for oil, contaminants, watercourse impoundments and diversions, habitat loss for prey
Indus River population of the Indian river dolphin: <i>Platanista gangetica</i> (Roxburgh, 1801)	Aquatic in main channels and tributaries of the Ganges, Indus, Brahmaputra, and Karnaphuli Rivers of India, Bangladesh, Nepal, and Bhutan	Population structure unknown; populations in different watersheds are isolated from one another	No data; population is thought to be very small	Incidental take in fishing gear, directed harvest for meat and oil, contaminants, watercourse impoundments and diversions, habitat loss for prey
Gulf of Alaska population of beluga whale: <i>Delphinapterus leucas</i> (Pallas, 1776)	Coastal marine habitats and river mouths, primarily in Cook Inlet, Alaska; also along the coast of the central Gulf of Alaska	One population	Approximately 350	Hunting by indigenous peoples, commercial whaling, contaminants, habitat disturbance
Gulf of St. Lawrence population of beluga whale: <i>Delphinapterus leucas</i> (Pallas, 1776)	Estuarine and coastal, St. Lawrence River estuary and Gulf of St. Lawrence, Canada	One population	1238 (standard error = 119) in 1997	Commercial whaling, culling to protect fisheries, contaminants
Mediterranean monk seal: <i>Monachus monachus</i> (Hermann, 1779)	Coastal in Mediterranean Sea and eastern warm-temperate North Atlantic; currently limited primarily to coastal areas of Turkey, Greece, western Sahara, and Mauritania	At least two populations likely: Cabo Blanco in Mauritania, western Sahara, and eastern Mediterranean	275–460 total	Directed harvest for meat and skins, illegal hunting, habitat destruction, overfishing of prey, contaminants
Lake Saimaa ringed seal: <i>Pusa hispida saimensis</i> (Nordquist, 1899)	Aquatic in Lake Saimaa, Finland; river connection to the Gulf of Finland is thought to be too swift to allow passage of seals	One population	Approximately 200	Incidental take in fishing nets

ring Sea to southern Baja California, Mexico. The eastern population may have suffered irreparable damage from the illegal Soviet whaling noted previously.

Failure of northern right whale populations to recover from the cessation of commercial whaling has been difficult to understand, excepting the obvious damage from recent illegal whaling in the eastern North Pacific. Southern right whale populations protected from exploitation have grown at rates estimated as high as 6 or 7% per year, and there is no evidence of major differences in fundamental vital rates between northern and southern populations. Thus, other factors must be retarding population growth in the north. The current consensus view is that two factors, incidental entanglement in fishing gear and inadvertent collisions with large commercial ships, are the primary causes for failure of North Atlantic populations to grow in recent decades. These right whales tend to be concentrated in regions that support productive and highly capitalized fisheries, facilitating damaging rates of incidental entanglement. Right whales are also concentrated in areas frequently transited by large ships. Vulnerability to ship strikes is enhanced by the apparent tendency of right whales to rest quietly at the surface for long periods. Given the long period of depletion, factors in addition to bycatch and shipstrikes may have contributed to failed recovery over the long term. For example, social dysfunction resulting from inability to find potential mates (the Allee effect) may diminish the growth potential of a small population.

Survival of northern right whale populations beyond the twenty-first century requires a major reduction of all forms of directed harvest and the largest possible reductions in rates of incidental taking and ship strikes. Eastern populations in the North Atlantic and North Pacific may be destined for extinction within the century regardless of actions taken.

### 3. Western North Pacific Gray Whale, *Eschrichtius robustus* (Lilljeborg, 1861)

The western North Pacific population of gray whales summers in the Sea of Okhotsk and probably migrates south to winter habitats in southern China. Very little is known about the ecology of the population, and focused studies have been done only within the past 5 years. The western gray whale population is among the smallest of large whale populations in the world and is one of the most vulnerable to extinction in the twenty-first century. The principal conservation concerns for western gray whales are the very small population size, resulting from both commercial whaling and subsistence harvests in previous centuries and risks posed by

offshore oil development in the current summer feeding range of the population.

Most of the available information on the western gray whale population comes from recent studies summarized by Weller *et al.* (1999). Western gray whales forage on benthic invertebrates during summer months in shallow coastal waters off northeastern Sakhalin Island, Russia. Foraging activity is particularly intensive in a small nearshore area off the entrance channel to Piltun Lagoon (52° 50' N, 143° 20' E). Data from photo-identification studies indicate high fidelity of foraging gray whales to the area north and south of the lagoon, within and among years, during the period 1995 to 1999.

Analyses of photographic data indicate that as few as 90 individual whales regularly use the summer foraging area off northeastern Sakhalin Island. Based on current data, the population numbers near 100 individuals. Data from the recent studies indicate that the crude birth rate in the Sakhalin summering aggregation was 4.3% in 1997 and 13.2% in 1998. Adult females with calves show particularly high fidelity to the feeding area near Piltun Lagoon, suggesting that the area is important to calf rearing and weaning.

The summer feeding area near Piltun Lagoon has been subject to intensive exploration in recent years and production of oil began in 1999 at offshore petroleum production facilities. Offshore petroleum exploration involves frequent use of intensive low-frequency sounds, and development of located petroleum resources involves increased shipping activity, placement of structures, modification of adjacent sediments and benthic communities, and the risk of unintended spills of drilling fluids, vessel and machinery fuels, and extracted crude oil.

The western North Pacific population of gray whales is in imminent peril of extinction. Recovery planning for the population will benefit from acquisition of significant additional data. The winter range, calving grounds, and migratory corridors of the population must be determined, and the associated risk factors must be identified and analyzed for significance with regard to extinction. The affinity of summering animals, and particularly females with dependent calves, for the feeding grounds off Piltun Lagoon must be better understood. Activities associated with development of petroleum resources in the coastal zone of northeastern Sakhalin Island must be carefully evaluated in the context of risks to gray whales using the area for feeding. All forms of harvest of western gray whales must be prevented indefinitely if extinction is to be avoided.

#### 4. Gulf of Alaska Beluga Whale, *Delphinapterus leucas* (Pallas, 1776)

The Gulf of Alaska population of beluga whales has always concentrated primarily in Cook Inlet on the southern mainland coast of Alaska. Animals have been seen on occasion in coastal waters of the Kodiak Archipelago, in Prince William Sound and Yakutat Bay, and along the outer coastal waters of the central Gulf of Alaska. However, it is not known if these animals are part of the Cook Inlet population. Historically, belugas have been widely distributed in Cook Inlet, a well-mixed, turbid, highly productive marine inlet that receives runoff from several large river systems. The whales concentrate seasonally near and in the river mouths to forage on migrating anadromous fish.

Belugas in Cook Inlet were hunted at low levels in previous decades by commercial whalers, although commercial harvest no longer occurs. Passage and implementation of MMPA in 1972 ensured the rights of native peoples in Alaska to pursue the traditional practice of hunting beluga whales, including those in Cook Inlet. Hunting of belugas in Cook Inlet by natives increased in the 1980s and 1990s compared to earlier decades. During the 1990s, it was recognized that the geographic range of beluga whales, based on opportunistic observations from research vessels, was much smaller than it had been in the 1970s. Few sightings of belugas in other locations of the Gulf of Alaska have been made in recent years. Intensive surveys in the 1990s confirmed that the population was declining. The population estimate for 1998, determined by NMFS, was 347 (CV = 0.29). Based on known whale kills by native hunters and evaluation of survey data, it was determined that native harvest of belugas may be excessive, possibly placing the Cook Inlet beluga population at risk of extinction. As a result, a temporary moratorium on native harvest of belugas in Cook Inlet was implemented in 1999. Native tribal leaders and U.S. federal agencies are now planning for recovery actions and a more carefully managed harvest.

The Cook Inlet beluga population is the first known modern case in which excessive harvest by indigenous North American peoples has placed a population of marine mammals in jeopardy. Other populations of belugas are exploited regularly by natives in Alaska, but none other than the Cook Inlet population is in peril. The tenuous status of Cook Inlet belugas, compared to other Alaskan populations, probably results from three factors. First, the Cook Inlet population has probably always been the smallest of the Alaskan beluga populations. Second, Cook Inlet belugas were subject to mod-

est commercial whaling in previous decades. Third, the cultural characteristics of native harvest of belugas in Cook Inlet differ from those of all other locations in Alaska where native whaling occurs. In all locations except Cook Inlet, native whaling is done from single coastal villages using practices consistent with lengthy tradition, subject to stringent oversight by village elders, and based on the subsistence needs of the village. In contrast, most native whalers working in Cook Inlet have moved from their home villages to the relatively large and secular city of Anchorage. Thus, it appears that cultural norms and limits on whaling activity characteristic of native Alaskan villages have been lost in the case of Cook Inlet. The consequence has been loss of traditional control over the harvest.

There is concern about risks of contaminants and marine oil activity for belugas in Cook Inlet. However, despite substantial offshore oil production activity, there is no firm evidence of negative effects of oil production on belugas in Cook Inlet, nor is there substantive evidence of negative effects of chemical contaminants. We suggest that the primary conservation concern for Cook Inlet belugas at the present time is the risk of excessive harvest by native hunters disconnected from the cultural regulation of their respective home villages. In our opinion, harvests of belugas in Cook Inlet must remain suspended until population recovery is apparent and a pragmatic harvest comanagement plan supported by tribal and agency authorities is in place. Otherwise, extinction is likely within the century.

#### 5. Vaquita, *Phocoena sinus* Norris and McFarland, 1958

The vaquita is a small porpoise limited to the northern part of the Gulf of California, Mexico. The vaquita has the smallest natural geographic range of any marine cetacean, and the single population probably has never contained a large number of individuals. Numbers were estimated at about 600 individuals based on a comprehensive survey in 1997 (Table III). The population biology of the vaquita is poorly known, and the very best surveys of the vaquita population have low precision. Thus, trends in the population are difficult to discern with confidence, but available data indicate that the population may be declining. Studies of DNA in sampled vaquitas indicate the complete absence of polymorphisms in the hypervariable region of the mitochondrial genome.

During the last two decades, the primary concerns for vaquita conservation have been high rates of incidental take in gillnet and shrimp trawl fisheries, effects of contaminants, effects of reduced genetic diversity,



and effects of diversion of the Colorado River away from the northern Gulf of California on regional rates of biological productivity. The significance of the listed risk factors has been evaluated by Rojas-Bracho and Taylor (1999). Losses due to incidental take were identified as the primary conservation problem for the vaquita. Concentrations of organochlorines, including PCBs and DDT congeners, were found to be low in vaquita tissues, low in other consumer species within the vaquita range, and low in the vaquita habitat. Organochlorine levels in the region are generally below minimum levels considered harmful to human health and are presumed to be innocuous for vaquitas we well. The levels of reduced genetic variability in vaquitas do not necessarily result in genetically based reduction of reproductive rates, particularly in the context of populations such as vaquitas that probably have always been small. Biological productivity in the upper Gulf of California is high compared to other coastal marine ecosystems despite diversion of the Colorado River. Although other risk factors may be detrimental to vaquita conservation over the long term, mortalities due to incidental take are the primary current problem.

Continued incidental take of vaquitas will cause extinction of the species during the twenty-first century, given even the most conservative estimates of the current rate of take. Elimination of the risk of anthropogenic extinction requires significant reduction in the level of fishing effort or changes in gear design or deployment strategy to reduce take rates in fisheries responsible for incidental take.

#### 6. Mediterranean Monk Seal, *Monachus monachus* (Hermann, 1779)

The Mediterranean monk seal was found originally in the western Black Sea, throughout the Mediterranean Sea, and along the coast of northwestern Africa from the Strait of Gibraltar to about 21°N latitude. Currently, there are thought to be no more than 275–460 individuals, occurring primarily in two populations. Prior to 1997, the largest group of approximately 300 seals occurred in a small population at Cabo Blanco, at the border of the western Sahara and Mauritania, on the outer coast of northwestern Africa. A second population of unknown size occurs in the eastern Mediterranean, primarily in the coastal waters of Turkey and Greece.

Mediterranean monk seals probably have been subject to directed subsistence harvest for meat, oil, and hides for several millennia. The precarious status of modern populations seems to result from many factors associated with the large, multicultural human populations of southern and eastern Europe and northern

Africa. For years, monk seals have been perceived as direct competitors of fisheries and have been harassed and killed in substantial numbers, often illegally, as a result. Harassment has included directed destruction of caves and other shoreline locations favored by seals for breeding and resting. Monk seals likely have also been affected by loss of prey due to overfishing and to various forms of contamination of the habitat and food webs. In 1997, a mass mortality event was observed in the colony at Cabo Blanco, reducing the local seal population to about 100 individuals. The cause and magnitude of the event have not been determined to our knowledge. Although comprehensive demographic and population survey data are lacking, the consensus opinion is that the total number of Mediterranean monk seals is probably declining over time.

In addition to the small size of the two known populations, two factors add great difficulty to the prospects for implementation of a successful recovery strategy for Mediterranean monk seals. First, the habitat of the monk seal is bounded by many culturally disparate political jurisdictions. Historically, the political and cultural diversity has interfered with cooperation across jurisdictions. Thus, the attainment of consistent, broadly supported conservation priorities for monk seals may be an unrealistic political objective. Second, ongoing damage to the monk seal populations apparently results from many factors acting in concert rather than one clearly predominant problem. Thus, agreement on conservation priorities and actions may be difficult even within jurisdictions.

Mediterranean monk seals appear to be destined for extinction, possibly within the twenty-first century, unless marine conservation authorities in countries bordering seal habitat can agree on two issues. First, risk factors for the seals must be evaluated dispassionately and placed in order of significance. Second, involved authorities must agree on a plan for recovery of seal populations based on the assessment of risk factors and convince the human populations of their respective jurisdictions that seal conservation is a worthwhile objective.

### D. Species, Subspecies, and Populations of Significant Concern with Regard to Extinction

#### 1. Synopsis

Here, we consider a group of taxa and populations that, we believe, are of significant concern with regard to extinction (Table IV). In contrast to the discussion in

TABLE IV  
Species, Subspecies, and Populations of Significant Concern with Regard to Extinction

Taxon or population	Range/habitat	Population identity/structure	Estimated population size	Primary risk factors
Blue whale: <i>Balaenoptera musculus</i> (Linnaeus, 1758)	Pelagic	Three subspecies—two in Southern Hemisphere, one in Northern Hemisphere; five populations in North Pacific, two in North Atlantic, one in north-west Indian; number in Southern Hemisphere unclear	2000 in eastern North Pacific, 400–2000 in eastern North Atlantic; others unknown	Commercial whaling, including illegal Soviet whaling, ship strikes
Hawaiian monk seal: <i>Monachus schauinslandi</i> (Matschie, 1905)	Coastal leeward Hawaiian Archipelago	Five primary breeding sites on different islands; some exchange of individuals among islands	Approximately 1500	Directed harvest for meat, oil, and skins; disturbance by military conflict and peacetime human activity; entanglement in marine debris; declining ecosystem productivity; aberrant breeding behavior associated with anomalous sex ratio; natural toxins in prey
Baltic Sea ringed seal: <i>Pusa hispida botnica</i> (Gmelin, 1788)	Coastal, in areas with pack or shorefast ice at least part of the year	Probably one population	Unknown	Pollution and contaminants
Lake Ladoga ringed seal: <i>Pusa hispida ladogensis</i> (Nordquist, 1899)	Coastal, in areas with pack or shorefast ice at least part of the year	Probably one population	Unknown	Pollution and contaminants
Western North Pacific harbor seal: <i>Phoca vitulina stejnegeri</i> Allen, 1902	Coastal along the shores of the western Aleutian and Commander Islands, southeastern Kamchatka Peninsula, Kuril Islands, Sea of Okhotsk, and Hokkaido	Population structure unknown; island breeding colonies probably somewhat isolated, although exchange among colonies may occur	Unknown	Incidental take in fishing nets
Western North Pacific Steller's sea lion: <i>Eumetopias jubatus</i> (Schreber, 1776)	Coastal and pelagic, islands and isolated rocky shores of the western Gulf of Alaska, western North Pacific, and Bering Sea	Thought to be a single population; some isolation of breeding colonies probably occurs; exchange among breeding colonies is thought to be minimal	Approximately 40,000 individuals; decline in numbers of approximately 65% since the mid-1970s	Unknown; primary possibilities are large-scale declines in ocean productivity, changes in the species composition and diversity of prey species, and competition for prey with commercial fisheries
Australian sea lion: <i>Neophoca cinerea</i> (Péron, 1816)	Coastal and continental shelf waters of southern and southwestern Australia	Population structure unknown; long-distance movements are uncommon, so separate breeding colonies may be isolated	Approximately 5000 individuals	Commercial sealing, incidental take in fishing nets, illegal directed killing, entanglement in anthropogenic debris

continues

Continued

Taxon or population	Range/habitat	Population identity/structure	Estimated population size	Primary risk factors
Hooker's sea lion: <i>Phocarcctos hookeri</i> (Gray, 1844)	Coastal waters of New Zealand	Probably one population, breeding primarily at the Auckland Islands	10,000–12,000 individuals	Commercial sealing, incidental take in fishing nets
Guadalupe fur seal: <i>Arctocephalus townsendi</i> Merriam, 1897	Coastal and continental margin habitats off Baja California, Mexico, and southern California	Probably one population, breeding only at Guadalupe Island, Mexico	3000–4000 individuals	Commercial sealing
Juan Fernandez fur seal: <i>Arctocephalus philippii</i> (Peters, 1866)	Coastal and pelagic habitats of the Juan Fernandez Archipelago, Chile	Population structure unknown	7000–10,000 individuals	Commercial sealing
Atlantic walrus: <i>Odobenus rosmarus rosmarus</i> (Linnaeus, 1758)	Coastal habitats with extensive pack ice in the Arctic North Atlantic, from eastern Canada to the Kara Sea	Four separate populations are recognized	Unknown	Commercial hunting, illegal hunting
Laptev Sea walrus: <i>Odobenus rosmarus laptevi</i> Chapskii, 1940	Coastal habitats of the eastern Kara, Laptev, and western Eastern Siberian Seas	Probably one population	Unknown	Commercial hunting
Amazonian manatee: <i>Trichechus inunguis</i> (Natterer, 1883)	Aquatic habitats of the Amazon watershed, South America	Population structure unknown	Unknown; thought to be declining	Commercial hunting, hunting by indigenous peoples, watercourse impoundment and diversion, sedimentation associated with forest modification, contaminants associated with mining, incidental take in fishing nets
West African manatee: <i>Trichechus senegalensis</i> Link, 1795	Warm shallow coastal marine habitats and rivers of West Africa from Senegal to Angola	Population structure unknown	Unknown	Commercial hunting, hunting by indigenous peoples, killing to reduce damage to fishing gear and rice crops
West Indian manatee: <i>Trichechus manatus</i> Linnaeus, 1758	Warm shallow coastal marine habitats and rivers of the Georgia Bight, Florida, the Caribbean, and northeastern South America southward to central Brazil	Two subspecies—one in the southeastern United States, especially in Florida, and the other along the mainland east coast of central America from northern Mexico to central Brazil, and in the islands of the Caribbean Sea	Approximately 2000–3000 in Florida; trend in Florida populations unclear; numbers elsewhere unknown	Habitat destruction, modification, and disturbance associated with growing human populations in Florida; commercial hunting; hunting by indigenous peoples; contaminants; incidental take in fishing nets; ingestion of debris
California sea otter: <i>Enhydra lutris nereis</i> (Merriam, 1904)	Shallow marine habitats along exposed outer coast of California	One primary mainland population, plus small separate colony at San Nicolas Island	2000–2500 individuals; numbers declining at 1 or 2% per year	Commercial hunting, incidental take in fishing nets, contaminants, disease, parasites

continues

Continued

Taxon or population	Range/habitat	Population identity/ structure	Estimated population size	Primary risk factors
Marine otter: <i>Lutra felina</i> (Molina, 1782)	Shallow marine habitats and coastal aquatic habitats along the ex- posed outer coasts of Peru, Chile, and south- ern Argentina	Population structure un- known; most individu- als occur in southern Chile; locally extinct in Argentina	Unknown	Commercial harvest for pelts, killing to reduce competition with fish- eries, incidental take in fishing gear, con- taminants

<sup>a</sup>Some taxonomists consider the southern right whale a separate species, *Balaena australis* (Desmoulins, 1822). Here we follow the convention of Rice (1998), regarding the northern and southern right whales as one species.

Section II,C, we do not regard entries in this group to be in imminent peril of extinction during the twenty-first century. In most cases, population sizes are large enough and conservation issues tractable enough that less dire predictions seem reasonable. However, we suggest that vigilance and positive action will be required to prevent taxa and populations in this group from falling to a more precarious status. We list 17 taxa or populations (Table IV), providing more detailed summaries for four arbitrarily selected examples.

## 2. Blue Whale, *Balaenoptera musculus* (Linnaeus, 1758)

Blue whales occur in all the world's coastal and pelagic marine habitats. Currently, three subspecies are known. The pygmy blue whale (*Balaenoptera musculus breviceauda*) occurs in southern cool-temperate and subpolar latitudes. The "true" southern blue whale (*B. m. intermedia*) summers in the Antarctic Ocean, and the northern blue whale (*B. m. musculus*) is found in the North Pacific and North Atlantic. Eastern and western populations are known in the North Atlantic, and at least five populations have been described in the North Pacific. The population structure in the Southern Hemisphere is unclear. However, "true" blue whales must have different populations in the southern Indian, Atlantic, and Pacific oceans, respectively. Available data suggest that blue whales migrate seasonally, utilizing higher latitude habitats in summer for feeding, primarily on euphausiid crustaceans, and lower latitude habitats during winter for courtship, breeding, and parturition.

Blue whales are an obvious target of choice for commercial hunting because of their great body size. However, blue whales are swift swimmers and are negatively buoyant postmortem. Prior to exploitation, blue whales were most abundant in the Antarctic Ocean and other

marine habitats distant from human population centers. Thus, they were beyond the technological capabilities of commercial whalers prior to the twentieth century. Blue whales became priority targets of whalers only after the development of the steam engine, factory ships, explosive harpoons, and air compressors to inflate carcasses after killing. Thus, the harvest and depletion of blue whales occurred primarily during the twentieth century. At least 360,000 blue whales were killed in the Antarctic region before commercial whaling declined in the 1960s, and other populations were exploited as well. Illegal harvests by Soviet whalers occurred after a moratorium was imposed, taking at least 8000 additional pygmy blue whales.

The current consensus opinion is that blue whale populations are now but a fraction of preexploitation size, but there are few data available to defend the perception. Recent data indicate that the eastern North Pacific population numbers approximately 2000 individuals and may be increasing. The eastern North Atlantic population has been estimated at 400–2000 individuals. A circumpolar survey of the Antarctic Ocean estimated 710 (CV = 0.64) individuals. There are published arguments that blue whale populations off southern Japan and in the eastern Gulf of Alaska are locally extinct or very small. Because of small population sizes and large CVs associated with surveys, it is not possible to identify trends in most blue whale populations. Thus, the status of the world's blue whale populations is generally unknown, and prospects for confidently understanding the size of populations in the foreseeable future are virtually nil.

The current status of blue whale populations seems to be entirely the result of excessive past commercial harvests. There is evidence that ship strikes may cause some mortality of blue whales off the California coast,

but current anthropogenic mortalities are probably minimal. Recovery of small blue whale populations will require indefinite suspension of all forms of harvest and prompt detection and elimination of emerging sources of anthropogenic mortality.

### 3. Western North Pacific Steller's Sea Lion, *Eumetopius jubatus* (Schreber, 1776)

Steller's sea lions occur in coastal waters of the North Pacific Rim from southern California to northern Japan and in the Bering Sea and Sea of Okhotsk. Recent genetic data are the basis for dividing the species into two populations, western and eastern, with the boundary at Cape Suckling, Alaska (144°W longitude). The eastern population is dispersed along the west coast of North America, numbers more than 20,000 individuals, and is increasing gradually, particularly in southeastern Alaska. The western population occupies the Bering Sea, Aleutian and Commander Islands, and remote locations of the Russian Far East. The western population numbered approximately 150,000 animals in the 1950s but has since declined precipitously, with current numbers estimated at approximately 39,500. The rate of decline has varied over time, with highest rates (approximately 15% per year) from 1985 until 1990. The decline currently continues.

The cause or causes of decline in the western population of Steller's sea lions are not understood. Possible risk factors include incidental take in fishing gear, competition with fisheries for prey in common, hunting by indigenous peoples, illegal hunting or harassment, inadvertent rookery disturbance, disease or parasitism, predation by killer whales, contaminants, and changes in the structure and productivity in the marine ecosystems of which Steller's sea lions are a part. Based on extensive research since the decline was first recognized, the current consensus opinion is that ecosystem change or competition with fisheries are the most likely factors driving the decline. Resolution of the question of cause has become the focus of intensive political interest because of the potential economic consequences. The groundfish fisheries of the Gulf of Alaska and the Bering Sea are the most valuable and highly capitalized of the fisheries in the coastal waters of the United States. Steller's sea lions feed extensively on groundfish species, such as walleye pollock, targeted by fisheries. Determination that competition with fisheries is contributing to the decline could result in forced reduction of fishing effort, with great economic loss and political discord.

Several lines of evidence favor the argument that ecosystem change has contributed to the decline in sea

lion numbers. Prey species composition has changed, dietary diversity has declined, and there have been measurable shifts in the spatial distributions of preferred prey during the past three decades. Numbers of other piscivores in the habitat, including seabirds and harbor seals, have declined over a similar timescale in many parts of the sea lion range. However, fishing activity may be intensive near important sea lion breeding locations, and it has not been possible to eliminate competition with fisheries as a potential cause of the decline. Recent rulings in U.S. courts have restricted fishing activity in some locations and seasons for the purpose of reducing the rate of decline, thus intensifying the associated political debate.

Several viability analyses have been applied to Steller's sea lion population data. Model results lead to extinction of the western population in all cases. Median estimates of time to extinction range from 62 to 160 years. The combination of ongoing decline, projected extinction risks, and uncertainty about primary risk factors leads to significant concern about the persistence of the western population. Because natural oceanographic changes can neither be predicted nor controlled, management authorities have no choice but to focus on understanding and minimizing anthropogenic risk factors, despite the political consequences, in order to reduce the probability of eventual extinction.

### 4. West Indian Manatee, *Trichechus manatus* Linnaeus, 1758

West Indian manatees occur in coastal habitats and the lower reaches of rivers in the southeastern United States, the islands of the Caribbean Sea, and the mainland shores of the Gulf of Mexico, Central America, and northeastern South America. Two subspecies are recognized. The Florida manatee [*Trichechus manatus latirostris* (Harlan, 1824)] is found in U.S. coastal waters, especially in Florida. The Antillean manatee (*T. manatus manatus* Linnaeus, 1758) occupies the remainder of the range of the species. The current population in Florida is the largest of the species, numbering 2000–3000. Despite a broad perception of increasing numbers in Florida, population data lack the statistical power to document a positive trend with confidence. Population structure, numbers, and trends in other locations are not well-known, but most populations probably number no more than a few hundred animals.

Manatees have been hunted by indigenous peoples for meat and other products for centuries. Commercial hunting probably contributed to a reduction of population sizes, but apparently there are few data available to assess the rates or significance of commercial harvest.

Manatees have been protected from all forms of directed harvest in Florida since the 1960s, but subsistence harvest for meat and oil and for ceremonial purposes continues in other populations. During the twentieth century many risk factors for manatees emerged, all in association with an expanding human population in immediate proximity to manatee habitat. The primary problems are a variety of modifications and ongoing disturbances of habitat. The latter include watercourse diversions and impoundments in aquatic manatee habitat and disturbance and collision risk associated with recreational boating in aquatic and marine manatee habitats. There is also continuing concern about increased levels of contaminants.

Long-term studies of stranded carcasses indicate three major sources of mortality in Florida manatees. Perinatal mortalities are newborn animals with the proximate cause of death uncertain but possibly linked to contaminants. Significant mortality rates are also associated with entrapment in dam floodgates and collision with recreational powerboats. Manatees clearly prefer areas without significant powerboat traffic. The ongoing expansion of human populations and associated demand for recreational opportunities inevitably leads to continuing reduction in the size of available manatee habitat.

Manatees are sensitive to water temperature and typically congregate in warm-water refugia during winter, especially when ambient sea surface temperatures drop below approximately 18°C. Refugia include natural warm aquatic springs and lagoons associated with thermal effluent from electrical generating plants. The limited size and number of natural refugia may be threatened by the encroaching effects of human development and activity. Use of power plant lagoons may be risky if plant operations dictate precipitous shutdown, resulting in cutoff of thermal effluent and rapid chilling of lagoons.

Manatees in Florida have experienced several significant mass mortalities in recent years. The mortality events apparently result from an interaction of compromised immune systems, disease, and natural toxins associated with phytoplankton blooms.

Habitat loss and disturbance is the primary conservation problem for all populations of West Indian manatee, and subsistence or illegal harvest remain significant risk factors for populations outside of Florida. Over the long term, avoidance of extinction will require cessation of all forms of human harvest and an effective, broadly supported strategy for balancing the habitat needs of manatees with the consequences of human population growth, especially in Florida.

#### 5. California Sea Otter, *Enhydra lutris nereis* (Merriam, 1904)

Sea otters originally ranged throughout the coastal waters of California, including San Francisco Bay and the southern California islands. The population was originally contiguous with other otter populations ranging from the central outer coast of Baja California, Mexico, through the North Pacific Rim to northern Japan. Sea otters in California comprise one of three subspecies currently recognized.

Sea otters have probably been hunted by indigenous peoples of the North Pacific Rim for several millennia for meat and pelts and for ceremonial purposes. Observations of the Bering Expedition of 1741 and 1742 and other voyages of exploration in the North Pacific found abundant sea otters throughout their range. Commercial harvest of sea otters for pelts began with the Bering Expedition. In California the hunt was pursued by Russians, often utilizing enslaved Aleut hunters, and by hunters from Spain, Mexico, and the United States. All commercial hunts for sea otters were terminated in 1911 with approval of the Treaty for the Preservation and Protection of Fur Seals (37 Stat. 1542, T.S. No. 542) by Japan, Russia, Great Britain, and the United States. The treaty included a passage affording protection to sea otters. However, in California and Mexico the sea otter populations were depleted commercially by the 1860s. By 1900, only two small populations survived. One was along the Big Sur coast south of the Monterey Peninsula, numbering approximately 50–100 animals. The second, of unknown size off the San Benito Islands, Mexico, was extinct by the 1920s. The Big Sur population grew at a rate of approximately 5% per year through much of the twentieth century and now numbers approximately 2000–2500 animals, ranging from San Francisco southward to Point Conception. The observed rate of growth in California has been much lower than rates commonly observed in more northerly populations with protection from harvest and available adjacent vacant habitat. In the late 1980s a new colony was established by translocation at San Nicolas Island off southern California. The colony now numbers approximately 20 animals and persistence remains uncertain.

Several risk factors are known for sea otters in California. Nearshore net fisheries are responsible for significant rates of incidental take. Changes in fishery regulations have reduced but not eliminated incidental take. Sea otters are known to compete with nearshore marine shellfish fisheries in California, particularly for abalones, sea urchins, clams, and crabs. Concern about

interactions of sea otters and shellfisheries is sufficient that it was until recently the policy of the U.S. federal government to actively remove and relocate individual otters found along the California coast south of Point Conception in order to minimize damage to shellfisheries. The management effort was done despite the small size of the sea otter population, its precarious status, and various state and federal legal protections including a listing as "Threatened" under the auspices of the U.S. Endangered Species Act of 1973 (ESA) [16 U.S. Code §§1531-43 (Supp. IV 1974)] as amended. Relocation efforts were inefficient and costly, and alternative management protocols are now in active development. There is also concern about illegal killing of sea otters to protect shellfisheries, but little direct evidence of a significant problem. The California sea otter population is also regarded to be at high risk of potential damage from oil spills, although oil spills have yet to cause sea otter mortality in California. Oiling mats the pelage and eliminates thermoregulatory function, rapidly leading to chilling and death. Oiling is also known to cause other respiratory and physiological pathologies.

During the 1990s, the growth of the population slowed, and currently the population is declining at 1 or 2% per year. A clear consensus on the cause of the decline has not emerged, but primary risk factors are thought to be ongoing incidental taking in fishing nets, effects of contaminants, and the emergence of mortalities from diseases and exotic parasites not previously associated with sea otters. The long-term survival of the sea otter population in California will require ongoing research to more effectively characterize current risk factors and the development of strategies to minimize associated mortalities. The large and growing human population in California is a major underlying cause of the jeopardy status of sea otters, and broad popular support, including economic compromise by shellfishery and marine oil interests, likely will be required if recovery is to be successful.

## E. Species, Subspecies, and Populations Once Thought to Be Near Extinction But Now Showing Evidence of Recovery

### 1. Synopsis

Here, we summarize data for eight species, subspecies, or populations of marine mammals that have been near extinction in the recent past but are now either recovered or en route to recovery (Table V). We regard the taxa or populations in this group as unlikely to become extinct in the foreseeable future as long as risk factors

currently unknown do not appear. We present case summaries for five taxa, selected arbitrarily, to illustrate patterns typical of the group.

### 2. Western Arctic Population of the Bowhead Whale, *Balaena mysticetus* Linnaeus, 1758

Bowhead whales occur in at least five populations in northern habitats characterized by frequent sea ice. Four of the five populations are small and at risk of extinction (Table III). The western Arctic population is by far the largest of the bowhead populations, numbering approximately 8000 individuals with annual growth rates averaging 2%. The population ranges in the Bering, Chukchi, and Beaufort Seas, migrating northeasterly to the Beaufort Sea in spring and returning to the Bering and Chukchi Seas in the fall.

The western Arctic population of bowhead whales probably numbered approximately 23,000 individuals when first exploited by U.S. whalers during the late 1840s. By the demise of the commercial harvest in 1919, approximately 20,000 whales had been taken, and the population was thought to include approximately 1000 individuals.

Bowhead whales have been hunted for subsistence and ceremonial purposes by the indigenous peoples of Alaska, western Canada, and Russia for many centuries. Russian and Alaskan native villages continue to harvest whales from the western Arctic population based on a quota approved by the International Whaling Commission (IWC) and, in Alaska, managed jointly by representatives of native villages and U.S. federal agencies. Management policies have been successful, allowing continued growth of the population despite annual hunts for whales by several villages. In recent years, the annual native harvest has been 20–50 whales in Alaska. For many years, there has been concern about effects of marine petroleum exploration and development activities in the habitat of bowhead whales, especially in the Beaufort Sea. Offshore oil activity is known to influence movement patterns of whales during migration, but demographic effects have not been demonstrated to our knowledge.

The western Arctic bowhead whale population appears to be recovering while other bowhead whale populations are not. The most likely reason is that other populations were pushed much closer to extinction by commercial whaling, limiting the capacity for recovery. It appears that the western Arctic population will continue to recover as long as harvests by native villages are regulated conservatively and other significant risk factors do not emerge.

TABLE V

Species, Subspecies, or Populations Once Thought to Be Near Extinction But Now Showing Evidence of Increasing Numbers

Taxon or population	Range/habitat	Population identity/structure	Estimated population size	Primary risk factors
Western Arctic population of the bowhead whale: <i>Balaena mysticetus</i> Linnaeus, 1758	Coastal and pelagic, Arctic and subarctic	One population	Approximately 8000	Commercial whaling, disturbance from offshore petroleum exploration and development
Humpback whale: <i>Megaptera novaeangliae</i> (Borowski, 1781)	Coastal and pelagic in all oceans and seas except the Arctic polar region	Five populations in the Southern Hemisphere, three in the North Pacific, one in the North Atlantic, and one in the Arabian Sea	Recent surveys in the North Pacific indicate 6000–8000 individuals, with aggregate population growth rates of approximately 7% per year; numbers are not well-known for other populations	Commercial whaling, incidental take in fishing nets, habitat disturbance from recreational vessels
Eastern North Pacific gray whale: <i>Eschrichtius robustus</i> (Lilljeborg, 1861)	Coastal from the Bering and Chukchi Seas southward to the southern Gulf of California, Mexico, along the west coast of North America	One population	25,000–30,000 individuals; population growth rate may be declining as the population approaches carrying capacity	Commercial whaling, incidental take in fishing nets, habitat disturbance from industrialization and recreational boating in and near breeding lagoons in Mexico and along migration route
Northern elephant seal: <i>Mirounga angustirostris</i> (Gill, 1866)	Breeding and molting seasons: coastal, on island and mainland haul-outs along the Pacific coast of northern Mexico and California; Other seasons: Pelagic in cool-temperate and subpolar latitudes of the North Pacific	Separate breeding colonies at each of the major haulouts; exchange among colonies is known to occur	Approximately 200,000 individuals, increasing at approximately 6% per year	Commercial sealing, disturbance of breeding colonies
Galapagos fur seal: <i>Arctocephalus galapagoensis</i> Heller, 1904	Coastal and pelagic habitats of the Galapagos Archipelago	Separate breeding colonies at each of the major haulouts; exchange among colonies is known to occur	Approximately 30,000–35,000 individuals; thought to be increasing in numbers	Commercial sealing, habitat disturbance
Subantarctic fur seal: <i>Arctocephalus tropicalis</i> (Gray, 1872)	Subantarctic islands north of the Antarctic convergence zone in the Antarctic Ocean	Separate breeding colonies at each of the major haulouts; population structure is otherwise unknown	Approximately 300,000–400,000 individuals; thought to be increasing in numbers	Commercial sealing, habitat disturbance
Antarctic fur seal: <i>Arctocephalus gazella</i> (Peters, 1875)	Subantarctic islands south and, in a few cases, slightly north of the Antarctic convergence zone in the Antarctic Ocean	Separate breeding colonies at each of the major haulouts; population structure is otherwise unknown	Approximately 1,200,000 individuals; thought to be increasing in numbers	Commercial sealing, habitat disturbance, entanglement in marine debris
Northern and Russian sea Otter: <i>Enhydra lutris kenyoni</i> Wilson, 1991, and <i>E. l. lutris</i> (Linnaeus, 1758)	Shallow marine habitats along the exposed outer coast	Northern subspecies ranges from western Aleutian Islands to the Columbia River mouth; Russian population ranges through the Commander and Kuril Islands, southern Kamchatka, and northern Hokkaido	Approximately 100,000 individuals in Alaska and approximately 20,000 individuals in Russia; data for many Alaskan locations are obsolete; populations in the Aleutian Archipelago may be declining rapidly	Commercial hunting, incidental take in fishing nets, contaminants, increased predation rates by killer whales



### 3. Humpback Whale, *Megaptera novaeangliae* (Borowski, 1781)

Humpback whales are distributed globally in marine habitats. At least five populations are recognized in the Southern Hemisphere, three in the North Pacific, and two in the North Atlantic. An unusual nonmigratory population occurs in the Arabian Sea. Humpback whales often are found near shore. Lengthy seasonal migrations between high-latitude summer feeding areas and low-latitude winter breeding areas are well documented for all populations except as noted.

Humpback whales were among the earliest target species of commercial whalers because of their abundance, large size, and coastal distribution. During twentieth-century commercial hunts in the southern ocean, several hundred thousand individuals were taken. Illegal harvest by whalers from the USSR was extensive and damaging to populations following the decline of large-scale legal commercial whaling. Subsistence harvest by indigenous peoples was common in previous centuries. Additional modern risk factors include incidental take in fishing nets and habitat disturbance by large recreational cruise ships.

Although trends in most populations have not been confidently documented, there is a broad perception that humpback whale numbers are increasing on a global scale. Recent estimates are 9000–12,000 individuals in the North Atlantic and a combined total of 6000–8000 individuals in the North Pacific. Data from the Southern Hemisphere and the Arabian Gulf are not sufficient to develop confident population estimates. There is concern that the population wintering near Tonga, in the western South Pacific, may have experienced excessive subsistence harvests in recent decades, although the hunt has been stopped. The eastern and western North Pacific population remain small enough to be of ongoing concern. The relatively small Arabian Gulf population came to be known because of extensive Soviet whaling in the 1960s. Recent concerns have been expressed over sustainability of current levels of incidental taking by local fisheries. We suggest that prospects for survival of humpback whales, as a species, are good as long as commercial harvests are not resumed, incidental taking is reduced, and habitat disturbance by human activities continues to be restricted and carefully monitored.

### 4. Eastern North Pacific Gray Whale, *Eschrichtius robustus* (Lilljeborg, 1861)

The eastern North Pacific population of gray whales ranges from the Bering and Chukchi Seas of the Arctic

region during summer to the shallow protected coastal lagoons of Baja California, Mexico, during winter months. The annual migration of this population is among the lengthiest of the world's mammals and is the best documented and most familiar of the cetaceans.

Eastern gray whales were hunted intensively by commercial whalers from the eighteenth through the early twentieth centuries. Preexploitation estimates of population size are inconsistent and difficult to interpret. The population may have reached a minimum of approximately 4000 individuals in the late 1890s. The population numbered approximately 11,000 individuals when regular quantitative surveys were begun in 1967. Based on 19 surveys over a span of 30 years, the current population is thought to be about 27,000 individuals. Patterns in population trend, observed adult mortality rates, and the number of whales in unusual locations during summer all support the perception that the population is approaching carrying capacity. In 1994, the eastern gray whale was the first marine mammal to be removed from the ESA "List of Endangered Species."

Several issues are of current concern for eastern gray whales. Modest rates of incidental take in fishing nets are reported. Disturbances by boats supporting "whale-watching" activities are of concern, both in migratory corridors along the heavily populated west coast of North America and in the breeding and calving lagoons in Mexico. Recently suspended plans to expand salt extraction industries at Mexican lagoons drew substantial opposition from those interested in gray whale conservation. Possible demographic damage to gray whales by existing saltworks remains a matter of speculation.

Eastern gray whales are subject to ongoing annual subsistence harvests by indigenous peoples in Russia, Alaska, and Washington state. The subsistence harvests are authorized by IWC quota and, in the United States, are managed by village elders and government agencies. Subsistence hunts have taken over 100 whales per year for several decades, with 95% of the harvest occurring along the Russian coast of the Bering and Chukchi Seas during summer months. In 1998, the Makah tribe of northwestern Washington state was allocated a harvest quota of 5 whales per year for 4 years in exchange for a comparable reduction of the quota for Russian and Alaskan hunts. The allocation of harvest quota to the Makah people was consistent with the Treaty of Neah Bay of 1855 (12 Stat. 939) between the Makah tribe and the U.S. government. One whale was taken in 1999 by Makah hunters.

In our opinion, the eastern gray whale population is no longer at risk of extinction in the foreseeable

future. The status of the population should remain good as long as subsistence harvests are conservative and carefully managed in the context of emerging knowledge of population structure. Increasing problems with incidental taking, disturbance by boats, and industrialization at the calving/breeding lagoons in Mexico must be monitored and regulated to ensure that these issues are specifically managed.

#### 5. Northern Elephant Seal, *Mirounga angustirostris* (Gill, 1866)

Northern elephant seals breed and molt on coastal islands of Baja California, Mexico, and southern and central California. Since 1975, at least four mainland breeding and molting colonies have developed in central California. When not hauled out, northern elephant seals forage in pelagic habitats of the temperate and subpolar North Pacific. Two complete migrations are made each year between hauling sites and foraging habitats. Seals pup, nurse, and wean young and breed from December through mid-March at the hauling sites, then they swim north to forage. They return to haulouts to molt during spring and early summer, and then return again to foraging areas. Schedules for migration vary among age and sex categories. When hauled out, elephant seals are concentrated at high density. Hauled seals are easily approached by humans, even more so than other land-breeding pinnipeds of North America.

Northern elephant seals have been utilized for food and oil over the centuries by native peoples of North America. Intensive commercial harvests for oil began early in the nineteenth century. Most colonies of seals were severely depleted by 1850, but commercial harvests continued until 1884, at which point the species was considered extinct. Subsequent discoveries of small numbers led to continued harvests by scientific collectors working for museums of natural history, including the Smithsonian Institution. Scientific collecting continued until at least 1911. The total number of surviving seals is thought to have been as low as 20–100 in 1890. At the time harvests finally ended, elephant seals survived only at Guadalupe Island off the west coast of Baja California.

Northern elephant seals numbered approximately 127,000 individuals in 1991. The total is probably now approaching 200,000. They occupy at least 16 major hauling sites in Mexico and California. Population growth has been approximately 6% per year through the latter half of the twentieth century. Recent studies indicate a marked lack of genetic diversity at examined loci, almost certainly a result of the severe population “bottleneck” associated with commercial and museum

harvests. In contrast, several of the southern elephant seal populations currently in decline have much higher levels of measured genetic diversity than those of the northern species.

The elimination of commercial harvest has allowed northern elephant seals to recover fully from near extinction despite a loss of genetic diversity. Most current management concerns for the species involve perceived problems of overabundance rather than rarity. Indefinite survival should be ensured as long as harvests and disturbances to habitat can be monitored and controlled.

#### 6. Northern and Russian Subspecies of the Sea Otter, *Enhydra lutris kenyoni* Wilson, 1991, and *E. lutris lutris* (Linnaeus, 1758)

The northern subspecies of sea otter ranges from the Aleutian Archipelago eastward and southward along the coasts of Alaska, British Columbia, and Washington to the mouth of the Columbia River. The Russian subspecies ranges from northern Hokkaido, Japan, through the Kuril Archipelago to southern Kamchatka and the Commander Islands. Preexploitation estimates of population size are not available. As noted in Section II,D,6, large-scale commercial hunting of sea otters for pelts began in 1741. By the time legal hunting ended in 1911, the combined number of individuals in both subspecies was probably less than 2000 individuals, scattered among 10 isolated remnant populations from the Queen Charlotte Islands, Canada, to the Kuril Islands. The Queen Charlotte population was extinct soon after. The two subspecies of sea otter have been subject to ongoing subsistence harvest by native peoples for meat and pelts, probably for many centuries.

Populations from Prince William Sound westward have largely recovered without assistance, other than prohibition of harvest, during the twentieth century. Observed annual population growth rates have been as high as 10–15% in some cases. From the 1950s through the early 1970s, several translocation projects were attempted, moving groups of animals from Prince William Sound and Amchitka Island eastward to the Pribilof Islands, southeastern Alaska, British Columbia, Washington, and Oregon. The projects in the Pribilofs and Oregon failed, but all others succeeded, producing large populations in southeastern Alaska and small but rapidly growing populations off British Columbia and Washington.

Illegal commercial harvest of sea otters has been an occasional problem throughout the range. Poaching increased in Kamchatka with the development of eco-

conomic crises following the fall of the Soviet Union in 1991. The current scope of the poaching problem is unknown. Legal subsistence harvest by native villages in Alaska has averaged approximately 500 animals per year since the mid-1980s. The harvest is concentrated primarily in Prince William Sound and southeastern Alaska. The *Exxon Valdez* oil spill of 1989 killed several thousand sea otters in Prince William Sound and nearby coastal areas. Despite the intensive public interest and media coverage that facilitated apocalyptic scenarios, it appears that sea otter numbers in the sound did not experience long-term reduction, with the possible exception of a few local areas most heavily affected by the oil spill. Some effects on sea otter prey populations and habitats may persist to the present, but sea otter numbers are large and increasing in most areas affected by the spill.

Currently, Russian sea otter populations include approximately 20,000 individuals and are viewed as stable and near carrying capacity. Alaskan populations number approximately 100,000 individuals, although many of the survey data are obsolete. Recently, some populations in the Aleutian Islands have declined rapidly for unknown causes. There has been speculation about possible effects of contaminants and of increased rates of predation on sea otters by killer whales. A clear understanding of the scope and causes of the apparent declines will require new survey data and more definitive studies of population risk factors.

We are confident that northern and Russian sea otters have largely recovered from the excessive harvests of the eighteenth and nineteenth centuries and are free of significant extinction risk for the twenty-first century. For sea otters to remain out of jeopardy, legal harvests must be conservatively regulated, and all possible efforts must be made to reduce or eliminate incidental taking, poaching, contaminants, and oil spills. The recent population declines in the Aleutian Islands are a concern and must be evaluated carefully.

### III. DISCUSSION

#### A. Synopsis of Factors and Processes Known to Be Facilitating Modern Anthropogenic Extinctions of Marine Mammals

In our review of taxa and populations of marine mammals currently in jeopardy of anthropogenic extinction, there are two major, recurring categories of vulnerability. The first is prolonged excessive harvest, primarily in

the past and usually for commercial purposes, reducing numbers to a small fraction of preexploitation status. In this category we find taxa or populations from all marine locations on Earth, including many that are distant from human population centers. The second is a combination of risk factors strongly linked to a proximate and encroaching human population. The factors include habitat loss or chronic habitat disturbance, incidental taking in fishing gear, and contaminants as well as directed harvests. The second category primarily includes taxa or populations restricted ecologically to a limited geographic range in nearshore marine or aquatic habitats. Here, we consider the vulnerabilities of taxa or populations in the two categories.

The first category primarily involves excessive directed harvest. All species of mysticetes, the larger odontocetes, nearly all pinnipeds, the marine otters, and the polar bear have been hunted extensively at least during the past few centuries, in most cases in order to obtain articles of commerce. Most exploited taxa lack the necessary demographic features to sustain viable populations at the level of harvest experienced (see Section I,C,6). There are two important results. First, such exploitation has often reduced populations to small sizes. Second, populations so affected require decades or even centuries to recover to levels free of the risks of extinction. Directed, commercial-scale harvests of marine mammals ended for most species during the twentieth century, and many species are now subject to rigorous protection. However, the risk of extinction persists for reduced populations despite relaxation of harvest activity.

Small populations are vulnerable to any factor that reduces survival or collective reproductive success. Survival and reproduction can be impaired by anthropogenic factors, such as contaminants or disturbance to critically important breeding locations, or by natural fluctuations in the biological habitat. Anthropogenic factors should be controllable in principle by the appropriate management actions, but in reality effective risk management is difficult and often fails. Natural fluctuations are effectively stochastic in timing, duration, and intensity and cannot be anticipated or controlled by any form of management authority. Precautionary management can, however, reduce risks associated with stochastic events in some cases.

Once marine mammal populations are reduced, they recover slowly and are therefore at risk of the damaging effects of anthropogenic or natural disturbances over an extended period even under the most rigorous protection. For example, North Atlantic right whales have not been harvested for decades, but the harvest reduced them to low numbers. Now, even modest rates of ship-

strike mortality or incidental take in fishing gear are adequate to hold the population at a dangerously small size. A major change in ocean productivity, however brief, could easily push the populations to virtual extinction without any possibility of effective human intervention.

The second category involves a group of factors associated with increasing human populations in coastal regions, interacting with taxa or populations constrained to life in the coastal zone. Species in this category include the sirenians, river dolphins, and several coastal odontocetes. The essential problem here is that increasing human populations produce a suite of effects, each damaging to proximate marine mammal populations. Reduction of the effects often requires a deliberate curtailment of economic enterprise such as fishing or of institutional infrastructure such as waste disposal, flood control, or the provision of drinking water. The emergent dilemma is the perception by political institutions that there must be a choice between human welfare and the welfare of nearby marine mammals. Our case studies suggest that, given the choice, human cultures of major population centers act in favor of human needs. Thus, for example, recreational boating activities continue to crowd needed habitat for manatees in Florida despite 20 years of documentation that manatees do not tolerate powerboat activities. Fishing interests continue to set nets in vaquita habitat despite general recognition that incidental take is driving the population to extinction. The latter case is complicated by the lack of economic alternatives for the artisanal fishers of the region.

Species in our first category have reasonable probability of survival, given some luck. Sea otters and northern elephant seals have escaped the window of vulnerability associated with small population size, and other taxa seem well on their way. Some taxa or populations probably will not persist. Northern right whales and western gray whales will survive the new century only with good fortune and the most rigorous imaginable protection. We are less optimistic about species in our second category. Their ultimate survival depends on conscious economic restraint by human cultures and a possible reevaluation of values regarding the survival of marine mammals and other species in habitats also used or coveted by people.

Excessive subsistence harvests, anthropogenic noise, contaminants, oil spills, and depletion of genetic diversity are issues that have at least occasionally been invoked as risk factors for extinction of marine mammals. We find that there are relatively few taxa or populations clearly falling toward extinction as the direct result of any one of these factors. Subsistence harvest by native

peoples is without question a serious risk factor for Cook Inlet beluga whales, and it may have been a crucial precursor to the extinction of Steller's sea cow. However, the western Arctic bowhead whale population has been increasing steadily for years despite regular annual subsistence harvest. Thus, subsistence harvests are manageable risk factors and need not be regarded universally as unacceptable practice. None of the other listed factors are alone causing widespread extinction risk, although there are isolated examples for each. Of greater concern here is the problem of significant effects from the interaction of multiple factors. The best known cases involve mass mortalities that result from disease outbreaks. Such outbreaks often result from immunosuppression, which in turn may result from contaminants or from natural disturbances such as the toxic by-products of certain phytoplankton. Interacting factors are often a problem near human population centers and are difficult, if not impossible, to manage. Thus, we believe that the danger of many risk factors discussed here is not the direct effect of a single factor but rather the synergistic effects of multiple factors that may be less damaging when separated from one another.

The detection of extinction risks for both categories of species will continue to be confounded by the inherent uncertainties of population data for marine mammals. Several kinds of errors are possible in the future, although predictions of error rate are beyond the scope of this review. Populations that are numerically stable may be categorized incorrectly as declining and will receive unwarranted research effort. Declining populations may be categorized incorrectly as stable and may be denied attention from survey efforts necessary to detect problems and plan recovery actions. Recovery efforts may be directed to one component of life history, such as juvenile survival, when similar efforts toward other components, such as adult female survival, could produce far greater return for the same effort. All the error types will be more likely to influence perceptions of small populations in remote locations, where the incremental cost of survey effort is prohibitively large.

## B. Geographic Regions of Greatest Concern with Regard to Anthropogenic Extinctions of Marine Mammals

As previously suggested, marine mammal taxa and populations constrained to life near human population centers are in general most vulnerable to anthropogenic extinction. Extinction risks will be greatest where human cultures have the fewest economic options when

confronted with the need for restraint in order to solve conservation problems. Such circumstances are most likely in “developing” countries at lower and middle latitudes in which large concentrations of people face ongoing economic shortfalls. During the next century, we anticipate the greatest extinction risks for coastal marine or aquatic marine mammals off southern and southeastern Asia, eastern Europe, Central America, and central Africa. High rates of per capita resource consumption in the “developed” countries are also linked to risk factors for marine mammal populations. Thus, extinction risks are a significant concern for coastal marine mammals off North America and western Europe. Example cases include manatees in Florida, sea otters off California, and ringed seals in Lake Saimaa, Finland.

Marine mammals in the Southern Hemisphere generally should be at lower risk of extinction during the next century than those in the north. Most major human population centers and most cases of coastal habitat degradation occur in the Northern Hemisphere. However, some southern populations remain small because of past excessive harvests. Thus, distantly located southern taxa and populations will be removed from the risk of extinction only to the extent that all forms of harvest are regulated with extreme caution and conservatism. Moreover, small distant populations are at risk of incorrect conclusions about number of individuals, trend in numbers, or demographic characteristics because of the statistical limitations associated with survey and demographic data.

### C. General Approaches toward Minimizing the Rate of Anthropogenic Extinctions of Marine Mammals

Excessive directed harvests have caused more cases of jeopardized marine mammal taxa and populations than any other single factor. Thus, the most direct and straightforward approach to the control of extinction risk for marine mammals is a precautionary approach to the concept of marine mammal harvest on a global scale. Fortunately, this is the approach currently adopted by many governments, and by international regulatory cooperatives such as IWC, for some species. In this context, we offer three points of caution. First, the protections provided by international treaties and conventions such as the IWC, and by individual governments, do not necessarily extend to all marine mammals. There are high-profile protective protocols, with active ongoing oversight, for larger cetaceans, some

pinnipeds, sea otters, and polar bears. Many small cetaceans, some pinnipeds, and some sirenians are not actively and explicitly protected at the national or international level. Second, many of the populations subject to active protection are very small and as a consequence will be subject to risks associated with stochastic events, both natural and anthropogenic, for many decades. Thus, some extinctions are possible because not all species are equally protected, and because the most aggressive protection cannot eliminate all risk factors. Third, the crucial process of detecting trends in small distant populations is so costly that errors are likely in determining which populations are most seriously jeopardized. Thus, despite the best of human intentions, protective effort may be directed inadvertently to the wrong taxa or populations.

The group of risk factors associated with human population growth will almost certainly cause some extinctions of coastal marine mammals in the next century. Here, we believe that the outcome is more certain, and the methods of prevention more intractable, than in the simpler cases of small populations distant from major concentrations of people. Effective protection of imperiled species requires that human cultures forego economic benefits for the good of jeopardized marine mammals. Such sacrifices must extend indefinitely to be effective, given the demographic limitations of most marine mammals. The acceptance of foregone benefits is most needed in cultures least able (“developing” countries) or most unwilling (“developed” countries) to accommodate the loss. Cultural acceptance of economic loss motivated by a conservation ethic will require education, reorientation, and provision of meaningful economic alternatives at a level of effectiveness that we find difficult to imagine.

We conclude with the reminder that both evolution and extinction of species have been characteristic features of the history of marine mammal taxa on Earth since the early Eocene. Although the anthropogenic loss of species is both regrettable and inevitable, there is no reason to deny that new species will evolve as well. For example, killer whales off Washington and British Columbia occur in three distinctive social configurations that correlate with subtle but reliable morphological differences. Individuals in the different categories have entirely different diets, different acoustic repertoires, rarely co-occur in space, and do not interbreed. Some have characterized these patterns as a step in the process of speciation, although there are alternative viewpoints. The human mind can easily perceive extinction as a finite event, but not speciation because of differing timescales and imprecise definitions of the

latter. Thus, we are biased toward greater conscious awareness of extinction than of speciation. Although anthropogenic extinctions should not be accepted without exhaustive attempts at recovery, it is perhaps more reasonable to view such events as parts of a dynamic process of evolution rather than as catastrophic failures of human behavior.

### See Also the Following Articles

EXTINCTIONS, CAUSES OF • EXTINCTIONS, MODERN  
EXAMPLES OF • MAMMALS, BIODIVERSITY OF • MARINE  
ECOSYSTEMS, HUMAN IMPACT ON

### Acknowledgments

We thank Carlos Alvarez-Flores, Amanda Bradford, Allan Fukuyama, Laura Litzky, Josh London, Robert Suydam, and Alexandre Zerbini for technical guidance and constructive reviews of the draft manuscript. We received financial support from the Biological Resources Division of the U.S. Geological Survey, the School of Fisheries of the University of Washington, the National Center for Ecological Analysis and Synthesis at the University of California, Santa Barbara, and the Southwest Fisheries Science Center of the National Marine Fisheries Service (U.S.).

### Bibliography

Aguilar, A., and Borrell, A. (1996). *Marine Mammals and Pollutants, an Anotated Bibliography*. Fundació pel Desenvolupament Sostenible, Barcelona.

Beddington, J. R., Beverton, R. J. H., and Lavigne, D. M. (Eds.) (1985). *Marine Mammals and Fisheries*. George, Allen & Unwin, London.

Berta, A., and Sumich, J. L. (1998). *Marine Mammals. Evolutionary Biology*. Academic Press, San Diego.

Clapham, P. J., Young, S. B., and Brownell, R. L., Jr., (1999). Baleen whales: Conservation issues and the status of the most endangered species. *Mammal Rev.* 29, 35–60.

Dizon, A. E., Chivers, S. J., and Perrin, W. F. (Eds.) (1997). *Molecular Genetics of Marine Mammals*, The Society for Marine Mammalogy Special Publ. No. 3. Allen Press, Lawrence, KS.

Garner, G. W., Amstrup, S. C., Laake, J. L., Manly, B. F. J., McDonald, L., and Robertson, D. G. (Eds.) (1999). *Marine Mammal Survey and Assessment Methods*. Balkema, Rotterdam.

Gerber, L. (1998). Seeking a rational approach to setting conservation priorities for marine mammals. *Integrative Biol.* 1, 90–98.

Jefferson, T. A., Leatherwood, S., and Webber, M. A. (1993). *Marine Mammals of the World*. Food and Agriculture Organization of the United Nations, Rome.

Loughlin, T. R. (Ed.) (1994). *Marine Mammals and the Exxon Valdez*. Academic Press, San Diego.

Perrin, W. F., Brownell, R. L., Jr., and DeMaster, D. P. (Eds.) (1984). *Reproduction in Whales, Dolphins, and Porpoises*, Special Issue No. 6. International Whaling Commission, Cambridge, UK.

Reynolds, J. E. III, and Rommel, S. A. (Eds.) (1999). *Biology of Marine Mammals*. Smithsonian Institution Press, Washington, D.C.

Rice, D. W. (1998). *Marine Mammals of the World. Systematics and Distribution*, The Society for Marine Mammalogy Special Publ. No. 4. Allen Press, Lawrence, KS.

Richardson, W. J., Thomson, D., Greene, C., and Malme, C. I. (1995). *Marine Mammals and Noise*. Academic Press, San Diego.

Riedman, M. (1991). *The Pinnipeds. Seals, Sea Lions, and Walruses*. Univ. of California Press, Berkeley.

Rojas-Bracho, L., and Taylor, B. L. (1999). Risk factors affecting the vaquita (*Phocoena sinus*). *Mar. Mammal Sci.* 15, 974–989.

Twiss, J. R., Jr., and Reeves, R. R. (Eds.) (1999). *Conservation and Management of Marine Mammals*. Smithsonian Institution Press, Washington, D.C.

Weller, D. W., Würsig, B., Bradford, A. L., Burdin, A. M., Blokhin, S. A., Minakuchi, H., and Brownell, R. L., Jr. (1999). Gray whales (*Eschrichtius robustus*) off Sakhalin Island, Russia: Seasonal and annual patterns of occurrence. *Mar. Mammal Sci.* 15, 1208–1227.





# MARINE SEDIMENTS

Paul V. R. Snelgrove  
*Memorial University of Newfoundland*

---

- I. Benthos and the Benthic Environment
  - II. Marine Sedimentary Habitats
  - III. Global Estimates of Species Numbers
  - IV. Global Patterns of Biodiversity
  - V. Regulation of Diversity
  - VI. Ecosystem Services and Sedimentary Diversity
  - VII. Threats to Sedimentary Diversity
  - VIII. Summary
- 

## GLOSSARY

**abyssal plains** Relatively flat areas of the ocean bottom below ~4000 m depth.

**benthic** Aquatic bottom habitat, encompassing areas on or below the interface between the water and the bottom.

**benthos** Bottom-living organisms, including those that reside on hard and soft bottom surfaces and others that reside between sediment grains.

**continental rise** An area at the base of the continental slope between 3000 and 4000 m where the bottom slope is slight and sediments often accumulate.

**continental shelf** A region of ocean bottom extending from the low water mark at the edge of continents to a depth (~200 m) at which the incline increases markedly and the continental slope begins.

**continental slope** Ocean bottom extending from the edge of the continental shelf at an ~4° incline to a depth (3000–10,000 m) at which the slope decreases and the continental rise begins.

**deposit feeders** Organisms that feed primarily by ingesting organic material occurring on or between sediment grains.

**emergent vegetation** Plants that are attached to the bottom but extend up through the water column above the ocean surface.

**infauna** Organisms living below the sediment–water interface and between sediment grains.

**macrofauna** Animals large enough to be retained on a 300- or 500- $\mu\text{m}$  sieve.

**megafauna** Animals that are sufficiently large to be identified from bottom photographs.

**meiofauna** Animals small enough to pass through a 500- $\mu\text{m}$  sieve but large enough to be retained on a 63- $\mu\text{m}$  sieve.

**phytoplankton** Unicellular and chain-forming algal plants that are suspended and transported in the water above the bottom.

**suspension feeders** Organisms that feed primarily on organic material suspended in the water above the bottom.

**trench** A steep-sided depression in the ocean floor ranging from 6000 to 10,000 m depth and associated with areas of tectonic plate subduction.

---

**MARINE SEDIMENTS COVER ALL** but a few percent of the ~70% of the earth's surface that is ocean and as such comprise the largest single ecosystem on Earth in area. Given the vast global coverage, it is not surprising



that marine sediments are species-rich systems, particularly considering the range of sedimentary habitats that occur. Sediments range from rubble and coarse gravel to fine clays. They occur from tidally exposed areas, where temperature, salinity, and sediment water content fluctuate greatly, to the greatest depths of the oceans, where no light penetrates, temperatures approach freezing, environmental variables are largely invariant, and pressure exceeds 100 times that experienced at sea level. The species that inhabit these environments, how they are regulated, the ecological roles that they play, and activities that threaten them vary considerably from one habitat to the next and are the basis for this article.

## I. BENTHOS AND THE BENTHIC ENVIRONMENT

Depending on the specific habitat and species, different benthic (bottom-dwelling) organisms may reside just above the bottom but closely associated with it (hyperbenthos), on the sediment surface (epifauna), or among the sediment grains (infauna). Because sediments have limited permeability, very little oxygenation (penetration of millimeters) is achieved through diffusion. Greater oxygen penetration is achieved only when bottom currents mix sediments or, more commonly, when bioturbation (biological mixing of pore water and sediments by benthos) facilitates transfers of oxygenated water from the water column into the sediment pore water. Bioturbated sediments are usually oxygenated within the top few centimeters, although active bioturbation can lead to deeper pockets of penetration. Light does not penetrate well through sediments and most productivity is provided by phytoplankton detritus sinking from surface waters above, macrodetritus transported from coastal areas (e.g., kelps and seagrasses), or from primary producers living attached to the sediment surface in shallow areas where light penetrates all the way to the bottom. Thus, most of the primary production or detritus in benthic systems is concentrated near the sediment surface rather than within the sediments. This combination of limited food and oxygen penetration results in the vast majority of organisms being confined to the upper few centimeters of sediment near the sediment–water interface. Smaller organisms that can tolerate anoxia and larger organisms that maintain a burrow or appendage to the surface can live deeper, but even then distributions are generally limited to 10–20 cm or less.

As in other environments, feeding modes include herbivores, predators, scavengers, parasites, and omnivores. However, most organisms that reside in marine sediments are either deposit feeders or suspension feeders, and some species are capable of either feeding mode. Deposit feeders ingest particles associated with sediments or, in many cases, they ingest the sediment particles themselves. Nutrition is provided by detritus associated with the sediment grains and also by microbes decomposing the detritus. In shallow areas, benthic algae (diatoms) may provide a food source for some deposit-feeding species. Deposit feeders are important to sediment geochemistry because, as they move through the sediments, they facilitate pore water movement in sediments and also move around sediment grains (bioturbate). Suspension feeders are organisms that remove particles such as phytoplankton and detritus from the water column. Because they rely on suspended particles, suspension feeders tend to be most abundant in energetic environments and absent from areas in which currents are weak and horizontal flux of food particles is reduced. Differences in the physical energetics of an environment will also influence the stability of the benthic habitat and the organisms that are able to live there. Quiescent areas usually have fine-grained sediments that are relatively stable except where storms pass through. High-energy environments are more dynamic in terms of sediment movements; fine-grained sediments are usually absent and the coarse-grained sands that remain are often moved around by ambient flow conditions. Not surprisingly, the most dynamic environments are generally in shallow areas, whereas much of the deep sea is relatively quiescent because wave energy rarely penetrates beyond 50–60 m with sufficient strength to move sediment grains.

For reasons of sampling and taxonomic practicality, benthic researchers routinely focus on just one of several size groupings. Megafauna are defined as organisms that are sufficiently large to be identified in bottom photographs and include organisms such as flatfish, crabs, and scallops. Macrofauna are defined as those that are retained on a 300- or 500- $\mu\text{m}$  sieve (standards vary). This size grouping includes polychaete annelids, crustaceans, bivalves, and many other phyla (Fig. 1). Meiofauna refers to organisms that pass through a 300- or 500- $\mu\text{m}$  sieve but are retained on a 44- or 63- $\mu\text{m}$  sieve and include nematodes, foraminiferans, tiny crustaceans, and many others (Fig. 2). Microorganisms include living organisms that pass through a 44- or 63- $\mu\text{m}$  sieve and include bacteria and protists. These size groupings are not absolute in that the larval or juvenile stages of one group may be comparable in size

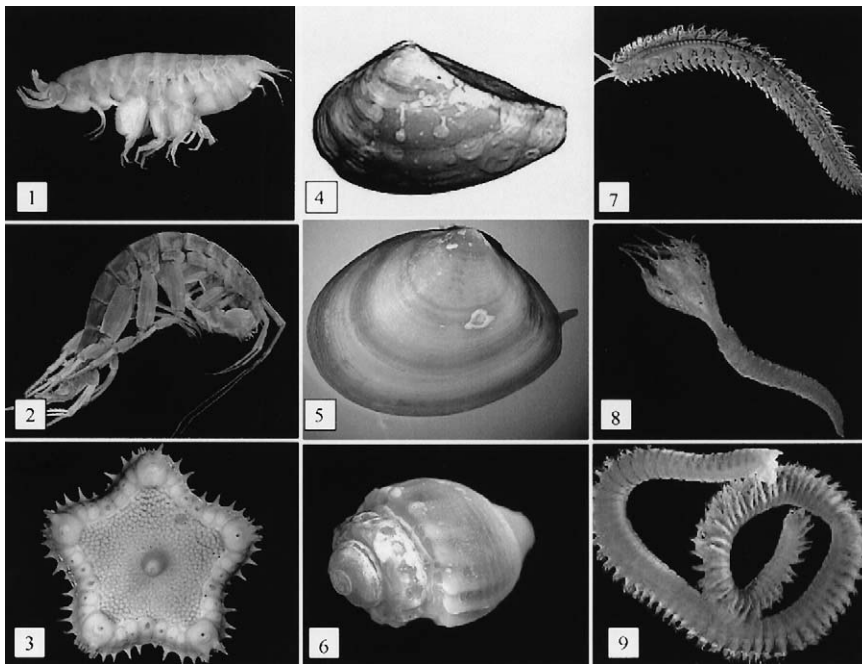


FIGURE 1 Typical sedimentary macrofauna, including amphipod crustaceans (1 and 2), echinoderms (3), bivalve (4 and 5), and gastropod (6) mollusks and annelid polychaetes (7–9). Magnification in the photos varies but individuals range from approximately 1 to 10 mm in length (photographs by P. Ramey).

to individuals from a smaller group, but this division of organisms is necessary because of sampling logistics. The sample size appropriate for enumerating megafauna, for example, is inappropriate for sampling meiofauna that are orders of magnitude more abundant. Moreover, the taxonomic challenges even within each of these size groupings are considerable, and synthesis across groups is therefore rare in a single study.

These groups of organisms do not operate independently of one another. Megafauna prey on infauna, and macrofauna prey on meiofauna. There is also evidence that meiofauna prey on juvenile macrofauna. Microbes

form an important dietary component for many meiofaunal and macrofaunal taxa, and macrofaunal bioturbation impacts microbial composition because it oxygenates sediments and thereby regulates the relative importance of aerobic versus anaerobic bacterial activity. Microbes also initiate breakdown of food material that would otherwise be undigestible by metazoans and play a major role in cycling of organic matter. Thus, there is considerable interaction between these groups of organisms.

Many factors are thought to influence the spatial distribution of benthic sedimentary organisms, includ-

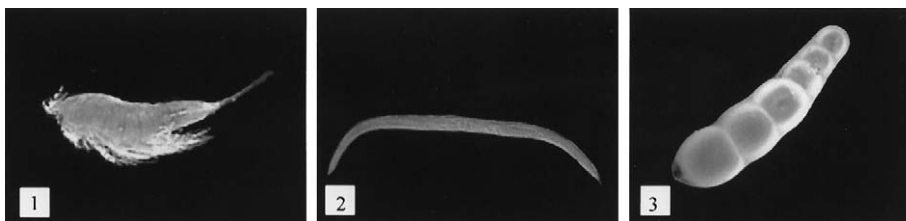


FIGURE 2 Typical sedimentary meiofauna, including harpacticoid crustaceans (1), nematodes (2), and foraminiferans (3). Magnification in photos varies but individuals range in size from approximately 100  $\mu\text{m}$  to 300  $\mu\text{m}$  (photographs by P. Ramey).

ing sediment type, productivity, temperature, salinity, pressure, depth, oxygen, sediment stability, air exposure, current speeds and wave action, and biological interaction with other species.

## II. MARINE SEDIMENTARY HABITATS

Oceans range in depth from intertidal habitats to ocean trenches at 10,000-m depth (Fig. 3). Light can penetrate to 1000 m in the clearest oceanic waters, but in coastal waters penetration may be limited to from 200 m to only a few meters. Like most ecosystems, marine sedimentary communities are fueled by photosynthesis, and because many marine sediments occur below photosynthetic depths they rely on phytoplankton sinking from surface waters above and macrophyte detritus transported from nearshore environments. Not surprisingly, water depth is a major variable in categorizing marine sedimentary habitats.

### A. Intertidal Areas

The shallowest sedimentary habitats are those that occur at the land–sea interface. They include habitat with emergent vegetation such as mangroves and salt

marshes as well as sandflats and mudflats without emergent vegetation that may or may not include diatom mats (Table I). Mangroves and salt marshes occur intertidally, and as such they experience wide ranges of salinity and temperature. They are also extremely productive, and organic matter in the form of decaying vegetation is abundant. Because of the high productivity of these systems, bacterial respiration is extremely high and sediments are often hypoxic or even anoxic unless they are right at the sediment–water interface. The organic matter produced by vascular plants is also relatively refractory and difficult for most species to digest. In combination with oxygen limitation and tidal exposure, this refractory material contributes to the low diversity of organisms residing in the sediments. Unvegetated sandflats and mudflats tend to be much lower in productivity unless they are immediately adjacent to vegetated areas, and they have slightly higher diversities. Nonetheless, intertidal areas experience high variation in environmental variables and are characterized by a low-diversity fauna consisting of organisms able to cope with this variability.

### B. Subtidal Areas

At depths below low tide, species diversity increases relative to that of intertidal systems. In the shallowest

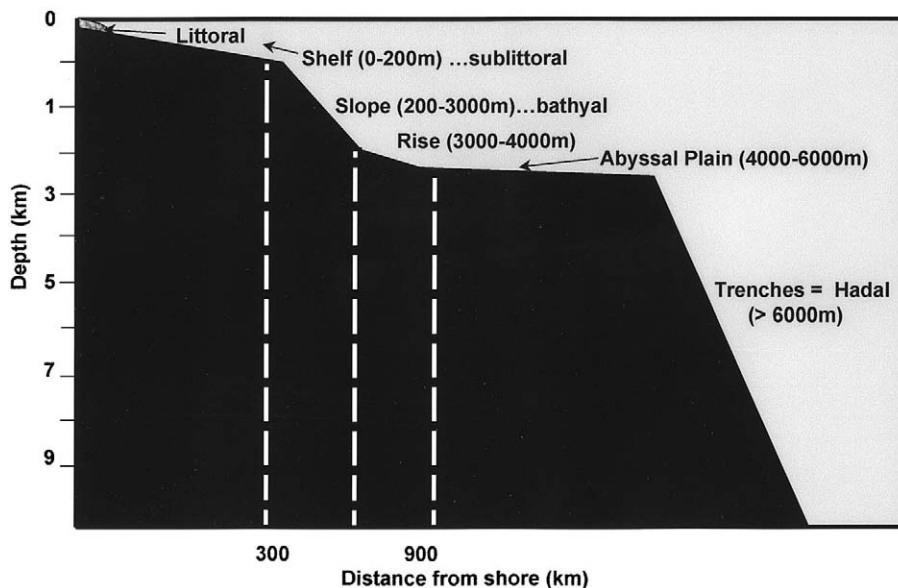


FIGURE 3 Schematic diagram of the depth zones of the oceans. Dark area indicates the ocean bottom. The distances and depths given are only examples of a range or values observed throughout the world. The horizontal axis is greatly exaggerated in size relative to the vertical axis. See text for more detailed explanation of depth zones.

TABLE I  
Summary of Sedimentary Habitats, Local Diversity, Ecosystem Services Provided, and Current Threats<sup>a</sup>

Sedimentary habitat	Diversity	Ecosystem services	Current threats
Salt marshes	Low	Nutrient cycling, critical habitat, shoreline stability, filtration, productivity, pollutant cycling	Habitat destruction, pollution, global climate change
Mangroves	Low	Nutrient cycling, critical habitat, shoreline stability, filtration, productivity, pollutant cycling	Habitat destruction, pollution, global climate change, exotic species
Sandflats	Low to modest	Nutrient cycling, filtration, productivity	Habitat destruction, pollution, exotic species
Mudflats	Low to modest	Nutrient cycling, productivity, pollutant cycling	Habitat destruction, pollution, exotic species
Estuaries	Low	Nutrient cycling, critical habitat, filtration, productivity, pollutant cycling	Habitat destruction, pollution, global climate change, exotic species
Coral reef sediments	Very high	Nutrient cycling, productivity	Habitat destruction, pollution, fishing
Continental shelf	Modest to very high	Nutrient cycling, critical habitat, filtration, productivity, pollutant cycling	Habitat destruction, pollution, fishing, exotic species
Continental slope	Very high	Productivity, nutrient cycling (global)	Habitat destruction, fishing
Abyssal plains	High to very high	Nutrient, element cycling (global)	—
Hydrothermal vent sediments	Low	Nutrient, element cycling (global)	—
Deep ocean trenches	Low	?	—

<sup>a</sup> The table summarizes generalities for which there are exceptions. In the case of global climate change, the large-scale circulation changes predicted in some models would threaten all habitats, and only more immediate impacts are included here.

subtidal areas, vegetation in the form of seagrass and kelp may occur, and some shallow sediments may be covered in diatom mats. However, all these plant forms are confined to depths of tens of meters or less, and most subtidal sedimentary habitats rely on sinking plant material as the base of their food chain. Sedimentary environments may range from gravel to fine clays. Fine-grained sediments are characteristic of low-energy environments and coarse-grained sediments are characteristic of high-energy areas. The latter are often dynamic environments, with sediments often moving around in response to bottom flow. Continental shelves are largely covered in sediments and are inhabited by a wide variety of organisms. The shallower portions of these environments (up to tens of meters) can experience strong seasonality in productivity and temperature, and where offshore runoff is significant (e.g., estuaries) salinity may vary considerably. The greater the environmental variability in these systems, the lower the diversity. With increased depth, the variability in salinity and temperature is generally reduced, although productivity can still be seasonal. Continental shelf areas are large, productive areas of the oceans and support the majority of marine commercial fisheries.

### C. Deep-Sea Areas

The deep sea, defined here as habitats beyond the edge of the continental shelf at ~200 m, encompasses continental slopes, continental rises, abyssal plains, midocean ridges, and trenches. Although the sediments that cover the bottom can vary in terms of grain size composition, environmental variability is considerably reduced. Temperatures are largely invariant, salinity does not change, and sediments are rarely moved around other than by bioturbation. Depending on the region, productivity is generally less seasonal than in shallow water, but for the most part deep-sea benthic environments have much less productivity entering the system than their shallow-water counterparts. Most deep-sea areas are below light penetration and productivity consists of whatever material sinks through the thousands of meters of water column to the bottom. As a result of this reduced productivity, densities of organisms are very low and individuals are relatively small; surprisingly, species diversity is quite high. A very notable exception to low productivity in the deep sea is hydrothermal vents and seeps, for which chemoautotrophic bacteria

provide the basis for a very productive localized community.

Not all deep-sea communities are diverse. Because hydrothermal vent environments are characterized by fluids rich in compounds such as hydrogen sulfide and heavy metals, most species are unable to tolerate these toxic compounds and diversity is quite low. Deep-sea trenches are subject to mud slumping and poor circulation. Not surprisingly, species diversity is very low. Upwelling regions are also generally low in diversity, likely because large amounts of organic matter accumulate on the ocean floor and decompose, leading to hypoxia. A few deep-sea areas are subject to intensive "storms," in which currents become intense and sediment resuspension occurs. Evidence suggests that macrofaunal diversity is depressed in such areas, but surprisingly meiofaunal diversity is not. Presumably, the meiofauna are able to cope with the disturbance more effectively than are the macrofauna.

### III. GLOBAL ESTIMATES OF SPECIES NUMBERS

Several attempts have been made to estimate the total number of species that live in the ocean. The Danish marine scientist Thorson estimated in the early 1970s that 100,000 species would eventually be described from the marine environment, and until recently this was the ballpark figure most thought appropriate. In the late 1960s, deep-sea biologists Sanders and Hessler documented that the deep sea was not the low-diversity habitat that many had previously thought and could contain a significant portion of global marine species diversity. The debate on species estimates heated up considerably when Grassle and Maciolek published data from a collection of samples along a 176-m-long depth contour along the east coast of the United States, constituting the most extensive quantitative data set ever assembled for a single area of the deep sea. Based on the rate that species were added with increased area sampled, and the total area of the deep sea, they extrapolated to 10 million macrofaunal species. May pointed out that approximately half of the species in their study were new to science, and one could extrapolate from known to unknown species using a similar ratio. Given that approximately 250,000 marine species have already been described, May's extrapolation suggests ~500,000 total species living in the oceans. Recent work from the Pacific suggests that only 1 in 20 species has been described for that area of the deep sea, in which case May's approach would yield ~5 million species.

Although most of the discussions on species number have centered on macrofauna, Lambshead, a nematode ecologist, extrapolated upward from Grassle and Maciolek's estimate based on the fact that nematodes are more abundant and species rich than macrofauna, at least on local scales. Scaling up from their number, he estimates that there may be 100 million species of nematode alone. Currently, we know little about how widely distributed individual nematode species may be, making extrapolation even more tenuous than that for macrofauna. For microbial groups we know very little. There is evidence that protists may not be very diverse on a global scale, but studies on water column bacteria indicate vast amounts of undocumented diversity.

Clearly, more data are needed to improve estimates of global species numbers. Current estimates for metazoans range from 500,000 to 100 million species, but vast areas of the ocean and some taxonomic groups remain unsampled or poorly sampled. The South Pacific and tropical latitudes are among the least known areas. Even many shallow environments remain poorly known, and recent evidence (described later) suggests that some of these areas may be very species rich. A United States National Research Council committee summarized recent surveys and showed that more than two-thirds of the polychaetes in Hawaiian coral reef sediments were previously undescribed; even Georges Bank, a relatively well-studied area, was found to have only two-thirds of its polychaete species previously described. The situation is even more severe in deep-sea areas and for meiofaunal taxa such as nematodes, for which proportions of known species can sometimes be less than 10–20% of total species present. Even for taxa thought to be well-known, there is increasing evidence based on genetic data that "cosmopolitan taxa" are often species complexes that are difficult to distinguish based on morphological characters. Regardless of which projection of total species number is appropriate, there remains a huge amount of undescribed species diversity in marine sediments.

### IV. GLOBAL PATTERNS OF BIODIVERSITY

The historical perception of the deep sea as an azoic or species-poor environment changed markedly with a series of papers published by Sanders and Hessler. From data collected along a transect running from Martha's Vineyard, Massachusetts, to Bermuda, they demon-

strated that diversity in deep-sea sediments exceeded that in most shallow areas and rivaled that observed in shallow tropical areas. These findings contradicted a previous notion that diversity would generally decrease with depth. Rex followed up this work with extensive analysis of large-scale pattern with depth in the North Atlantic and found that diversity is highest at intermediate depths. Diversity in shallow water is low in estuaries and seasonal shallow areas, is somewhat higher on the continental shelf, increases down the continental slope, and peaks on the lower slope. At greater depths it declines again. Other workers have observed a similar pattern for other taxa, although the exact depth of the diversity peak is not always consistent. Moreover, diversity in the abyssal plains of the Pacific appears to exceed that observed at shallower depths, suggesting that a parabolic diversity–depth pattern is not universal. The generality of lower diversity in shallow water has been questioned by Gray and colleagues, who presented convincing data that the local diversity of sedimentary fauna of Bass Strait, Australia (11–51 m deep) may rival or exceed that of the deep sea. The pattern observed by Rex and others has been based largely on data from the North Atlantic, and Gray's findings emphasize the fact that patterns described from one area may not be easily extrapolated globally.

Many studies have documented latitudinal gradients in shallow-water benthic diversity, with decreasing diversity toward the poles. In the deep sea, the existence of such a trend is the subject of debate. The pattern has been proposed for North Atlantic deep-sea macrofauna, but isopod data from the South Pacific do not indicate such a pattern.

The inconsistency of conclusions regarding global patterns in diversity indicates a need for more complete sampling coverage with latitude and depth. Improved coverage would allow more effective comparison of different oceans and reduce the impact of "noise" added by regional differences in variables such as productivity that are bound to obscure any global patterns. Within this coverage, areas that have been poorly sampled historically, such as tropical sediments and the Southern Hemisphere, must be emphasized to reduce the North Atlantic bias. An additional problem with these large-scale comparisons is that most studies focus on just one of the size groupings of organisms and generally use different sorts of sampling techniques. For example, we must ask whether described differences in large-scale patterns in different taxa and ocean basins reflect differences in biology or sampling coverage. Studies that are more comprehensive in taxonomic and geographic coverage are necessary to resolve these debates.

## V. REGULATION OF DIVERSITY

### A. Shallow Water

Our understanding of diversity regulation in marine systems is limited. In environments with extremes of temperature (e.g., intertidal sediments), salinity (e.g., estuaries), geochemistry (e.g., sediments near hydrothermal venting), or oxygen (e.g., eutrophied coastline, mangrove sediments, and fjords with poor circulation), there are obvious physiological limitations for many species that allow only the most robust species to exist. The same may be said for physically disturbed environments such as sandflats, which tend to be relatively low in diversity. Similarly, highly seasonal environments are reduced in diversity, presumably because relatively few organisms are able to cope with the variability in the physical environment. However, beyond these extreme habitats, there are few obvious conclusions on how diversity is regulated.

Historical events and modern conditions both play major roles in determining global pattern in diversity. Certainly, the evolutionary history of a region must be considered. Jablonski's analysis of fossil invertebrates suggests that the tropics provide evolutionary novelty and as such contribute to the higher diversity noted in tropical latitudes. Glaciation effects are well-known to have played a role in establishing current marine patterns. For example, the reduced diversity observed in Arctic relative to Antarctic fauna likely relates to extinctions during Arctic glaciation and the slow reestablishment of the Arctic community. Thus, historical events set the stage for current ecological processes, which may subsequently play a role. In terms of modern conditions, the high level of total energy input in the tropics has been proposed to explain shallow-water mollusk diversity gradients, and the high degree of seasonality with increasing latitude may be a factor that depresses diversity.

The role that ecological interactions play in sedimentary diversity is less well understood. In shallow-water experiments, it has been observed that the exclusion of predators seems to enhance diversity. This finding is the direct opposite of Paine's classic findings in rocky intertidal communities and suggests that it is inappropriate to generalize from the relatively well-studied rocky intertidal system to the less understood sedimentary environment. Whether generalizations regarding predators depressing diversity can be extended to the deep sea remains to be tested. There is a dearth of experiments linking diversity in the water column to sedimentary faunal diversity and little data on how

predator diversity might impact benthic diversity. One might predict, for example, that increased diversity in food sources (e.g., phytoplankton) might lead to enhanced diversity in the infauna that feed on those food sources. Increased diversity of predators might also allow tighter species packing in an environment.

## B. Deep Sea

Regarding the deep sea, there has been considerable debate on how an environment that appears so physically homogeneous is able to support such a broad array of species. Sanders, who played a pivotal role in recognizing the high diversity of marine systems, suggested that the high level of stability over evolutionary time in the deep sea has allowed greater specialization and niche diversification. If this were the full explanation, then diversity should be highest on abyssal plains; the depth patterns described above do not consistently support this notion. There is also little evidence of niche specialization in deep-sea organisms. Most species appear to be relatively nonselective deposit feeders. In the early 1970s, several contrasting theories were put forward. First, cropping by predators could prevent competitive equilibria from being attained by infauna. If this were the case, however, then fast growth, early reproductive maturity, and an age structure dominated by young individuals might be expected, but this pattern is not typical for the deep sea. Moreover, evidence from shallow water suggests that predators decrease sedimentary diversity. In any case, the role that predators may play in maintaining deep-sea diversity remains largely unknown. A contrasting theory held that small patches may form disequilibria habitats that promote certain species. Tests of this patch mosaic model have included sampling of natural patches and creation of experimental patches. Both approaches have yielded a similar result: The species that occur in many patch types are usually rare or absent from background sediments, although diversity in patches is usually reduced. Such a pattern is consistent with the patch mosaic model, but experiments so far have demonstrated that patches promote only a modest number of species. Whether it is necessary to sample more patch types, or whether other explanation needs to be invoked, remains unclear.

The large area of the deep sea almost certainly impacts its species richness in that species richness of most environments increases with area. However, the vast rolling plains of the deep sea do not seem to add habitat heterogeneity with area as seen in other species-rich habitats, such as tropical rain forests or coral reefs.

Given that abyssal plains are the largest deep-sea habitat in area, the parabolic diversity–depth relationship described earlier is also inconsistent with a simple species–area relationship.

Huston's intermediate disturbance hypothesis is another explanation for the high diversity of deep-sea systems. The idea is that at one end of the spectrum, high levels of disturbance depress diversity by eliminating sensitive species. At the other end of the spectrum, very benign habitats allow superior competitors to eliminate weaker species and thus depress diversity. Highest diversity would then be expected in habitats with levels of disturbance that prevent competitive dominants from taking over but not so frequent and harsh that few species can cope. The midslope peak in diversity described earlier is consistent with this hypothesis but does not test the hypothesis explicitly. Most disturbed deep-sea environments, such as hydrothermal vents, low-oxygen areas, environments with benthic storms, and slumping areas, have reduced diversity. However, such extreme examples are far from conclusive in drawing comprehensive generalities.

Efforts to understand regulation of biodiversity in the deep sea have been hampered by the considerable logistical challenges of deploying and recovering experiments in deep, oceanic environments. In shallow-water environments, diversity regulation has received considerably less attention, and there is a pressing need to move some of the deep-sea framework and experimental effort into shallow water where logistics are simplified and outcomes are bound to produce interesting comparisons with the deep sea. Our understanding of diversity maintenance in sedimentary systems is generally very poor, and we are unable to predict, for example, whether loss of pelagic diversity will impact benthic diversity. The problem is that humans are impacting diversity of marine benthic organisms in many areas through fishing, habitat destruction and modification, pollution, exotic species introductions, and global climate change, but we have little idea how different components of the ecosystem will be affected.

## VI. ECOSYSTEM SERVICES AND SEDIMENTARY DIVERSITY

Although the huge fraction of the earth's surface covered by marine sediments and the large numbers of species residing within sediments provide strong motivation to understand the pattern and regulation of benthic diversity, there are also good ecological motivators.

The oceans provide a variety of ecosystem services (Fig. 4), and although we understand little about the role that biodiversity plays in maintaining these services there is very clear evidence that benthic species play critical roles. There is a very real chance that ongoing losses of sedimentary biodiversity will result in loss of ecosystem services.

### A. Nutrient Cycling

Benthic marine organisms play a critical role in global cycles of nitrogen, sulfur, and carbon. As organic material sinks to the sediment surface, it carries with it organic carbon and nitrogen. This material may be directly ingested by benthic organisms or microbes may colonize it. In areas with high sedimentation rates or low densities of organisms, there may be substantial burial loss of this material, but most is consumed or decomposed. Even the organic material that passes through organisms undigested will be colonized and broken down by microbes. Organisms that ingest the material respire, are preyed on, or die and decompose in the sediment. Depending on how decomposition and digestion occur, nitrogen may be released as ammonia, nitrite, nitrate, or nitrogen gas that diffuses or is physically mixed into the water column above. In coastal sediments, this cycle is a critical part of regenerating the dissolved nitrogenous compounds that are critical for primary producers; without this regeneration the cycle would end and primary production would cease. Carbon is cycled in a similar way. Material sinks to the

bottom and may eventually be buried (e.g., fossil fuels) or it may be decomposed into carbon dioxide or other carbon compounds. The relative rates at which carbon is buried, tied up in living organisms, or respired to carbon dioxide are closely tied to microbial activity and thus to the animals that influence microbial activity through feeding and oxidation of sediments. As a result, benthic organisms play a key role in global carbon budgets. Sulfur cycling through marine sediments is also dependent on sediment oxygenation and therefore bacterial activity. Whether sulfur is buried and stored in sediments or recycled is determined in part by metazoan and microbial activity.

### B. Pollutant Cycling

Because benthic organisms move around and ingest particles, and may themselves be ingested by predators, they can greatly impact the burial fate and mobility of pollutants. As sediment particles are bioturbated, so are any pollutants linked to them. As a result, benthic organisms can dilute pollutants at the sediment–water interface by mixing them downward. By the same process, the continual remixing of sediments by infauna may increase burial time of pollutants that might otherwise be buried more quickly by sedimentation. Bioturbation also tends to destabilize sediments and increase the likelihood that they will be resuspended. Some sedimentary organisms can also stabilize sediments by excreting mucous (e.g., bacterial mats) or creating tubes that they inhabit. Benthos may also increase effective

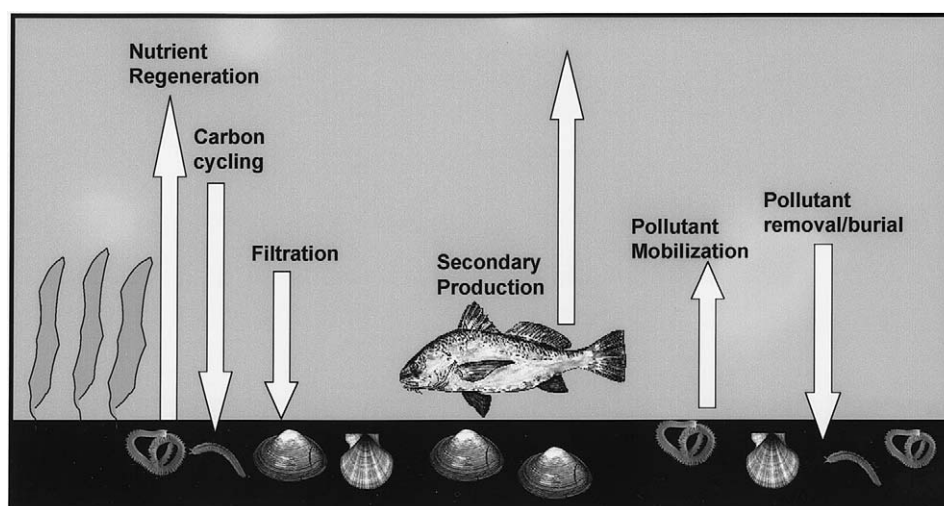


FIGURE 4 Schematic representation of key ecosystem services associated with sedimentary organisms. Arrows indicate transfer of material between the water column of the ocean and the sediment.



grain size by repackaging grains as fecal pellets. All these processes impact whether sediments are likely to resuspended or not. Infauna that ingest pollutants may also provide a conduit up the food chain if they concentrate the pollutant in their tissues. For example, fish that feed on polychaetes with high levels of a heavy metal may also have high concentrations of that metal. Finally, benthic organisms have some capacity to break down pollutants. Microbes in particular may have the capacity to break down some toxic compounds and render them harmless.

### C. Sediment and Shoreline Stability

Because benthos can destabilize (bioturbation) and stabilize (microbial mats, mucous excretion, and tube building) sediments, they can have a major impact on sediment transport and coastal geology. Mucous excretion, for example, binds sediment grains together and therefore increases the amount of energy required to move sediments. However, as some organisms produce products that bind sediment grains together, other organisms move among the sediment grains, breaking apart binding structures, increasing the water content, and increasing the likelihood of resuspension.

The vegetation in seagrass beds, mangroves, and salt marshes acts as flow baffles and therefore traps fine sediments, a process that is further aided by root structures that add further stability. The presence of these plants protects the shoreline and results in accretion rather than erosion. Indeed, rapid shoreline growth has been documented in areas with mangroves and marshes, where land is extended seaward on scales of kilometers over timescales of hundreds to thousands of years.

### D. Filtration

The benthos can also impact water clarity in coastal areas. The sediment-trapping capacity of emergent vegetation reduces sedimentation to the coastal waters beyond the marsh, mangrove, or seagrass bed. In tropical areas, for example, mangroves and seagrasses can trap sediments that might otherwise be transported onto coral reefs, where the suspended sediment would reduce water clarity and productivity and potentially smother corals. In temperate latitudes, marshes are well-known not only for their ability to trap sediments but also for their capacity to take up excess nutrients. Agricultural runoff to the coastal environment can result in eutrophication (discussed later), but the presence of salt marshes can significantly offset this problem

and provide a natural water filtration system. Mangroves have a similar capacity.

Suspension-feeding benthos can also have a major impact on water clarity by filtering out suspended particles. The reduction of oysters in estuaries along the east coast of the United States through overfishing, destruction of oyster reef habitat, and disease has resulted in a marked decrease in water clarity. Impacts of this magnitude are generally confined to shallow nearshore habitats with very high densities of suspension feeders, but the effect is nonetheless important.

### E. Secondary Production

Benthic megafauna and macrofauna provide an important source of secondary production. Some of this production, such as clams, scallops, shrimp, and crab, is directly consumed by humans and forms the basis of important commercial fisheries. However, even groups such as the polychaetes, which have no obvious economic importance, form an important food source for benthic-feeding fishes and crustaceans. Because sinking organic material collects on the bottom, the secondary production in the benthos can often be considerably higher than that in the water column above.

### F. Linking Biodiversity and Ecosystem Services

Although it is inarguable that benthic organisms provide key ecosystem services, there is no evidence that species diversity is important in maintaining these services. This deficiency does not necessarily mean that species diversity is unimportant but rather that we have not studied it. In virtually all the categories listed previously, there is a particular functional group that is critical to the service, but in most cases there is more than one species of a particular functional group in a given habitat. For example, deposit feeders are the major bioturbators in marine sediments, and bioturbation has major impacts on nutrient cycling, pollutant burial and mobilization, and sediment stability. However, most sediments that contain deposit-feeding organisms will contain multiple species. At least superficially, it appears that different deposit feeders often do the same sorts of things, although there are species that feed on particular grain sizes and at particular depth horizons in the sediment. What we do not know is whether those species that remain after others are eliminated can provide the same ecosystem services as the more diverse community. Will the removal of a given species

result in fewer total numbers of deposit feeders or will other species increase in number and/or activity to compensate? Elmgren and Hill provide evidence from the Baltic Sea that complete removal of a functional group (in this case, suspension feeders) will indeed fundamentally impact trophic linkages. Experiments to test these ideas are sorely needed. It is clear that changes in functional groups will alter the way in which ecosystem services are carried out, but any linkage between regional or local biodiversity and ecosystem services in marine systems remains to be demonstrated.

In some instances, there are likely to be keystone species that provide services disproportionate to their abundance. Clearly, the loss of these species will have a major impact. In other instances, there may be only a single species providing a service so that loss of that species will have an easily predicted outcome. Seagrasses and mangroves, for example, are often composed of only one or two species of emergent plant in a given area, and the loss of that species will have a clear effect. The loss of seagrasses to disease in Europe earlier this century had very clear effects on coastal sediments and shoreline erosion. Similarly, the loss of oysters described previously had an obvious impact. Unfortunately, we are not always able to predict keystone species and we are even less able to predict the cumulative impacts of multiple species that belong to a given functional group.

## VII. THREATS TO SEDIMENTARY DIVERSITY

Documented extinctions in the marine environment are relatively few, and many of these are birds or mammals. However, there are certainly biodiversity losses occurring in the sedimentary fauna. Particularly in coastal areas, where human impacts have been most severe, large areas of habitat are being damaged or lost. At the very least, genetic diversity is disappearing with these habitats. Species that live in coastal habitats are often adapted to local conditions and may therefore have a specific tolerance to variables such as salinity and temperature. Once this gene pool has been lost, there is no guarantee that conspecifics from other areas will be able to invade that specialized environment. There is also increasing evidence, summarized earlier, that a significant portion of sedimentary diversity is undescribed and may subsequently be lost without our realizing it even existed. With these losses, changes in ecosystem services could also be missed or misinterpreted.

It is possible to generalize that with increasing proximity to the shoreline, and therefore to human populations, the number and magnitude of threats to sedimentary diversity increase (Fig. 5). Thus, salt marsh and mangrove habitats are being reduced at a rapid rate, continental shelf habitat is being damaged in many although not all areas, and deep-sea habitats are currently being impacted in mostly isolated instances. The types of impacts are far from uniform between systems.

### A. Fishing Impacts

Most of the major marine fisheries in the world occur on continental shelf areas, and they can impact sedimentary habitats in several different ways. Among the most destructive forms of fishing are dredges and trawls that are dragged across the bottom to collect benthic fishes and invertebrates. These types of fishing gear are often heavy and specifically designed to dig into sediments, and in doing so they alter the sediment composition and geochemistry, damage infaunal organisms, and destroy any benthic epifauna and plants and the habitat they create for other species, including some commercial fish. Many fisheries are also notoriously nonselective, and subsequently remove (as by-catch) large numbers of species other than those intended. By-catch, which is often 50–80% of the total catch, is often thrown overboard; most organisms discarded in this manner will be unable to recover, and pockets of decaying organisms then create localized areas that are similar to eutrophied habitat.

Fisheries also tend to remove large and abundant predators, and because these are large and abundant taxa they are species that have a major impact on their ecosystem. There are well-documented cases of shifts in species composition and trophic mode that occur as dominant predators are fished out.

In general, fisheries tend to reduce benthic diversity and homogenize habitat, although the specific means by which diversity loss occurs is only well documented in the case of physical destruction of habitat. Indirect effects, such as predator removal, require additional study.

### B. Habitat Destruction, Degradation, and Shoreline Modification

Fishing activities are certainly the greatest source of degradation of bottom habitat beyond the immediate coastline, but coastal development has resulted in considerable “reclamation” of intertidal habitat ranging

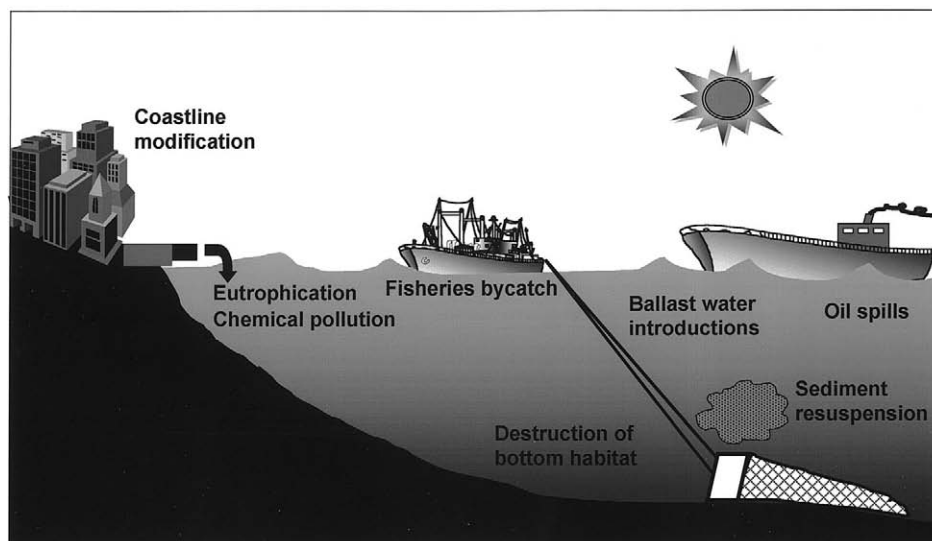


FIGURE 5 Schematic representation of major threats to sedimentary organisms and the services they provide.

from salt marshes to mangroves and mudflats. A significant portion of The Netherlands and Boston's Logan Airport are just two examples in which shoreline has been engineered by filling or erecting dykes to allow development at the land-sea interface. Globally, large areas of marsh habitat have been modified to allow building, and large areas of tropical mangroves have been eliminated to allow development of shrimp aquaculture farms. Clearly, the loss of large areas of habitat changes local species composition and likely reduces genetic diversity of individual species. Extinctions are likely occurring in some areas.

### C. Pollution and Eutrophication

The large human populations and industries that occur along coastlines throughout much of the world create problems of eutrophication, pollution, and habitat degradation. Eutrophication occurs when excess nutrients are supplied to coastal water from agricultural runoff and sewage. Phytoplankton and shallow vegetation respond to increased nutrients by growing, and as they die and sink to the bottom microbial decomposition occurs. Where large amounts of organic matter are available, microbial respiration exceeds photosynthetic production of oxygen and sediments, and bottom waters may become hypoxic or anoxic.

Pollutants are also problematic in coastal environments. Industrial and storm drain runoff deliver heavy metals, hydrocarbons, synthetic compounds (PCBs,

DDT, dioxin, etc.) and countless other toxic compounds to sedimentary communities. Some of these compounds result in increased respiration, reduced reproductive output, and countless other physiological responses that may ultimately lead to death. Alternatively, the physiological stress caused by contaminants may weaken an organism and increase its susceptibility to disease. Transfer of contaminants may occur by diffusion, ingestion of contaminated particles, or feeding on contaminated organisms.

Problems of eutrophication and pollution are most intense in industrialized and densely populated coastal areas where circulation with the open ocean is limited. Thus, partially enclosed estuaries and harbors provide no mechanism for rapid dilution and removal of contaminants. Not surprisingly, the deep ocean is largely buffered from most of these impacts by the size of the continental shelf, given that most contaminants arrive as runoff of some sort from land or rivers. Nitrogenous compounds may also be delivered over broader distances by atmospheric deposition. For the coastal areas that are impacted by eutrophication and pollution, many benthic species are unable to survive under these conditions, and bottom habitats become low in diversity and dominated by weedy species that thrive under such conditions.

### D. Exotic Species Introductions

Within recent years, there has been increasing concern regarding the impact that introduction of nonnative

(exotic) species can have on benthic communities. Although many people assume that the oceans are open systems and lack barriers to dispersal, there are actually many discontinuities in temperature, land barriers, depth, and ocean currents that limit the dispersal of reproductive propagules from different areas of the world. Historically, ships have provided a mechanism by which exotic species are transported to nonnative habitat. In the case of hard-substrate fauna, fouling of ship hulls allows adults to cross deep ocean basins that they could not otherwise cross because of a lack of suitable shallow habitat. However, for sedimentary invaders, ship ballast water is the main culprit. A tanker may take on large volumes of ballast water in one port, sail to a new port, and dump the ballast water before loading goods. The dumped ballast water may contain adults and/or reproductive propagules that have been carried through unfavorable open-ocean habitat.

In recent years, the increasing interest in developing marine aquaculture has also helped spread nonnative species. Nonnative species are transported into new environments in which adults or reproductive propagules can escape into the wild, invade nonnative habitat, and displace native forms. An additional problem is that parasites and pathogens can be transported along with species imported for aquaculture, creating additional problems. Another transfer mechanism that is less common, but embarrassing, is the transport of organisms throughout the world for scientific study; although scientists are usually conscientious about preventing individuals or reproductive propagules from escaping into the natural environment, there are instances of accidental release.

In many instances, nonnative individuals will not successfully colonize the new environment, but particularly when conditions are stressful for native organisms (e.g., pollution and unusually low salinity or high temperatures), the exotic species may not only invade but also displace native organisms. When displacement occurs, even after conditions return to normal, the invasive species may remain the dominant taxon. San Francisco Bay and Long Island Sound are two good examples in which it is well documented that many of the dominant species today are nonnative. In some instances, the exotic and native species may be able to coexist, resulting in slightly higher biodiversity, but the complexity of species interactions can sometimes result in elimination of native species. In other words, the introduction of nonnative species is a very dangerous and unpredictable disturbance whose potential impacts are only beginning to be understood.

## E. Global Climate Change

Perhaps the least easily predicted impact of human activity on sedimentary biodiversity is the long-term effects of global climate change. In a simplistic sense, species distributions might be expected to shift toward the poles to compensate for warming of the oceans. However, community response is likely to be far more complicated. Because the habitat of a given species is more complex than just temperature, it will often not be possible to simply shift distributions because suitable sediment and geomorphic features (e.g., productive banks) may be unavailable. As sea level rises, intertidal communities will not simply encroach landward because in populated areas we will prevent such encroachment to preserve expensive coastal development. Moreover, models of global warming predict potential circulation changes in the oceans, and if this indeed happens then a whole suite of ecological processes ranging from primary productivity to dispersal of reproductive propagules will be impacted. It is very likely that coastal habitats and species will be lost if significant ocean warming occurs, and it is possible that deep areas will also be impacted if these changes alter large-scale circulation patterns.

## VIII. SUMMARY

Although it is likely that ocean sediments contain in excess of 500,000 species, we know little of how this diversity is maintained and how loss of biodiversity may impact the ecosystem services that benthos provide. It is critical that benthic ecologists working in sedimentary habitats ride the current wave of interest in biodiversity to improve on these deficiencies in our knowledge. Marine sedimentary systems are widespread, rich in biodiversity, and ecologically important in terms of the processes that occur within them. It is therefore reasonable to expect exciting discoveries regarding biodiversity in these habitats in the new millennium.

## See Also the Following Articles

COASTAL BEACH ECOSYSTEMS • ESTUARINE ECOSYSTEMS • EUTROPHICATION AND OLIGOTROPHICATION • INTERTIDAL ECOSYSTEMS • MARINE ECOSYSTEMS • RESOURCE EXPLOITATION, FISHERIES

## Bibliography

Botsford, L. W., Castilla, J. C., and Peterson, C. H. (1997). The management of fisheries and marine ecosystems. *Science* 277, 509–514.

- Carlton, J. (1985). Transoceanic and interoceanic dispersal of coastal marine organisms: The biology of ballast water. *Oceanogr. Mar. Biol. Annu. Rev.* **23**, 313–371.
- Dayton, P. K., Thrush, S. F., Agardy, M. T., and Hofman, R. J. (1995). Environmental effects of marine fishing. *Aquatic Conserv. Mar. Freshwater Ecosyst.* **5**, 205–232.
- Elmgren, R., and Hill, C. (1997). Ecosystem function at low biodiversity—The Baltic example. In *Marine Biodiversity: Patterns and Processes* (R. F. G. Ormond, J. D. Gage, and M. V. Angel, Eds.), pp. 319–336. Cambridge Univ. Press, Cambridge, UK.
- Giblin, A. E., Foreman, K. H., and Banta, G. T. (1995). Biogeochemical processes and marine benthic community structure. In *Linking Species and Ecosystems* (C. G. Jones and J. H. Lawton, Eds.), pp. 29–36. Chapman & Hall, New York.
- Grassle, J. F., and Maciolek, N. J. (1992). Deep-sea species richness: Regional and local diversity estimates from quantitative bottom samples. *Am. Nat.* **139**, 313–341.
- Gray, J., Poore, G., Uglund, K., Wilson, R., Olsgard, F., and Johannesen, Ø. (1997). Coastal and deep-sea benthic diversities compared. *Mar. Ecol. Prog. Ser.* **159**, 97–103.
- Hessler, R. R., and Sanders, H. L. (1967). Faunal diversity in the deep sea. *Deep-Sea Res.* **14**, 65–78.
- Jablonski, D. (1993). The tropics as a source of evolutionary novelty through geological time. *Nature* **364**, 142–144.
- Lamshead, P. J. D. (1993). Recent developments in marine benthic biodiversity research. *Oceanis* **19**, 5–24.
- May, R. (1992). Bottoms up for the oceans. *Nature* **357**, 278–279.
- National Research Council (1995). *Understanding Marine Biodiversity*. National Academy Press, Washington, D.C.
- Norse, E. (Ed.) (1993). *Global Marine Biodiversity Strategy: Building Conservation into Decision Making*. Center for Marine Conservation, Washington, D.C.
- Pearson, T. H., and Rosenberg, R. (1978). Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. *Oceanogr. Mar. Biol. Annu. Rev.* **16**, 229–311.
- Peterson, C. H. (1979). Predation, competitive exclusion, and diversity in the soft-sediment benthic communities of estuaries and lagoons. In *Ecological Processes in Coastal and Marine Ecosystems* (R. J. Livingston, Ed.), pp. 223–264. Plenum, New York.
- Poore, G. B. C., and Wilson, G. D. F. (1993). Marine species richness. *Nature* **361**, 597–598.
- Rex, M. A. (1983). Geographic patterns of species diversity in the deep-sea benthos. In *The Sea* (G. T. Rowe, Ed.), pp. 453–472. Wiley, New York.
- Rex, M. A., Stuart, C. T., Hessler, R. R., Allen, J. A., Sanders, H. L., and Wilson, G. D. F. (1993). Global-scale latitudinal patterns of species diversity in the deep-sea benthos. *Nature* **365**, 636–639.
- Sanders, H. L., and Hessler, R. R. (1969). Ecology of the deep-sea benthos. *Science* **163**, 1419–1424.
- Snelgrove, P. V. R. (1998). Getting to the bottom of marine biodiversity: Sedimentary habitats. *BioScience* **49**, 128–138.
- Snelgrove, P. V. R., Blackburn, T. H., Hutchings, P. A., Alongi, D. M., Grassle, J. F., Hummel, H., King, G., Koike, I., Lamshead, J., Ramsing, N., Solis-Weiss, V., and Freckman, D. (1997). The importance of marine sediment biodiversity in ecosystem processes. *Ambio* **26**, 578–583.
- Watling, L., and Norse, E. (1998). Physical disturbance of the seabed by mobile fishing gear: A comparison to forest clear-cutting. *Conserv. Biol.* **12**, 1180–1197.



# MARKET ECONOMY AND BIODIVERSITY

R. David Simpson and Pamela Jagger  
*Resources for the Future*

---

- I. Biodiversity and Market Economics
  - II. Biodiversity in Existing Markets
  - III. Making Markets
  - IV. The Market as a Mechanism for  
Conserving Biodiversity
- 

## GLOSSARY

**certification** Procedure of determining whether a particular process is conducted in an ecologically benign fashion (also referred to as eco-labeling).

**economic efficiency** State of affairs in which all goods in an economy are allocated in such a way that no one can be made better off without making someone else worse off; also known as Pareto optimality.

**hedonic price** Price of a good expressed as a function of the particular attributes it embodies.

**marginal benefits and costs** Additional benefit or cost incurred in response to a small incremental change in the quantity of some variable (in the limit, the first derivative of the benefit or cost function).

**market-based incentives** Policy designed to change behavior by altering the price paid, or cost borne, by a person engaged in an activity that affects biodiversity.

**market economy** Organization of economic activity in which private individuals buy and sell goods at prices that balance supply and demand.

**open access resource** Resource that can be exploited by any person, regardless of the consequences of

that exploitation for other would-be users; a resource for which *property rights* (see below) are not well defined.

**perfectly competitive economy** Ideal organization in which all goods are owned by perfectly rational, perfectly informed individuals, all goods can be bought and sold in markets, and no one individual can influence the prices at which transactions can take place. A perfectly competitive economy is efficient.

**property right** Legal and social construct under which the owner of any good cannot be deprived of it without the payment of compensation that is acceptable to her.

**public good** A good which, if one person consumes it, everyone else does as well, and for which one person's consumption does not diminish another's.

---

**A MARKET ECONOMY IS A SYSTEM** in which private individuals make decisions regarding how much to produce and how much to consume so as to best meet their own wants and needs. Market economies typically do not result in the most socially desirable production and consumption of all goods, as there are some instances in which the individual self-interest that drives the market economy should be tempered by a concern for the welfare of others. This is the case with biodiversity. Often one person can make herself better off by diminishing the biodiversity around her, but in doing

so she ignores the effects of this reduction on others. Ideally, people would be required to incorporate the effects of their actions on others in making their production and consumption decisions. There are, however, some activities that are consistent with both improving the welfare of those conducting them and preserving biodiversity. These market-based approaches to preserving biodiversity may be promising in some contexts, but they are not without a number of difficulties. It seems unlikely that private markets will ever be able wholly to reflect the many values to which biodiversity gives rise.

## I. BIODIVERSITY AND MARKET ECONOMICS

Biodiversity confers a great many benefits on society. These benefits are most easily recognizable when embodied in a commodity of market value. In such instances, social benefits can be proxied by the prices people are willing to pay. Many other benefits arising from biodiversity are not bought and sold in markets. The market values of biodiversity, in combination with less tangible values that cannot easily be measured by market prices, represent the total value of biodiversity to society. The challenge to the natural or social scientist, policymaker, or resource manager is to recognize both the market and the nonmarket values associated with biodiversity and to direct social policy accordingly.

The full range of biodiversity values is discussed elsewhere in this Encyclopedia. The task here is to perform a more detailed analysis of the relationship between biodiversity and the market economy. To begin with, the phrase “market economy” can be used in two very different ways. The first is the textbook abstraction more often referred to in the economics literature as “perfectly competitive general equilibrium,” or a “perfectly competitive economy.” In this sense a free market economy is an idealized state in which all goods and services are owned by perfectly rational individuals, and assets change hands in voluntary trades. Under these conditions (and some additional assumptions) an ideal free market would allocate goods and services efficiently. Economists use the word “efficiency” in the sense of “Pareto optimality”: a state of affairs in which no one could be made better off without making someone else worse off. While proving that a perfectly competitive economy is efficient is decidedly not trivial, it makes a certain intuitive sense: when everyone is free to trade with everyone else, the process of trading will

continue until no mutually beneficial trades can be made. Efficiency does not embody everything one might desire of an economic system, but a basic result in economic theory establishes that systems can be made both “fair” and efficient by first distributing wealth fairly and then allowing the market economy to function without hindrance.

The second sense of the phrase “market economy” is one in which *certain*—but, importantly, not all—goods and services are owned and can change hands in voluntary transactions. Because only some goods and services are subject to private ownership, this “free market economy” is decidedly *not* efficient. Those who suffer as a result of the economic choices of others are not afforded an opportunity either to offer payment to dissuade others from harming them or to demand compensation for such harm.

We will use the phrase “market economy” in the second sense in this article. In particular, we will concentrate on the relationships between biodiversity and a real-world economy that is clearly imperfect in the ways in which it produces and allocates some goods and services. In particular, we will ask to what extent “the market” can be relied on to provide incentives for the preservation of biodiversity, and to what extent a more active public policy may be required.

It is important to note that “active public policy” may be a more effective and efficient thing in theory than it is in practice. The “benevolent social planner” who appears in economic analyses to make such policy is every bit as much a *deus ex machina* as is the idealized perfectly competitive market itself. Choices must be made between relying on an imperfect market economy on one hand and, on the other, government officials who can only be expected to advance the general welfare imperfectly. Which option is better is not clear in general, and is likely to depend on particular circumstances.

Despite our focus on the existing market economy and its imperfections, it can still be useful to compare it with the benchmark provided by the textbook abstraction of the perfectly competitive economy. It is revealing to see how and why actual economies fall short of this ideal. In the following section we discuss the conditions that characterize a perfectly competitive economy, and why these conditions are unlikely to be met with respect to biodiversity. We turn next to a discussion of why and how biodiversity can be undervalued in an imperfect economy. Following this discussion, we consider some of the ways in which some biodiversity values are reflected even in imperfectly organized real-world economies. Next, we consider steps that could be taken to rectify some of the “market failures” that result in the

undervaluation of biodiversity, and review some evidence on attempts that have been made to correct these failures. We conclude by noting that while markets should be relied on to preserve biodiversity to the extent to which they are able, private markets will probably not soon, if ever, provide completely adequate incentives for that purpose.

### A. Biodiversity in an Ideal Economy

A good place to begin is to ask what conditions would characterize the amount of biodiversity preserved in an economy that recognizes the total value of all aspects of biodiversity. An important principle in economics is that the maximization of *total* value is achieved by meeting conditions on *marginal* values. The marginal value of something is its incremental contribution to total value (in mathematical models, it is the first partial derivative of total value with respect to the good in question). In a perfectly competitive economy, all goods would be provided in such a way as to afford equal marginal value. The intuition behind this notion is straightforward. Finite resources place constraints on what we as a society can have and use. An optimal allocation of goods would be one in which we cannot have a little more of one thing without losing a quantity of something else that is at least as valuable. If one of the “goods” in question is the amount of biodiversity we maintain, the value an ideal market places on it would be equal to its marginal contribution to welfare. This would, in turn, be equal to the marginal contribution afforded by any of the other things that we could have instead of that incremental amount of biodiversity. In short, the value of biodiversity—its price in an ideal economy—would, as with every other good, be set in accordance with its marginal contribution to the general welfare. The marginal contribution made by biodiversity to welfare might be realized directly, in enjoyment of the preservation of biodiversity for its own sake, or indirectly, via its contribution to the generation of other goods and services.

### B. Practical Problems in Preserving Biodiversity in Markets

Biodiversity differs in an important respect from bread, automobiles, or other goods that might be better candidates for allocation in an unregulated economy. Biodiversity is (at least in some aspects) an example of what is known in economics as a *public good*; it is something that, if provided to one person, is necessarily available

to others. Inasmuch as many people can benefit from the preservation of the same biodiversity, its marginal contribution to welfare should properly be calculated as the *sum* of the incremental contributions to each affected person’s welfare. Closely related to this consideration, however, is the fact that no one can be excluded from the enjoyment of biodiversity. If a rare species is saved from extinction, all those people who value its continued existence benefit, regardless of whether they have made any sacrifice themselves to maintain it. In short, if one person pays for the conservation of biodiversity, many benefit. This is what is known as the “free rider” problem in economics.

The existence of public goods typically means that market economies do not efficiently allocate goods. This means that the real-world economy, in contrast to the perfectly competitive ideal, provides insufficient incentives for preservation. In theory, one could create appropriate incentives for the provision of public goods through elaborate rules for the punishment of free riders and for revealing individuals’ true attitudes toward resources. Given the formidable practical impediments to applying such fixes to a problem as complicated as the preservation of biodiversity, however, it seems unlikely that private markets could ever fully handle the problem in practice.

It should be noted, if only for completeness, that the preceding passage assumes that there is a problem. If the nonmarket aspects of biodiversity prove to be negligible, then by definition there would be little problem with allocating what we might label “biodiversity-related” goods in markets. One should also bear in mind what was said earlier concerning the difficulties inherent in correcting “market failures” with public policy options that are themselves inadequate or poorly designed. While there is considerable debate about the extent of the problem, a conclusion that the values to which biodiversity gives rise are negligible is surely premature. Current evidence suggests that the issue is not so much whether biodiversity conservation is a challenge facing society but rather, given that it is a significant challenge, how it compares in relation to other social and economic issues. We will continue to assume that this is, in fact, an important problem.

The observation that markets do not allocate public goods efficiently is not tantamount to saying that markets cannot, and for that matter do not, affect the preservation of biodiversity. In the following section we consider some of the ways in which biodiversity enters into market transactions. The focus will be on determining the extent to which market forces may be counted on to alleviate the biodiversity conservation problem.



## II. BIODIVERSITY IN EXISTING MARKETS

It is important to avoid a logical fallacy in thinking about the relationship between biodiversity and the market economy. Private markets tend to undersupply biodiversity because of the free rider problem. This does not necessarily imply that data from market transactions are irrelevant in estimating the benefits that biodiversity provides society. To give an example, a farmer may not be required to pay for the services that bees from an adjoining forest provide by pollinating his crops, but the value of those services is reflected in the earnings the farmer receives from his crops and, by extension, in the market value of the farmer's land. Thus, the social value of the bees is not incorporated in the forest landowner's decision to fell her trees, but is reflected in the change in earnings the farmer experiences as a result.

This section considers the importance of biodiversity in a number of sectors of the economy. Some values of biodiversity are reflected in a number of market transactions. The contributions of biodiversity to many economic activities are the subject of dispute among experts, however, and in the opinion of some of those experts they are negligible. From this observation the reader should draw two conclusions. The first is simply a restatement of what we have just said: Experts disagree as to how effective market incentives can be for preserving biodiversity.

The second conclusion to be drawn from an overview of the contributions of biodiversity in the market economy is that, whether the accurately and fully measured contributions of biodiversity to the observable and quantifiable output of the economy prove to be large or small, they may still provide something less than complete guidance to policymakers. It would still be important to consider the incentives to preserve biodiversity for ecological, ethical, esthetic, or even spiritual reasons. Few economists would go so far as to say that everything of value is recorded in the measured economy.

Let us now consider some of the ways in which biodiversity enters into the measured economy.

### A. Agriculture

The market for agricultural produce provides us with an example of biodiversity integrated into the market economy. All agricultural products are derived from what were, at one time, wild organisms. Today, though,

agricultural productions relies on only a relatively small handful of species; Wilson (1992), for example, noted that the vast majority of human nutritional needs are met by only about 20 species. This is despite the fact that many thousands of species are edible and potentially cultivable.

The observation that this base is narrow is not necessarily a compelling argument that diversity is undervalued, however. Specialization in particular crops would not occur if there were not some savings in production or marketing costs. In addition, consumer preferences with respect to variety, and how these preferences vary geographically and culturally, also factor into the issue of the degree of specialization in agricultural production. The optimal degree of product diversity is an open question in economics more generally. About the only thing that can be said definitively is that the answer depends on the specifics of the situation, and therefore that different results will obtain under different circumstances.

There are, however, some reasons to suppose that existing markets for agricultural products do not generate sufficient variety. One such reason has to do with the inability of producers of differentiated products to realize the full benefits from their provision. Consumers receive what is called in economics "consumer surplus" from the purchase of products. If I only have to pay \$1 for a product that I would have been willing to pay \$2 for, I receive  $\$1 \equiv \$2 - \$1$  in consumer surplus from my purchase. Because this surplus cannot, by definition, be appropriated by the seller of the product, the seller does not consider it in making the decision of whether or not to introduce a new product.

Another way in which unregulated markets may offer too limited an agricultural product line concerns the spread of pests and disease. Every farmer, by making his decision of what to plant, affects not only what is going to be available to consumers, but also what may befall other farmers. Mile upon mile of uninterrupted fields of nearly identical crops may comprise a low-resistance passageway for the spread of pests or disease. A farmer who plants a different crop might expect to achieve greater profits in the event that others' crops are wiped out, but he or she would probably have inadequate incentives to provide such a pest-break for several reasons. First, as noted earlier, the farmer would not receive the full benefits of his contribution, as consumers would also receive some additional surplus. Second, by providing the barrier to the spread of pests, the farmer is not only benefiting himself and consumers, but also neighboring farmers who would be less vulnerable to the spread of pests.

Finally, there are circumstances in which political interference with what might otherwise be free markets prevents desirable outcomes. Farmers would have greater incentives to plant different, disease-resistant varieties to the extent that they believe they would be able to achieve high profits by charging high prices should their crops be successful while their neighbors' crops failed. Yet ample historical evidence suggests that governments generally step in to prevent such "price gouging." It is a situation in which *de facto* restrictions on profiting after the event of a catastrophic disturbance create perverse incentives for protecting against such occurrences in the first place (we are grateful to Professor Brian Wright of the University of California for making this point).

We might also remark in passing that while agricultural markets represent an instance in which one component of biodiversity is integrated into the economy, agricultural practices themselves affect the status of biodiversity more generally. Each farmer may be better off over the short term by applying more pesticides and fertilizers to her fields, but these actions affect adjoining, and sometimes even distant, ecosystems and the diversity they sustain. Perhaps most importantly, the sheer scale of agriculture affects patterns of land use and conversion. More land in agriculture typically means less land in forests, wetlands, grasslands, or other habitat more conducive to the support of indigenous biodiversity. As we will see, this reduction in surrounding biodiversity may also have an effect on the productivity of agriculture.

## B. The Harvest of Wild Organisms

Settled agriculture might be regarded as the culmination of a process that began with hunting and gathering. While the vast majority of food products and fiber of organic origin produced today comes from farms, hunting, gathering, and particularly fishing remain important in many areas. Unlike the case of settled agriculture, where many management practices (e.g., soil conservation) are intended to maintain production into the indefinite future, the harvest of organisms from their indigenous habitats can present a threat to the continued existence of the targeted population or species.

This can be true for two reasons. The first is that there is often a market failure that prevents private agents working in their own self-interest from behaving efficiently. Many wild populations are subject to open access. By this we mean that any or all of a number of individuals are able to harvest organisms without having to compensate the others for the privilege of doing

so. This leads to a situation that has come to be known as the "tragedy of the commons" (Hardin, 1968). If a resource base is subject to open access, benefits from its exploitation will be dissipated by excessive entry. So long as it is profitable to engage in harvesting wild organisms, more and more people will do so, until there is no further advantage to be gained from additional exploitation.

This outcome is perhaps most clearly seen in the case of fishing in international waters. Since no one owns the open ocean—and hence fishers from any nation cannot be excluded from fishing on the high seas by those of any other nation—international exploitation of fish stocks has often proved excessive. There are also terrestrial examples. Animals have been hunted to the point of extinction in various places and times. Similarly, in situations where property rights are nonexistent, poorly enforced, or subject to frequent abrogation, forests may be felled at excessive rates.

The solution to the "tragedy of the commons" is either to define property rights in the resources—make it illegal for me to catch a fish or shoot an elephant that "belongs" to you—or to achieve the same end by regulating timing, location, or other attributes of harvest. Regrettably, this is often more easily said than done. It can be difficult to define ownership in fugitive resources. If animals are not fenced in, property rights in real estate confer rights to the animals resident on it only during such periods as they are actually on the property. Extending rights to the animals themselves is difficult when individuals cannot be readily identified. The alternatives of selling rights to entire populations or of devising schemes for defining the ownership of particular animals may not be easily practicable. Despite these practical difficulties, some promising experiments are under way with assigning such property rights (see Section III,A). Regulatory approaches are also problematic. Outright prohibitions can be difficult to enforce. More flexible approaches, such as tradable fishing quotas, can result in unintended consequences, such as discarding "bycatch" of species that the fisher does not have the right to catch.

The "tragedy of the commons" gives rise to both static and dynamic problems. From a contemporaneous viewpoint, the more I harvest today, the less there is for you to harvest today. It is also true, however, that the more I harvest today, the less there will be for either you or me to harvest tomorrow. Now if I do not trust you to leave anything for me to harvest tomorrow, I will take as much as I possibly can today. If everyone behaves this way, the target population may be extinguished.

While imposing property rights may mitigate the “tragedy of the commons,” the assignment of property rights does not assure that whatever harvesting practices are then adopted will be sustainable. It can be shown (Clark, 1991) that the optimal path for exploiting a biological resource depends on the rate of interest. In this analysis, the target population is regarded as an asset, which must compete with other assets in the economy. The hypothetical owner of, say, a fishery, would manage it so as to equate the rate of return on owning fish with that from holding an alternative asset, such as a bank account. It can be shown that under some circumstances the profit-maximizing strategy of a private owner would be to completely extinguish the fishery and invest the proceeds of this effort in an interest-bearing bank account.

This finding is, understandably, regarded as morally repugnant by many. The fault lies not with the logic driving the conclusions, but with the assumptions underlying the analysis. To suppose that a species *should* be extinguished if its rate of intrinsic growth does not compare favorably with the rate paid on Treasury Bills presumes that only the profits for those harvesting the species matter. Defining the problem more broadly to include the benefits to society of maintaining a species may yield a different answer. Thus we see that even a relatively well-functioning “free market economy,” one in which the rights to harvest organisms are completely defined and allocated, will not achieve truly optimal outcomes unless all those who care about the existence of biodiversity *per se* are afforded opportunities to purchase and retire such rights of harvest.

Before leaving the subject of hunting and gathering, we might briefly consider the special topic of its conduct in less-developed countries. Some research has suggested that biodiversity-rich natural habitats can be considerably more valuable as sources for the sustainable collection of nontimber forest products (NTFPs) than they would be if clear-cut and converted to pasture. Despite these findings that land would be more valuable in a more natural state, land is, in fact, being clear-cut and converted to pasture in many of the areas studied. There has been a great deal of criticism, on both conceptual and empirical grounds, of such studies. Moreover, some commentators have expressed doubts that the collection of forest products in commercial quantities can be truly “sustainable,” in the sense of preserving a natural, diverse forest ecosystem.

This is not to say that profitable, sustainable NTFP collection is impossible everywhere or that it does not occur in some areas. In most parts of the world, however, technological progress and increasing scarcity of

land motivate more concentrated production via *ex situ* cultivation. As with the other strategies for the conservation of biodiversity in developing countries, the important considerations in evaluating the efficacy of NTFP collection as a conservation incentive concern measuring its true profitability and assuring that it is, in fact, conducted in such a way as to preserve biodiversity-rich habitats.

### C. Biodiversity Prospecting

Perhaps the area in which the market economy has come closest to pricing biodiversity *per se*, as opposed to elements of it, is in “biodiversity prospecting.” Biodiversity prospecting is the search among wild organisms, both plants and animals, for new products of industrial, agricultural, or, especially, pharmaceutical value. Researchers from a number of industries have undertaken searches in many nations and ecosystems for valuable products. In some instances cash payments have been made for rights of access to indigenous biodiversity, while in other cases search is undertaken under contracts that specify royalties to be paid in the event that commercial products are developed from natural sources.

Despite the fact that such compensation has been paid or promised, the potential of these market forces to motivate conservation remains unclear. This is so for several reasons. First, it is not clear that the property rights necessary to motivate truly remunerative payments have yet been established. Some commentators have, in fact, suggested that fuller delineation of such property rights will lead to greater payments for access to biodiversity and, by extension, for its conservation. On the other hand, most existing arrangements offer compensation to individuals or organizations that do more than simply provide access to samples for testing. In many instances, the initial steps of collection, identification, preparation, and sometimes even preliminary testing are conducted by the sample seller. It is unclear, then, what share of observed payments and prospective royalties are allocated for the natural samples themselves as opposed to the labor and other inputs involved in their processing. Finally, many existing agreements specify royalties in the event of discovery. The probability with which discoveries will be made is unknown, however, and in many cases the rates at which royalties would be paid are not made public. Thus, it is impossible to infer the values attributed to the resources themselves.

For these reasons, valuation of genetic diversity remains highly controversial. Absent clear evidence from

market transactions, attempts to value biodiversity for its use in the development of new products have been based on indirect inferences. One thing is obvious, however. The aggregate value of biodiversity for its use in new product development is astronomical. We would not have food, nor many of the other necessities of life, were it not for the natural organisms from which they are derived. Even the value of extant biodiversity in the development of new products and the improvement of existing ones can be quite substantial.

Economic value, however, is determined by “value on the margin.” Whatever philosophical reservations one may have concerning this principle, it is the way in which businesses make decisions. Hence the value that private companies assign to biodiversity will be determined by the marginal contribution that additional biodiversity makes to their profitability. This marginal contribution is composed of two elements. The first concerns the expected contribution of an incremental component of biodiversity *if it proves to be the best source of a new or improved product*. The second concerns the probability with which other potential sources will prove to provide better leads for new product development. Simpson, Sedjo, and Reid (1996) argue that this marginal value is negligible. Thus the “free market” value of biodiversity with respect to the search for new products is neither a compelling reason for preservation nor a plentiful source of funds for habitat preservation (they also emphasize that this conclusion says nothing about the value of biodiversity in any of its myriad other uses and aspects). This analysis has been revisited by Rausser and Small (in press), who argue that earnings could be higher in exceptional circumstances. There seems to be little disagreement, however, that appreciable earnings could only be realized in such exceptional circumstances, and that biodiversity prospecting does not hold great promise as a tool for conserving all of the many areas in which biodiversity is now threatened.

#### D. Biodiversity Values Embedded in Property Prices

Economists often suppose that the price of some property reflects the advantages of its location. Statistical techniques exist for inferring the “hedonic prices” implicit in such properties. For example, if a hectare of land in a city sells for \$10,000, while another, similar in all ways save that it adjoins an area of parkland, sells for \$15,000, the value of proximity to natural amenities may be inferred to be \$5000.

This technique of hedonic pricing has been used in a variety of contexts. In addition to its common use in

the valuation of differentiated products, hedonic pricing studies have been employed to measure the market value of environmental amenities, such as clean air. There have been few hedonic pricing studies that attempt to measure the market value of proximity to biodiversity, however.

The reasons for this are related to the difficulties in defining and measuring biodiversity, and in separating its value from the values of undeveloped habitat more generally. Empirical economists must rely on natural experiments for the generation of their data. Controlled experiments, at least of a phenomenon as complicated as the valuation of biodiversity, cannot be conducted in the laboratory. Only the extreme points of the observed distribution provide unambiguous data. At one end of the spectrum, market economies typically place the highest value on land in densely populated, largely “unnatural” areas. At the other extreme, lands containing the greatest diversity of (relatively large, at least) organisms are often virtually valueless, as reflected in market prices. In many instances biodiversity continues to thrive in habitats far enough distant from concentrations of population and industry so as not to have been brought within the modern market economy.

Yet between these extremes it is undeniable that proximity to nature does contribute to the values of some properties. The city lot adjoining the park is the canonical example. However, one must be clear what it is one is considering. The park in the city, often laid out on land cleared decades or centuries earlier, cannot compare in diversity to truly pristine habitat. The evidence that markets value open space is considerably stronger than that they value diversity per se.

#### E. Tourism and Conservation Incentives

One way in which biodiversity may be reflected in the market economy is through its impact on patterns of, and expenditures on, tourism. Earnings arising from rain forest excursions, river rafting, photo safaris, hunting, and other such nature-based tourism options have proved more lucrative in many areas than are alternative, less ecologically benign activities. Travel to destinations that derive their appeal from their proximity to living assets but that does not degrade such assets (one might define this as “ecotourism,” although the precise definition of the term is a matter of contention) is a promising option in many parts of the world.

Such nature-based tourism is not without its pitfalls, however. Encouraging tourism in ecologically sensitive areas can have unintended consequences: hotel, road, and trail construction, the physical passage of tourists

through sensitive habitats, and the depletion of forests for firewood are examples. Making a comfortable and attractive destination for tourists can be a different thing than preserving an area's full biodiversity.

The relationship of even innocuous ecotourism to the market economy may be complicated as well. If there were sufficient money to be made from the operation of tourism facilities or related activities, one would expect that more agents in the market economy would identify and exploit the opportunities for doing so. It is problematic, then, that some of the impetus behind nature-based tourism comes from government, international, or private donors. If there is not money to be made in tourism, one has to ask if conservation-related funds might not be better spent on other activities. Moreover, a number of analyses suggest that a disproportionate share of the expenditures related to tourism are often received by the providers of transportation and other services rather than by those who "maintain" the natural habitats of the destinations, commonly assumed to be the local populations who forgo other use of the resources. Incentives for conservation at the local level may, then, be attenuated.

A fundamental principle of economics is that value is related to scarcity. As was the case with biodiversity prospecting, money might be made at truly spectacular and unique locations. Responsibly conducted tourism in such places is clearly desirable. The truly spectacular and unique is inherently scarce, however. Thus the potential of tourism to conserve biodiversity must also be limited. Tourists can choose from a variety of spectacular destinations: there are rain forests, coral reefs, high mountains, and arid deserts on most of the continents. Inasmuch as all of these destinations are in competition with one another, this competition will reduce willingness to pay for travel to a particular destination. Moreover, there are any number of potential tourist destinations that could be developed. We return again to our earlier points. Private incentives may be adequate for the establishment of tourism locations where they are economically justified, but more effective conservation strategies might be pursued when ecotourism is not justified.

## F. The Diffuse Benefits of Biodiversity

The benefits of biodiversity may be widespread and diffuse, and for that reason not easily identified in the market prices of particular properties. This begs the question as to whether such values can be identified at all within the prices of the existing market economy.

The answer to the question depends on the nature of the values.

On one hand, a host of services are provided by diverse natural ecosystems that, while they themselves are not bought and sold in market transactions, are essential inputs into things that are. Ecosystems purify water, cycle nutrients, capture and break down pollutants, harbor pollinators, moderate local and global climate, reduce the frequency and intensity of flooding, retard erosion, and provide many other services. In the case of most of these services, however, natural ecosystems provide these services to some particular set of properties. If, for example, the supply of ecosystem services makes particular properties more agriculturally productive, then the measurable product of the agricultural properties should reflect the contributions of the ecosystems. Although it may be difficult to determine which properties benefit by how much, total effects would appear in aggregate statistics.

Thus national income accounting would reflect, if imprecisely and indirectly, the services of natural ecosystems. On this score, one can read the evidence in several ways. The advanced industrial economies that have achieved the greatest measured economic performance are typically not the best endowed with biodiversity. Conversely, the less-developed countries are often rich in biodiversity and other natural resources. To suppose that the historical "success" of the industrialized nations indicates the insignificance of biodiversity to measurable economic performance may be premature, however. First, such a statement may not be true in a broader historical—or even prehistorical—and geographical context. Jared Diamond (1997) has argued that the achievements of the wealthiest nations are attributable in part to their inhabitation of environments with greater exposure to broader arrays of organisms. Second, some would argue that the apparently wealthier nations are incurring an unrecognized debt by straining their natural asset base to a point at which decline may be inevitable or recovery long and expensive. Third, countries differ in any number of respects. The advanced industrial economies differ from the less-developed countries in the education of their populations, the vintage of their technologies, the nature of their institutions, and in innumerable other respects over and above the state of their biological resources. To say that the nations of the "first world" have achieved their status despite the initial paucity or subsequent degradation of their biological resources is not so say that their performance might not have been better still had their economies grown less profligately.

Even if one were to claim that the leading industrial

economies have gotten where they are today either because of or despite their relative lack of biodiversity, and even if one were to project that this leadership would continue, it would not provide a satisfying answer concerning the importance of biodiversity to society. There may be very important things that are not, and may never be, traded in markets.

Foremost among these important things are those that go under the rubric of nonuse, or existence, values. These are benefits that are wholly unrelated to any actual or potential current or future consumption. It is posited that people derive satisfaction merely from the continued existence of some elements of the natural world, regardless of whether they, their descendants, or anyone else will ever benefit directly from such elements. Philosophical debate continues on the possibility and magnitude of values that are, by construction, totally divorced from the price of any good bought or sold in markets. It seems reasonable to suppose, however, that the relationship between some of the things people care about and market transaction is so tenuous as to make inferring values from market price data impossible. At best, then, the market economy can provide only a partial guide to the social benefits of biodiversity.

### III. MAKING MARKETS

The conclusion to the previous section notwithstanding, the incentives provided by the market economy should be aligned as closely as possible with the social and economic values afforded by biodiversity. This can be done in two ways. First, while the establishment of property rights in wild organisms and their habitats cannot be expected to solve all problems relating to the preservation of biodiversity, it can at least amplify the incentives now provided by private markets. Second, “market-based incentives” in biodiversity preservation can be established by government action, and may achieve the socially desirable level of biodiversity preservation at less cost to society than would be incurred under more direct regulation. This section treats these two themes in turn.

#### A. The Extension of Property Rights and the Preservation of Biodiversity

We noted earlier that the extension of property rights is not a panacea for the preservation of biodiversity. Vesting an owner with *some* property rights encourages

her to maximize the value she can obtain from the resource, but unless ownership is assigned to absolutely *all* the things people care about, inefficient choices may be made in resource uses. For example, providing legal title to a forest may remove the incentive to cut it down before somebody else does, but does not mean that the forest will remain standing if its owner would make more money by converting the land to residential or agricultural use.

Traditional societies have often been able to avoid “the tragedy of the commons”—the tendency of all users of a resource to overexploit it—by developing societal rules for sharing a resource. A community in which a fishery, game population, or forest has managed to survive for centuries is probably one in which rules have evolved over time for allocating rights to exploit the resource within the community. This may take the form of *de facto* ownership of certain areas, or of restrictions on the amount of the given resource that any particular individual is allowed to harvest. As traditional social structures are supplanted, however, establishment of *de jure* property rights may again result in sustainable management being undertaken.

Some recent developments in property rights and conservation policy can be interpreted as efforts to replicate the procedures that traditional societies commonly establish. The Community Areas Management Programme for Indigenous Resources (CAMPFIRE) in Zimbabwe gives local people the right to manage the wildlife that crosses their land. CAMPFIRE illustrates an interesting paradox that arises in many biodiversity management situations. If local people are not given the right to manage herds—including the right to benefit from hunting—they may have no incentive to preserve endangered species at all. In fact, to the extent that large animals such as elephants can be extremely destructive of farmer’s fields, local people would otherwise have a clear incentive to eradicate them. By providing an incentive to cull animals in a sustainable fashion, CAMPFIRE removes the incentive to eradicate the entire population. This is not to say, of course, that a program like CAMPFIRE will always be successful—CAMPFIRE itself has received some criticism. However, this case may demonstrate that “exceptions prove the rule.” Areas in which such programs are not successful are often those in which community involvement and cohesion are low.

In the United States, Defenders of Wildlife, a non-profit organization devoted to conservation, has initiated programs to encourage the reintroduction of wolf populations. The “Wolf Reward Program” makes payments to landowners who can demonstrate that wolves

have reproduced on their properties. A related program compensates livestock owners for losses from predation by wolves. In each instance, Defenders of Wildlife encourages protection of the wolf population by making payments for occurrences that had previously not been the subject of market transactions: the reproduction of wolves in the first case, the loss of livestock in the second.

Efforts such as the Defenders of Wildlife programs and CAMPFIRE may be the prototypes for more widespread future imitation. One might expect to see increasing extension and enforcement of property rights as time passes. The economic theory of property rights holds that property rights come into existence when the benefits of their definition exceed the costs of their enforcement. When populations are large it makes little sense to sell the right to cull individuals; because there is little scarcity, such rights would sell for less than the cost of initiating and monitoring the transaction. As populations grow smaller, however, the benefits from allowing hunting may grow very large.

## B. Market-Based Incentives

Traditional regulatory approaches to environmental matters have often taken the form of so-called “command and control” instruments. Under these approaches a regulator “commands” people to do certain things (e.g., reduce pollution or preserve wetlands) and “controls” them by imposing civil or criminal penalties. National and international regulation pertaining to biodiversity has generally taken this form. The Convention on International Trade in Endangered Species (CITES) outlaws trade in certain listed species. The United States Endangered Species Act prohibits harming of listed species.

Command and control regulations are generally not favored by economists. Their argument consists of two parts. The first is that regulations typically should not totally forbid activities. The depletion of biodiversity is clearly not desirable *per se*, but our lives would indeed be “nasty, brutish, and short” if our ancestors had not reconfigured the natural world to at least some extent. Given this fact, the second component of the argument against command and control regulation is that the regulatory burden should fall most heavily on those individuals or firms who can most easily bear it. This is exactly what “market-based incentives” are intended to do. If some biodiversity loss is believed to be the necessary consequence of economic development, we should at least design programs to get the most development in exchange for the least loss.

Some such programs are now in place for biodiversity-related matters. In the United States, for example, private parties can buy and sell wetlands and the obligations to establish wetlands. A number of suggestions have been made for modifying the Endangered Species Act by introducing market-based incentives. Perhaps the most prominent of these suggestions concern “tradable development rights” (TDRs). Under such plans a certain number of permits for the conversion of endangered species habitat would be established. Any conversion of habitat would require the purchase of a permit, and with it the obligation on the permit seller’s part to maintain a specified area of land as habitat for the endangered species. The advantage of such a plan would be that it would allocate the preservation obligation in the most cost-effective fashion. Owners of endangered species habitat with high commercial value would buy permits from those whose lands have low commercial value, and the species would be preserved at the lowest overall cost.

Of course, real-world complications can generate more controversy than is suggested by this thumbnail sketch. The problem with any program that involves a trade-off of one unit of habitat for another (or, perhaps in the future, one unit of biodiversity for another) lies in defining the “unit.” A hectare of land in one location may be “worth” more or less than that in another with respect to its capacity for supporting biodiversity. Such problems of commensurability grow even more complex if, instead of evaluating habitat with respect to its capacity to support particular species, the issue becomes one of which species to protect. Until an operational consensus emerges as to how to measure and trade off biodiversity, all but the most rudimentary tradable development right programs will be impossible to implement.

Another way in which market-based incentives for biodiversity preservation can be established is through “certification” or “eco-labeling.” In these programs certain products often associated with biodiversity loss (most prominently timber) are certified to have been produced in such a fashion as to minimize biodiversity loss (ideally, on sustainable plantations so that no further clearing of natural habitat would be required). Because the process of growing, harvesting, and processing products is typically more expensive if conducted sustainably, certification programs presume that consumers are willing to pay higher prices for certified products. Moreover, an effective certification program must necessarily be one in which consumers are able to distinguish between certified and uncertified products, as well as between truthfully certified and untruth-

fully certified products. The more effort that must be expended in order to credibly certify sustainable production, of course, the more expensive will be the final product. Thus, there is some concern that certification programs may not yet be financially feasible.

Nevertheless, certification programs have begun under the auspices of conservation organizations such as the Rainforest Alliance and industry groups such as the Forest Stewardship Council, and by the beginning of 1999 it was estimated that some fifteen million hectares of forestlands were covered under independent certification programs. Agreements by major purchasers, such as home improvement chains, as well as smaller “alternative” outlets, to purchase certified wood may also indicate an increasing willingness on the part of distributors, as well as producers and consumers, to participate in certification programs.

The ultimate success of market-based incentive programs will require both additional information on the biodiversity trade-offs that society faces and the emergence of a social consensus that these trade-offs should be resolved in a fashion more conducive to biodiversity preservation. This is not to say that these problems constitute an insurmountable impediment to market-based incentive programs. A generation ago there was little support for market-based incentive programs for the reduction of industrial emissions, but in recent years programs such as sulfur dioxide trading have reduced the costs of environmental compliance by millions of dollars. We are likely to see increased experimentation with, and eventually reliance upon, market-based incentives for biodiversity preservation in the future.

#### IV. THE MARKET AS A MECHANISM FOR CONSERVING BIODIVERSITY

We have seen that biodiversity enters into a number of aspects of the free market economy, even when the phrase “free market economy” is used to describe our existing, imperfect mechanism for deciding what to produce, what to consume, and what to preserve. The services that biodiversity provides are partially reflected in agricultural production, land prices, and aggregate measured economic production. Biodiversity can also generate values via biodiversity prospecting, nontimber forest product collection, and tourism, although expectations for these activities must be measured.

The real-world “free market economy” still falls short of how an ideal free-market economy, the perfectly competitive economy of textbook abstraction, would value biodiversity. We can approximate that ideal somewhat more closely by making broader use of market-based incentives. We must also realize that an ideal economy cannot be achieved in a real world of imperfect institutions and behaviors. Given these constraints, the pivotal question of how much biodiversity we can and should save will likely remain unresolved for many years to come.

#### See Also the Following Articles

AGRICULTURE, INDUSTRIALIZED • BIOPROSPECTING • COMMONS, THEORY AND CONCEPT OF • ECONOMIC VALUE OF BIODIVERSITY, MEASUREMENTS OF • ECONOMIC VALUE OF BIODIVERSITY, OVERVIEW • PROPERTY RIGHTS AND BIODIVERSITY • RESOURCE EXPLOITATION, FISHERIES • TOURISM, ROLE OF

#### Bibliography

- Barzel, Y. (1989). *The Economic Theory of Property Rights*. Cambridge University Press, New York.
- Clark, C. (1991). *Mathematical Bioeconomics*. John Wiley & Sons, New York.
- Daily, G. C. (ed.). (1997). *Nature's Services: Societal Dependence on Natural Ecosystems*. Island Press, Washington, D.C.
- Diamond, J. (1997). *Guns, Germs, and Steel: The Fates of Human Societies*. Norton, New York.
- Godoy, R., Lubowski, R., and Markandya, A. (1993). A method for the economic valuation of non-timber tropical forest products. *Econ. Botany* 47, 3.
- Hardin, G. (1968). The tragedy of the commons. *Science* 162, 1243–1248.
- Peters, C. M., Gentry, A. H., and Mendelsohn, R. O. (1989). Valuation of an Amazonian rainforest. *Nature* 339, 655–656.
- Ostrom, E. (1990). *Governing the Commons*. Cambridge University Press, Cambridge, United Kingdom.
- Rausser, G. C., and Small, A. A. In press. Valuing research leads: Bioprospecting and the conservation of genetic resources. *J. Political Econ.*, in press.
- Reid, W. V., Laird, S. A., Meyer, C. A., Gamez, R., Sittenfeld, A., Janzen, D. H., Gollin, M. A., and Juma, C. (1993). A new lease on life. In *Biodiversity Prospecting: Using Genetic Resources for Sustainable Development* (W. V. Reid, S. A. Laird, C. A. Meyer, R. Gamez, A. Sittenfeld, D. H. Janzen, M. A. Gollin, and C. Juma, eds.), Chap. 1. World Resources Institute, Washington, D.C.
- Simpson, R. D., Sedjo, R. A., and Reid, J. W. (1996). Valuing biodiversity for use in pharmaceutical research. *J. Political Econ.* 104(1), 163–185.
- Southgate, D. (1998). *Tropical Forest Conservation: An Economic Assessment of the Alternatives for Latin America*. Oxford University Press, New York.
- Wilson, E. O. (1992). *The Diversity of Life*. Belknap, Cambridge, Massachusetts.







# MASS EXTINCTIONS, CONCEPT OF

J. John Sepkoski, Jr.<sup>†</sup>  
*University of Chicago*

---

- I. History of the Concept
  - II. Models of Mass Extinction
  - III. Interpreting Data From the Fossil Record
  - IV. Magnitudes of Mass Extinctions
  - V. Hypothesis of Periodicity
  - VI. The Kill Curve and Self-Organized Criticality
  - VII. Selectivity of Mass Extinction: Victims and Survivors
  - VIII. Recoveries from Mass Extinction
  - IX. The Modern Biodiversity Crisis
  - X. Summary
- 

## GLOSSARY

- benthos** Organisms living on or in sediment, below the water.
- extinction** The disappearance of a species upon death of its last surviving individual. In the fossil record, this is treated as the last fossil occurrence of individuals of a species.
- foraminifera** An order of animal-like protists, many of which secrete calcareous skeletons (“tests”).
- mass extinction** The simultaneous extinction of a disproportionate number of species over timescales of  $10^0$  to  $10^6$  years resulting in loss of biodiversity.
- phanerozoic** The geological interval of abundant animal fossils, beginning approximately 545 years ago.
- stratigraphic section** An outcrop of rock (with fossils

in this case) or a drilled core. The term can also refer to a composite for a region in which fossil ranges and stratigraphic events have been summarized into a synthesized rock column.

**tetrapod vertebrates** Vertebrate animals with four limbs (or vertebrates that have evolved from such animals, such as snakes).

---

**MASS EXTINCTION** refers to the disappearance of large numbers of organisms over relatively short geologic spans of time. The result is diminished biodiversity, which can take millions of years to recover, depending on the magnitude of the extinction event. This chapter presents topics related to this concept, including its history, current measurements of the magnitude and timing of mass extinctions, and consequences for the recovering biota.

## I. HISTORY OF THE CONCEPT

The concept of extinction of species goes back at least several centuries. The extirpation of aurochs (wild relatives of cattle) and disappearance of lions from Europe were well known in the era of enlightenment and ascribed to human interference. The fact that species could become extinct from nonhuman causes was promoted by Cuvier at the end of the 18th century through his exquisitely detailed studies of mammalian fossils of the Paris Basin. His arguments were not accepted by

---

<sup>†</sup> Deceased.

all intellectuals at the time, and, in fact, Thomas Jefferson, the third president of the United States, doubted species could disappear before humans; he assigned Lewis and Clark a secondary mission in their explorations of the American northwest to search for living mammoths and mastodons.

The division of sequences of sedimentary rock into geologic systems by British geologists and paleontologists in the first half of the 19th century reflected a concept of major changes in marine faunas between these still-used time periods. But the first quantitative depiction of mass extinctions—major declines in biodiversity followed by recovery—appears to be Phillips's (1860) count of known numbers of fossil species and interpretive graphing of massive drops in diversity between the Paleozoic and Mesozoic eras and the Mesozoic and Cenozoic eras (terms he coined; see Fig. 1).

The study of mass extinctions rested largely in limbo from Phillips's pioneering work into the mid-20th century. This was perhaps because of emphasis on documenting evolutionary continuity in the fossil record and an assumption of substantive uniformitarianism, inherited from Lyell. However, with accumulation of paleontological data, the greatest of all Phanerozoic mass extinctions—the end-Permian, or “Permo-Triassic,” event—could not remain unnoticed. Schindewolf (1963) wrote a seminal paper discussing this event and invoking lethal radiation from an extraterrestrial catastrophe of a nearby supernova explosion. In response, Newell (1967) carefully counted fossil taxa (mostly described families) and argued that there were at least five events of mass extinction in addition to the Permo-Triassic. These papers set the stage for modern studies: examining detailed biostratigraphic data on local species disappearance and global compilations of taxonomic ranges.

Despite the contributions of Schindewolf and Newell, work on mass extinctions remained largely a “cottage industry” among paleontologists until 1980. Workers would examine one of Newell's events (usually in isolation of others) and posit some associated physical event as the cause, such as fall of sea level, or invent ad hoc hypotheses, such as heavy metal poisoning in the oceans as a result of mountain building.

Maturation of the study of mass extinction came with the bold hypothesis of Alvarez *et al.* (1980) that the Mesozoic-Cenozoic event, recognized 120 years before by Phillips, was caused by impact of a 10-km meteorite. The initial evidence of Alvarez and coworkers was concentration of the rare terrestrial element, iridium, at solar-system abundances in a clay layer at the Cretaceous-Tertiary boundary. But the hypothesis implied

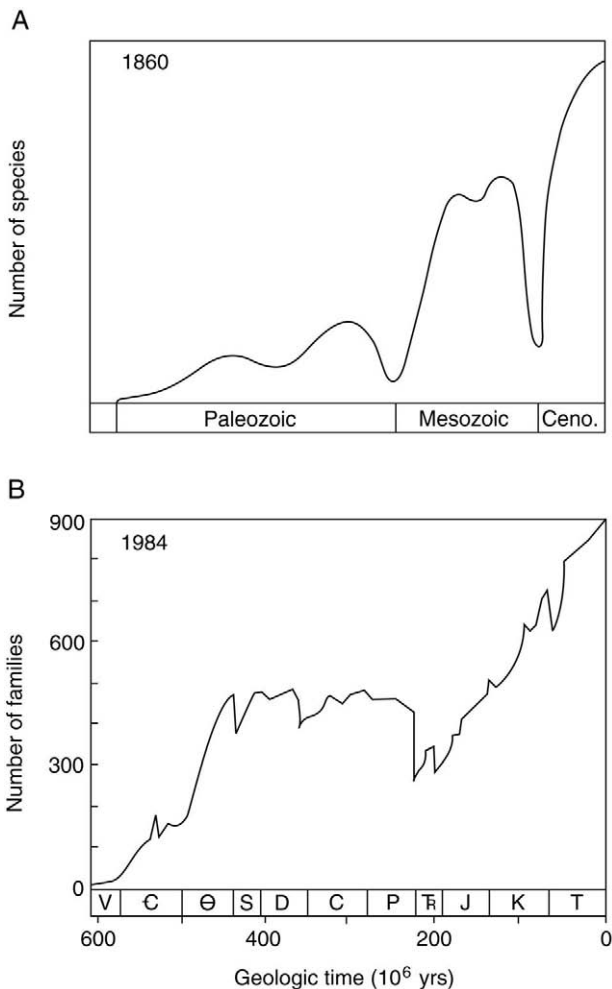


FIGURE 1 The first diversity curve for fossil organisms (a) compiled by Phillips (1860) compared to a late-20th century diversity curve for fossil marine animals (b). While Phillips' curve was generalized and based on comparatively few described taxa, the end-Permian mass extinction (between the Paleozoic and Mesozoic eras) and end-Cretaceous mass extinction (between the Mesozoic and Cenozoic ["Ceno"] eras) were recognized. The lower graph illustrates three additional mass extinctions among those recognized by Newell (1967): the end Ordovician, Late Devonian, and end-Triassic. Abbreviations along the abscissa for geologic time in (b) and in other figures are V = Vendian; C = Cambrian; O = Ordovician; S = Silurian; D = Devonian; C = Carboniferous; P = Permian; Tr = Triassic; J = Jurassic; K = Cretaceous; T = Tertiary. From Sepkoski and Schopf (1992).

other testable questions: (a) Is there additional physical evidence of an impact at the Cretaceous-Tertiary boundary and (b) Is the abruptness of biological extinctions at the event consistent with a catastrophe induced by meteorite impact? Impressive evidence affirming the first question has been assembled, including global identification of the iridium-rich clay layer (in both

marine and terrestrial stratigraphic sections); presence of shocked mineral grains (e.g., quartz), microspherules of shock-melted minerals, and abundant soot in the global clay layer; and the presence at 65 Ma of the largest impact crater known from the Phanerozoic in the Yucatan, Mexico. Answers to the second question have proved more difficult and will be considered later.

The recognition of extinction events in the geologic past demands comparison to modern extinctions, from the auroch to the myriad other species documented to have disappeared from human interference. Questions include not only how modern rates of extinction compare to those in the fossil record but also the consequences of extinction: What kinds of organisms are most susceptible to extinction and what are the patterns of biotic recovery after the pressure of extinction has been removed? These questions will be considered further later in this discussion.

## II. MODELS OF MASS EXTINCTION

A variety of patterns of extinction have been observed in the fossil record around events, ranging from abrupt termination of species at an extinction level to gradual disappearance up to it, and perhaps beyond, the extinction level. Three scenarios for mass extinction have been proposed based on empirical observations: abrupt extinction, gradual extinction, and stepwise extinction.

### A. Abrupt Extinction

This is the pattern hypothesized by Alvarez *et al.* (1980) of species disappearing in a geologic instant (which could in fact be  $10^1$  to  $10^3$  years). Observed declines in diversity before the event (such as seen in detailed records of foraminifera or broader records of dinosaurs before the Cretaceous-Tertiary) are a result of the Signor-Lipps sampling effect (discussed later).

### B. Gradual Extinction

The fossil record is taken on face value, especially if there has been extensive sampling around the horizon of extinction. Slow attrition of species up to the end of a mass extinction has been claimed for extensively sampled foraminifers around the Cretaceous-Tertiary event, where several large, but rare foraminifers seem to disappear below the stratigraphic boundary and some small, generalized foraminifers occur above the boundary.

### C. Stepwise Extinction

The fossil record is again taken at face value but exhibits a series of pulses of species terminations. This pattern has been hypothesized for situations such as the Late Devonian mass extinction ("upper Kellwasser Event") where intensively sampled taxa of different groups appear to disappear in small pulses separated by  $10^4$  to  $10^5$  years. The model can be expanded to intervals such as the end Ordovician (the second largest marine mass extinction of the Phanerozoic) when many trilobites and other marine animals of tropical areas appear to become extinct at the onset of major glaciation, and then, perhaps  $10^6$  years later, surviving deep-shelf benthos disappear as normal conditions of low oxygen return with the end of glaciation. (In this case, each pulse of extinction could be dissected to determine if it had been abrupt, gradual, or stepwise at finer timescales.)

These scenarios need to be distinguished from the low levels of extinction that are observed in all geologic intervals. These levels are normally termed "background extinction" to distinguish them from events of mass extinction. Background extinction for marine animals appears to decline through Phanerozoic time. Thus, smaller extinction events are more obvious in Newell-type data over the Mesozoic and Cenozoic eras than during the early Paleozoic when background rates were high.

## III. INTERPRETING DATA FROM THE FOSSIL RECORD

The fossil record provides direct evidence of previous mass extinction but only an incomplete accounting because of differences in preservability of organisms (e.g., bivalve mollusks versus polychaete worms) and in scientific sampling (e.g., Europe and North America versus Antarctica). With an incomplete record, observed last occurrences of fossil species are only a minimum estimate of actual times of extinction. This consideration was formalized for mass extinction by Signor and Lipps (1982) who modeled how observed terminations of species would appear around an abrupt extinction (Fig. 2). With less intensive sampling or less complete preservation, the expectation is a pattern of gradual disappearance of fossil species up to a boundary of abrupt mass extinction. This sampling pattern holds true whether one is examining detailed stratigraphic sections of fossils or analyzing compilations of fossil taxa, like those of Newell.

Raup presented a very intuitive example of the Sig-

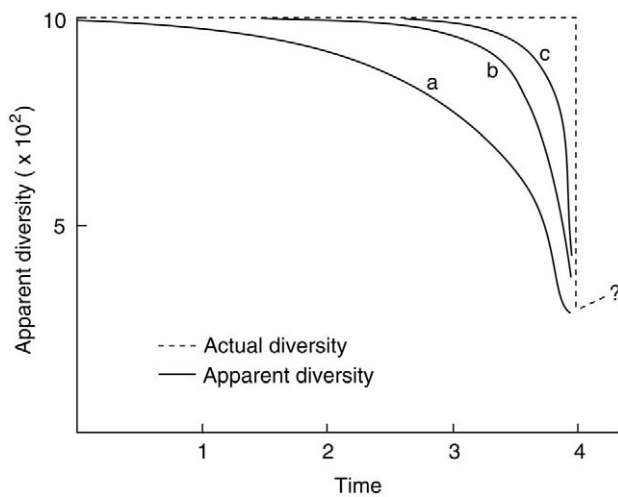


FIGURE 2 Hypothetical and calculated diversity curves reflecting the “Signor-Lipps effect”—that is, imperfect sampling around a catastrophic mass extinction. The dashed curve represents true diversity, which is treated as constant until time unit 4 at which there is a catastrophic mass extinction reducing diversity by three-quarters. If fossil taxa are not sampled up to the times of their true extinctions, declines in diversity will appear more or less gradual, depending on the intensity of sampling. Curve a represents the poorest sampling, and curve c the best (three times better than a). From Signor and Lipps (1982 with permission from the Geological Society of America).

nor-Lipps effect (Fig. 3). He used data collected for ammonites below the Cretaceous-Tertiary boundary to investigate how the fossil record of extinction would have appeared if the mass extinction had occurred at an interval 100 m lower in the stratigraphic section. (A meteorite impact or other unpredictable catastrophe potentially could occur anywhere in a stratigraphic section.) The decline in observed diversity up to the artificial event does not appear substantially different from the gradual decline of diversity actually observed below the true Cretaceous-Tertiary boundary. This suggests that problems of variable preservation and incomplete sampling of fossils can indeed influence empirical patterns of species disappearance around extinction events, hindering discrimination among the three models of mass extinction.

More powerful statistical methods have been developed to investigate how apparent last occurrences of fossil species may relate to actual times of extinction. These methods involve calculating “confidence intervals” on the time of last sampled occurrence of a species (Fig. 4). The basic concept is that the last fossil occurrence of a species that is rarely sampled is a poorer predictor of actual time of extinction than the last occurrence of a species that is densely sampled. Using various

models of the density distribution of sampling of a fossil species, probability statements can be made about how far true extinction lies above the last observation of a species.

## IV. MAGNITUDES OF MASS EXTINCTION

Questions of abrupt, gradual, or stepwise extinction involve patterns in the fossil record resolved over  $10^3$  to  $10^5$  years (encompassing the range from the late Pleistocene extinctions of large mammals to the historical extirpation of species). On larger timescales, general magnitudes of mass extinction can be measured from global fossil data. These data are best for the marine record of animals from continental shelves and seas and fall into roughly three classes of magnitude (Fig. 5).

### A. The End-Permian Mass Extinction

This class of magnitude stands alone in its effects on the biota (Erwin, 1993). Compilations of taxa lost indicate that more than 50% of animal families and 80% of genera in the oceans became extinct. Extrapolations of species loss have been attempted, using ecological rarefaction (how many species would be lost given measured declines in genera or families, assuming some distribution of species within higher taxa); results range from 90 to 96% loss of marine species. This loss of marine biodiversity at the end-Permian is unprecedented. Recent work suggests that on land important groups, including insects, tetrapod vertebrates, and plants, also experienced substantial declines in diversity.

### B. Four Other Events of Marine Mass Extinction

This class of magnitude eliminated substantial proportions of marine animals and seem to have had nearly equal magnitudes: the end-Ordovician, Late Devonian, end-Triassic, and end-Cretaceous events. (The occurrence of these events at or near the end of geologic periods reflects the use of faunal change to define intervals of geologic time.) The four events have measured family extinction in the oceans of 15 to 25% and extrapolated species extinctions of 64 to 85%.

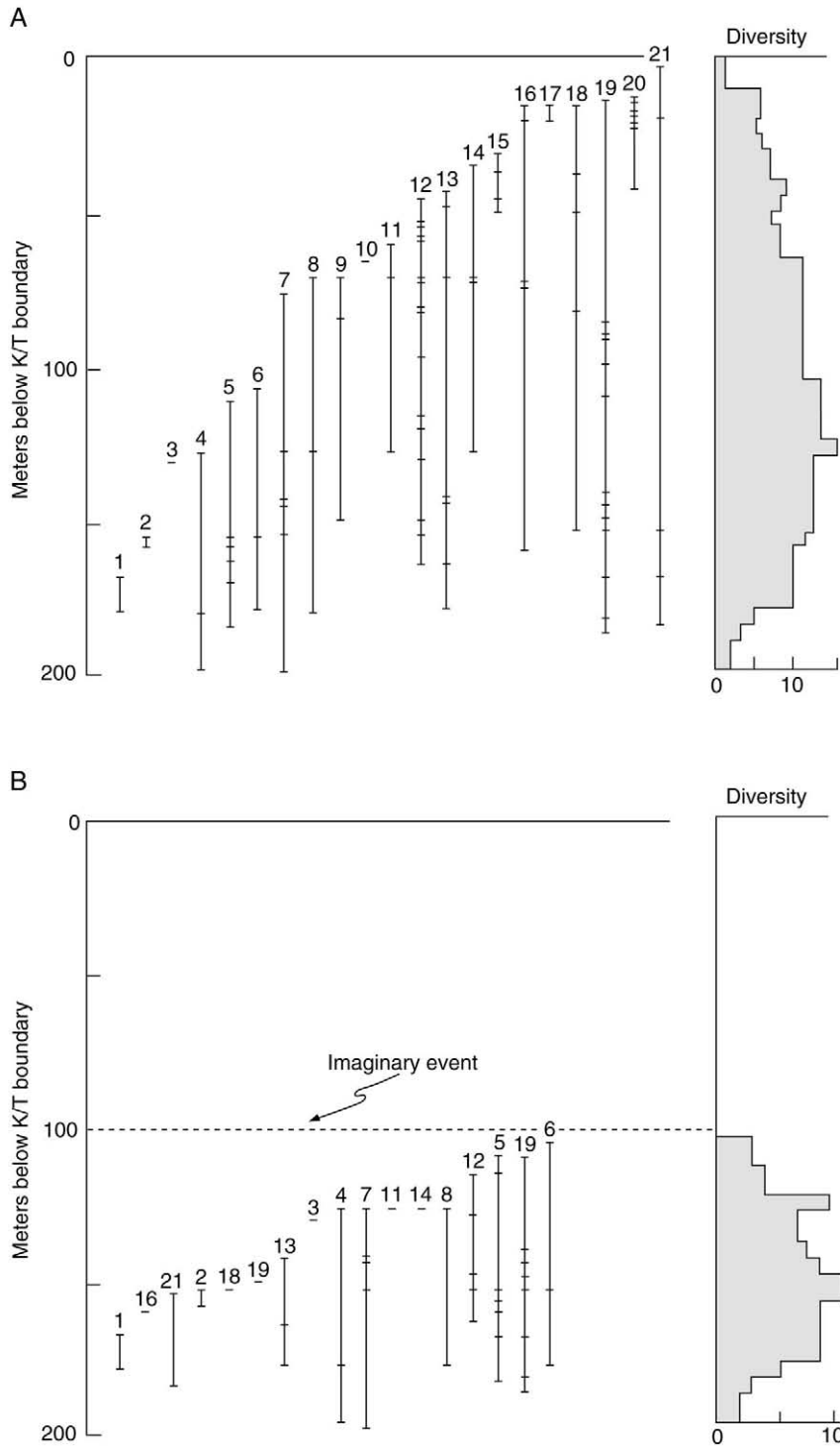


FIGURE 3 Apparent gradual extinction at an imaginary sudden extinction event. (a) Actual observed geologic ranges of 20 fossil species. The y-axis is stratigraphic position, in meters, below the Cretaceous/Tertiary boundary. Vertical lines show the total observed ranges of the species and horizontal ticks indicate horizons at which they were actually sampled. The curve at the right, labeled "Diversity," is the sum of the ranges of the species. (b) Resultant ranges if an imaginary catastrophic extinction were imposed at 100 m. Apparent last occurrences of species again are graded below the mass extinction and diversity appears to decline, both as a result of species being irregularly sampled through the geologic interval. (Note that the species in b have been reordered based on their highest geologic occurrence.) From Sepkoski and Koch (1996), based on Raup.

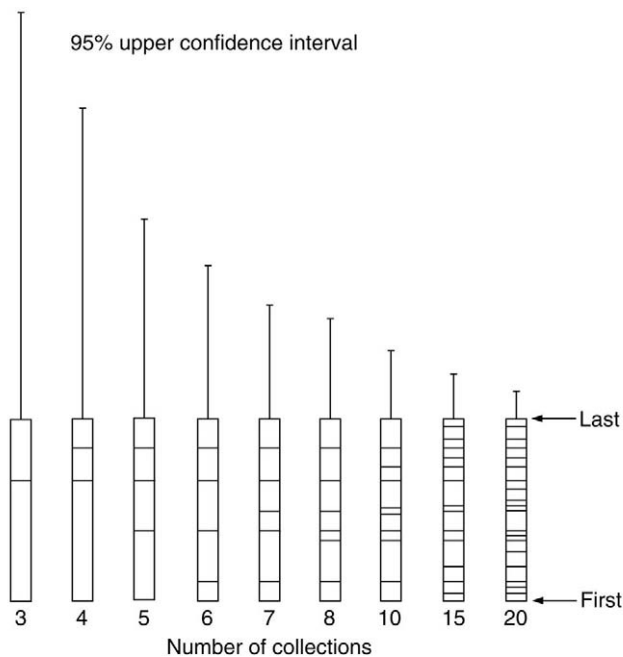


FIGURE 4 Confidence intervals on times of extinction based on the density of sampling of fossil species. The bars represent observed ranges of fossil species with horizontal ticks (including bottom and top of the bars) representing samples of the species. Vertical lines capped by ticks are calculated relative intervals during which there is a 95% probability that extinction actually occurred (thus allowing a 5% probability that real extinction actually occurred higher). The probable position of true extinction is closest to the position of last observation when species are densely sampled, but true extinction can still be expected to be somewhere above the position of last observation. From Sepkoski and Koch (1996) after Strauss and Sadler, with permission from the Geological Society of America.

### C. Other Intervals of Unusual Amounts of Extinction

This class of magnitude is now termed “extinction events” (Fig. 6). These appear in more detailed compilations of the kind Newell made as well as in precise biostratigraphic and paleogeographic analyses. Paleogeographic analyses suggest that many of these third-order events were not global in extent or taxonomic effects, unlike the two previous categories. Examples include the Cambrian bioturbation events, marked by nearly complete extinction of North American trilobites but with uncertain effects in other parts of the world or on other taxa (although some events are recognized in Australia, China, Siberia, and Europe); the lower Toarcian event affecting mollusks in Europe but not around the Pacific basin; and the Pliocene event, again documented for mollusks, in the south temperate Northern

Atlantic Ocean and Mediterranean Sea but not evident in the tropics or in the Pacific.

There are extinction events that could be classified at lower ranks. For example, it has been demonstrated that marine animal communities persist on order of  $10^6$  years during the middle Paleozoic, punctuated by rapid changes in faunal composition (Miller, 1997). This raises the question of Raup (1991a) as to whether extinction events represent a continuum between rare major events and frequent small events (discussed later).

## V. HYPOTHESIS OF PERIODICITY

Given the varying magnitudes and geographic extents of extinction events, one would expect differing forcing agents. With a plethora of forcings acting independently over geologic time, it would be expected that extinction events would be distributed at random through the fossil record. Thus, it came as a surprise when Fischer and Arthur (1977) and Raup and Sepkoski (1984) observed that extinction events of first through third rank appeared regularly distributed through time. Raup and Sepkoski performed extensive statistical analyses of Newell-type data and concluded there was a strong periodicity of 26 myr for events during the Mesozoic and Cenozoic eras (Fig. 7). This suggested some sort of clocklike mechanism behind mass extinction with a periodicity unknown in terrestrial processes. Because one of the periodic events was the end-Cretaceous mass extinction, Raup and Sepkoski speculated that the clocklike mechanism might be extraterrestrial.

This speculation engendered both intriguing hypotheses from astronomers and geologists and critical scrutiny of data and statistical methods from paleontologists, geologists, and statisticians (Raup, 1991b; Sepkoski, 1989). The best-known hypothesis is Nemesis, sometimes called “Shiva”; this is an hypothesized small binary companion to the sun with a large and eccentric orbit that brings it through the Oort Cloud of comets beyond the planets every 26 myr. Nemesis’ gravitational perturbation scatters up to  $10^9$  comets of which  $10^1$  to  $10^2$  might impact the earth, disrupting climate and causing mass extinction.

Nemesis has not been observed, and astronomical models cast doubt on the possibility that a small star could maintain a stable orbit at large distance from the sun through the 4.6 billion-year history of the solar system. Also, there is direct evidence of extraterrestrial impact for only a few of the periodic extinctions, despite intensive investigation: Shocked quartz has been found

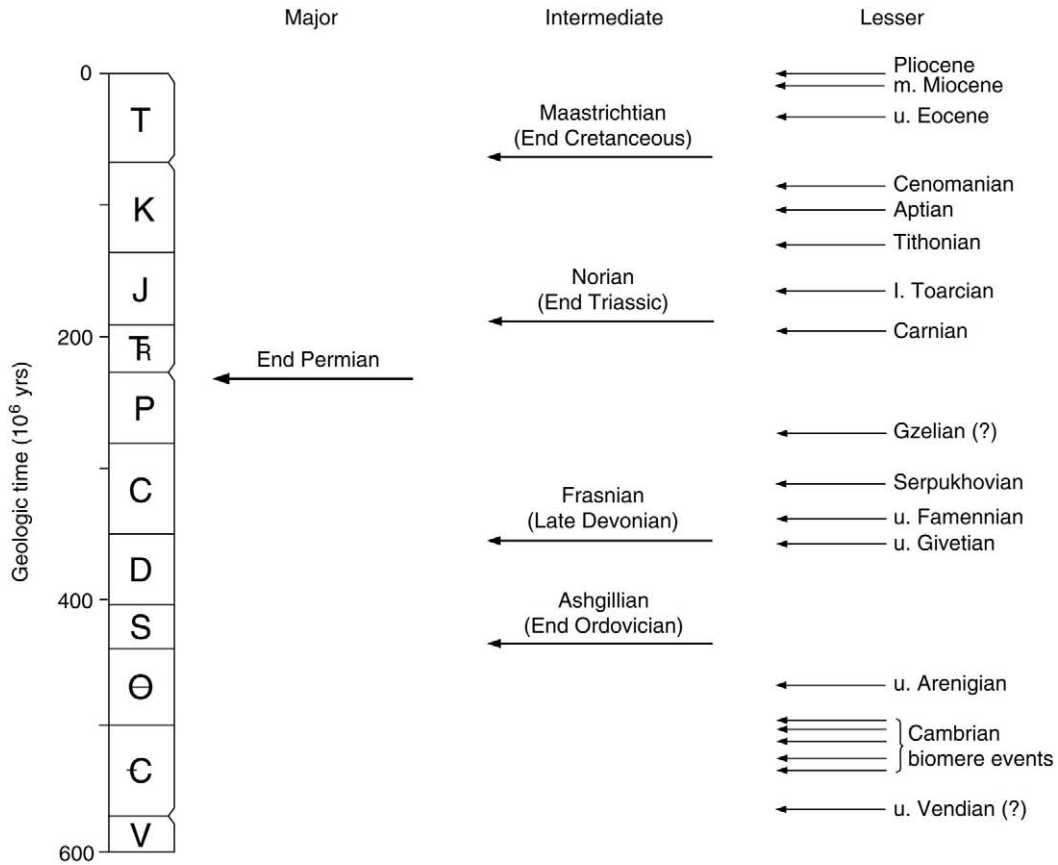


FIGURE 5 Time distribution of notable extinction events in the marine fossil record, classified as major (“first order”), intermediate (“second order”), and lesser (“third order”) events. Only major and intermediate events are normally referred to as “mass extinctions.” Abbreviations are mostly as in Figure 1; l = lower, u = upper. Names refer to stratigraphic intervals (mostly form stages). (?) indicates an event that has not been well documented. From Sepkoski (1992).

at several end-Triassic localities, and concentrated iridium, microtektites (annealed globules of impact melt), and a crater have been found in the Late Eocene, although their temporal correspondence to extinctions remains unclear. Searches of this nature are difficult and time consuming, and one should not be surprised if more discoveries are made.

Questions about data and analyses with respect to the hypothesis of periodicity have been largely technical (Sepkoski, 1989). Questions of data concern the veracity of fossil families and genera as indicators of extinction in the geological past and the accuracy of geochronological dating of intervals of extinction. Statistical questions involve problems of analyzing data that are not sampled evenly through time (geologic intervals vary in duration; Fig. 7) and problems of significance when performing multiple tests when searching for periodicities of best fit. Some of these questions entered

new analytic ground, leading to new kinds of analyses, but not all analytical issues have yet been resolved.

Thus, a nonrandom, periodic distribution of extinction events remains a hypothesis on the table, with neither a clear forcing mechanism nor a definitive analytical test.

## VI. THE KILL CURVE AND SELF-ORGANIZED CRITICALITY

Raup (1991a) treated extinction in the fossil past not as a hierarchy of events but as a continuum, in which the obvious events of large magnitude are rare relative to less obvious events of small magnitude. An analogy is to the historical record of river floods, in which the 1000-year event stands out but is really part of a spec-



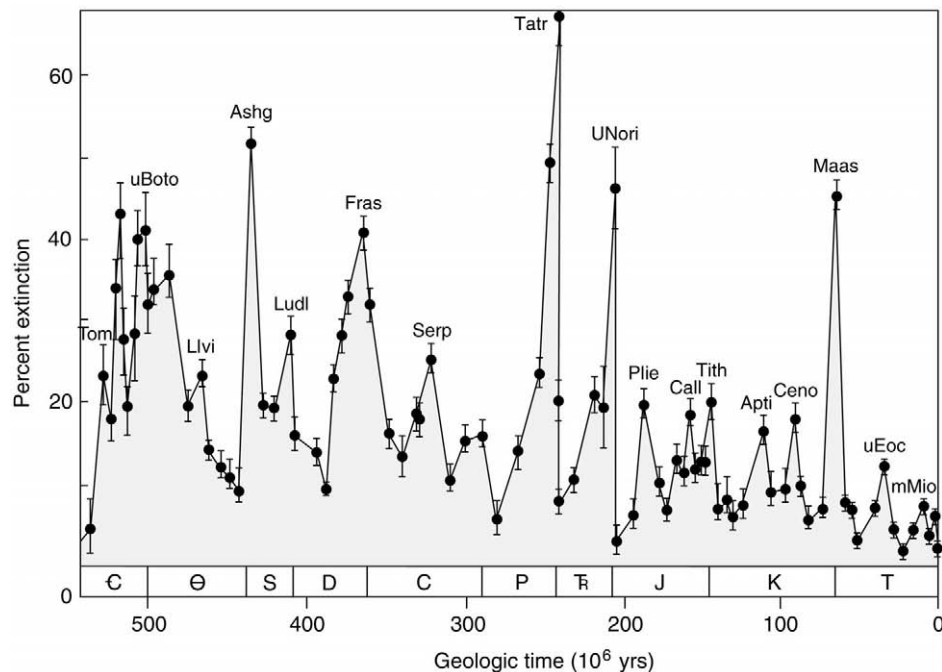


FIGURE 6 Extinction intensities for marine animal genera through the Phanerozoic. Intensity is measured as percent extinction ( $= 100 \times \text{number of extinctions/observed diversity}$ ) in geologic intervals averaging about 5.5 myr in duration. Estimated errors of estimate are indicated by error bars ( $\pm 1$  standard deviation). Significant events are labeled: Tom = Tommotian; uBoto = upper Botomian; Llvi = Llanvirnian; Ashg = Ashgillian (end-Ordovician mass extinction); Ludl = Ludlovian; Fras = Frasnian (Late Devonian mass extinction); Serp = Serpukhovian; Tatr = Tatarian (end-Permian mass extinction); uNori = upper Norian (= end-Triassic mass extinction); Plie = Pliensbachian; Call = Callovian; Tith = Tithonian; Apto = Aptian; Ceno = Cenomanian; Maas = Maastrichtian (end-Cretaceous mass extinction); uEoc = upper Eocene; mMio = middle Miocene. From Sepkoski (1996) with permission.

trum with smaller decadal or even yearly events. The 1000-year event does not occur regularly but rather has a chance of 0.001 of occurring during any given year. Raup used a standard function for these kinds of events to analyze probabilities of species extinction of a given magnitude during some duration of geologic time (Fig. 8): the longer the duration, or “waiting time,” the larger a pulse of extinction that can be expected. The limitation of Raup’s analysis is that it was based on generalized geological intervals, the same as illustrated in Figs. 6 and 7, and cannot distinguish between sudden events, such as observed at the end of the Cretaceous, and cumulative small events, such as may also have occurred during the last stratigraphic stage of the Cretaceous during the 9 million years prior to the impact-induced mass extinction.

Raup’s analysis was posited on an assumption that external perturbations of varying magnitude caused most extinction through geologic time, and he in fact demonstrated a linear relationship between his empiri-

cally based kill curve and the similar curve established for the flux of meteorites of varying magnitude (Raup, 1992). Other approaches have posited extinction to be a result of internal dynamics of the biota over time. Solé *et al.* (1997), for example, analyzed Phanerozoic time series of extinction and diversity fluctuation to determine if there was evidence of “self-organized criticality.” The prediction was that a power series from fourier transforms of the data would exhibit a linear  $1/f^x$  relationship when power was plotted logarithmically against the logarithm of frequency,  $f$ . The exponent,  $x$ , is expected to be between 1 and 2, which is indeed what was found.

Self-organized criticality refers to systems of interactive components that grow to a degree of complexity that leads to cascades into chaos or collapse. The standard analogy is to a pile of sand onto which one grain is added at a time: As the pile approaches critical size, one more grain can cause a small shuffle of sand while, far less frequently, another grain can cause a major

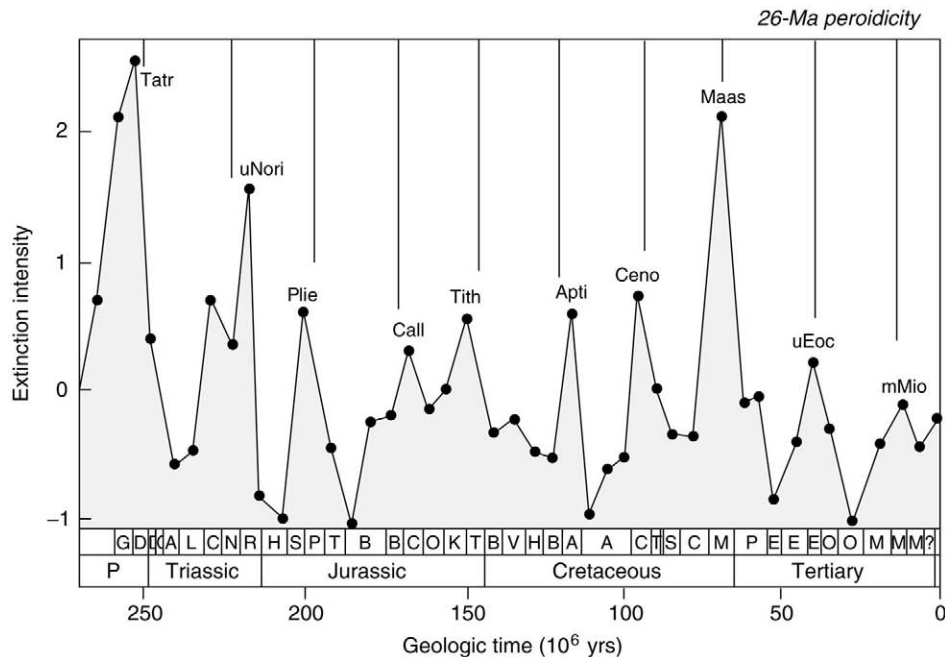


FIGURE 7 Time series for extinction intensity of marine animal genera through the Mesozoic and Cenozoic showing the putative 26-myrr periodicity of mass extinction (vertical bars). The abbreviations are as in Fig. 6. Extinction intensity here is measured as an average of percent extinction in major taxa that have been standardized to zero means and unit standard deviations. From Sepkoski (1990).

avalanche. The sizes of the avalanches follow the  $1/f$  power law. For species, one more new population packed into an interacting community might lead to extinction of a variable number of species if the community were self-critical; usually, extinction numbers would be small, but occasionally there would be a major avalanche of extinction.

There are three criticisms of this argument. First, some mass extinctions are clearly associated with physical disturbances, such as the meteorite impact at the end of the Cretaceous. Second, many other processes can produce  $1/f$  power laws when analyzed in the way organized self-criticality is; for example, the secular decline in background extinction over time, perhaps induced by changing taxa with different characteristic rates of extinction (Sepkoski, 1984), can cause power series to decline relative to increasing frequency. Finally, there is no independent evidence that species within marine communities, for which data have normally been analyzed, are as interactive as posited.

There still may be some interesting avenues of inquiry to pursue. The state of the biota is not constant, given variations in the earth's climates, changing positions of continents, and varying tectonic events and their effects on environments and the time over which

the biota evolves to these varying conditions. A perturbation at one stage of the earth's conditions and the biota's development could have very different effects than an identical perturbation at a different time (Sepkoski, 1989). Thus, an expectation of some chaotic element in the record of extinction cannot be ruled out.

## VII. SELECTIVITY OF MASS EXTINCTION: VICTIMS AND SURVIVORS

An obvious effect of the end-Cretaceous mass extinction is that all (nonavian) dinosaurs disappeared whereas some mammals survived. At other mass extinctions there are also cases of major taxonomic groups disappearing and others surviving. This observation has led to a search for rules as to what makes some kinds of organisms more vulnerable to mass extinction than others. Patterns that have been found are largely probabilistic, and all seem to have unexplained exceptions:

1. Taxa that have high rates of extinction during times of background extinction suffer disproportion-

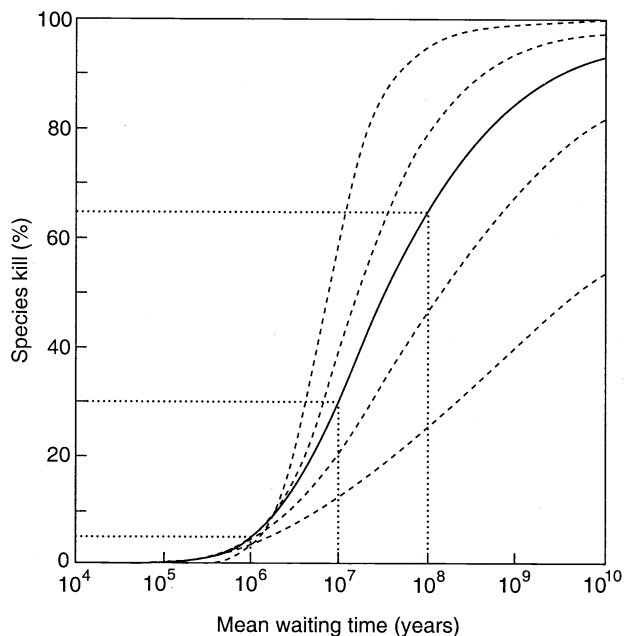


FIGURE 8 Raup's "kill curve" showing the expected extinction of marine species, as a percentage of standing diversity, over intervals of time ("mean waiting time"). During intervals of less than  $10^6$  years, percent of diversity expected to become extinct is low, but over longer intervals, especially greater than  $10^8$  years, extinction events of large magnitude become quite probable. The solid line is the best fit curve to fossil data; the dotted curves fall at bounds of uncertainty for empirical data. From Raup (1991a).

ately at mass extinctions. Rates of extinction measured for higher taxa (e.g., classes) during background times tend to be good predictors of magnitudes of extinction at major events. Thus, trilobites, with higher rates of background extinction than articulate brachiopods, suffered more loss at the end-Ordovician, and brachiopods, with higher characteristic rates of extinction than bivalves, suffered greater proportional declines at the end-Permian and even the end-Cretaceous, when brachiopods were relatively marginalized (Sepkoski, 1984). It remains a major question as to why major taxonomic groups have different characteristic rates of extinction during background times and why these rates are conserved over vast stretches of macroevolutionary history and species turnover.

2. Animals, particularly terrestrial vertebrates, with large body size seem to be particularly prone to extinction. Dinosaurs perished at the end of the Cretaceous whereas many of the small mammals survived; many large mammals and birds that had evolved by the Eocene disappeared during the Eocene-Oligocene transition; and the late to end-Pleistocene extinctions that

afflicted all continents other than Africa (and Antarctica) are often referred to as the "Pleistocene large mammal extinctions." Explanations for this selectivity involve (a) lower absolute population numbers of large animals, exposing them to greater chance of all individuals perishing during perturbations and (b) longer gestation times of large mammals, lengthening the time to restoration of pre-perturbation population numbers. Size-selectivity may not apply to marine invertebrates. Jablonski (1996) examined marine mollusks (gastropods and bivalves) for size-selectivity at the end-Cretaceous and found no differential extinction among large-sized species. On the other hand, Fischer and Arthur (1977) argued that marine vertebrate "megacarnivores" were very vulnerable to smaller extinction events.

3. It is often claimed that tropical plants and animals in the fossil past have been more susceptible to extinction than organisms living at higher latitudes. This is based in part on the observation that tropical reef communities disappear at several of the major mass extinctions, including the Late Devonian, end-Permian, and end-Cretaceous events, and millions to tens of millions of years intervene before diverse reefs re-evolve. The pattern for level bottom marine communities is not so clear: Jablonski and Raup (1995) analyzed the biogeography of bivalve extinction at the end of the Cretaceous and found no difference in extinction between tropical and temperate genera that were not associated with reefs.

4. Taxa with restricted geographic ranges appear to have been more prone to extinction than widespread taxa. Jablonski (1995) documented this for Late Cretaceous molluscan species during normal times of background extinction. At the end of the Cretaceous, however, the geographic range of individual species did not seem to be a factor, with widespread species suffering as much extinction as restricted species. At the rank of the genus, however, genera with geographically dispersed species tended to survive preferentially relative to genera with species restricted to one or a few biogeographic provinces. Similar results have been obtained for Cambrian trilobites at various extinction events.

5. Environment seems to have played only a minor role in selectivity of extinction among marine animals at mass extinctions. Rates of extinction among Paleozoic marine animals tended to increase toward shallower waters during background times but to be equal across continental shelves at mass extinctions. There is some suggestion, however, that deep-water animals fared better: Marine shelves seem to have been repopulated by descendants of deep-water trilobites after Cambrian extinction events; shelf communities contained dispro-

portionate numbers of deep-water sponges and corals after the Late Devonian mass extinction (McGhee, 1996); and descendants of mollusks typical of low-oxygen deep-water communities of the late Paleozoic are common after the end-Permian mass extinction (Erwin, 1993).

These reports of taxonomic selectivity at mass extinctions are largely confined to analyses around a few events of extinction, and much larger comparative studies need to be conducted across all mass extinctions for both marine and terrestrial organisms. However, if there are consistent probabilistic biases in terms of the properties of species and higher taxa that survive mass extinctions, then extinction events on frequencies of  $2.6 \times 10^7$  years, eliminating 30 to 90% of the biota, could be a major factor in the history of life.

## VIII. RECOVERIES FROM MASS EXTINCTIONS

Global data on mass extinctions exhibit not only the geologically rapid declines of the biota at mass extinctions but also subsequent recoveries of biodiversity in the aftermaths. With modern data, these recoveries are seen to encompass longer intervals of time than the extinction events, ranging up to about 5 myr for third-order extinction events to around 10 to 15 myr for second-order mass extinctions; recovery from the great end-Permian mass extinction took even longer but was interrupted by the end-Triassic event, some 30 myr later.

On long geologic timescales, recoveries from mass extinctions follow simple patterns expected for a diversifying biota. At more detailed scales, there is great complexity, with some patterns repeated after every mass extinction.

### A. Large-Scale Rebound of Diversity

To a first approximation, diversification in the oceans can be described as a system with equilibrium constraints imposed by limitations of resources and their utilization (Sepkoski, 1984, 1992). This is most obvious in the long plateau of diversity during the Paleozoic era, spanning approximately one-quarter billion years. This interval witnessed two second-order mass extinctions, the end-Ordovician and Late Devonian. Significantly, after each, animal diversity rebounded to previous, unperturbed levels and then continued the

Paleozoic plateau (Fig. 1b). Also, the rate of rediversification after these mass extinctions was approximately the same as during the Ordovician radiations that established the Paleozoic plateau of global marine diversity.

Such a system can be modeled if it is assumed that Paleozoic animals were constrained by an equilibrium diversity, conditioned on the environment and the way these animals subdivided resources (Fig. 9). The simplest mathematical description of diversification is a logistic model, assuming decreasing rates of origination as diversity increases. A mass extinction is a perturbation to the environment that increases extinction rates (Fig. 9b). Following the perturbation, when extinction rates return to background levels, taxa rediversify to the normative level (Fig. 9c).

Patterns are more complex through the quarter-billion years after the Paleozoic era, as diversity in the oceans generally increased. The increase reflects expansion of the modern marine fauna, which may have subdivided resources in ways different from the fauna of the Paleozoic era. Nonetheless, rebounds after mass extinctions during the Mesozoic and Cenozoic eras were much more rapid than the long-term increase in marine diversity when measured over millions of years and were congruent with initial rates of diversification of the taxonomic groups involved (cf. Miller and Sepkoski, 1989).

### B. Delayed Diversification

When rebounds from mass extinction are analyzed at finer timescales, many complications become apparent (Erwin, 1998). One is delayed recovery: rediversification does not commence immediately after perturbations. For example, there are only slow rates of rediversification for  $10^5$  years after the end-Cretaceous event among planktonic foraminifers (actually, fast at first and then slow for some four myr; D'Hondt *et al.*, 1996), benthic mollusks (Jablonski, 1998), and terrestrial mammals (Maas and Krause, 1994). Recovery seems to be delayed for nearly 5 myr after the much larger end-Permian event, with only depauperate faunas in the oceans and on land (Erwin, 1998).

Explanations for this pattern have varied. Some workers have suggested that the extended post-extinction intervals of low diversity reflect lingering effects of the external perturbation—that is, continued environmental stress or instability. Others have hypothesized that substantial amounts of time are required for the evolution of new species that can reestablish normal ecosystem function which, in turn, can support high diversity. An example is the evolution of planktonic

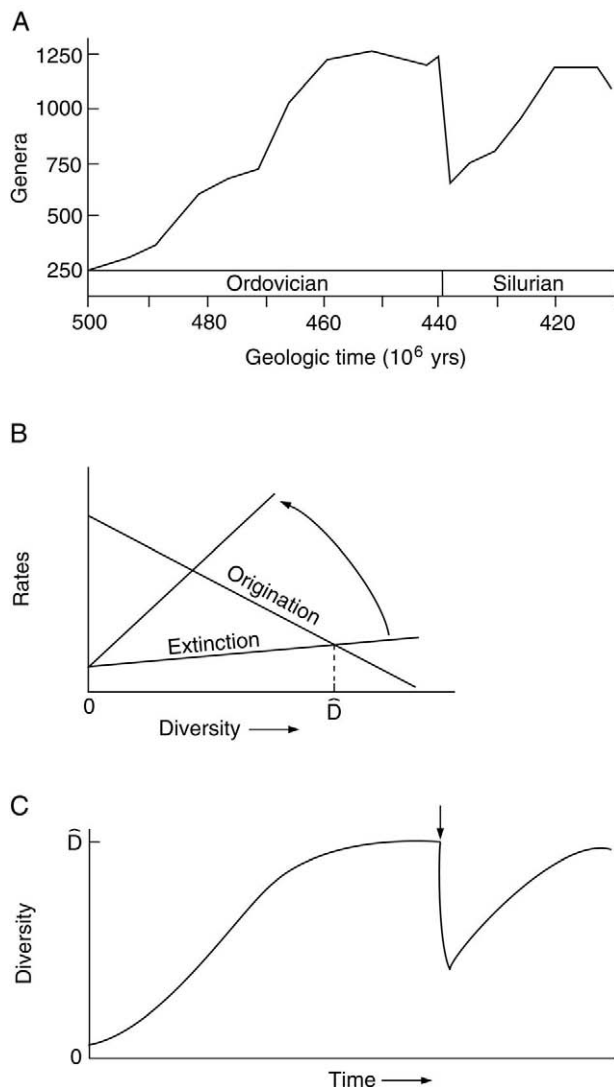


FIGURE 9 Observed and modeled loss and recovery of diversity around a mass extinction. (a) Observed numbers of genera through the Ordovician and Silurian periods, illustrating the major decline in diversity at the end of the Ordovician and subsequent recovery during the Silurian. Note that the rate of recovery is not very different from the rate of radiation during the Ordovician (cf. Fig. 1) and that diversity during the Silurian recovers to the same level as during the late Ordovician. (b) Structure of a logistic model to describe diversification, mass extinction, and recovery. It is assumed that diversity increases multiplicatively and that per capita rates of origination decrease with increasing diversity, leading to sigmoidal growth of diversification. A mass extinction is a temporary steepening of the extinction function (or, alternatively, increasing of its elevation).  $\bar{D}$  signifies the diversity at which origination rate equals extinction rate during normal times and diversity no longer grows. (c) A solution to the logistic model using rates measured from Ordovician marine animals, with a perturbation (arrow) applied at a model time corresponding to the observed end-Ordovician mass extinction. Diversity declines precipitously when extinction rate is increased, but then rebounds when extinction is returned to normal. Diversity in the model solution rebounds at the same rate as the Ordovician radiations and equilibrates at the same level as previously. From Sepkoski (1992).

microherbivores after the end-Cretaceous mass extinction (D'Hondt *et al.*, 1998). The delayed interval of recovery after the severe end-Permian event could even be conceived as a long lag in a logistic system after diversity had been reduced by an order of magnitude.

### C. "Disaster Species" and "Lazarus Taxa"

A feature that appears consistent during early phases of recovery from mass extinction is the appearance of "disaster species." These are remarkably abundant and geographically widespread species that appear in the waning stages or aftermaths of mass extinctions. Examples include the small foraminifer, *Guembelitra*, that spread out from marginal environments to form nearly monospecific assemblages immediately after the end-Cretaceous extinction and the terrestrial "fern spike" observed at the same time on land. Another example is the terrestrial synapsid ("mammal-like reptile"), *Lystrosaurus*, that produced monospecific assemblages in Gondwanaland following the end-Permian mass extinction.

Disaster species are characterized by not only great abundance in fossil deposits but also short geologic durations. Marine disaster species seem to flourish for around  $10^5$  years and then disappear, often to be replaced by another "disaster species" with similarly short duration. This appears like an evolutionary "boom and bust" cycle. Thus, the earliest Tertiary foraminifer, *Guembelitra*, declined and was replaced in dominance by a succession of bursts of *Paravularugoglobigerina*, "*Eoglobigerina*," and finally *Woodringina* over timescales of  $10^5$  years as the planktonic ecosystem seemed to regain stability.

Perhaps the antithesis of disaster species are "Lazarus taxa." These are lineages that disappear around mass extinctions—seemingly to have died—only to reappear in the fossil record some  $10^6$  to  $10^7$  years later (Erwin, 1998; Jablonski, 1986). These can be conspicuous but usually never dominant members of the biota prior to a mass extinction. Presumably, these taxa survived in some environmental or geographical refugium until the ecosystem regained sufficient function so that their dispersal—and recovery of sufficient abundance to be encountered by paleontologists—was possible.

## IX. THE MODERN BIODIVERSITY CRISIS

The collapses of ecosystems in the fossil past can inform thinking and actions with respect to the contemporary

loss of biodiversity. The fossil record is incompletely known, but the data seem little worse than estimates of living biodiversity and current rates of species extinction. Estimates of modern global biodiversity range over more than an order of magnitude, from 5 to  $100 \times 10^6$  species. Estimates of present-day loss of species range over nearly two orders of magnitude, from 5 to 150 species extinction *per day*. Most of these extinctions are terrestrial animals (largely insects), whereas the best data from the fossil record are for marine species. Still, in the geologic past, mass extinctions appear to have occurred largely contemporaneously both on land and in the seas (except for the end-Pleistocene event), so some simple calculations can be made.

If a median contemporary extinction rate of 41 species per day is assumed and species attrition is treated as a negative exponential, then it would take only about 16,000 years for 96% of the modern biota to become extinct (Sepkoski, 1997). This is a long time by human standards, but it is beyond the limits of geologic resolution at 250 Ma (the end-Permian), the only other time when 96% of Earth's biota disappeared.

This discussion has been largely in terms of fossil diversity. Another variable that can be measured in the geologic record is ratios of stable carbon isotopes. Because organisms tend to utilize  $^{12}\text{C}$  in slight bias over the heavier  $^{13}\text{C}$ , the ratio of the two isotopes preserved in the rock record can indicate, among other things, the activity of primary producers. For example, after the end of the Cretaceous, there was no difference in isotopic ratios of carbon in the skeletons of surface planktonic foraminifers and deep-sea benthic foraminifers. This indicates that organic carbon stopped sinking to the deep ocean. This change must have resulted from either (a) a major decline in oceanic productivity or (b) a major collapse of community structure in marine producers or consumers (D'Hondt *et al.*, 1998).

## X. SUMMARY

Exploration of the fossil record has demonstrated that the earth's biota is fragile at timescales of  $10^7$  years, suffering numerous declines in diversity. The magnitudes of these declines have been variable, but at least five times in the last 500 myr animal diversity was reduced by more than 50%, with the most severe event, at the end of the Permian, eliminating more than 90% of animal diversity. Recoveries from these events have been slower than the mass extinctions, often taking 5 to 10 myr or more. Although the long-term rebounds of biodiversity are predictable, the detailed patterns of recovery are complex, involving outbreaks of disaster

species and considerable ecological instability over timescales of  $10^5$  to  $10^6$  years. Current rates of species extinction could eliminate as many species as seen in past mass extinctions in a geologically short interval, and biotic recovery could be long and unpredictable.

## See Also the Following Articles

DINOSAURS, EXTINCTION THEORIES FOR • EXTINCTIONS, CAUSES OF • EXTINCTIONS, MODERN EXAMPLES OF • FOSSIL RECORD • MARINE MAMMALS, EXTINCTIONS OF • MASS EXTINCTIONS, NOTABLE EXAMPLES OF

## Bibliography

- Alvarez, L. W., Alvarez, W., Asaro, F., and Michel, H. V. (1980). Extraterrestrial cause for the Cretaceous-Tertiary extinction. *Science* 208, 1095–1108.
- D'Hondt, S., Donaghy, P., Zachos, J. C., Luttenberg, D., and Lindinger, M. (1998). Organic carbon fluxes and ecological recovery from the Cretaceous-Tertiary mass extinction. *Science* 282, 276–279.
- D'Hondt, S., Herbert, T. D., King, J., and Gibson, C. (1996). Planktic foraminifera, asteroids, and marine production: Death and recovery at the Cretaceous-Tertiary boundary. *Geological Society of America Special Paper* 307, 303–317.
- Erwin, D. H. (1993). *The Great Paleozoic Crisis*. Columbia University Press, New York.
- Erwin, D. H. (1998). The end and the beginning: recoveries from mass extinctions. *Trends in Ecology and Evolution* 13, 344–349.
- Fischer, A. G., and Arthur, M. A. (1977). Secular variation in the pelagic realm. In *Deep-Water Carbonate Environments* (H. E. Cook, and P. Enos, Eds.), pp. 19–50. *Society of Economic Paleontologists and Mineralogists Special Publication* 25.
- Jablonski, D. (1986). Causes and consequences of mass extinctions: A comparative approach. In *Dynamics of Evolution* (D. K. Elliot, Ed.), pp. 183–229. Wiley Interscience, New York.
- Jablonski, D. (1995). Extinctions in the fossil record. In *Extinction Rates* (J. H. Lawton and R. M. May, Eds.), pp. 25–44. Oxford University Press, Oxford.
- Jablonski, D. (1996). Body size and macroevolution. In *Evolutionary Paleobiology* (D. Jablonski, D. H. Erwin, and J. Lipps, Eds.), pp. 256–289. University of Chicago Press, Chicago.
- Jablonski, D. (1998). Geographic variation in the molluscan recovery from the end-Cretaceous extinction. *Science* 279, 1327–1330.
- Jablonski, D., and Raup, D. M. (1995). Selectivity of end-Cretaceous marine bivalve extinctions. *Science* 268, 389–391.
- Maas, M. C., and Krause, D. K. (1994). Mammalian turnover and community structure in the Paleocene of North America. *Historical Biology* 8, 91–128.
- McGhee, G. R., Jr. (1996). *The Late Devonian Mass Extinction*. Columbia University Press, New York.
- Miller, A. I. (1997). Coordinated stasis or coincident relative stability? *Paleobiology* 23, 155–164.
- Miller, A. I., and Sepkoski, J. J., Jr. (1989). Modeling bivalve diversification: The effect of interaction on a macroevolutionary system. *Paleobiology* 14, 364–369.
- Newell, N. D. (1967). Revolutions in the history of life. *Geological Society of America Special Paper* 89, 63–91.
- Phillips, J. (1860). *Life on Earth: Its Origin and Succession*. Macmillan, Cambridge.

- Raup, D. M. (1989). The case for extraterrestrial causes of extinction. *Philosophical Transactions of the Royal Society of London B* **325**, 421–435.
- Raup, D. M. (1991a). A kill curve for Phanerozoic marine species. *Paleobiology* **17**, 37–48.
- Raup, D. M. (1991b). Periodicity of extinction: A review. In *Controversies in Modern Geology* (D. W. Muller, J. A. McKenzie, and W. Weisert, Eds.), pp. 193–208. Academic Press, London.
- Raup, D. M. (1992). Large-body impact and extinction in the Phanerozoic. *Paleobiology* **18**, 80–88.
- Raup, D. M., and Sepkoski, J. J., Jr. (1984). Periodicity of extinctions in the geologic past. *Proceedings of the National Academy of Sciences, USA* **81**, 801–805.
- Schindewolf, O. H. (1962). Neokatastrophismus? *Geologische Gesellschaft, Zeitschrift Jahrg.* **114**, 430–445.
- Sepkoski, J. J., Jr. (1984). A kinetic model of Phanerozoic taxonomic diversity. III. Post-Paleozoic families and mass extinctions. *Paleobiology* **10**, 246–267.
- Sepkoski, J. J., Jr. (1989). Periodicity in extinction and the problem of catastrophism in the history of life. *Journal of the Geological Society of London* **146**, 7–19.
- Sepkoski, J. J., Jr. (1990). The taxonomic structure of periodic extinction. *Geological Society of America Special Paper* **247**, 33–44.
- Sepkoski, J. J., Jr. (1992). Phylogenetic and ecologic patterns in the Phanerozoic history of marine biodiversity. In *Systematics, Ecology, and the Biodiversity Crisis* (N. Eldredge, Ed.), pp. 77–100. Columbia University Press, New York.
- Sepkoski, J. J., Jr. (1996). Patterns of Phanerozoic extinctions: A perspective from global data bases. In *Global Events and Event Stratigraphy* (O. H. Walliser, Ed.), pp. 35–52. Springer, Berlin.
- Sepkoski, J. J., Jr. (1997). Biodiversity: past, present, and future. *Journal of Paleontology* **71**, 533–539.
- Sepkoski, J. J., Jr., and Koch, C. F. (1996). Evaluating paleontologic data relating to bio-events. In *Global Events and Event Stratigraphy* (O. H. Walliser, Ed.), pp. 21–34. Springer, Berlin.
- Sepkoski, J. J., Jr., and Schopf, J. W. (1992). The Proterozoic fossil record: Special problems in analyzing diversity patterns. In *The Proterozoic Biosphere: A Multi-Disciplinary Study* (J. W. Schopf and C. Klein, Eds.), pp. 525–527. Cambridge University Press, Cambridge.
- Signor, P. W., and Lipps, J. H. (1982). Sampling bias, gradual extinction patterns, and catastrophes in the fossil record. *Geological Society of America Special Paper* **190**, 291–296.
- Solé, R. V., Manrubia, S. C., Benton, M., and Bak, P. (1997). Self-similarity of extinction statistics in the fossil record. *Nature* **388**, 764–767.



# MASS EXTINCTIONS, NOTABLE EXAMPLES OF

Douglas H. Erwin  
*National Museum of Natural History*

---

- I. Great Mass Extinctions
  - II. Lesser Mass Extinctions
- 

## GLOSSARY

**carbon cycle** Photosynthetic organisms preferentially use the carbon-12 isotope instead of carbon-13, enriching living organisms in C-12. This differentiates the organic from inorganic carbon reservoirs. Quantifying shifts between the two isotopes of carbon reveals shifts between the two reservoirs.

**mass extinction** A rapid loss of a large fraction of biodiversity on timescales of  $10^0$ – $10^6$  years, generally involving a variety of unrelated groups

**supercontinent** The amalgamation of many continental masses into a single mass through continental drift. The supercontinent of Pangea (from 300 to approximately 190 million years ago) included most continental areas other than those that comprise east Asia.

---

**THE 3.6-BILLION-YEAR FOSSIL** record is interrupted by numerous mass extinctions, but only the 600 million years since the appearance of animals is sufficiently dense to provide an adequate record of these great biotic crises. Five great mass extinctions punctuate the record, several of which had major impacts on the course of

evolution. Additionally, there are many smaller biotic crises, but not all of these are well studied.

## I. GREAT MASS EXTINCTIONS

The five most severe biotic crises occurred during the Late Ordovician Period, the Late Devonian, the Late Permian (two closely spaced episodes), at the end of the Triassic Period, and at the end of the Cretaceous [the great Cretaceous–Tertiary (K/T) mass extinction that eliminated the dinosaurs] (Figs. 1 and 2). Despite numerous attempts to find common mechanisms for these events, a variety of different causes appear to have been involved, from global cooling during the end-Ordovician event to the impact of an extraterrestrial object with the earth during the K/T mass extinction.

### A. Late Ordovician

The second largest of the great mass extinctions occurred during the Late Ordovician [439 million years ago (Ma)] in two pulses separated by approximately 1 million years. With few, if any, organisms on land, this marine extinction of approximately 25% of all families and 60% of genera affected most marine groups. Despite the magnitude of the extinction, it had few lasting ecological effects. With some minor exceptions, Silurian faunas look much like Ordovician ones, in contrast to the profound differences between Triassic and Permian faunas.



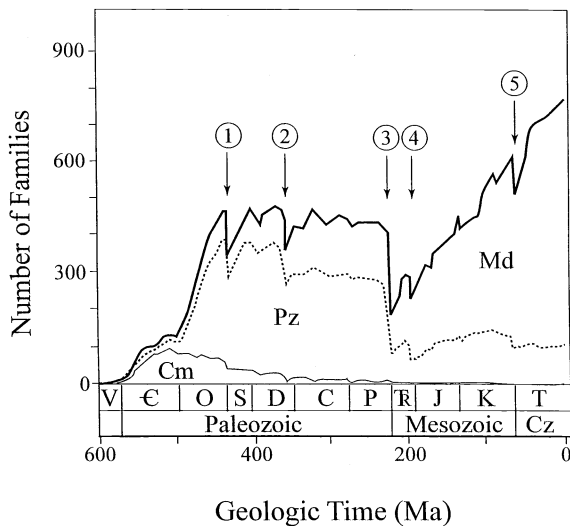


FIGURE 1 Diversity of marine families through the Phanerozoic (the past 600 million years). The thick black line shows the total number of marine families in each of 83 geologic stages. The circles correspond to the five major mass extinctions: 1, Late Ordovician; 2, Late Devonian; 3, Late Permian; 4, end-Triassic; and 5, end-Cretaceous. The lower solid line shows the proportion of family diversity assignable to the Cambrian evolutionary fauna (Cm); the dotted line shows the proportion of family diversity assignable to the Paleozoic evolutionary fauna (Pz), and the area between the dotted and thick solid line is the proportion of family diversity assignable to the modern evolutionary fauna (Md). V, Vendian; C, Cambrian; O, Ordovician; S, Silurian; D, Devonian; C, Carboniferous; P, Permian; Tr, Triassic; J, Jurassic; K, Cretaceous; T, Tertiary; Cz, Cenozoic; Ma, million years ago [after Sepkoski, J. J., Jr. (1984), A kinetic model of Phanerozoic taxonomic diversity. III Post-Paleozoic families and mass extinction. *Paleobiology* 10, 246–267; used with permission].

Graptolites, a group of colonial, floating marine organisms, were commonly found on the outer shelves and were almost completely eliminated by this extinction, with only a few species surviving. Among other swimming and floating groups, conodonts, nautiloids, and planktonic groups all appear to have suffered considerable extinction. The articulate brachiopods were one of the major components of Ordovician ecosystems, and approximately 83% of brachiopod genera became extinct, primarily during the first of the two crises. A brachiopod fauna associated with the brachiopod *Hirnantia* developed in many areas between the two extinction pulses. At a single locality, fewer than 10 species are present and all appear to be adapted to cold water. This *Hirnantia* fauna appears to have reached equatorial latitudes during this interval, providing one index of the extent of cold-water conditions. The second phase of the extinction wiped out this fauna. Trilobites display a similar pattern of extinction followed by the spread of a low-diversity, broadly distributed group between the extinctions; however, the second extinction appears to be more significant than the first. Trilobites living in deeper water and floating forms suffered greater extinctions than those occupying shallow-water habitats. Bivalves and echinoderms also experienced considerable extinction, whereas bryozoans, a colonial group of filter feeders, were not affected as severely. Approximately 70% of rugose corals disappeared, but they recovered quickly so the extinction did not have a major impact. In fact, this is the only major mass extinction that did not trigger an extensive turnover in reef communities. The decrease in sea level associated with glaciation eliminated deposition of marine rocks in many areas, of course, thus complicating analysis of changes in biodiversity during this interval.

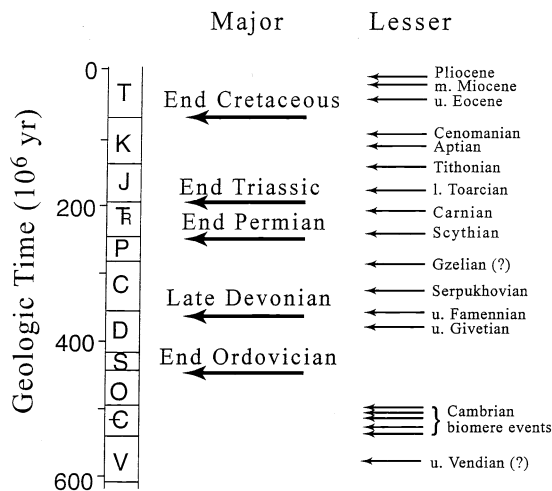


FIGURE 2 Major and lesser mass extinctions during the Phanerozoic. Not all the lesser mass extinctions are discussed in the text. Some only appear in compilation studies, as in Fig. 1, and may be statistical artifacts. See the legend to Fig. 1 for abbreviations.

The causes of the Late Ordovician mass extinction are relatively clear, although there is disagreement about the relative significance of different aspects. Global climates during the Ordovician were warm; geologists term this a greenhouse climate. One consequence of this was a decline in vertical circulation in the oceans, leading to a depletion of oxygen in the deep oceans. As a large amount of continental area moved near the South Pole, the climate cooled and glacial conditions set in; as ice developed, sea level decreased, eventually exposing the continental shelves and leading to the first phase of the extinction. Deep-water groups were particularly affected. A low-diversity, broadly distributed community adapted to relatively cool conditions developed in the wake of this extinction. The end of the glaciation 500,000–1 million years later led to an increase in sea level and ocean temperatures. The rising

seas are associated with the spread of low oxygen, or anoxic, conditions, leading to pervasive extinctions among shallow-water species. Changes in the carbon cycle, chronicled as a shift in the ratios of the isotopes of carbon-12 to carbon-13, coincide with the major events of this scenario.

Although rocks recording this extinction are common in Europe and China, few deposits in North America span the extinction interval other than an important sequence in Anticosti Island, Canada, and a newly described sequence in central Nevada. The rocks in Nevada preserve an excellent record of the decline in sea level, the biological extinction, and consequent perturbation of the carbon cycle (Finney *et al.*, 1999). As the onset of glaciation lowered sea level, graptolites suffered the first extinctions, and the extinctions can be tracked from near-shore regions into deeper water habitats. Finally, as the decrease in sea level exposed most of the shelves, most marine communities were affected. The Nevada sections cover a diversity of environments within a small region, allowing the extinctions to be tracked between different environments. This is rarely the case in other areas; therefore, the difference in timing of extinctions has been difficult to resolve. The Nevada sections also reveal that a sharp shift in the carbon cycle was related to the onset of glaciation and the resulting decrease in sea level but was not directly related to the extinction.

## B. Late Devonian

Although extraterrestrial impact is commonly associated with the Cretaceous/Tertiary mass extinction today, it was first proposed in 1970 for the Late Devonian mass extinctions. These biotic crises have been recognized since the 1870s, but the rapidity of these events remains unclear. For the Givetian and Frasnian Stages of the Late Devonian, Sepkoski (1996) records a 57% extinction of marine genera (47% for filtered data) and approximately 22% extinction of marine families. Particularly in European rocks, a sequence of events has been distinguished: a Taghanic event at the end of the Givetian Stage, the Kellwasser event at the end of the Frasnian, and a subsequent extinction episode near the Devonian-Carboniferous boundary (the Hangenberg event), although it is unclear whether the last event is part of the same series or has a different cause. Terrestrial plants had evolved and evidently experienced their first biotic crisis during this interval as well, although the available data are too spotty to reveal much information (McGhee, 1996; Hallam and Wignall, 1997).

Perhaps the most geologically obvious effect of the

mass extinction is the elimination of the widespread Devonian reefs. These reefs were dominated by rugose and tabulate corals (the corals of the Paleozoic) and stromatoporoids, a group of calcareous sponges, in addition to some brachiopods and large foraminifera. Major Devonian reef complexes are known from Australia, western North America, Russia, and Europe. The extinctions suffered by these groups essentially eliminated true reefs until the Middle Permian. Brachiopods were one of the most diverse members of Devonian marine communities and illustrate the magnitude of the extinction. At least six families disappeared during the Taghanic crisis. McGhee (1996) indicates that 75% of brachiopod genera disappeared during Frasnian, including at least two major orders; many of the extinct genera were from tropical regions, suggesting, as for other groups, that higher latitude, cold-water groups were less susceptible to extinction. Brachiopods recovered quickly during the Famennian and were relatively unaffected by the Hangenberg event. Bryozoans have perhaps the strangest pattern of extinction. They suffered one of their largest biotic crises at the end of the Givetian, but no other groups are known to be affected at this time; it is suspected that the data may be faulty in this case.

Trilobites, although much reduced in diversity from the Cambrian and Ordovician, were still important during the Devonian. The Givetian assemblage included eight orders and 13 families, but only a single family and eight genera survived to the Famennian. Although there is heightened extinction from the Givetian through the Frasnian, the only discrete crisis occurred during the Kellwasser event when numerous important groups disappeared. Deep-water and some shallow-water trilobites diversified immediately after the extinction during the Famennian, but trilobites suffered considerable extinction again at the Devonian-Carboniferous boundary and never regained their previous diversity.

Two other arthropod groups, the eurypterids and the ostracods, also belonged to Devonian marine communities. Eurypterids were relatively rare, and their diversity appears to be declining throughout the interval, although one study suggests the extinction of several families during the last two stages of the Devonian. The ostracod record is far better. The pelagic ostracods suffered considerable extinction during the Kellwasser event but little extinction during the Hangenberg crisis. Among the benthic ostracods significant extinctions occurred during the Kellwasser, with groups tolerant of low oxygen showing greater survival.

Ammonoids, the coiled cephalopods that were distant cousins of the modern chambered *Nautilus*, almost disappeared, with an estimated 88% species extinction.

Ammonoids display a boom-and-bust pattern throughout their history, and this includes massive extinctions during the Taghanic event, in which only two genera survived, and the Kellwasser event, with only few surviving genera, mostly from deep waters. Famennian ammonoids diversified considerably, producing some bizarre morphologies, but only a single family, and perhaps a single genus, survived the Hangenberg event at the end of the Devonian. Ammonoids once again diversified strongly in the early Carboniferous and again became important members of marine ecosystems. Nautiloids, the other major group of fossilizable cephalopods, encountered only a minor perturbation during the Kellwasser event but suffered perhaps their greatest crisis during the Hangenberg event when 11 of 19 families disappeared. In contrast to the ammonoids, early Carboniferous nautiloids did not expand greatly in diversity. The other major molluscan groups are the bivalves and gastropods. Neither was particularly important during the Devonian and each suffered only minor losses in diversity.

The only prominent vertebrate group was the fish, which were amazingly diverse during the Devonian (hence the nickname "the age of fishes"). Fish have proven to be very resistant to mass extinction, but all fish groups other than sharks did poorly during this crisis. Among the jawless, armored fish all nine families disappeared. More advanced fish suffered as well during both the Kellwasser and Hangenberg crises. Rediversifications occurred after each extinction, however, and that following the Hangenberg eventually led to the teleosts, the modern bony fish. Conodonts are tiny phosphatic elements that formed part of the jaw of small eel-like chordates. Because of their diversity, conodonts are important tools in correlating rocks from different parts of the world. The Kellwasser represents a major biodiversity crisis for the group, at least at the generic and species level. Less is known of their history during the Hangenberg crisis.

What patterns of ecological selectivity are apparent from the variations in extinction intensity among different groups? Often, patterns of differential extinction are among the best clues to the causes of mass extinctions and other biotic crises. The patterns of change are complex, with long-term decline in many groups as well as during the discrete Taghanic, Kellwasser, and Hangenberg events. The first of the three biotic crises primarily affected shallow-water benthic groups. The Kellwasser is the best known of the three events (McGhee, 1996). Warm-water groups, particularly brachiopods and reef forms, suffered greater extinction, whereas groups adapted to high latitudes and colder

waters seemed to have done well. Most swimming groups either became extinct or declined greatly in diversity. Whether the Hangenberg crisis qualifies as a mass extinction is unclear. Many swimming groups, including placoderm fish, nautiloids, and ammonoids, as well as trilobites and stromatoporoids experienced considerable extinction. The overall pattern is the reverse of that of the Taghanic event (Hallam and Wignall, 1997). The causes of the first and third events, the Taghanic and Hangenberg, are poorly understood, so emphasis will be placed on the better studied Kellwasser episode.

Extraterrestrial impact, sea-level fluctuations with attendant spread of anoxic waters, and climatic changes and global cooling are among the more prominent suggested causes of the biotic crisis (McGhee, 1996). The evidence for impact can charitably be described as spotty. Such evidence would include demonstration that extinctions are rapid and virtually simultaneous; signs of enrichment of elements found in extraterrestrial objects, particularly iridium; indications of impact including shocked quartz; fused glass from the impact site; and geochemical anomalies, including a shift in the carbon cycle. Although evidence of impact is present in Late Devonian rocks, including the 52-km Siljan crater in Sweden and heightened levels of iridium, these are not closely associated with extinction horizons. The most recent advocates of impact have placed considerable emphasis on shifts in the carbon cycle, but such shifts can arise through a variety of causes and are not diagnostic of impact. This does not imply that no impact was associated with the extinctions, only that no compelling evidence for such impact has been advanced (Copper, 1998).

Global cooling due to the closure of an equatorial seaway with an associated decrease in sea level has been advanced based on the heightened survival of cold-water groups, the presence of glacial sediments in Brazil, and the rapid sea level fluctuations, reminiscent of Pleistocene glacially induced fluctuations (Copper, 1998). The correlation between the physical events and the extinction is doubtful, however, and the enhanced survival of cold-water taxa can also be interpreted as survival of deep-water forms. Given the time span over which these changes operate, this model also seems to require a fairly prolonged extinction. In McGhee's (1996) excellent analysis of the extinction he concludes that global cooling best explains the pattern of extinction and survival, but he argues that the extinction was relatively sudden and thus glaciation is not an adequate explanation. Although acknowledging the difficulties with the im-

pact hypothesis, he is sympathetic to it as a trigger for the global cooling.

This alternative explanation of the patterns of biotic survival leads to the most widely discussed models, all of which invoke sea level, shifts in the carbon cycle, and low-oxygen waters (Hallam and Wignall, 1997). The scenarios differ in the triggering mechanism, which ranges from rapid glaciations to oceanic volcanism and the effects of increased flux of nutrients to the oceans caused by the first development of land plants. The evidence for warming and anoxia has been questioned, with Copper (1998) noting that black shales are not unambiguous indicators of warming and may form by global cooling, and that no clear correlation has been found between black shales and low-oxygen waters. Furthermore, he notes that many of the atrypid brachiopods adapted to low-oxygen environments became extinct while other brachiopods from the same communities survived. Investigations of these events continue, but currently there does not appear to be sufficient evidence to provide definitive tests of the various competing models for the causes of this extinction. The most reasonable explanation, however, appears to involve a multitude of interacting causes rather than a single critical cause. This is a feature shared with the greatest mass extinction, at the end of the Paleozoic.

### C. Late Permian

The two pulses of the end-Permian mass extinction collectively extinguished approximately 95% of all marine species and perhaps 70% of species on land. The destruction caused the most pervasive reorganization of marine communities since the beginning of the Paleozoic, and modern marine communities continue to reflect the heritage of this event (Erwin, 1993). The only mass extinction ever suffered by insects occurred during this interval, and increasing evidence suggests that it accelerated an ongoing, climatically driven change in land plants. Two discrete intervals of extinction are apparent in the marine realm with most of the terrestrial effects evidently associated with the second of the two pulses (Hallam and Wignall, 1997). The first pulse came at the end of the Middle Permian, or Guadalupian, when a decrease in sea level resulted in the drying out of marine basins in many parts of the world. The second, and probably more significant, extinction occurred approximately 10 million years later at the end of the Permian. Recent evidence from China indicates that this extinction occurred in less than 500,000 years, or essentially catastrophically (Bowring *et al.*, 1998). Most of the detailed studies have focused on this second

event, as have competing hypotheses about the causes. Less is known about the first pulse, and no detailed field studies have been published.

A great diversity of reefs occurred during the Permian; at least seven major types have been identified, ranging from small masses of sponges and algae to the massive bryozoan-sponge-calcareous algal reefs of west Texas that now harbor Carlsbad Caverns. Many reefs, including those in west Texas, disappeared during the first extinction pulse, but other reefs continued until the very end of the Permian. Large reefs are found in Sichuan Province, China, immediately below the Permo-Triassic boundary. Reefs are missing from all but the end of the Early Triassic, and when they reappear in the Middle Triassic scleractinian corals, a new group of corals, are prominent members, as they are today. The Paleozoic tabulate and rugose corals became completely extinct at the end of the Permian and recent study has shown that the rugose corals persisted at least until the very end of the interval, contradicting earlier reports of a gradual decline.

Brachiopods are perhaps the most diagnostic group of Paleozoic marine organisms, much as clams and snails are today. However, they are unfamiliar to many people because they never recovered from the extensive extinctions they suffered during the Late Permian. As with several other groups discussed later, they suffered considerable tropical extinction at the end of the Guadalupian, during which 53 of 82 genera died out. A small increase in diversity occurs after this extinction, followed by the second phase of extinction, with the loss of 90% of families and 95% of the genera. The colonial bryozoa suffered catastrophic extinctions, although only the lacy fenestrates became completely extinct. Echinoderms suffered pervasive extinction as well, with the demise of several major groups of stalked echinoderms, including an entire class, the Blastoidea. Although echinoids (sea urchins) are ubiquitous members of marine communities today; only a single genus (*Miocidaris*) survived the end-Permian extinction. Echinoderms are present in the Early Triassic, but most fossils are fragmentary and appear to represent very few species that were highly abundant. They were essentially weeds, and are found in many parts of the world.

A group of microfossils that produced a calcareous test, or shell, were a major part of Permian marine communities. Known as the Foraminifera, they have been well studied, in part because they evolved rapidly and are useful in determining the age of rocks. They were also ecologically diverse and provide some clues to the nature of the extinction. One important group of shallow-water forams almost completely disappeared

in the first extinction pulse and the few remaining species became extinct during the second phase of the extinction. Many new species appeared after the first wave of extinction. The second phase of extinction eliminated many architecturally complex forms found in the tropics.

Among molluscs, the extinction was more muted. Gastropods suffered considerable extinction at the generic level, but this eventually led to new groups of gastropods becoming important in the Triassic. Broad geographic range and occupation of a wide range of environments evidently aided survival at the end of the Permian. Bivalves experienced relatively minor extinction, but took magnificent advantage of the postextinction possibilities, becoming dominant members of modern communities. The ammonoids, as usual, suffered near-catastrophic extinction.

Extensive biotic crises occurred on land as well, providing a critical clue to the cause. Clearly, any proposed cause which affects the oceans to the exclusion of land is ruled out. Twenty-two orders of insects had appeared by the end of the Permian, and 9 became extinct and another 10 suffered serious diversity declines. This is the only mass extinction ever suffered by insects. Many of the groups that disappeared could not fold their wings back over their body (as do flies and butterflies) but rather held them straight out to the side; dragonflies are a relic of these Permian groups. In contrast to studies in the 1980s which suggested that plants experienced relatively little extinction, recent work has revealed massive perturbations, including a relatively sudden extinction of plants in Australia. In many parts of the world a sharp spike of pollen from fungi has been documented at the Permo-Triassic boundary. What caused this remains unclear, but it suggests a profound and rapid disruption of terrestrial ecosystems.

Earliest Triassic sediments in the Karoo Basin of South Africa contain abundant specimens of the pugnosed therapsid *Lystrosaurus* and occasional specimens of a few other species. In contrast, the Permian rocks below have a diverse fauna of several different groups of carnivores and herbivores. Although the exact position of the Permo-Triassic boundary in these terrestrial sequences is unclear, as is the correlation to the marine Permo-Triassic boundary, abundant evidence demonstrates a considerable extinction of vertebrates. Approximately 78% of tetrapod families apparently became extinct at about this time.

The end-Permian mass extinctions occurred during an interval of considerable change in tectonics, climate, and other aspects of the physical environment. Virtually all these changes have been invoked by one or another

extinction model at some time. Perhaps the most common link has been between the extinction and the formation of the supercontinent of Pangea. During the Permian virtually all the continents, with the exception of some pieces that today comprise China and east Asia, collided to form a single supercontinent. The climatic and biological effects of the formation of Pangea have figured prominently in many extinction scenarios, but geologic data have shown that it formed approximately 20–25 million years before the extinction, broke apart slightly, and then reformed in the Late Triassic. Neither event is associated with any biodiversity crisis. Additionally, the rapidity of the extinction is much too fast to be explained by the slow dance of the continents. Glaciation has also been invoked and may have been involved in the first extinction pulse, but there is scant evidence for glaciation in the latest Permian.

Deposits of rock spanning the Permo-Triassic boundary extinction interval are relatively sparse, and this led several generations of workers to suggest a profound decrease in sea level at this point, and many extinction scenarios have focused on this issue. Recent studies of many regions have shown that the regression ended well before the onset of the extinction, and sea level was actually rising during the extinction (Hallam and Wignall, 1997). As with many of the extinction events, a sharp change in the carbon cycle has been well documented at the extinction horizon. Unfortunately, several different extinction scenarios can produce the observed shift.

More significant, however, is the close connection in time between the eruption of extensive flood basalts in Siberia and the extinction. These comprise one of the two largest known continental flood basalts and represent at least 11 large lava flow complexes and more than 45 individual flows. The flood basalt covers an area of 2–3 million cubic kilometers, with a depth of more than 4 km in some places. The entire complex of flows appears to have erupted in approximately 1 million years—a phenomenal rate of volcanic production. The link to the extinction is more obscure, however.

The cause of the end-Permian extinction remains ambiguous, despite the tremendous research effort since 1990. The following facts are clear, however (Bowring *et al.*, 1998): Low-oxygen levels occurred in deep and shallow waters during the extinction interval; sea level was rising; the extinction coincides with an abrupt shift in the carbon cycle, although the magnitude of the shift is less than previously thought; the extinction pattern is consistent with poisoning from carbon dioxide; and there is increasing evidence for a sudden climatic warming immediately after the extinction, per-

haps associated with the extinction. Several extinction scenarios are consistent with this scenario, including the impact of an extraterrestrial object (despite many searches, no evidence for impact has been discovered), release of CO<sub>2</sub> and sulfur dioxide from the Siberian flood basalts triggering greenhouse warming and acid rain, and a shift in ocean circulation.

#### D. End-Triassic

The end-Triassic extinction is the second smallest event in most analyses and probably the most poorly known of any of the five great mass extinctions. Sepkoski's (1996) data show a 53% extinction for all marine genera (40% when the data are filtered) and 22% for marine families, similar to the K/T mass extinction. Among marine groups, Sepkoski's database shows significant extinctions among ammonoids, bivalves, gastropods, and brachiopods, but the most reliable studies have focused on ammonoids and bivalves in northwestern Europe; little data are available from elsewhere. Whether this extinction is entirely at the end-Triassic or includes a substantial event earlier, during the Carnian Stage, has been the subject of active debate.

Ammonoids always experienced a boom-and-bust pattern, and this event marks one of their largest crises. The group was quite diverse during the latest Triassic, but at least six superfamilies became extinct near the boundary, followed by a rapid rediversification in the Early Jurassic. Some estimates suggest that only a single ammonite genus survived the crisis. Although no family-level extinctions occurred among bivalves, approximately half the genera disappeared and turnover at the species level was particularly high. Endemic genera and those in deep offshore facies appear to have experienced greatest extinction. Conodonts, a group of phosphatic microfossils formed by simple chordates, survived the end-Permian mass extinction in fine form but experienced a sharp decrease in diversity only 5 million years later, during the Early Triassic. This biostratigraphically important group then became extinct at the end of the Triassic. Extinction of foraminifera and ostracods was relatively minor. Reefs dominated by scleractinian corals first appear in the Middle Triassic and become prominent in the Carnian. These corals, associated sponges, and other reef taxa suffer a sharp reduction in diversity during this episode, followed by an interval during which reefs are absent.

Some terrestrial groups experienced extinction at this time as well, although less so than during the end-Permian or K/T events. Pollen records show that many seed ferns became extinct, as did some other plant

groups, but a lack of a detailed floral record for the interval has hampered analysis in Europe. In eastern North America the boundary is well preserved in the Newark Supergroup and approximately 60% of the pollen species disappeared within approximately 500,000 years. Similar patterns have been found in Arctic Canada but not in Australia or Siberia, possibly indicating considerable regional variation in extinction intensity. Of particular relevance is a sharp increase in the abundance of fern spores at the extinction, reminiscent of the fern spike associated with the K/T extinction. Whether or not any vertebrate extinctions occurred remains highly contentious. One analysis of the Newark group vertebrates indicated extensive extinctions, but this conclusion has been challenged by many vertebrate paleontologists who have argued that the Newark Supergroup record is sparse and there are no other data supporting an extinction of terrestrial vertebrates.

The causes of this extinction remain enigmatic, with most attention focused on flood basalt volcanism, possible bolide impact, and marine anoxia. Recent age dates reveal that the Central Atlantic Magmatic Province, an extensive region of flood basalts extending across eastern North America, Europe, northwest Africa, and northeastern South America, erupted coincident with the end-Triassic mass extinction (Marzoli *et al.*, 1999). As with the end-Permian and K/T extinctions, the causal link between the volcanism and extinction remains unclear. The assumption is that some combination of rapid climatic shifts, introduction of volcanic gases (carbon dioxide and sulfur dioxide), and acid rain would be sufficient to cause the extinctions. The available terrestrial data are consistent with a sharp increase in carbon dioxide and a disruption of the carbon cycle. The extinction had previously been associated with the Karoo flood basalts in South Africa, but these are now dated as Early Jurassic.

Changes in sea level, perhaps induced by uplift of the supercontinent of Pangea during the initial breakup that produced the Atlantic Ocean, have been invoked as well. Hallam and Wignall (1997) invoke a sharp decrease in sea level and then a transgression of low-oxygen waters, with the extinction coincident with the anoxia. Such a mechanism fails to explain the widespread extinctions of terrestrial plants, however.

The final possibility is the impact of an extraterrestrial object. The evidence is less secure than for the K/T extinction, but it includes the fern spike and suggestions of shocked quartz (an indicator of impact). The Manicouagan impact crater in Quebec was once associated with the end-Triassic event, but new dating reveals that it is at least 13 million years older. Cur-

rently, the evidence for impact remains equivocal but cannot be completely dismissed.

### E. End-Cretaceous

The end of the dinosaurs has been both a popular and a scientific question almost since their discovery more than 150 years ago. When the leading explanation became a collision with an extraterrestrial object, both popular and scientific interest exploded. Along with the extinction of nonavian dinosaurs, many other large vertebrates disappeared as well as a wide variety of marine organisms. Among marine organisms, the extinction claimed 16% of the families and 47% of the genera, making this the smallest of the five great mass extinctions. Approximately 18% of terrestrial vertebrate families disappeared.

Considering the marine organisms first, planktonic foraminifera, a group of floating single-celled organisms, experienced a fairly catastrophic extinction. There has been considerable debate regarding the abruptness of these disappearances (MacLeod *et al.*, 1997), which can only be resolved by appropriate statistical analysis. Preservation of fossils is so uneven that even a catastrophic mass extinction will produce an apparently gradual pattern of disappearance leading up to the extinction horizon. Statistical analysis of fossil occurrences can correct this problem, however. Such an analysis of ammonites from the coast of France and Spain demonstrated at most three phases of extinction (Fig. 3; Marshall and Ward, 1996). Of the 28 species of ammonite, 25% probably became extinct during the several million years leading up to the mass extinction; a decrease in sea level just before the extinction may have been responsible for the extinction of up to 35% of species, but 40–75% of the species went extinct at the K/T boundary. The disappearance of planktic foraminifera is likewise abrupt, as emphasized by the recent recovery of a core from the ocean bottom in the western Atlantic (Norris *et al.*, 1999). Another group of calcareous microfossils also suffered species extinction of perhaps 85%.

Among other marine groups, the benthic foraminifera suffered much less extinction, as did phytoplankton. The various mollusc groups vary greatly in extinction pattern. The ammonites finally succumbed, as did another group of cephalopods, the belemnites. The bizarre, reef-forming rudistid bivalves disappeared, but evidently did so well before impact event, as did a group of large (up to 1 m), flat clams, the inoceramids. As noted previously, the ammonite extinction is abrupt and coincides with the boundary. The diverse record

of bivalves and gastropods has been exploited for some of the most detailed studies of the evolutionary implications of mass extinctions.

Different groups display very different patterns of extinction selectivity. Despite many claims for increased extinction in the tropics, a detailed study of bivalve extinction patterns found no such pattern (Jablonski and Raup, 1995). Earlier, however, Jablonski showed that survival of bivalves and gastropods, two groups with relatively low rates of extinction, was dependent on the broad geographic range of an entire lineage. During the interval preceding the extinction, the broad geographic range of species (rather than multispecies lineages) and the type of larvae best predicted long-term survival. This pattern suggests that the factors controlling survival during mass extinction events may be very different than those controlling evolution between mass extinctions.

Despite the attractiveness of dinosaurs, the scarcity of bones makes them largely unsatisfactory for studies of the pace of the extinction. Northeastern Montana contains one of the best records of vertebrate diversity. Different groups experienced very different extinction patterns. Sharks, marsupial mammals, lizards, and the dinosaurs suffered very high extinction, whereas frogs, salamanders, turtles, placental mammals, and crocodiles suffered virtually none. Both large and small vertebrates are part of the list of extinct forms, countering earlier claims that only large vertebrates became extinct. Clearly, animals in freshwater settings encountered much less extinction than did fully terrestrial forms: Approximately 90% of freshwater forms survived but only 12% of the land species survived. Dinosaurs had high extinction throughout their history, although this is not generally appreciated. The last dinosaurs disappear about 1 m below the extinction horizon. Although claims for dinosaurs persisting after the end of the Cretaceous have generally been dismissed, there is still considerable uncertainty regarding whether dinosaurs were in decline prior to the impact. Recent work indicates that the fossil record of vertebrates, particularly dinosaurs, is simply too poor to resolve the question in a statistically rigorous fashion.

A remarkable increase in the abundance of fern spores coincides with the boundary at sites throughout North America and coincides with the iridium spike and the occurrence of shocked quartz. An extinction of almost 80% of plant species has been documented in western North America, but this declines away from this area so that in Antarctica and New Zealand the extinction is almost unrecognizable.

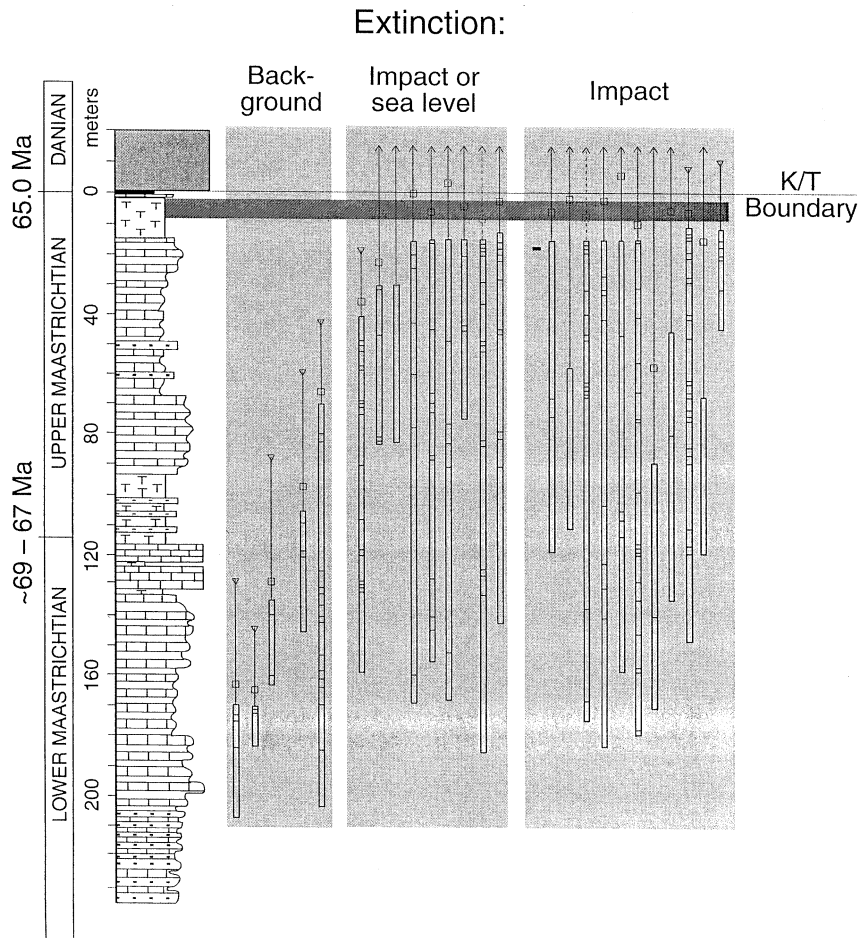


FIGURE 3 Graph showing the observed temporal ranges and occurrences (short horizontal lines) of 28 species of ammonites at a K/T boundary section in Spain. Since the observed ranges underestimate true ranges, they can be corrected by statistical analysis. The 50 and 95% confidence intervals on the final occurrences are represented by the small squares and inverted triangles, respectively. Species that extend beyond the boundary are indicated by arrows on their ranges. The species are classified into three categories: (i) those that became extinct prior to the boundary, (ii) those that may have become extinct due to either impact or a decrease in sea level shortly before the K/T boundary, and (iii) those that became extinct at the K/T boundary. Discussion and further details are provided in Marshall and Ward (1996) [reproduced with permission from Pope, K. O., d'Hondt, S. L., and Marshall, C. R. (1999). Meteorite impact and the mass extinction of species at the Cretaceous/Tertiary boundary. *Proc. Natl. Acad. Sci. USA* 95, 11028. Copyright 1999 National Academy of Sciences, U.S.A.].

Scientists have proposed extraterrestrial causes of mass extinctions for at least 250 years, but only recently have methods of testing such suggestions been available. In 1980, Luis Alvarez and colleagues provided the first definitive test when they announced the discovery of an increased amount of iridium associated with the Cretaceous/Tertiary boundary at Gubbio, Italy. Iridium is more common in material from space than on the surface of the earth so is a good index of the rate of

delivery of extraterrestrial material. This iridium anomaly has since been discovered at many other K/T boundary sections, both marine and terrestrial. Other indicators of an extraterrestrial influx have also been recovered, including shocked quartz and other chemical anomalies. The discovery of a crater approximately 200 km in diameter at Chicxulub in the Yucatan gave a significant boost to the impact scenario. The object is believed to have come from the southeast and vaporized



about 3 km of rock, creating a massive cloud of water vapor and sulfur dioxide. This cloud would cool the earth for a long period and produce extensive acid rain. North America probably felt the greatest effect of the blast, accounting for the higher terrestrial extinctions there. Although smoke and dust would have been produced as well, recent studies indicate that the effects would have ameliorated within a few weeks to months and probably played a minor role in the extinction.

Although most attention has focused on the impact at Chicxulub in the Yucatan, the latest Cretaceous was a time of considerable environmental change and many other causal mechanisms have been proposed. Like the end-Permian and Late Triassic mass extinctions, the K/T coincides with a very large flood basalt eruption—in this case, the Deccan traps in India. As with the earlier events, the eruptions could generate acid rain and climatic change. Proponents of this mechanism have been unable to explain many of the features of the K/T boundary that are explained by the impact scenario. The extinction also coincides with a major marine regression, which has also been invoked as a cause (Hallam and Wignall, 1997). Low oxygen levels in the oceans have been proposed, although Hallam and Wignall, generally advocates of such anoxia hypotheses, acknowledge that this is unlikely for the K/T. Currently, the existence of the impact seems indisputable, as does the disappearance of most species at that point. Of continuing interest, however, is the influence of other environmental changes on extinction leading up to the impact. In other words, is the impact solely responsible for the Cretaceous/Tertiary mass extinction, or are other processes also responsible?

## II. LESSER MASS EXTINCTIONS

During the past decade, paleontologists have identified many smaller mass extinctions (Hallam and Wignall, 1997; for postextinction recoveries, see Erwin, 1998). Many of these events are not well studied, and a few only appear on large-scale, synoptic analyses and may be statistical artifacts. Several of these events are significant however, and they provide a useful perspective on the larger mass extinctions. They also emphasize the continuum in biotic disturbances between the great end-Permian mass extinction and the common biotic crises which do not qualify as mass extinctions. Only reasonably well-characterized events are noted here.

### A. Neoproterozoic and Paleozoic

What happened to the enigmatic, soft-bodied Ediacaran biota of the late Neoproterozoic (575–543 Ma)? This

diverse assemblage of probable metazoans is found worldwide and includes a diverse assortment of cnidarian-grade organisms and a few probable ancestors to the clear metazoans of the Cambrian. Virtually all the Ediacaran forms disappear from the record approximately 543 Ma at the base of the Cambrian. Whether this represents a mass extinction, a change in preservational style so that Ediacaran fossil were no longer preserved, or a gradual replacement of these forms by diversifying animals is unclear. A very sharp shift in the carbon cycle occurs at this point, similar to but larger than that which marked the end-Permian mass extinction. This geochemical anomaly is consistent with a mass extinction, although whether one actually occurred remains unclear.

The Early Cambrian is characterized by a diverse assemblage of small shells, spicules and tubes, trilobites, and the sponge-like archaeocyathids. Two pulses of extinction near the end of the Early Cambrian each eliminated approximately 40–50% of marine genera. Reefs, small shelly fossils, and some brachiopods were particularly affected. These extinctions mark the end of this characteristic fauna and paved the way for the trilobite-dominated marine communities which dominate the remainder of the Cambrian. The causes of these events are unclear, but marine anoxia, a decrease in sea level, and possibly changes in atmospheric composition have been suggested. These crises mark the end of reefs until the Ordovician, when the origination of the Paleozoic corals (tabulates and rugosans), bryozoans, and some sponges produced a new suite of reef architectures.

Equally puzzling are a series of crises affecting near-shore trilobites and other marine benthos during the Middle and Late Cambrian in North America. Each event begins with a radiation of new trilobite families across marine shelves, evidently from deeper water ancestors. These new forms are initially widespread, with broad environmental tolerances. The new forms rapidly produce a diverse, endemic trilobite assemblage, in which individual taxa have relatively narrow environmental tolerances, that is then wiped out by a sharp extinction event that cuts across different marine environments. Brachiopods, conodonts, and other marine taxa are also affected. Although the extinctions were initially linked to decreases in sea level, with the diversifications occurring as sea level increased, other paleontologists have suggested temperature decreases or incursions of low-oxygen waters. At least four of these events occurred from the Middle Cambrian into the earliest Ordovician, with the interval between extinctions about 4 Ma, which is approximately half the time thought only a few years ago.

## B. Mesozoic

Many Mesozoic mass extinctions have been proposed, but most are small and not particularly well characterized. Hallam and Wignall (1997) note a late Early Triassic event which occurred 5 Ma after the end-Permian mass extinction and particularly affected ammonites, conodonts, and some bivalves; some vertebrates may also have disappeared. Interestingly, the recovery after the end-Permian mass extinction does not really begin until after this episode. Other crises occurred in the early Late Triassic (Carnian) in both marine and terrestrial realms, but these are not well characterized.

The Cenomanian/Turonian (C/T) mass extinction during the Late Cretaceous (93.4 Ma) eliminated approximately 8% of marine families and 26% of marine genera. Statistical analysis of extinction patterns from the western United States suggests the extinction occurred in many discrete steps. Benthic organisms were particularly affected, whereas cephalopods and other groups living high in the water column suffered relatively little extinction. The crisis occurred during warm, equable climates, relatively high sea level, and is associated with marine anoxia and the formation of black shales. In England, for example, the C/T boundary forms a well-known "black band" in quarries. The greenhouse climate evidently led to sluggish oceanic circulation and warm, saline bottom waters, which stratified the oceans and aided in the formation of anoxic conditions (Harries and Little, 1999).

Harries and Little (1999) noted some remarkable similarities between the C/T event and an earlier mass extinction during the early Toarcian (eTo; Early Jurassic). The extinction intensities were essentially identical, as were the environmental settings. Northwestern Europe was particularly affected by this event, although this may reflect the extent of sampling in this area. As with the C/T event, the eTo primarily eliminated benthic taxa, particularly bivalves, gastropods, and brachiopods, as well as many ammonites. In both extinctions diverse faunas were replaced by a much less diverse, stressed fauna dominated by epifaunal bivalves and ammonites.

## C. Cenozoic

The Cenozoic experienced many smaller biotic crises, each associated with climatic shifts. The causes of the most recent extinction, that of large mammals during the late Pleistocene, are also associated with human hunting, and whether climate or hunting was the primary cause is the subject of continuing debate.

The late Eocene to early Oligocene was an interval of heightened extinction both on land and in the oceans, although correlation between the two realms remains difficult. Approximately 80% of the foraminifera disappeared, with most of the extinctions coinciding with a sharp shift in the carbon cycle. Both molluscs and echinoids (sea urchins) display a prolonged period of increased extinction, evidently due to a period of global cooling, and this appears to have been the primary factor in the marine extinctions. The same is true on land, where the latest Eocene or earliest Oligocene marks the disappearance of many primitive mammal groups and the rise of modern mammals. Primates and rodents suffered considerable extinction, as did aquatic crocodiles and turtles; the terrestrial tortoises did well, however (Hallam and Wignall, 1997). Tektites, melted and then cooled terrestrial rock ejected from an impact, are well-known in the upper Eocene, leading to suggestions of possibly several extraterrestrial impacts. Re-study of this data has revealed that there is probably only a single layer of tektite material and that no increase in extinction is associated with it.

The most recent mass extinction occurred within the past 100,000 years, and some claim it continues today. Many large mammal species disappeared from Europe, North and South America, and Australia; only in Africa were they unaffected. In North America, mammoths, mastodons, camels, saber-tooth cats, ground sloths, and a host of other animals disappeared approximately 10,000–12,000 years ago. A similar pattern is found in northern Europe. The leading explanation for this extinction is the overkill hypothesis, with human hunting the most important factor. Proponents of this view note that humans became important ecological factors in each area about the time that the pace of extinction increased. Opponents have favored climatically driven extinction, although the nature of the climatic change has been continuously modified to accommodate new information. The overkill hypothesis has also been extensively criticized.

## See Also the Following Articles

CARBON CYCLE • DINOSAURS, EXTINCTION THEORIES FOR • EXTINCTION, RATES OF • EXTINCTION, CAUSES OF • MAMMALS, LATE QUATERNARY, EXTINCTIONS OF • MASS EXTINCTIONS, CONCEPT OF

## Bibliography

Bowring, S. A., Erwin, D. H., Jin, Y. G., Martin, M. W., Davidek, K., and Wang, W. (1998). U/Pb zircon geochronology and tempo of the end-Permian mass extinction. *Science* 280, 1039.

- Copper, P. (1998). Evaluating the Frasnian–Famennian mass extinction: comparing brachiopod faunas. *Acta Palaeo. Pol.* **43**, 137.
- Erwin, D. H. (1993). *The Great Paleozoic Crisis*. Columbia Univ. Press, New York.
- Erwin, D. H. (1998). The end and the beginning: Recoveries from mass extinctions. *Trends. Ecol. Evol.* **13**, 344.
- Finney, S. C., Berry, W. B. N., Cooper, J. D., Ripperdan, R. L., Sweet, W. C., Jacobson, S. R., Soufiane, A., Achab, A., and Noble, P. J. (1999). Late Ordovician mass extinction: A new perspective from stratigraphic sections in central Nevada. *Geology* **27**, 215.
- Hallam, A., and Wignall, P. B. (1997). *Mass Extinctions and Their Aftermath*. Oxford Univ. Press, Oxford.
- Harries, P. J., and Little, C. T. S. (1999). The early Toarcian (Early Jurassic) and the Cenomanian–Turonian (Late Cretaceous) mass extinctions: Similarities and contrasts. *Palaeogr. Palaeoclimatol. Palaeoecol.* **154**, 39–66.
- Jablonski, D., and Raup, D. M. (1995). Selectivity of end-Cretaceous marine bivalve extinctions. *Science* **268**, 389.
- MacLeod, N., Rawson, P. F., Forey, P. L., Banner, F. T., Boudagher-Fadel, M. K., Bown, P. R., Burnett, J. A., Chambers, P., Culver, S., Evans, S. E., Jeffery, C., Kaminski, M. A., Lord, A. R., Milner, A. C., Milner, A. R., Morris, N., Owen, E., Rosen, B. R., Smith, A. B., Taylor, P. D., Urquhart, E., and Young, J. R. (1997). The Cretaceous–Tertiary biotic transition. *J. Geol. Soc. London* **154**, 265.
- Marshall, C. R., and Ward, P. D. (1996). Sudden and gradual molluscan extinctions in the latest Cretaceous of western European Tethys. *Science* **274**, 1360.
- Marzoli, A., Renne, P. R., Piccirillo, E. M., Ernesto, M., Bellieni, G., and De Min, A. (1999). Extensive 200-million-year-old continental flood basalts of the Central Atlantic magmatic province. *Science* **284**, 616.
- McGhee, G. R. Jr., (1996). *The Late Devonian Mass Extinction*. Columbia Univ. Press, New York.
- Norris, R. D., Huber, B. T., and Self-Trail, J. (1999). Synchronicity of the K–T oceanic mass extinction and meteoritic impact: Blake Nose, western North Atlantic. *Geology* **27**, 419.
- Sepkoski, J. J. Jr., (1984). A kinetic model of Phanerozoic taxonomic diversity. III. Post-Paleozoic families and mass extinction. *Paleobiology* **10**, 246–267.
- Sepkoski, J. J. Jr., (1996). Patterns of Phanerozoic extinction: A perspective from global data bases. In *Global Events and Event Stratigraphy* (O. H. Walliser, Ed.), pp. 35–51. Springer-Verlag, Berlin.



# MEASUREMENT AND ANALYSIS OF BIODIVERSITY

Wade Leitner and Will R. Turner  
*The University of Arizona*

---

- I. Introduction
  - II. Definition of Symbols
  - III. Theoretical Properties of Richness Estimators
  - IV. Practical Sampling Considerations
  - V. Estimators Based on Sampling Theory
  - VI. Estimators Based on Extrapolation
  - VII. Which Estimation Method to Use?
- 

the number of species in an area and/or the distribution of their abundances.

**species richness** The number of species present in a region.

**temporal heterogeneity** In community sampling, variation in species capture probabilities over time; may arise from temporal environmental variation, migration, speciation, or other factors.

---

## GLOSSARY

**abundance** The number of individuals of a given species in a region.

**estimator** A statistic calculated from data to estimate the value of a parameter.

**incidence** The number of samples containing at least one of a given species in a census.

**sampling scope** The area, interval of time, and taxonomic grouping over which sampling takes place.

**spatial heterogeneity** In community sampling, variation in species capture probabilities across space; may be due to habitat variation, intraspecific clumping, interspecific association, or other factors.

**species abundance distribution** The number of species found in each interval of abundance in a community.

**species accumulation curve** A plot of the total number of species observed in a census against some measure of cumulative sampling effort.

**species diversity** Any of many measures concerning

*THE MEASUREMENT AND ANALYSIS OF BIODIVERSITY* encompasses sampling strategies and statistical estimation methods for assessing the diversity of species in biological communities and provides the basis for comparison of these measurements across time and space.

## I. INTRODUCTION

### A. Why Measure Diversity?

Biological diversity pervades every aspect of community ecology. As the key distinguishing feature of communities, it forms the basis of many ecological studies. Many studies have used a wide range of techniques to quantify diversity. These have included studies focused on patterns of species number in both time and space. Projects with more mechanistic aims have employed diversity measures to draw inferences about processes influenc-

ing diversity (e.g., migration, habitat selection, and competition at ecological timescales; and speciation, colonization, and extinction over evolutionary scales). The measurement of diversity applies to studies of the potential consequences of diversity (e.g., stability and resistance to invasion). In addition, diversity measurement plays a critical role in the study of human impacts on biological systems. Its uses in conservation include estimation of extinction rates due to habitat loss, development of conservation strategies, indication of effect due to disturbance, and use as a barometer of ecosystem status. These and other fields of inquiry require adequate methods for quantifying species diversity.

## B. Measures of Diversity

Diversity measures are classified into two broad categories: species richness (SR) and shape of the species abundance distribution. The first simply reflects the number of species present in an area. The second effectively measures the probability distribution of population sizes of the species in an area. Evenness and equitability measures belong to the second category. Evenness is highest when all abundances have equal probability in a certain range, whereas equitability is highest when all species have the same population size. Depending on the question under investigation, both categories of diversity measures may be used to document patterns. Many hypotheses concerning patterns in diversity use the relatively simple concept of SR to make concise statements connecting pattern and process. Although richness simply refers to the number of species, it encompasses many sophisticated issues in sampling design and estimation. For these reasons, most studies use species richness. We restrict our focus to species richness as well.

## C. The Problem

Species richness, the number of species present, is conceptually the most straightforward of diversity measures. Measuring richness, however, is not so simple. The number of species observed in a sample,  $S_{\text{obs}}$ , will always tend to underestimate the true species richness because we lack the resources for exhaustive sampling of communities. Even if we had such resources, properties of community structure may change during the course of sampling. This problem is exacerbated by the fact that many biological communities have many rare species, which are unlikely to be detected by sampling efforts. These issues are not unique to ecology.

The problem of determining the number of classes of objects in a collection has a long history. Many different

disciplines have sought ways to estimate the number of undiscovered classes from the classes already observed. For example, how many dies were used to mint a collection of ancient coins? How many undiscovered bugs remain in a large computer program? No one method has been successful with all such problems because we often find that we have the most information where it is the least useful. That is, a few classes account for most of the observations, whereas a few observations are scattered over most of the classes. Therefore, we wind up with small sample sizes for the rare observations where we need to characterize variation the most. In the most extreme case, imagine a community in which every species is represented by just one individual. Then, for any sample size below the true number of species we shall always observe fewer than the true number of species. Of course, in this case we could observe that the number of species was a linear function of the number of individuals. This might suggest to us a way to estimate the number of undetected species if we knew how many individuals were present in the study area. This example illustrates the strategy followed by richness estimators: Model the regularity in the behavior of the number of species detected as a function of sample size. Then, use this model to predict the number of species when the sample size becomes large. The regularity could involve a clear relationship between the average sample size and the average number of species observed; this is the basis of extrapolation-based estimators. On the other hand, the regularity could mean that we can replace a complicated sampling process with a more tractable model; estimators based on sampling theory follow this approach.

## D. Article Overview

An exhaustive review of all methods used for estimations of species richness is beyond the scope of this article. Therefore, this article reviews the most widely used methods for the estimation of species richness. Section II defines the symbols used in this article. Section III lists some theoretical properties of richness estimators. Section IV discusses practical considerations of using community sampling to estimate species richness. The next two sections present the estimators: richness estimators based on fine-scale, theoretical models of sampling are detailed in Section V, whereas those based on coarse-scale, global modeling of species accumulation (extrapolation techniques) are discussed in Section VI. Section VII addresses the practical problem of evaluating and selecting estimators for use on new data sets.

TABLE I  
Symbols Used

$s$	Number of species in a community
$n$	Number of individuals in a community
$x_i$	Number of individuals of species $i$ in a community
$r_i$	Number of species with $i$ individuals in a community
$S_{\text{obs}}$	Number of species in a census
$S_{\text{est}}$	Estimate of the number of species in a community using given method
$N$	Number of individuals in a census
$C$	The sample coverage of the census
$t$	Number of samples in a census
$X_i$	Number of individuals of species $i$ in a census
$X_{i,j}$	Number of individuals of species $i$ in sample $j$
$Y_i$	Number of samples containing species $i$ in a census
$Y_{i,j}$	The presence (1) or absence (0) of species $i$ in sample $j$
$R_i$	Number of species containing exactly $i$ individuals in a census
$F_i$	Number of species occurring in exactly $i$ samples in a census
$P_i$	Probability of detecting species $i$ in a sample
$q_i$	Probability that an individual is of species $i$
$I(\omega \in A)$	1 if $\omega \in A$ and 0 otherwise
$E[X]$	The expectation of the random variable $X$
$\text{VAR}[X]$	The variance of the random variable $X$
$\text{COV}[X, Y]$	The covariance of the random variables $X$ and $Y$
$\text{CORR}[X, Y]$	The correlation of the random variables $X$ and $Y$
$\gamma(Z)$	Coefficient of variation of the random variable $Z$

## II. DEFINITION OF SYMBOLS

We adopt the convention that random variables are written as uppercase symbols, whereas lowercase is reserved for deterministic variables or fixed constants. The definitions we shall need are given in Table I.

## III. THEORETICAL PROPERTIES OF RICHNESS ESTIMATORS

Salient features of richness estimators include the following:

**Type of data:** Estimators differ in whether they use incidence or abundance data. The incidence of species  $i$  is the number of samples in which  $i$  occurs. Formally,  $Y_i = \sum_{j=1}^t Y_{i,j}$ . The abundance of species  $i$  is the number

of individuals of  $i$  contained in the census. Formally,  $X_i = \sum_{j=1}^t X_{i,j}$ .

**Sampling assumptions:** Many estimators make assumptions about the sampling process (e.g., invariance of capture probabilities across all samples in a census).

**Parametric/nonparametric:** The estimator  $S_{\text{est}} = \psi(X)$  is nonparametric if the function  $\psi$  does not depend on a given distribution of  $X$ .

**Bias:** Bias measures average deviation from the true value. We define bias  $\equiv E[\psi(X)] - s$ . An estimator of  $s$  is unbiased in the strict statistical sense if  $E[\psi(X)] - s = 0$  for every  $s$ .

**Variance:** An estimator is a random variable. Therefore, we can use variance as a measure of the uncertainty in an estimate.

**Sufficiency:** Let  $w = \psi(x)$  be an estimator of  $s$ . The statistic  $W = \psi(X)$  is sufficient for  $s$  if  $P\{X \in A|W, s\} = P\{X \in A|W\}$ . Suppose  $S_{\text{est}} = \psi(X)$  is sufficient for  $s$ . Then there are no other estimators (other than functions of  $\psi$ ) that we could compute using  $X$  that could increase or decrease our confidence about our estimate of  $s$ .

## IV. PRACTICAL SAMPLING CONSIDERATIONS

### A. Limiting the Influence of Sampling Biases

We lack the resources for complete censusing of communities. Thus, sampling is the window through which we view the ecological world. The species we observe when we sample are determined both by underlying ecological processes and by biases associated with sampling. The simplest of these is bias due to sample size. If we plot the number of species observed in a census against some measure of sampling effort, such as individuals counted, we get a species accumulation curve as shown in Fig. 1. Sample size bias accounts for much of the increase in a species accumulation curve with increasing sampling effort. This bias may be reduced by increasing sample size or by estimating the number of species actually present from incomplete census data. An increase in the accumulation curve also arises from sampling over different habitats and at different times. Spatial and temporal heterogeneity are difficult to quantify, as are their effects on the performance of richness estimators. It may be possible, however, to reduce such effects through careful sampling design. As a general rule, one should minimize sources of variation in capture probabilities over sampling. Where possible, the spatial and temporal scale of a census (see Section IV,

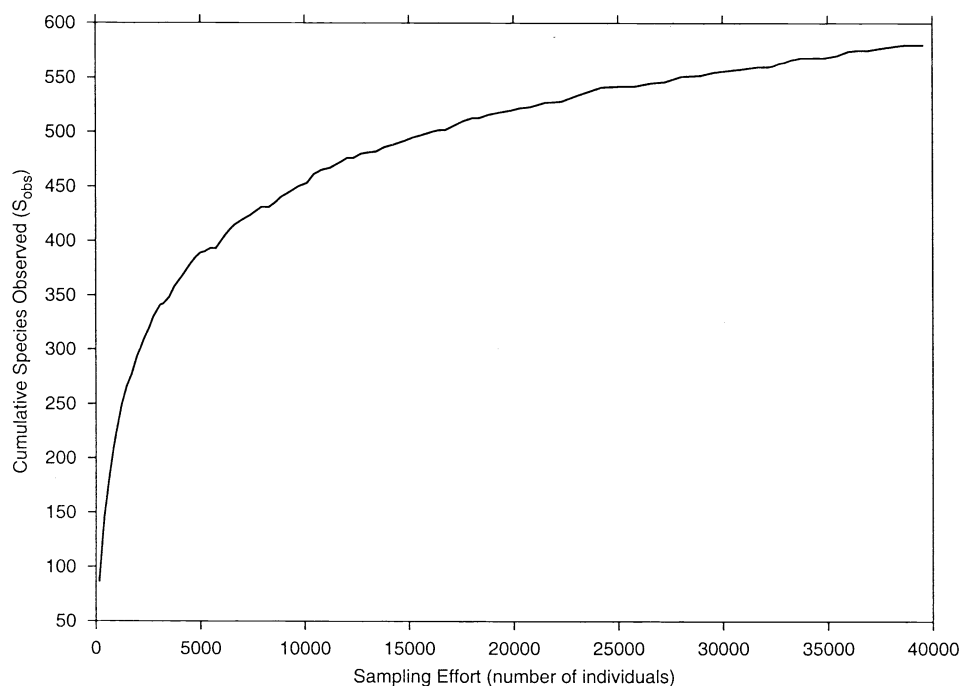


FIGURE 1 The species accumulation curve for a simulated data set. Note the nonlinear trend in the increase of species observed with increasing sample size.

B) should be small relative to the scale of the processes generating diversity.

### B. Sampling Scope

Any sampling effort can be described as taking place within some area, over an interval of time, and involving a particular taxon or group of taxa. Even within the area sampled, some habitats may not be sampled. Likewise, within an interval of time, some periods are often not considered (e.g., sampling that takes place over a week but never at night). We define sampling scope as the area (and habitats) sampled, the interval of time (and periods within that interval) over which sampling occurred, and the group of taxa sampled. The consideration of sampling scope is important because it means that the sampling design must be tailored to the question of interest.

The number of species predicted by richness estimators reflects the scope of sampling. In general, these methods estimate only the number of species in the taxa, area, and time interval sampled.<sup>1</sup> For example,

<sup>1</sup> There are exceptions. Methods based on extrapolation of accumulation curves, in particular, might be used to estimate species richness at larger spatial (and temporal) scales than that sampled. There are many scaling considerations to be made that are unique to this type of extrapolation.

estimates from ground-level plant surveys do not predict epiphyte diversity. Similarly, daytime sampling does not predict nocturnal animal species, estimates based on beetle abundance do not predict ant diversity, summer bird diversity does not inform us of winter migrants, and so on.

### C. Abundance and Incidence

With quadrat-based sampling, the area to be sampled is partitioned into many subsamples (quadrats) of equal size. A subset of  $t$  of these quadrats is then sampled using capture or observation techniques appropriate for the taxon of interest. The data recorded may take the form of either species incidences  $Y_{i,j}$  or abundances  $X_{i,j}$ .

Abundance data may not be possible to collect in all cases. For some taxa (e.g., some fungi), individuals are not readily recognizable as such. Additionally, some sampling methods may not allow abundances to be recorded. For example, point counts based on bird vocalizations are a practical method of recording species in bird communities but do not allow accurate counting of individuals. Even when abundances can be observed, obtaining abundances may require more effort than obtaining incidences alone. Nonetheless, a wider range of estimators may be used if abundance information is available. Thus, despite the greater effort involved in

collecting abundance data, it may be useful to record species abundances so that abundance-based estimators may be used as well.

## D. Other Sampling Issues: Surrogates for Quadrats and Species

Two other sampling issues bear mention here. First, it is important to note that the use of areal units (quadrats) for sampling units is not required. Indeed, although quadrats are often appropriate for plants and other sessile organisms, they are less so for mobile organisms. For mobile taxa, it is useful to substitute a more practical measure—such as observer-hours, trap-days, or volume of substrate—as a sampling unit. Additionally, all is not lost if sampling units are of unequal sizes. Equal-sized samples will meet the assumptions for a larger number of estimators. However, some estimators exist which may be applied when it is not possible to keep sample sizes approximately constant over a census.

Second, techniques for estimating species richness are potentially applicable to taxonomic levels other than species. For any of a variety of reasons, it is often impractical or impossible to key all organisms censused to the species level. For example, in studies of paleontological data, it may only be possible to place an individual to the generic level. In other cases, the presence of undescribed species or other lack of knowledge of species status may necessitate the use of a morphospecies concept. In any case, it is important to note that richness estimators will estimate the number of classes (genera and morphospecies in the previous examples) at whatever taxonomic level was used in sampling.

## V. ESTIMATORS BASED ON SAMPLING THEORY

Species richness estimators are all alike in the obvious way: They all estimate species richness. However, SR estimators take different approaches to estimation. This section presents approaches involving fine-scale, detailed models of the sampling process. Section VI discusses extrapolative methods based on coarse-scale, global models of species accumulation.

### A. Models of the Sampling Process: General Concepts

For the fixed collection from which we sample,  $(r_0, \dots, r_n)$  specifies the distribution of abundances among

the species. Sampling introduces randomness into our data so that when we census the area, we record  $(X_0, \dots, X_{S_{\text{obs}}})$  from which we compute  $(R_0, \dots, R_N)$ . When we sample a set of  $t$  quadrats, we may keep track of the abundance of the  $i$ th species within the  $j$ th quadrat,  $X_{i,j}$ . Then, we compute the abundance of species  $i$  by  $X_i = \sum_{j=1}^t X_{i,j}$ . However, in many cases, we may be unable to distinguish individuals. In these cases, we record  $Y_{i,j}$  from which we compute the incidence of species  $i$  by  $Y_i = \sum_{j=1}^t Y_{i,j}$ . The distribution of  $X_{i,j}$  depends on how we sample individuals.

Suppose we sample from a region (an urn) containing  $n$  individuals (marbles) distributed among  $s$  species (colors). Write  $x_i$  for the number of individuals contained in species  $i$ . We collect  $N$  individuals out of  $n$  and  $S_{\text{obs}}$  species out of  $s$ . If we sample with replacement, we find that the number of observations of species  $i$ , written  $X_i$ , has probability

$$P\{X_i = a\} = \binom{N}{a} \left(\frac{x_i}{n}\right)^a \left(1 - \frac{x_i}{n}\right)^{N-a} \quad (1)$$

If we sample without replacement

$$P\{X_i = a\} = \frac{\binom{x_i}{a} \binom{n-x_i}{N-a}}{\binom{n}{N}} \quad (2)$$

Under certain conditions, both of these distributions can be approximated by the Poisson distribution according to

$$P\{X_i = a\} = \frac{(N\lambda_i)^a}{a!} e^{-N\lambda_i} \quad (3)$$

where  $\lambda_i = x_i/n$ . Strictly speaking, the approximation of Eq. (1) is exact only for communities containing infinite species richness and infinite abundance. Practically speaking, the Poisson approximation should work best on large, species-rich communities. A similar argument holds for the approximation of Eq. (2). The requirement that  $n$  be considered infinite is sufficient to get to Eq. (3), but it is not necessary. For example, if the successive times between discoveries of an individual of species  $i$  are exponentially distributed and the number of individuals found in disjoint samples are independent, then we will get Eq. (3). There is a natural way in which this can happen. First, detections of individuals must be independent events. Second, once we know the current value of  $X_i$ , the distribution of future values



$X_i$  must be completely determined. Then, we can expect that  $X_i$  can be modeled with a Poisson process.

## B. Species Observed ( $S_{\text{obs}}$ )

As a starting point, we consider the number of observed species,  $S_{\text{obs}}$ , as our first estimate of  $s$ . Our presentation of  $S_{\text{obs}}$  serves as the template for the more complicated estimators to follow. First, we consider the type of data we record. Then, we state the assumptions made in deriving the estimator and its properties. We next turn to theoretical properties of the estimator, for example, its bias and variance. Finally, we discuss issues related to the application of the estimator to real data.

Estimators may be applied to two types of data: species abundance and species incidence. If we record abundance, denoted  $X_{i,j}$ , our data set will be a matrix of the number of individuals of species  $i$  captured on sampling occasion  $j$ .

We define the abundance statistics,  $R_i$ , to be the number of species containing exactly  $i$  individuals. Then,

$$R_i = \sum_{j=1}^s I(X_j = i) \quad (4)$$

We define the incidence statistics,  $F_i$ , to be the number of species occurring in exactly  $i$  samples in a census. Then,

$$F_i = \sum_{k=1}^s I(Y_k = i) \quad (5)$$

When we wish to refer to the vector of  $R_i$  or  $F_i$  values for a given data set we will write  $\mathbf{R}$  and  $\mathbf{F}$ , respectively. The number of species observed in a census  $S_{\text{obs}}$  may be written in terms of either the abundance or the incidence statistics:  $S_{\text{obs}} = \sum_{i=1}^N R_i = \sum_{i=1}^t F_i$ . In some cases, we use only the lower order statistics to correct estimator bias. We can visualize the difficulties inherent in estimator design by considering how information accumulates over the  $R_i$ . Let the number of species having fewer than  $k$  individuals be  $S_k$ . Similarly,  $N_k$  is the number of individuals representing these  $S_k$  species. Therefore,

$$S_k = \sum_{i=1}^k R_i \quad (6)$$

$$N_k = \sum_{i=1}^k iR_i \quad (7)$$

Figure 2 shows a plot of  $N_k$  and  $S_k$  vs  $k$  for real data collected on mist-netted birds. In Fig. 3, we plot  $N_k$  as a function of  $S_k$ . This plot reveals an apparently exponential relationship between  $N_k$  and  $S_k$ . These figures show that information may be distributed very unevenly in the data set. This means that we shall have few data points (e.g.,  $N_1$  and  $N_2$ ) where we wish to concentrate our estimation effort ( $S_1$  and  $S_2$ ). This is the worst possible distribution of information if we want to estimate the number of rare species from a simple transformation of the data. For example, an estimator using only  $E[N]$  and  $E[S]$  will probably do poorly for many real data sets.

Suppose we model sampling in terms of the probability of detection of individuals of each species,  $\mathbf{q} = (q_1, \dots, q_s)$ , and the number of individuals observed,  $N$ . We assume that each individual of a species has the same capture probability throughout the entire census. Thus,

$$E[S_{\text{obs}} | (q_1, \dots, q_s)] = s - \sum_{i=1}^s (1 - q_i)^N \quad (8)$$

We record the incidence,  $Y_{i,j}$ , of species  $i$  during sample  $j$  giving a data matrix whose entries are either 0, for no detection, or 1 when any individuals of  $i$  get detected on occasion  $j$ . For an entire census, the number of detections of species  $i$  is  $Y_i = \sum_{j=1}^t Y_{i,j}$ . Then we may define  $\mathbf{p} = (p_1, \dots, p_s)$  to be the vector of species detection probabilities. After  $t$  samples, we compute

$$E[S_{\text{obs}} | (p_1, \dots, p_s)] = s - \sum_{i=1}^s (1 - p_i)^t \quad (9)$$

Equations (8) and (9) show that  $S_{\text{obs}}$  is biased over finite sampling. In the following, consideration of incidence and abundance versions of  $S_{\text{obs}}$  parallel on another. Therefore, we focus on the incidence data.

As we increase sample size we have for any  $i$ ,  $\lim_{t \rightarrow \infty} P\{Y_i > 0 | \mathbf{q}\} = 1$ . However, in practical terms, we would like to know just how fast  $S_{\text{obs}}$  gets close to  $s$ . Suppose that the  $p_i$  are independent and identically distributed (iid) with distribution  $H(p)$  and probability density  $h(p)$  and consider  $E[s - S_{\text{obs}}]/s$ . The relative rate of convergence of the mean error is given by this expectation taken over all  $\mathbf{p}$ :

$$\frac{s - E_t[S_{\text{obs}}]}{s} = \int_0^1 (1 - p)^t h(p) dp \leq (1 - E[p])^t \quad (10)$$

When we have few rare species,  $E[p]$  will be close to 1 and  $S_{\text{obs}}/s$  will converge quickly to 1. Note that in absolute terms if  $s$  is large we may still have many rare,

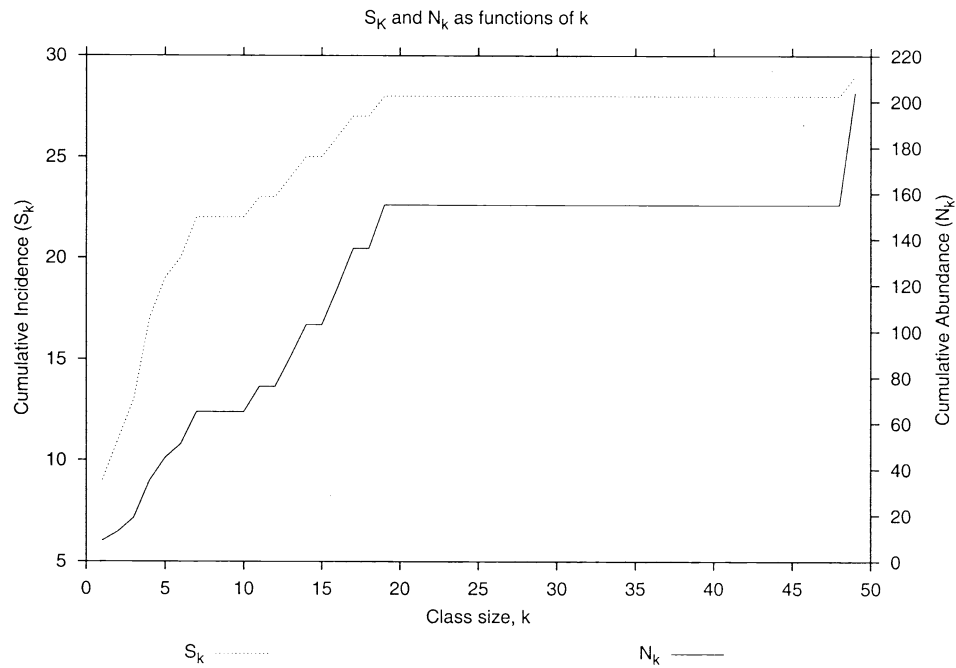


FIGURE 2 The accumulation of species and individuals with increasing abundance class size for a bird community in southeastern Arizona (W. Leitner, unpublished data). Note that  $N_k$  changes relatively slowly where  $S_k$  changes quickly and vice versa.

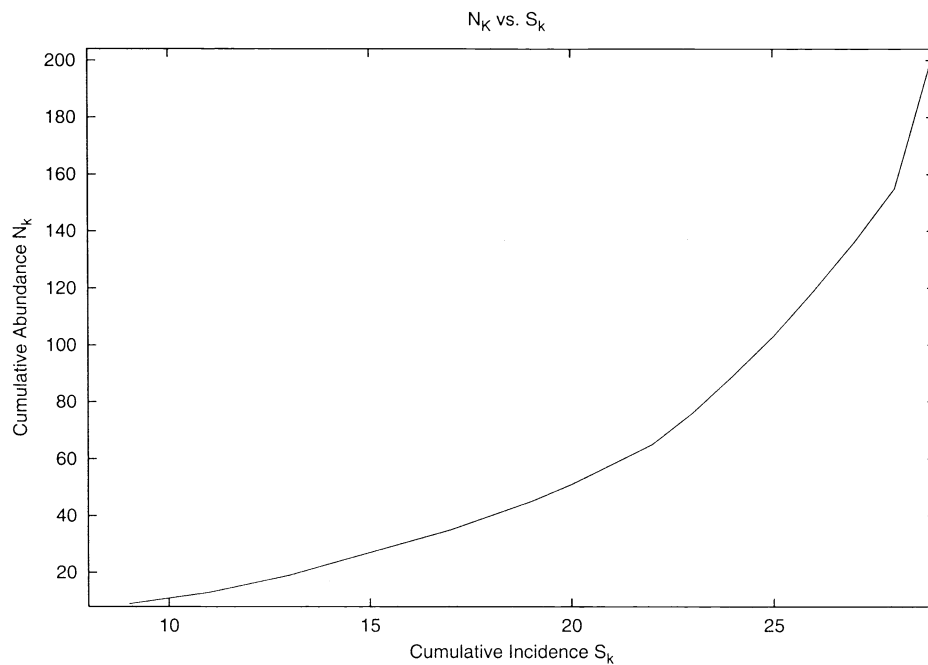


FIGURE 3 The accumulation of individuals as a function of accumulated incidence for a bird community in southeastern Arizona (W. Leitner, unpublished data). Note that  $N_k$  is nearly exponential with  $S_k$ . We shall have few data points (e.g.,  $N_1$  and  $N_2$ ) at which we wish to concentrate our estimation effort ( $S_1$  and  $S_2$ ).

undetected species. When we have a large proportion of rare species, the convergence will of course be much slower. Now,  $S_{\text{obs}}$  is a random variable and the variance in  $S_{\text{obs}}$  is

$$\text{VAR}(S_{\text{obs}}) = s \int_0^1 ((1-p)^t - (1-p)^{2t})h(p)dp \quad (11)$$

The root mean square error will then decrease approximately like the square root of the error in the mean. Perhaps one of the indices of diversity that measure evenness could be used to sharpen confidence intervals in some way. Before undertaking the design of such a method, it would be beneficial to know if any function of  $F$  could be used to increase our confidence in our estimate of  $E[S_{\text{obs}}]$ . Burnham and Overton (1978) show that  $(F_1, \dots, F_t)$  are sufficient for  $\mathbf{p}$  under the assumption of iid  $p_i$ . Now, the distribution of  $S_{\text{obs}}$  is determined by the distribution of  $\mathbf{p}$  because  $S_{\text{obs}} = \sum_{i=1}^t F_i$ . Therefore, to increase our confidence in our estimator we should focus directly on reducing its variance. Thus, our goals in estimating SR should include getting estimators that decrease the bias faster than  $(1 - E[p])^t$  and whose standard errors decrease faster than  $\sqrt{(1 - E[p])^t}$ . Often, these goals will conflict with one another. However, knowing that the bias decreases like  $(1 - E[p])^t$  tells us that for large enough  $t$ , the bias decreases faster than  $t^{-k}$ ;  $k = 1, 2, \dots$ . Resampling methods can use this fact to accelerate the rate of bias reduction.

### C. Resampling Methods ( $S_{\text{jk}}$ , $S_{\text{B}}$ )

We can very rarely expect

$$E[f(X)] = f(E[X]) \quad (12)$$

When  $f(x)$  is linear in  $x$  then Eq. (12) holds. Suppose that  $f(x)$  is nonlinear. Then as  $x$  gets spread away from  $x = E[X]$  the function  $f$  can over- or underweigh the contribution of  $X$  to the mean. Thus, both the nature of  $f$  and the distribution of  $X$  contribute to bias. For this reason, outliers (points far from  $E[X]$ ) can severely alter the approximation  $E[f(X)] \approx f(E[X])$ . Most estimators derive from just such an approximation. Therefore, when we have outliers, we can sometimes improve our estimate by deleting them from the data set. A version of this approach applied to data not so clearly identified as extreme is the intuitive basis for the jackknife and bootstrap estimators. If the order in which data are recorded does not matter, then we should expect that the relationship between the distribution of

$t$  and  $t - 1$  data points should inform us about the relationship between the distribution of  $t$  and  $t + 1$  data points. Thus, the common property that underlies both techniques is the exchangeability of the data. There are two resampling approaches we consider: jackknifing and bootstrapping. In each case, we describe the resampling strategy that gives rise to the estimator. However, in each case it turns out that actual resampling need not be done. The  $F$  statistics contain the same information we would obtain from resampling.

#### 1. Jackknife Estimator ( $S_{\text{jk}}$ )

Jackknifing involves computing the average value of a statistic on a reduced data set. One removes each combination of  $k$  data points in a data set and computes the statistic of interest to give a set of new pseudostatistics. Then, by taking a suitably weighted average of these statistics we get a reduced-bias version of the original statistic. The number of data points removed at a time gives the order of the jackknife. The most obvious statistic to which we can apply the jackknife is the number of observed species,  $S_{\text{obs}}$ . As discussed previously,  $S_{\text{obs}}$  is a biased estimator of  $s$ . To apply the jackknife technique to  $S_{\text{obs}}$  we need three assumptions:

1. The capture probabilities may vary across species,
2. The capture probabilities do not change during the census,
3. The bias in  $S_{\text{obs}}$  decreases at least as fast as  $1/t$ .

Under these assumptions, Burnham and Overton (1978) propose a generalized jackknife estimator for species richness, simplified by Smith and Van Belle (1984):

$$S_{\text{jk}} = S_{\text{obs}} - \frac{1}{k!} \sum_{j=1}^k F_j \sum_{i=j}^k (-1)^i \binom{k}{i} \frac{\binom{t-j}{i-j}}{\binom{t}{i}} (t-i)^k \quad (13)$$

where  $k$  gives the order of the estimator.  $S_{\text{jk}}$  does not require that we actually do the resampling. To better understand the properties of  $S_{\text{jk}}$ , we consider how the first-order jackknife estimate of  $S_{\text{obs}}$ , written  $S_{\text{J1}}$ , would be produced.

Suppose we compute a statistic,  $D(X_1, \dots, X_t) = D_t$  from  $t$  observations of  $X$ . Now, we want to consider the effect of each observation in turn by removing it from the data set. However, we do not want the order in which we do the removal to matter. Within a given sampling occasion, there may be time of day or other

local differences. For example, during bird banding operations we may consistently catch flycatchers later in the day than warblers because it may take time for flying prey to become active. Then, resampling within a trapping occasion may introduce heterogeneity. Thus, we shall conduct jackknife resampling by adding or removing entire quadrats.

Let us remove the  $i$ th quadrat and compute  $D$  without this quadrat. Call this value  $D_{(i)}$ . Do this for each  $i$ . The mean of these  $D_{(i)}$ ,

$$D_{t-1} = \frac{\sum_{i=1}^t D_{(i)}}{t} \quad (14)$$

gives rise to the first-order jackknife estimate

$$D_{j1} = tD_t - (t-1)D_{t-1} \quad (15)$$

Now,  $D_{j1}$  will be less biased than  $D_t$  if

$$E[D_t] = s + \sum_{i=1}^{\infty} \frac{b_i}{t^i} \quad (16)$$

where the  $b_i$  do not depend on  $t$ . If the bias of  $D_t$  decreases like  $1/t$ —that is,  $b_1 \neq 0$ —then the bias in  $D_{j1}$  decreases like  $1/t^2$ .

Suppose  $D = S_{\text{obs}}$ . If assumption 3 holds, then we can easily get

$$S_{j1} = S_{\text{obs}} + \frac{t-1}{t} F_1 \quad (17)$$

The higher order estimates are considerably more complicated but depend on the assumptions in the same way. As we increase the order of the jackknife, we get greater bias reduction. Unfortunately, we can expect that removing data to reduce bias might increase estimator variance. We see from Eq. (13) that  $S_{jk}$  is a linear combination of the  $F_i$  with constant coefficients,  $a_{i,k}$ , given  $t$  and  $k$ . Burnham and Overton (1978) give the unconditional variance as

$$\text{VAR}(S_{jk}) = \sum_{i=1}^t a_{i,k}^2 E[F_i] - \frac{E[S_{jk}]^2}{s} \quad (18)$$

Note that as we increase the order from  $k$  to  $k+1$  the coefficient of  $F_{k+1}$  becomes nonzero and therefore increases the variance. Further more, a quick check of the lower order coefficients reveals that they also increase with  $k$ . Therefore, as we expected, the variance increases with  $k$ . Thus, we tradeoff bias reduction

against increased noise. Two methods have been proposed for choosing the order of the jackknife. The first method tests the hypothesis that  $E[S_{jk} - S_{j,k+1}] = 0$  using the test statistic

$$T_i = \frac{S_{j,k+1} - S_{jk}}{\sqrt{\text{VAR}(S_{j,k+1} - S_{jk}|D)}} \quad (19)$$

If we can reject this hypothesis for  $k=1$  then we proceed to test successive values of  $k$  until we fail to reject. The difficulty with this test is that quadrat information comes in discontinuous jumps so that the order used in estimation may jump around during the course of a census. Burnham and Overton (1978) suggest an interpolation technique that smoothes this behavior.

The power of the jackknife technique is that it requires very little knowledge about the distribution of the data. It is thus nonparametric. An argument like the one presented previously on species observed tells us that assumption 3 is reasonable for finite collections of species. The first two assumptions, however, require biological information. For example, suppose we sample a site containing migratory animals. Then, when we sample determines, to a large extent the distribution of number of individuals observed. If we sample across a productivity gradient then we would not expect the distribution of a given species' abundance,  $(X_{1,1}, \dots, X_{1,t})$ , to be symmetric. That is, the quadrats would not be exchangeable. This changes the importance of each quadrat on removal, making some have a deterministically larger effect. Consequently, the effective rate of bias reduction may become much smaller than  $1/t^2$ . Nonetheless, with attention to census design, such issues can be minimized. If in addition we assume that the  $p_i$  are iid random variables, then Burnham and Overton (1978) show that  $F$  is sufficient for the distribution of the capture histories. This result does not depend on the resampling protocol. This assertion will apply to some of the other estimators we examine.

## 2. Bootstrap Estimator ( $S_B$ )

Bootstrap estimation begins by constructing a surrogate data set by sampling the original data set with replacement. We shall bootstrap to remove bias in  $S_{\text{obs}}$ , so we shall once again focus on incidence-based data. Smith and van Belle (1984) analyzed the bootstrapping process to derive

$$S_B = S_{\text{obs}} + \sum_{i=1}^{S_{\text{obs}}} \left(1 - \frac{Y_i}{t}\right)^t \quad (20)$$

Just as in the jackknife estimator, we analyze the resampling process to get the estimate that would result if we actually did the resampling. We would like to think of resampling as a surrogate for doing multiple replicate censuses. Let  $s = \text{bias} + E[S_{\text{obs}}]$ . If we knew the bias and  $E[S_{\text{obs}}]$  we would be done. To estimate the bias over all possible sampling outcomes, we would need to know the distribution of the  $X_i$ . Let  $S_{\text{obs}}$  play the role of  $s$ , and let  $(X_1, \dots, X_{S_{\text{obs}}})$  approximate  $(x_1, \dots, x_s)$ . Then, if we resample the  $X_i$  multiple times, we can estimate  $S_{\text{obs}}^*$ , the mean number of species found in the replicates. Then  $\text{bias}_r = S_{\text{obs}} - S_{\text{obs}}^*$  is an estimate of  $s - E[S_{\text{obs}}]$ , the true bias. Now, we have a single point estimate of  $E[S_{\text{obs}}]$ , namely,  $S_{\text{obs}}$ . Therefore, we estimate  $s$  by  $S_B = S_{\text{obs}} + \text{bias}_r$ . We just need to compute  $\text{bias}_r$ . We assume that

1. Detection probabilities,  $p_i$ , may vary between species,
2. The  $p_i$  remain fixed during sampling.

Then, on resampling we find that

$$\text{bias}_r = \sum_{i=1}^{S_{\text{obs}}} \left(1 - \frac{Y_i}{t}\right)^t \quad (21)$$

which leads us to Eq. (20).

The variance in  $S_B$  given by Smith and Van Belle (1984) is

$$\begin{aligned} \text{VAR}(S_B) = & \sum_{i=1}^{S_{\text{obs}}} \left(1 - \frac{Y_i}{t}\right)^t \left(1 - \left(1 - \frac{Y_i}{t}\right)^t\right) \\ & + 2 \sum_{i=1}^t \sum_{j=i}^t \left( \left(\frac{Z_{ij}}{t}\right)^t - \left(1 - \frac{Y_i}{t}\right)^t \left(1 - \frac{Y_j}{t}\right)^t \right) \end{aligned} \quad (22)$$

where  $Z_{ij}$  are the number of quadrats that jointly lack species  $i$  and  $j$ . For large  $t$  we can replace  $(1 - Y_i/t)t$  by  $e^{-Y_i}$ . Then, as  $t$  gets large the variance in the estimator goes to zero as long as the tendency of species to co-occur decreases sufficiently fast as  $t$  increases. Therefore, species that interact strongly, for example, core members of a mixed foraging flock, will tend to inflate estimator variance.

#### D. Moment Estimators ( $\alpha$ , $S_{\text{AM}}$ , $S_{\text{IM}}$ )

These methods all derive relationships between the moments of various distributions. The moments of a distribution, defined by  $E[X^r]$ , where  $r$  is a positive integer, are used to simplify the analysis. At a minimum, the

relationship between various moments reflects assumptions about the sampling process. In other cases, we require special assumptions about the abundance distribution.

##### 1. Fisher's $\alpha$

Fisher's  $\alpha$  is intermediate between species richness and species diversity. It measures species richness but does so in terms of the parameters of the abundance distribution and the sampling process. It is a parametric estimator of richness when applied as described later. Unfortunately,  $\alpha$  involves some rather technical arguments. This has led to two different ways to define  $\alpha$ . The first follows from Fisher's use of the gamma distribution to model the abundance distribution. The second results from a limit taken on the functional relationship between the mean sample size and the mean number of observed species in a sample. As we shall see, properties that are exact for one definition of  $\alpha$  are approximations for the other. We shall focus on two key properties of  $\alpha$ : (i) sample size dependence of  $\alpha$  and (ii) estimating species richness from  $\alpha$ . We begin by defining  $\alpha$ , as Fisher did, in terms of the sampling process and the abundance distribution under study. Then, we derive the second version of  $\alpha$  from the first by considering what happens when the distribution under study represents an arbitrarily large number of species. Finally, we consider the practical problem of calculating and interpreting  $\alpha$  when we know very little about the true nature of the abundance distribution.

##### a. Derivation of $\alpha$

Fisher's  $\alpha$  was first introduced (Fisher *et al.*, 1943) as a means for estimating the parameters of empirical abundance distributions that had a negative binomial shape. Fisher made three key technical assumptions in his derivation of  $\alpha$  that show up in its properties. First, Fisher assumed that the sample was drawn from a community of infinitely many individuals. As we shall see, this allowed Fisher to use the Poisson process to model the sampling of individuals. Second, he assumed that the community was composed of species whose mean abundances were distributed according to a gamma distribution. Finally, Fisher considered the limiting case of a community containing an infinite number of species.

If we assume that the order in which we discover species is unimportant (*i.e.*, individuals are exchangeable) then  $P\{N_1 = x\}$  specifies the probability that we observe a species represented by  $x$  individuals. With  $P\{N_1 = x\}$  we can compute  $E[S_{\text{obs}}]$  and  $E[N]$  because

$$E[S_{\text{obs}}] = s \sum_{x=0}^n P\{N_1 = x\} \quad (23)$$

$$E[N] = s \sum_{x=0}^n xP\{N_1 = x\} \quad (24)$$

Fisher begins with the assumption that  $P\{N_i = x|m_i\} = (m_i^x/x!)e^{-m_i}$ , where  $m_i$  is the number of individuals of species  $i$  expected in the sample. The number of individuals in a sample,  $N$ , or the sample size, is usually a random variable because trapping and other forms of detection are random processes. Suppose that detection is a Poisson process, or very nearly so, and that our sampling region has extent  $A$ . Then, we expect that  $m_i/A = \lambda_i$  where  $\lambda_i$  is the density of species  $i$ .

Fisher's second assumption was that the  $\lambda_i$  possess the gamma distribution with parameters  $1/p$  and  $k$ . The density of a gamma random variable,  $f(\lambda, p, k)$ , is given by

$$f(\lambda, p, k) = \frac{1}{\Gamma(k)} \frac{1}{p} \left(\frac{\lambda}{p}\right)^{k-1} e^{-\lambda/p}$$

where  $\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx$ . The gamma distribution has three features which made it an obvious first choice. First, gamma random variables must be nonnegative just like species' densities. Second, the gamma distribution can take on a wide range of shapes. Lastly, computations using the gamma distribution are often tractable. To spread out densities evenly over a community,  $k$  must be close to zero.

Suppose we sample a region by censusing areas of increasing size. The larger areas should contain more individuals. Let  $\mu_A = E[N|A, s]$  and  $\nu_A = E[S_{\text{obs}}|A, s]$  give the expected number of individuals and species, respectively, in a sample of size  $A$  drawn from a community containing  $s$  species. Then,

$$P\{N = x|A, s\} = \int_0^\infty \frac{(\lambda A)^x}{x!} e^{-\lambda A} f(\lambda) d\lambda \quad (25)$$

$$= (1 + pA)^{-k} \frac{\Gamma(x+k)}{\Gamma(k)x!} \left(\frac{pA}{1+pA}\right)^x \quad (26)$$

Using Eq. (26) in Eqs. (23) and (24) it can be shown that

$$\nu_A = s(1 - (1 + pA)^{-k}) \quad (27)$$

$$\mu_A = skpA \quad (28)$$

Now, if the relative abundance of species is the same at all sampling scales it must be that the distribution's shape and thus  $k$  are fixed. Although we expect smaller sample sizes in smaller  $A$ , the number of species from which we sample remains fixed at  $s$ . This means that  $\mu_A$  can only change as a function because of  $A$  or  $p$ . Therefore, we collect all the constants from Eq. (28) into  $\alpha$  and write  $\mu_A = \alpha pA$  and set  $\alpha = sk$ . This is Fisher's definition of  $\alpha$  (with the effect of sampling extent made explicit). When defined this way,  $\alpha$  depends on only the shape of the distribution and not on the expected sample size. Using this definition for  $\alpha$ , we can rearrange Eq. (27) and substitute Eq. (28) to get

$$\left(1 - \frac{\nu_A}{s}\right)^{-s/\alpha} = 1 + \frac{\mu_A}{\alpha} \quad (29)$$

Equation (29) defines  $\mu_A$  as a function of  $\nu_A$  with parameters  $\alpha$  and  $s$ . In fact, Eq. (27) is a species-area relationship. At  $A = 0$  we expect no species and for  $A \rightarrow \infty$  we expect  $s$  species. A little calculus shows that

$$\frac{d\nu_A}{d \ln(A)} = \frac{\alpha}{1 + \frac{\alpha}{\mu_A}} \frac{s - \nu_A}{s} \quad (30)$$

Consequently, if we plot  $\ln(s - \nu_A)$  vs  $\ln(A)$  we get a line of slope  $(\alpha/s)(1/(1 + (\alpha/\mu_A)))$ . Now,  $\mu_A$  quickly exceeds  $\alpha$  in many cases at very small areas. Thus, for most sample sizes we have constant slope  $\alpha/s$ . Therefore, we may estimate  $s$  if we know  $\alpha$ . Unfortunately, this will give only a lower bound on  $s$  because the sensitivity of the slope to changes in  $s$  decreases as  $s$  increases. Nonetheless, we need a method for estimating  $\alpha$  that does not involve  $s$ . To do this, we will want to consider communities that contain many species. Again, calculus tells us that if we let  $s \rightarrow \infty$  in Eq. (29) we get

$$e^{-\nu_A/\alpha} = 1 + \frac{\mu_A}{\alpha} \quad (31)$$

The solution to Eq. (29) gives us the second way of defining  $\alpha$ . In order to avoid confusion, we write  $\alpha^*$  for the solution of Eq. (31). We do not need to assume anything about the sampling procedure or the abundance distribution, except that  $\mu_A$  and  $\nu_A$  exist, in order to compute  $\alpha^*$ . Note that  $\alpha \rightarrow \alpha^*$  as  $s \rightarrow \infty$ . Unlike  $\alpha$ ,  $\alpha^*$  may depend on sample size. This can be seen from

$$\frac{\partial \alpha}{\partial A} = \frac{-\alpha^{*2}}{\mu_A \nu_A + \alpha^* \nu_A - \alpha^* \mu_A} \left( \frac{\partial \mu_A}{\partial A} - \frac{\mu_A + \alpha^*}{\alpha^*} \frac{\partial \nu_A}{\partial A} \right) \quad (32)$$

because, in general,

$$\frac{\partial \mu_A}{\partial A} - \frac{\mu_A + \alpha^*}{\alpha} \frac{\partial \nu_A}{\partial A} \neq 0 \tag{33}$$

If we apply the assumptions Fisher made about sampling from a gamma abundance distribution, then

$$\begin{aligned} \frac{\partial \mu_A}{\partial A} - \frac{\mu_A + \alpha^*}{\alpha^*} \frac{\partial \nu_A}{\partial A} &= p\alpha \\ \left( 1 - \frac{\alpha(\mu_A + \alpha^*)}{\alpha^*(\mu_A + \alpha)} \left( 1 + \frac{\mu_A}{\alpha} \right)^{-\alpha/s} \right) &\rightarrow 0 \text{ as } s \rightarrow \infty \end{aligned} \tag{34}$$

Therefore, the independence of  $\alpha$  from  $A$  carries over to  $\alpha^*$ , once again, in the limit of large  $s$ . Most often, however, computations of  $\alpha$  do not in fact use  $\nu_A$  or  $\mu_A$ . Rather than average values, the raw numbers of species and individuals observed serve as estimates of the respective expectations. This leads to  $\hat{\alpha}$ , an estimate of  $\alpha^*$ :

$$N = \hat{\alpha}(1 - e^{-S_{\text{obs}}/\hat{\alpha}}) \tag{35}$$

Unfortunately, Eq. (35) cannot be solved directly for  $\hat{\alpha}$  in terms of  $S_{\text{obs}}$  and  $N$ .

We can get an approximate solution to Eq. (35) by expanding the function  $\hat{\alpha}(N, S_{\text{obs}})$  about the point  $(\mu_A, \nu_A)$  in a second-order Taylor series using the fact that  $\hat{\alpha}(\mu_A, \nu_A) = \alpha$ . Then, we find the bias is approximately

$$\begin{aligned} E[\hat{\alpha}(N, S_{\text{obs}}) - \alpha] &\approx \frac{\alpha^2(\mu_A + \alpha)}{(\mu_A \nu_A + \alpha \nu_A - \alpha \mu_A)^3} \\ (\nu_A^2 \text{VAR}(N) + \mu_A^2 \text{VAR}(S_{\text{obs}}) - \mu_A \nu_A \text{COV}(N, S_{\text{obs}})) \end{aligned} \tag{36}$$

and the sample variance of  $\hat{\alpha}$  is

$$\begin{aligned} \text{VAR}(\hat{\alpha}) &\approx \frac{\alpha^4}{(\mu_A \nu_A + \alpha \nu_A - \alpha \mu_A)^2} \\ &\left( \text{VAR}(N) + \left( \frac{\mu_A + \alpha}{\alpha} \right)^2 \text{VAR}(S_{\text{obs}}) \right. \\ &\left. + 2 \text{COV}(N, S_{\text{obs}}) \right) \end{aligned} \tag{37}$$

To this point, we have needed no information regarding abundance distributions or sampling methods. Therefore, any conclusions we can draw from Eqs. (36) and (37) apply to any study system. Let  $\gamma(X)$  be the

coefficient of variation of  $X$ . Then, even for moderate sample sizes,

$$\begin{aligned} E[\hat{\alpha}(N, S) - \alpha] &\approx \frac{\alpha^2}{\nu_A} (\gamma(N)^2 + \gamma(S)^2 \\ &- \gamma(N)\gamma(S_{\text{obs}})\text{CORR}(N, S_{\text{obs}})) \end{aligned} \tag{38}$$

For moderate  $A$  the sampling process will be nearly Poisson. Therefore,  $N$  will be Poisson because it is a sum of Poissons. Thus,  $\gamma(N)$  will decline with sample size. As we increase sampling effort,  $A$ , it must be that  $\nu_A$  is nondecreasing because we cannot undiscover species. For most collections, the probability of finding new species decreases as we increase sampling effort because we generally find common species first and rare species after prolonged sampling. Thus,  $\gamma(S_{\text{obs}})$  must also decline with increases in  $A$ . From  $-1 \leq \text{CORR}(N, S_{\text{obs}}) \leq 1$  we can see that bias decreases as fast as  $\alpha^2/\nu_A$ . This means that  $\hat{\alpha}$  should depend weakly on sample size. Once a sample large enough to get good estimates of the various coefficients of variation has been gathered,  $\hat{\alpha}$  will show little further change. This holds for any abundance distribution for which the variances and means of  $N$  and  $S_{\text{obs}}$  exist. The fluctuations of  $\hat{\alpha}$  can be seen from Eq. (37) to decrease as  $\alpha/\nu_A$  as well so that  $\hat{\alpha}$  closely approximates  $\alpha^*$  for large  $A$ . However, we should remember that  $\alpha$  expresses a relationship between the means of  $N$  and  $S_{\text{obs}}$ . Therefore, we will need more information on the abundance distribution if we are to interpret the number of undiscovered species remaining in terms of  $\hat{\alpha}$ .

## 2. Chao's Abundance-Based Estimator ( $S_{\text{AM}}$ )

Abundance-based methods apply to data sets in which individuals can be readily identified and counted. From our data we compute the statistics  $(R_1, \dots, R_N)$  and seek an estimate of  $E[R_0]$ , the expected number of unobserved species. Chao (1984) proposed

$$S_{\text{AM}} = S_{\text{obs}} + \frac{R_1^2}{2R_2} \tag{39}$$

as a lower bound for  $s$  in the limit of infinite sample size.  $S_{\text{AM}}$  makes the following two assumptions:

1. Capture probabilities,  $q_i$ , may vary among species
2. Capture probabilities remain fixed during sampling

where  $q_j$  gives the probability of capture of an individual of species  $j$ . The starting point for the derivation of  $S_{AM}$  begins with a model of the sampling process.

Using the Poisson approximation as described previously, we get

$$E[R_i] \approx \sum_{j=1}^s \frac{(Nq_j)^i}{i!} e^{-Nq_j} \quad (40)$$

From Eq. (40) we see that this approximation works best when  $i$  is small relative to  $N$  and  $q_i$  is small. Thus, this method should perform best when applied to collections that have relatively few common or abundant species but many rare species. The exponential form of the Poisson approximation permitted Chao (1984) to do some clever analysis using a judiciously chosen distribution function,  $F_{AM}(x)$ . The key consequence of the Poisson approximation is that the  $i$ th moment of  $F_{AM}(x)$ , written  $\mu_i$ , can be used to estimate  $E[R_{i+1}]$  because

$$\mu_i \approx (i + 1)! \frac{E[R_{i+1}]}{E[R_i]}. \quad (41)$$

Next, Chao (1984) connects the seen with the unseen species through

$$E[R_0] \approx E[R_1] \int_0^N \frac{1}{x} F_{AM}(dx) \quad (42)$$

Once again, we have used the Poisson approximation to estimate the expected number of species. This integral representation reveals two pathways to further analysis: approximate the integrator or approximate the integrand. Chao (1984) approximated the integrator rather than the integrand. By constructing a suitable distribution function that possesses the same first and second moments as  $F_{AM}(x)$ , Chao shows that the smallest value taken by Eq. (42) must be  $E[R_1]/\mu_1$ . Using Eq. (41) and taking the  $R_i$  as estimates of  $E[R_i]$ , we find

$$S_{AM} = S_{obs} + \frac{R_1^2}{2R_2} \quad (43)$$

The derivation incorporates the following assumptions into  $S_{AM}$ :

1. Sampling is a Poisson process,
2. Third-order and higher moments of  $F_{AM}$  may be neglected,
3. The  $R_i$  are good estimates of  $E[R_i]$ .

These assumptions determine the statistical properties of  $S_{AM}$ . The first assumption could be used to get the

sufficiency of  $(R_0, \dots, R_N)$  if the  $q_i$  were iid random variables with distribution  $H(q)$ . Then,

$$P(r_0, \dots, r_N) = \binom{s}{r_0, \dots, r_N} \prod_{i=0}^N [\theta_i(H)]^{r_i} \quad (44)$$

where  $\theta_i(H) = \int_0^1 \binom{N}{i} q^i (1 - q)^{N-i} h(q)$ . Paralleling Burnham and Overton (1978), Eq. (44) can be rewritten to permit use of a factorization theorem to assert the sufficiency. If we relax the assumption that the  $q_i$  be identically distributed, then we need a model for selecting  $H_i(q)$ , the distribution function for  $q_i$ . For example, if the  $H_i$  corresponded to habitat heterogeneity, then the sufficient statistics would presumably incorporate information on variation in habitat type during sampling. The first and the last assumptions allow us to avoid assuming a form for the abundance distribution.  $S_{AM}$  is therefore nonparametric. Assumption 2 makes  $S_{AM}$  a lower bound for  $s$  that increases as sample size increases, indicating that  $S_{AM}$  is a biased estimator. Assuming that we have a large enough sample size, we expand the function  $S_{AM}$  as a first-order Taylor polynomial about  $E[S_{obs}]$ ,  $E[R_1]$ , and  $E[R_2]$ . Then we can use

$$\text{VAR}(aX + bY) = a^2 \text{VAR}(X) + b^2 \text{VAR}(Y) + 2ab \text{COV}(X, Y) \quad (45)$$

to estimate the variance. Working through the analysis and assuming the covariance terms are small gives

$$\begin{aligned} \text{VAR}(S_{AM}) \approx & \left(1 + \frac{R_1}{R_2}\right)^2 \text{VAR}(R_1) \\ & + \left(1 - \frac{R_1^2}{2R_2^2}\right)^2 \text{VAR}(R_2) + \sum_{i=3}^N \text{VAR}(R_i) \end{aligned} \quad (46)$$

Even when we have the abundance distribution, the indicated variances can be challenging to compute. Chao (1984) used bootstrapping techniques coupled with simulations and tests on well-characterized data sets to evaluate the performance of the  $S_{AM}$ . When applied to data sets with known  $s$ ,  $S_{AM}$  generally had smaller bias and narrower confidence intervals than  $S_{J1}$  or  $S_{J2}$ .

### 3. Chao's Incidence-Based Estimator ( $S_{IM}$ )

Incidence-based methods apply to data sets in which individuals are not readily identified or counted. We compute  $(F_1, \dots, F_N)$  from the data matrix  $X_{i,j}$  and seek an estimate of  $E[F_0]$ , the expected number of unob-



served species. Chao (1987) used the same approach as used for  $S_{AM}$  to derive

$$S_{IM} = S_{obs} + \frac{F_1^2}{2F_2} \tag{47}$$

for data sets of species incidence. Like the jackknife estimator, the original application of this method was to estimate population sizes. It produces a lower bound for  $s$ . The starting point for the derivation of  $S_{IM}$  begins as  $S_{AM}$  did with a model of the sampling process. Let  $p_i$  be the probability of detection of species  $i$ . In this case, sampling is as in the jackknife estimator. Therefore, we assume

1. Detection probabilities,  $p_i$ , may vary between species.
2. The  $p_i$  remain fixed during sampling.
3. The  $p_i$  are iid random variables with distribution  $H(p)$ .

As we saw for  $S_{JK}$ , if these assumptions are met then

$$(F_0, \dots, F_t)$$

will be sufficient for  $H$ . The last assumption allows us to write

$$E[F_i] = \int_0^1 \binom{t}{i} p^i (1-p)^{t-i} h(p) dp \tag{48}$$

From here we can once again use the Poisson approximation to simplify Eq. (48). In this case, the focus on data sets that concentrate probability over  $F_1$  and  $F_2$  is less critical. The Poisson approximation will still work well if the probability of large  $p_i$  is small. From here, Chao (1987) parallels the derivation of  $S_{AM}$  using an absolutely continuous version of  $F_{AM}(x)$ . This distribution function,  $F_{IM}(x)$  has moments that satisfy

$$\mu_i \approx (i + 1)! \frac{E[F_{i+1}]}{E[F_1]} \tag{49}$$

An integral relationship between  $F_0$  and  $F_1$  analogous to Eq. (42) once again leads us to approximate either integrand or integrator. While  $S_{IM}$  arises from approximating the integrator as before, we gain some insight into the relationship between  $S_{JK}$  and  $S_{IM}$  by considering the other alternative first. If we approximate the integrand,  $1/x$ , with a  $k$ th order polynomial then we find

$$E[F_0] \approx \sum_{i=1}^k (-1)^{i+1} \binom{k}{i} E[F_i] \tag{50}$$

Now, the bias should decrease as the expected number of undiscovered species decreases. This expectation, given by

$$E[F_0] = \int_0^1 (1-p)^t h(p) dp \tag{51}$$

must decrease at least as fast as  $1/t$  because  $(1-p)^t \leq e^{-tp}$  for  $0 \leq p \leq 1$ . Equation (50) has, as Chao (1984) points out, the same form of the  $k$ th-order jackknife estimator. The connection between  $S_{JK}$  and  $S_{IM}$  reflects the assumption made in any jackknife estimator that the bias can be expressed as a power series in  $1/t$ . Chao (1987) obtains  $S_{IM}$  by approximating the integrator rather than the integrand. Then, using an analog to Eq. (41) and taking the  $F_i$  as estimates of  $E[F_i]$ , Chao proposes

$$S_{IM} = S_{obs} + \frac{F_1^2}{2F_2} \tag{52}$$

As with  $S_{AM}$ , we have several assumptions built into the derivation of  $S_{IM}$ :

1. The  $p_i$  are small enough to justify the Poisson approximation to the binomial probabilities of Eq. (48).
2. Third-order and higher moments of  $F$  may be neglected.
3. The  $F_i$  are good estimates of  $E[F_i]$ ,

First, we note that  $S_{IM}$  is nonparametric. Assumption 2 makes  $S_{IM}$  a lower bound for  $s$  that increases as sample size increases, indicating that  $S_{IM}$  is a biased estimator. Chao (1987) estimates the variance of  $S_{IM}$  as

$$\text{VAR}(S_{IM}) = \frac{F_2}{4} \left( \left( \frac{F_1}{F_2} \right)^4 + 4 \left( \frac{F_1}{F_2} \right)^3 + 2 \left( \frac{F_1}{F_2} \right)^2 \right) \tag{53}$$

Equation (53) can be used to get approximate confidence intervals for  $S_{IM}$  provided that  $p_i t$  is not too large. As we increase  $t$ , we expect the lower order  $F_i \rightarrow 0$  if we are sampling a finite collection of individuals. This would seem to imply that the variance goes to zero with increasing  $t$ . However, large  $t$  means large  $p_i t$  and so makes Eq. (53) a poor approximation. Chao (1987) used real data sets with known  $s$  to evaluate  $S_{IM}$ , finding that it performed well relative to  $S_{JK}$ . For low  $t$ ,  $S_{IM}$

displayed smaller standard error and less negative bias than  $S_{J2}$ . However, as  $t$  was increased,  $S_{JK}$  outperformed  $S_{IM}$ . Notably, the data sets tended to have more recaptures with increasing  $t$ . Chazdon *et al.* (1998) found similar results.

### E. Estimates Derived from Sample Coverage ( $S_{AC}$ , $S_{IC}$ )

Sampling of a finite community must eventually capture each individual multiple times. Thus, we can expect that exhaustive sampling would lead to  $F_i = 0$  and  $R_i = 0$  for the lower order  $i$ . An alternative that measures representation of species in the data set while retaining analytic tractability is sample coverage. The sample coverage,  $C$ , is defined by

$$C = \sum_{i=1}^s q_i I(X_i > 0) \text{ for abundance data or}$$

$$C = \sum_{i=1}^s p_i I(Y_i > 0) \text{ for incidence data}$$

#### 1. Chao's Abundance-Based Coverage Estimator ( $S_{AC}$ )

Chao and Lee (1992) propose a nonparametric estimate of  $s$  using estimates of  $E[S_{obs}]$ ,  $E[C]$ , and  $\gamma(q)$ . They define the abundance coverage-based estimator  $S_{AC}$  as follows:

$$S_{AC} = \frac{E[S_{obs}]}{E[C]} + \frac{E[R_1]}{E[C]} \gamma(q)^2 \quad (54)$$

Intuitively, we see that  $S_{obs}$  is inflated by  $C$  to adjust for the fraction of the distribution left uncovered. Additional bias correction reflecting the shape of the distribution comes through  $R_1$  and the  $\gamma$  of the distribution. Thus, we are using information from more than just the lower order abundances and may expect  $S_{AC}$  to perform better than  $S_{AM}$  and  $S_{IM}$  when higher order abundances support significant probability. We start with the same two basic assumptions that apply to  $S_{AM}$ :

1. Individual detection probabilities,  $q_i$ , may vary between species.
2. The  $q_i$  remain fixed during sampling.

Thus, the  $q_i$  can be regarded as a sequence of random variables with mean  $\bar{q}$ , variance  $\sigma_q$ , and coefficient of variation  $\gamma(q)$ .

Then, we find that

$$E[S_{obs}] = s - \sum_{i=1}^s (1 - q_i)^N$$

$$E[C] = 1 - \frac{\sum_{i=1}^s (1 - q_i)^N}{\sum_{i=1}^s q_i}$$

Of course, if we knew the true values of the  $q_i$  we could compute  $s$  directly. Instead, we get estimates of  $E[S_{obs}]$  and  $E[C]$  from our data. Now,  $E[S_{obs}]$  and  $E[C]$  are functions of  $s$  and the  $q_i$ . Therefore, we would like to eliminate the dependence on the  $q_i$  in order to estimate  $s$ . If our estimates of  $E[S_{obs}]$  and  $E[C]$  are close to the truth, we can use Taylor's theorem to expand about the point  $(q_1, \dots, q_s)$  to get

$$s \approx \frac{E[S_{obs}]}{E[C]} + \frac{E[R_1]}{E[C]} \gamma(q)^2 \quad (55)$$

Now, all we need are estimates of the expectations and the coefficient of variation for the abundance distribution. Chao and Lee (1992) suggest

$$E[S_{obs}] \approx S_{obs} \quad (56)$$

$$E[R_1] \approx R_1 \quad (57)$$

$$E[C] \approx \hat{C} = 1 - \frac{R_1}{N} \quad (58)$$

for the various expectations. The coefficient of variation is somewhat more complicated. First, it must always be positive. Second, when the true  $\gamma$  is large, we can expect to need more data to estimate it. It can be easily shown that

$$\gamma^2 = \frac{1}{s} \frac{\sum_{i=1}^s q_i^2}{\bar{q}^2} - 1 \quad (59)$$

and

$$E \left[ \sum_{i=1}^N i(i-1)R_i \right] = n(n-1) \sum_{i=1}^s q_i^2. \quad (60)$$

Chao and Lee (1992) then suggest that

$$\hat{\gamma}^2 = \max \left\{ \hat{N}_0 \frac{\sum_{i=1}^N i(i-1)R_i}{(\sum_{i=1}^N iR_i)^2} - 1, 0 \right\} \quad (61)$$

where  $\hat{N}_0 = S_{\text{obs}}/\hat{C}$  when  $\hat{\gamma}^2 \leq 0.5$ . Then we have

$$S_{AC} = \frac{S_{\text{obs}}}{\hat{C}} + \frac{R_1}{\hat{C}} \hat{\gamma}^2 \quad (62)$$

In Eq. (61) the term  $\hat{N}_0$  is an initial estimate of  $s$ . If  $\hat{\gamma}^2 > 0.5$  the bias can be reduced by using  $S_{AC}$  in place of  $S_{\text{obs}}/\hat{C}$  to yield

$$\hat{\gamma}^2 = \max \left\{ S_{AC} \frac{\sum_{i=1}^N i(i-1)R_i}{(\sum_{i=1}^N iR_i)^2} - 1, 0 \right\} \quad (63)$$

Then, for larger  $\gamma$  we have

$$\tilde{S}_{AC} = \frac{S_{\text{obs}}}{\hat{C}} + \frac{R_1}{\hat{C}} \hat{\gamma}^2 \quad (64)$$

It turns out that  $S_{\text{obs}}/\hat{C}$  is biased low when the  $q_i$  differ from one another [when sample sizes are large enough to permit estimation of  $\gamma(q)$ ]. Thus,  $S_{AC}$  should be biased low. The variance of  $S_{AC}$  is approximately a linear combination of  $\text{COV}(R_i, R_j)$ . It can be shown that

$$\text{VAR}(R_i) = E[R_i] \left( 1 - \frac{E[R_i]}{s} \right) \quad (65)$$

$$\text{COV}(R_i, R_j) = -\frac{E[R_i]E[R_j]}{s} \quad (66)$$

Now  $R_i \leq s$  so the variance is bounded and decreases with increasing richness. However, for a fixed  $s$ , variance depends on how the  $R_i$  spread out as we increase  $t$ . This in turn depends on the sampling model we use. Here, we have assumed that the  $q_i$  remain fixed so that we sample with replacement. Therefore,

$$E[R_i] = \sum_{j=1}^s \binom{N}{i} p_j^i (1 - p_j)^{N-i} \quad (67)$$

Evidently, as  $N$  increases,  $R_i \rightarrow 0$  with probability 1. Now, the  $R_i$  represent  $N$  different abundance classes. However, we have only  $s$  species. Therefore, the probability that  $R_i > 0$  will be spread out over an increasing set of the  $R_i$  as  $N$  increases. Thus,  $E[R_i]$  should become small relative to  $s$  as  $N$  increases. The simulations performed by Chao and Lee (1992) indicate that at high sample coverage  $S_{AC}$  is nearly unbiased. However, Colwell and Coddington (1994) found that  $S_{AC}$  tends to overestimate  $s$  when sample sizes are small. The difficulty appears to occur for large values of  $\gamma$ . In this case, many of the individuals in a sample represent

common species, so there are a few dominant  $q_i$  in the collection. In this case,  $\hat{C}$  is a poor estimate of sample coverage. However, we can easily ascertain the number of common species in a collection. Chao *et al.* (1993) suggest truncation of the data set to species with  $k$  or fewer representatives. The new estimates for  $E[C]$  and  $E[S_{\text{obs}}]$  that result are

$$S_{\text{obs},k} = \sum_{i=1}^k R_i \quad (68)$$

$$\hat{C}_k = 1 - \frac{R_1}{\sum_{i=1}^k iR_i} \quad (69)$$

Similarly, we have

$$\hat{\gamma}_k = \max \left\{ \frac{S_{\text{obs},k}}{\hat{C}_k} \frac{\sum_{i=1}^k i(i-1)R_i}{(\sum_{i=1}^k iR_i)^2} - 1, 0 \right\} \quad (70)$$

so that

$$S_{ACk} = S_{\text{obs}} - S_{\text{obs},k} + \frac{S_{\text{obs},k}}{\hat{C}_k} + \frac{R_1}{\hat{C}_k} \hat{\gamma}_k^2 \quad (71)$$

## 2. Chao's Incidence-Based Coverage Estimator ( $S_{IC}$ )

The incidence-based coverage estimator,  $S_{IC}$ , applies to data sets that have the quadrat structure outlined for  $S_{IM}$ . Briefly, when we have the presence-absence of data then incidence-based estimators apply. Lee and Chao (1994) found ways to relax many of the assumptions present in  $S_{AM}$ ,  $S_{IM}$ ,  $S_{JK}$ , and  $S_{AC}$ . They follow Pollock (1976) by partitioning variability into between-species, within-species, and temporal components. The results involve more complex notation and more complicated proofs but use the same conceptual framework as  $S_{AC}$ . For brevity, we present only the case of unequal detection probabilities among species. The estimator for incidence-based coverage is

$$S_{IC} = \frac{E[S_{\text{obs}}]}{E[C]} + \frac{E[F_1]}{E[C]} \gamma(p)^2 \quad (72)$$

We define  $p_i$  as the probability of detection of species  $i$ . For  $S_{IC}$  we assume that

1. Detection probabilities,  $p_i$ , may vary between species.
2. The  $p_i$  remain fixed during sampling.

Thus, the  $p_i$  can be regarded as a sequence of random variables with mean  $\bar{p}$ , variance  $\sigma_p$ , and coefficient of variation  $\gamma(p)$ .

Then, we find that

$$E[S_{\text{obs}}] = s - \sum_{i=1}^s (1 - p_i)^t$$

$$E[C] = 1 - \frac{\sum_{i=1}^s (1 - p_i)^t}{\sum_{i=1}^s p_i}$$

$E[D]$  and  $E[C]$  are functions of  $s$  and the  $p_i$ . Therefore, we would like to eliminate the dependence on the  $p_i$  in order to estimate  $s$ . We can use Taylor's theorem to expand about the point  $(p_1, \dots, p_s)$  to get

$$s \approx \frac{E[S_{\text{obs}}]}{E[C]} + \frac{E[F_1]}{E[C]} \gamma(p)^2 \quad (73)$$

Using the approximations

$$E[S_{\text{obs}}] \approx S_{\text{obs}} \quad (74)$$

$$E[F_1] \approx F_1 \quad (75)$$

$$\hat{C} = 1 - \frac{F_1}{\sum_{i=1}^t iF_i} \quad (76)$$

$$\tilde{C} = 1 - \frac{(t-1)F_1 - 2F_2}{(t-1)\sum_{i=1}^t iF_i} \quad (77)$$

and

$$\hat{\gamma}^2 = \max \left\{ \hat{N}_0 \frac{\sum_{i=1}^t i(i-1)F_i}{(\sum_{i=1}^t iF_i)^2} - 1, 0 \right\} \quad (78)$$

where  $\hat{N}_0 = S_{\text{obs}}/\hat{C}$  when  $\hat{\gamma}^2 \leq 0.5$ , we get

$$S_{\text{IC}} = \frac{S_{\text{obs}}}{\hat{C}} + \frac{F_1}{\hat{C}} \hat{\gamma}^2 \quad (79)$$

If  $\hat{\gamma}^2$  is large, the bias can be reduced by using  $S_{\text{IC}}$  in place of  $S_{\text{obs}}/\hat{C}$  to compute  $\hat{\gamma}^2$ . Then, using  $\tilde{C}$  for  $\hat{C}$ , we have

$$\tilde{S}_{\text{IC}} = \frac{S_{\text{obs}}}{\tilde{C}} + \frac{F_1}{\tilde{C}} \hat{\gamma}^2 \quad (80)$$

As  $\hat{C}$  increases, the bias must decrease. Therefore, we can expect that the assumptions of fixed  $p_i$  will lead to  $S_{\text{IC}}$  underestimating  $s$  in most cases. The variance of  $S_{\text{AC}}$

is approximately a linear combination of  $\text{COV}(F_i, F_j)$ . It can be shown that

$$\text{VAR}(F_i) = E[F_i] \left( 1 - \frac{E[F_i]}{s} \right) \quad (81)$$

$$\text{COV}(F_i, F_j) = -\frac{E[F_i]E[F_j]}{s} \quad (82)$$

The variance is thus bounded and decreases with increasing richness because  $F_i \leq s$ . Now, as we increase  $t$  each of the species will be detected a different number of times with probability 1 (even if they have the same  $p_i$ ). Therefore, there will be at most  $s$  nonzero  $F_i$  among  $(F_1, \dots, F_t)$ . Thus,  $E[F_i] \rightarrow 0$  as  $t \rightarrow \infty$ . As with  $S_{\text{AC}}$ , the effects of a few dominant  $p_i$  can be adjusted for using the methods of Chao *et al.* (1993). If we truncate the data set to species with  $k$  or fewer representatives. The new estimates for  $E[C]$  and  $E[D]$  that result are

$$S_{\text{obs},k} = \sum_{i=1}^k F_i \quad (83)$$

$$\hat{C}_k = 1 - \frac{F_1}{\sum_{i=1}^k iF_i} \quad (84)$$

Similarly, we have

$$\hat{\gamma}_k^2 = \max \left\{ \frac{S_{\text{obs},k}}{\hat{C}_k} \frac{\sum_{i=1}^k i(i-1)F_i}{(\sum_{i=1}^k iF_i)^2} - 1, 0 \right\} \quad (85)$$

so that

$$S_{\text{IC}k} = S_{\text{obs}} - S_{\text{obs},k} + \frac{S_{\text{obs},k}}{\hat{C}_k} + \frac{F_1}{\hat{C}_k} \hat{\gamma}_k^2 \quad (86)$$

## F. Estimation on Heterogeneous Samples

Increasing sample size often improves estimator performance through the operation of the laws of large numbers. We get better estimates because the means of many summary statistics become closer to the truth and the variation in these estimates decreases with increasing sample size. Of course, increasing sample sizes increases the cost of performing the census. However, a potentially for more serious problem accompanies increases in sample size. In most cases, larger samples take more time or cover larger areas. Inevitably, this means increasing heterogeneity in the sampling process. For example, seasonal variation or habitat heterogeneity will cause detection probabilities to change within a species

as sampling progresses. A similar problem occurs in estimating the population size of a single species. Pollock *et al.* (1984) and Lee and Chao (1994) modeled the contribution of heterogeneity in individual capture probabilities to population size estimates. Just as most of the estimators presented so far have been adapted from population size estimators, Nichols *et al.* (1998) applied Pollock's robust estimation method to species richness estimation. These methods incorporate many parameters to characterize the effect of heterogeneity on capture probabilities. Consequently, the estimators are substantially more complicated. Although we may better adjust an estimator to the needs of a given census, the added complexity tends to obscure the overall relationship between sample size and species richness.

## VI. ESTIMATORS BASED ON EXTRAPOLATION

In order to approximate exhaustive sampling or to elucidate large-scale patterns, many studies extrapolate beyond the extent of a given census. These studies focus less on the sampling process and more on biological processes that operate over scales that are too large to survey in detail. As noted previously, studies that increase the scope of sampling usually encounter a wider range of heterogeneity. Heterogeneity increases due to the responses of individual species to the range of environmental and biological conditions as we increase the study area or the time span of the sample. As we increase sample size, we expect that the correlation in the abundance of arbitrarily selected species should decrease. This decreasing correlation permits us to average over the fine-scale, detailed distribution of the components of  $S_{\text{obs}}$  and focus on the coarse-scale, global behavior of  $S_{\text{obs}}$ . When most species in large communities interact weakly with most other species, our experience with the laws of large numbers suggests that there should be a trend relating species richness to area or time span. If the interactions are weak enough relative to sample size, the distribution of fluctuations about the trend may display regular, perhaps even Gaussian, behavior. Our models of species accumulation can then be built with two pieces of information. First, we shall need the average rate of increase in richness as a function of sample size. Second, we need the variability about the mean. The first piece gives us a trend we can extrapolate. The second allows us to fit the model to the data. Most extrapolation methods focus on the trend model and assume a form for the noise that makes curve fitting convenient.

Once we have a suitable model, we can then extrapolate beyond the areas in which we actually sample. We must keep in mind that this extension confounds the contribution of ecological process and sampling effort to diversity. At smaller scales, sampling effort dominates the accumulation of diversity. However, at larger scales, we encounter heterogeneity effects that may scale differently than the effects of sampling effort. Often, we have no general guidelines for selecting one of the several functional forms for species accumulation. Most of the difficulty in applying these models stems from fitting the nonlinear models to data. For brevity, we omit the technical details of curve fitting. We briefly consider a few representative forms here. These are species–area curves, birth chains, and the Michaelis–Menten equation.

To evaluate extrapolation models, we need at least three pieces of information. First, is the trend nondecreasing? Even though it may seem counterintuitive, there are models whose rate of species accumulation increases at small sample sizes (Leitner and Rosenzweig, 1997; Soberón and Llorente, 1993). Second, does the trend have an asymptote? Third, what curvature does the functional form of the model predict? The fitting of the model to data depends on the overall shape of the function. Less obviously, the model transforms the noise that gives rise to the residuals used to evaluate the goodness of fit.

### A. Species–Area Curves

Species–area curves may use samples from disjoint areas or nested areas. This approach has a long history due primarily to its flexibility. There are several functions that may be fit to the data in order to extrapolate to a large area. Chief among these are  $\log(S_{\text{obs}})$  vs  $\log(N)$  and  $S_{\text{obs}}$  vs  $\log(N)$ . Note that these curves have no asymptote. Furthermore, their curvature is often too shallow to give the best fits to the data. The key parameters of interest are the slope (the  $z$ -value) and intercept of the plot. The slope in particular has drawn much attention in both theoretical and applied circles. Given the abundance distribution, one may use methods similar to Fisher's approach in deriving  $\alpha$  to make theoretical predictions about the magnitude of the slope. Preston (1962) and May (1975) both claim theoretical predictions for  $z$ -values built solely on the assumption of the lognormal distribution abundances. This work implied that the species–area curve, at least for lognormal distributions, reflected nothing more than sampling. However, this conclusion rests on the tacit assumption that the abundance distribution be lognormal

at all scales. This assumption cannot be achieved without invoking ecological process. Leitner and Rosenzweig (1997) demonstrate that realistic species–area curves must include information about spatial heterogeneity. The curves obtained in this work did have finite asymptotes and variable curvatures. This work does not, however, provide insights into the selection of log–log or log–linear models.

## B. Birth Chains

Many different functions can exhibit similar asymptotic behavior and curvature. Suppose we select the candidate functions that fit the data best. Functions that give equally good fits to the data can make very different asymptotic predictions. One way to select among these candidate models would be to choose the model that best captures the mechanisms of species accumulation. Soberón and Llorente (1993) use the framework of birth chains for this purpose. Let  $\{S(t): t \geq 0\}$  be the number of species observed by time  $t$ . Then  $S(t)$  is a birth chain if

$$P\{S(t + dt) - S(t) = 1 | S(t)\} = \lambda(S(t), t)dt \quad (87)$$

$$P\{S(t + dt) - S(t) > 1\} = 0 \quad (88)$$

$$P\{S(0) = 0\} = 1 \quad (89)$$

This set of conditions means that the time between discoveries of new species is exponentially distributed. However, the rate at which new species occur depends on the time spent sampling and the number of species already found. The rate function,  $\lambda(S, t)$ , captures the details of the mechanism of species discovery. For example, suppose we assume that the rate of discovery declines linearly with species discovered. Then  $\lambda(S, t) = a - bS(t)$ . Soberón and Llorente (1993) show that

$$E[S(t)] = \frac{a}{b}(1 - e^{-bt}) \quad (90)$$

This method facilitates curve fitting because it allows us to derive expressions for the variance of  $S(t)$ . Existing methods that lack mechanistic foundations may be justified using this approach. For example, Soberón and Llorente (1993) provide insights into the popular Michaelis–Menten equation using this birth chain model.

## C. Michaelis–Menten

The Michaelis–Menten equation has often been fit to species accumulation data. It is a widely known func-

tion that has a finite asymptote, is nondecreasing, and accommodates a wide range of curvatures. The equation states that

$$S_{\text{obs}}(t) = \frac{at}{1 + bt} \quad (91)$$

Soberón and Llorente (1993) use a birth chain model to suggest a mechanistic basis for the Michaelis–Menten equation. The birth chain model assumes that the number of new species found is an inhomogeneous Poisson process with rate

$$\lambda(S, t) = a + \frac{b^2}{a} \frac{at}{1 + bt} - 2 \frac{b}{a} S \quad (92)$$

where  $j$  is the number of species found by time  $t$  and  $a$  and  $b$  are constants to be fit to the data. Soberón and Llorente (1993) argue that if the chances of adding a new species improve with effort, to some upper limit, then a form such as Eq. (92) obtains. The key drawback with this and other curve-fitting models is that we do not have a distribution for the changes in  $S_{\text{obs}}$ . This means that we cannot rationally assign different weights to the data during fitting.

## VII. WHICH ESTIMATION METHOD TO USE?

Many estimators have been put forth, leaving the investigator with the practical question of which estimators are appropriate for use with a given new data set. Two facts may give us confidence a priori in an estimator's applicability to a new data set. As mentioned previously, each estimator makes assumptions about properties of the community being sampled (e.g., variation in capture probabilities across species or samples) and the sampling strategy used. If the community sampled and the sampling procedures used satisfy the assumptions of the estimator, then we can have some confidence in applying that estimator to a new data set.

However, meeting a method's explicit assumptions is not always sufficient. Another difficulty with relying solely on the theoretical basis of estimators is the fact that one or more assumptions are often violated in the sampling of real communities. Thus, theoretical support alone may not justify the use of a particular estimation method.

A second way to have confidence a priori in an estimator is to know its performance on similar communities sampled using similar methods over similar scales.

Evaluating estimator performance on data allows testing implicit assumptions. Also, many estimation methods (e.g., extrapolation) have little biology built in: They exist because they seem to work in a few cases, and our confidence in them may come only from verification with relatively complete data sets. One should take care, however, not to conclude too much about an estimator based on its performance on single data sets. Rather, only through comprehensive testing on real data sets from a variety of taxa, locations, and spatial/temporal scales, as well as testing on simulated data, will we gain insight into estimator performance.

## A. Estimator Evaluation

### 1. Evaluation on Real Data and Simulated Data

The testing of estimators has used data from real communities and from computer simulation. Each type of data has advantages and drawbacks. Real data take considerable effort to obtain. As a result, replication of real data sets is difficult. Simulated data sets, on the other hand, can be turned out in comparatively large numbers. This allows replication of data and testing of various sampling strategies on the same data.

To test estimators effectively with data, we must have some idea of the true parameters underlying our data (otherwise, how do we know what estimators should predict?). With simulated data, each parameter's true value is known because it is either specified by the investigator or is directly measurable from the simulated community. With real data, the truth may be approximated by testing estimators on smaller sets of data from communities in which the parameters are reasonably well-known from separate, intensive sampling.

Finally, there are many properties of ecological communities which are likely to influence estimator performance. These include factors such as species richness  $s$ , various properties of the species abundance distribution, intraspecific clumping, and interspecific associations. Each of these can vary widely. Simulated data allow us to explore large parts of this parameter space. More important, with simulated data we can cover the range corresponding to anticipated field conditions. This allows us to test and select estimators before taking them into the field. Nevertheless, only real data can tell us which parts of the vast parameter space are relevant. Thus, we cannot ignore either approach: The biological reality of actual data is complemented by the versatility of simulated data.

Some progress has been made in the evaluation of estimation methods with data. Every such study cannot be detailed here, and not enough work has been done to allow generalizations to be made. Chazdon *et al.* (1998) provide one useful example of an approach combining both real and simulated data. They examined the performance of eight estimators using young woody regeneration in six tropical forest sites. Noticing that patchiness (intraspecific clumping) varied across sites, they then resampled the data sets to create simulated data sets with a range of patchiness levels. Testing estimators on these simulated data sets revealed the effect of patchiness on the performance of various estimators.

### 2. Estimator Evaluation Criteria

Clearly, one comprehensive study of the performance of all estimators over all possible data is impossible. Evaluation of estimators, then, will necessarily be done by many investigators with different communities and over different ranges of simulated data. To facilitate this cooperative approach, we must identify a common set of evaluation criteria that can be used to measure the performance of an estimator on a data set. These criteria are slightly different from theoretical properties of estimators. Numerous such criteria have been put forth; although they differ in calculation methods, they are often variants of a few common themes, which are listed as follows:

**Bias:** As defined previously, bias measures deviation from the true value. Common measures of bias include  $E[S_{\text{est}}] - s$ , or deviation from  $s$ , and  $(E[S_{\text{est}}] - s)/s$ , or relative deviation from  $s$ . In the evaluation of estimators with data, bias can be calculated as a mean across replicate data sets or can be assessed with increasing sample size within a data set.

**Variance:** As discussed previously, variance measures uncertainty in an estimate. If we have high variance, we can have little confidence that a single observation is a good indication of the mean. Variance may be calculated numerically over replicates, or an analytical estimator of variance may be used. These measures can also be expressed as confidence intervals and may be used in hypothesis testing when comparing estimated richnesses. Note that this measure does not depend on  $s$ , which is considered by other criteria.

**Sample size independence:** This measures the rate of convergence of an estimator's mean to its asymptotic value. An estimator that rapidly approaches its asymptotic value requires less sampling effort to obtain an equally good estimate. Such an estimator is relatively

sample size independent. This measure, too, does not depend on  $s$ .

Note that the previous three criteria may be calculated over multiple realizations of the same underlying parameters (e.g., the capture probabilities  $q_i$ ). The following two may only be considered among multiple data sets in which these parameters vary.

**Correlation with  $S$ :** Since  $s$  is a fixed parameter, we introduce  $S$  as a random variable representing the true richness of any of a number of data sets, between which  $S$  and capture probabilities  $q_i$  may vary. If the correlation between an estimator  $S_{\text{est}}$  and true richness  $S$  across these data sets is high,  $S_{\text{est}}$  may be useful in comparing SR between sites or census occasions (Palmer, 1990). Note that correlation with  $S$  does not require that an estimator have low bias. Rather, correlation reflects a relative deviation of  $S_{\text{est}}$  from  $S$  that is somewhat constant across data sets.

**Robustness:** A robust estimator is one for which performance (as measured by any of the previous criteria) changes little across a range of data sets.

It is important to note that the relative importance of each of these criteria depends on the problem of interest. In measuring absolute species richness, minimizing bias may be a priority. In comparisons across time or space, variance and correlation with  $S$  become more important, whereas bias may become more tolerable.

A final criterion represents a minimum requirement of sorts for estimator performance:

**Beating  $S_{\text{obs}}$ :** An estimator should perform better than  $S_{\text{obs}}$ , the number of species observed. Most estimators should have lower bias than  $S_{\text{obs}}$  most of the time. However, there are trade-offs. For example, the variance of some estimators may be higher than that of  $S_{\text{obs}}$ .

## B. Selecting an Estimator

Although it is tempting to think that an estimator may exist which is robust to all census conditions, such an estimator is unlikely given the difficulty of the problem. Therefore, estimators should be chosen based on the context of the problem of interest. To select the estimators that are most likely to perform well on a new data set, one might use the following approach. First, consider the anticipated properties of a data set—based on knowledge of the system's biology—in relation to

the modeling assumptions behind estimators. Select those estimators whose assumptions are best met by the expected data. If possible, select or modify the sampling strategy so that assumptions will be better met and/or more estimators may be used. Then, test estimators using data (simulated or previously existing) similar to that anticipated. The previous approach should indicate (i) which estimators might perform best on new data and (ii) whether these estimators perform well enough to meet the needs of the problem at hand. Although evaluating a number of estimators on multiple replicate data sets may seem a daunting task, computer programs (e.g., EstimateS, R. K. Colwell, <http://viceroy.eeb.uconn.edu/estimates>; WS2M, W. R. Turner *et al.*, unpublished) may be used to automate many of these calculations. However, use of such programs should complement, not replace, thoughtful consideration of modeling assumptions, sampling design, and the biology of the community under study.

## See Also the Following Articles

ECONOMIC VALUE OF BIODIVERSITY, MEASUREMENTS OF • ECOSYSTEM FUNCTION MEASUREMENT, AQUATIC AND MARINE COMMUNITIES • ECOSYSTEM FUNCTION MEASUREMENT, TERRESTRIAL COMMUNITIES • FRAMEWORK FOR ASSESSMENT AND MONITORING OF BIODIVERSITY • MICROBIAL BIODIVERSITY, MEASUREMENT OF

## Bibliography

- Burnham, K. P., and Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* 65, 625–633.
- Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.* 11, 265–270.
- Chao, A. (1987). Estimating the population size for capture–recapture data with unequal catchability. *Biometrics* 43, 783–791.
- Chao, A., and Lee, S. M. (1992). Estimating the number of classes via sample coverage. *J. Am. Stat. Assoc.* 82, 210–217.
- Chao, A., Lee, S. M., and Jeng, S. L. (1992). Estimating population size for capture–recapture data when capture probabilities vary by time and individual animal. *Biometrics* 48, 201–216.
- Chao, A., Ma, M. C., and Yang, M. C. K. (1993). Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika* 80, 193–201.
- Chazdon, R. L., Colwell, R. K., Denslow, J. S., and Guariguata, M. R. (1998). Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of northeastern Costa Rica. In *Forest Biodiversity Research, Monitoring and Modeling: Conceptual Background and Old World Case Studies* (F. Dallmeier and J. A. Comiskey, Eds.). Parthenon, Paris.
- Colwell, R. K., and Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philos. Trans. R. Soc. London B* 345, 101–118.



- Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relations between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**, 42–58.
- Lee, S. M., and Chao, A. (1994). Estimating population size via sample coverage for closed capture–recapture models. *Biometrics* **50**, 88–97.
- Leitner, W. A., and Rosenzweig, M. L. (1997). Nested species–area curves and stochastic sampling: A new theory. *Oikos* **79**, 503–512.
- May, R. M. (1975). Patterns of species abundance and diversity. In *Ecology and Evolution of Communities* (M. L. Cody and J. M. Diamond, Eds), pp. 81–120. Belknap/Harvard Univ. Press, Cambridge, MA.
- Palmer, M. W. (1990). The estimation of species richness by extrapolation. *Ecology* **71**, 1195–1198.
- Pollock, K. H. (1976). Building models of capture–recapture experiments. *Statistician* **25**, 253–260.
- Pollock, K. H., Hines, J. E., and Nichols, J. D. (1984). The use of auxiliary variables in capture–recapture and removal experiments. *Biometrics* **40**, 329–340.
- Preston, F. W. (1962). The canonical distribution of commonness and rarity. *Ecology* **43**, 185–215.
- Smith, E. P., and van Belle, G. (1984). Nonparametric estimation of species richness. *Biometrics* **40**, 119–129.
- Soberón, J. and Llorente, J. (1993). The use of species accumulation functions for the prediction of species richness. *Conserv. Biol.* **7**, 480–488.



# MEDITERRANEAN- CLIMATE ECOSYSTEMS

Philip W. Rundel

University of California, Los Angeles

---

- I. Introduction
  - II. Natural Disturbance Regimes
  - III. Patterns of Speciation
  - IV. Biodiversity in California
  - V. Biodiversity in Chile
  - VI. Biodiversity in the Mediterranean Basin
  - VII. Biodiversity in the Cape Region of South Africa
  - VIII. Biodiversity in Southwestern Australia
  - IX. Conclusions
- 

## GLOSSARY

**chaparral** Evergreen sclerophyllous-leaved shrublands that cover large areas of California.

**coastal sage scrub** Semiwoody shrublands dominated by drought-deciduous species and occurring in semi-arid areas along the coast of Southern California and the interior transition from chaparral to desert.

**drought-deciduous** Descriptive of plant species that lose their leaves during the dry season as soil moisture becomes limited.

**fynbos** Evergreen sclerophyll vegetation dominating the Cape Region of South Africa.

**garrigue** Low-growing secondary evergreen shrublands that dominate extensive areas of the Mediterranean Basin.

**Gondwanaland** Former supercontinent of the Southern Hemisphere from which South America, Africa, Australia, and India are derived.

**kwongan** Evergreen heathlands that cover extensive areas of southwestern Australia.

**maquis** Tall vegetation cover of the Mediterranean Basin dominated by evergreen sclerophyllous shrubs and trees.

**matorral** Evergreen sclerophyllous-leaved shrublands that dominate large areas of central Chile.

**Mediterranean-type ecosystem (MTE)** Habitat characterized by mild wet winters and warm dry summers. MTEs occur in California, central Chile, the Mediterranean Basin, the Cape Region of South Africa, and southwestern and South Australia, all at N and S latitudes of 30 to 35°.

**phrygana** Dwarf shrub land of the eastern Mediterranean Basin characterized by a dominance of species with seasonal leaf dimorphism.

**renosterfeld** Evergreen needle-leaved shrubland on richer soils of the Cape Floristic Region of South Africa and dominated by the resinous shrub *Elytropappus rhinocerotis* (Asteraceae); the name literally means rhinoceros bush.

**sclerophyllous** Descriptive of leaves with a leathery texture due to the presence of sclerenchyma with large amounts of lignin and cellulose in their tissues.

**seasonal leaf dimorphism** A phenological trait in which plant species change the morphology of leaves through the year to better adapt them to prevailing temperature and water stress.

---

**MEDITERRANEAN-TYPE ECOSYSTEMS (MTES)** are those habitats located in five regions of the world characterized by mild wet winters and warm dry summers

TABLE I  
Comparative Area, Topographic Heterogeneity, Climatic Heterogeneity, and Estimated Natural Fire Frequency for the Five Mediterranean-type Climate Regions of the World

Region	Area (10 <sup>6</sup> km <sup>2</sup> )	Topographic heterogeneity	Climatic heterogeneity	Natural fire frequency (yrs)
California	0.32	High	Very high	40–60
Central Chile	0.14	Very high	Very high	Fire-free
Mediterranean Basin	2.30	High	High	25–50
Cape Region	0.07	Moderate	High	10–20
Southwestern Australia	0.31	Low	Moderate	10–15

due to the influence of subtropical high-pressure centers. These five regions are located on the western margins of all of the major continental landmasses at latitudes of about 30 to 35° N and S. A variety of individual climate regimes occur within the MTEs under the influence of topography and marine influences. The characteristic form of vegetation in MTEs is a shrubland dominated by evergreen sclerophyllous shrubs, but woodlands are widespread as well in most MTEs.

## I. INTRODUCTION

Mediterranean-climate ecosystems (MTEs) with their characteristic and unique climatic regimes of mild wet winters and warm and dry summers occur in just five regions of the world. These regions are California, central Chile, the Mediterranean Basin, the Cape Region

of South Africa, and southwestern and South Australia (Table I). Biodiversity is particularly notable in the MTE regions for vascular plant species. Although the combined area of these five regions is little more than 2% of the land area of the earth, MTEs are home to approximately 50,000 species of vascular plants (Table II), 20% of the world's total. Nowhere outside of lowland tropical rain forests are there ecosystems with higher regional diversities of species, providing a strong justification for each of these regions being designated as a global "hot spot" of evolution.

The five Mediterranean-type ecosystems of the world share the characteristic climate regime of dry summers and cool wet winters. Typically 90% or more of the annual precipitation falls in the 6 months centered on winter. Mean annual precipitation is as low as 250 mm in coastal areas of the MTEs and reaches to as high as 900 mm at the upper margins of the classic evergreen

TABLE II  
Comparative Species Richness of Vascular Plants and Major Vertebrate Groups for the Five Mediterranean-type Climate Regions of the World

	Vascular plant species	Mammal species	Resident bird species	Reptile species	Amphibian species
California*	4,300	130	357	35	30
Chile**	4,600	99	285	87	42
Mediterranean Basin	25,000	197	366	165	63
Cape Region	8,550	127	288 <sup>#</sup>	109	38
Western Australia	8,000	82	190	144	36 <sup>##</sup>

\*Excluding desert areas.

\*\*All of Chile.

#Excluding seabirds.

##Frogs only.

Some caution must be used in interpreting these numbers because of varying contexts of what constitutes core Mediterranean-climate habitats.

shrub zone. Mediterranean-type climates in the broad sense, however, also include adjacent arid regions that maintain winter rainfall regimes (e.g., the Mojave Desert in California, Atacama Desert in Chile, and Succulent Karoo in South Africa), as well as montane areas with winter rainfall where the majority of precipitation falls as winter snow. Many areas in the Sierra Nevada of California, for example, experience 10 m or more of annual snowfall. Although frosts may occur throughout much of the MTE regions, these are infrequent and relatively mild in lowland areas. Generally there is an upper limit of 3% of the annual hours with temperature below freezing in these regions.

A characteristic ecological feature of MTEs is the widespread prevalence of evergreen shrublands dominated by species with sclerophyllous leaves. These shrublands are called chaparral in California, matorral in Chile, garrigue and maquis in the Mediterranean Basin, fynbos in South Africa, and kwongan or heathlands in southwestern Australia. The dense cover of these shrublands burns readily under dry summer conditions with low humidity, although with differing frequencies characterizing individual MTE regions as described later. Thus, morphological, ecophysiological, and phenological adaptations to post-fire regeneration of these stand through resprouting and fire-stimulated reseedling is characteristic. Adaptations in tolerance of low nutrient availability and low summer water potentials are common in MTE sclerophyll shrubs.

Despite the general characterization of MTE regions as having dominance by evergreen, sclerophyll shrublands, other vegetation forms are also present. Woodlands are widespread in most MTEs, particularly in areas with deeper or richer soils, or as riparian woodlands or gallery forests in wetter sites. Oak woodlands dominated by species of *Quercus* are widespread in California and the Mediterranean Basin, with both evergreen and deciduous species as dominants. These communities can take the form of closed canopy evergreen woodlands grading into shrublands as in live oak woodlands of Southern California and the maquis of Europe, or open savannas of deciduous oaks that are widespread in both regions. Central Chile once had widespread dry sclerophyll and wet sclerophyll woodlands, but the dominance of these has been dramatically reduced by human activities. Evergreen woodlands dominated by *Eucalyptus* are widespread in western Australia. Only the Cape Region of South Africa of all of the MTEs is largely lacking in woodlands, with such communities restricted to scattered stands of relict Afro-montane forest along the southern coast in areas lacking in strong seasonal drought.

The evergreen, sclerophyll shrublands of California and Chile commonly grade off at their arid interior margin and along drier coastal margins to a plant community dominated by drought deciduous shrubs. This community, which may also dominate early successional disturbance sites or arid microsites in evergreen shrublands, is termed coastal sage scrub in California and coastal matorral in Chile. Structurally similar communities with mixed dominance of low evergreen and deciduous shrubs are called phrygana in Greece and batha in the eastern Mediterranean Basin. Deciduous shrubs are largely lacking, however, from similar habitats in the Cape Region of South Africa and southwestern Australia.

There has been a long history of comparative ecological studies between Mediterranean-climate regions. Particularly prominent in this respect have been comparative studies of California chaparral versus Chilean matorral and of South African fynbos versus Australian kwongan. Regular international meetings of scientists representing all five Mediterranean-climate regions have led to a series of edited volumes investigating individual environmental themes across these regions. Such syntheses have included themes of fire, nutrient relationships, ecosystem resilience, water, plant-animal interactions, biodiversity, and landscape disturbance (see the bibliography).

## II. NATURAL DISTURBANCE REGIMES

Although the five Mediterranean-climate regions share many aspects of natural climatic and environmental disturbance regimes, there are significant differences as well. While all five regions experience characteristic summer drought, for example, the magnitude of this drought is particularly severe in California, Chile, and the much of the subarid portion of the Mediterranean Basin where 6 to 8 months or more may pass without measurable rainfall. Extreme drought such as this is rare in South Africa and southwestern Australia where summer months frequently have light showers.

The dense canopies of evergreen sclerophyllous shrubs, which form the major component of vegetation cover over large areas of MTEs, are highly flammable. These structures, combined with summer drought, make fire an important component of natural disturbance regimes. Natural fires, however, are significant ecological events in consuming above-ground vegetation in only four of the five MTEs. Chile, where natural fires are a rare event, is the exception among MTEs. The Mediterranean-climate region of central Chile is

protected from summer storms and lightning moving westward across Argentina by the high Andean Cordillera. The native flora of Chile shows little evidence that fire has been an important ecological disturbance regime in the evolution of life history characteristics.

Natural fire frequencies are quite different among the other four Mediterranean-climate regions (Table I). In South Africa, for example, fynbos vegetation in the Cape Region commonly burns at intervals of 10 to 15 years, while in California natural frequencies are thought to be 40 to 60 years or more.

There are other strong environmental differences in addition to potential drought stress that make the Cape Region of South Africa and southwestern Australia distinct from the other three MTEs. These two areas lie in geologically ancient and stable landscapes, resulting in highly leached and nutrient-poor soils. In contrast, earthquakes, volcanic activity, orogenic uplift, and other dynamic processes create natural disturbance regimes in California, Chile, and the Mediterranean Basin that are absent in South Africa and Australia. Unlike South Africa and southwestern Australia, the younger landscapes of these three regions have experienced tremendous changes in climate regime and landscape structure in Quaternary and even Holocene times, and these changes have had profound impacts on community structure and speciation. The remarkable patterns of speciation in fire-sensitive shrub lineages in the Cape Region of South Africa and southwestern Australia likely has resulted from a combination of relatively mild and stable Quaternary climatic conditions coupled with high fire frequencies in these nutrient-poor habitats.

### III. PATTERNS OF SPECIATION

Although there has been widespread speculation for many years on the causal factors responsible for the high levels of vascular plant biodiversity in MTEs, few clear generalizations have emerged. What is known is that there are conditions promoting the coexistence of seemingly ecologically equivalent shrubs, graminoids, and geophytes in frequently burned shrublands on nutrient-poor soils of fynbos and kwongan communities. Slow growth rates and diverse strategies of post-fire regeneration and reestablishment appear to promote this coexistence. MTEs on less infertile soils and with longer intervals between fires (chaparral, matorral, garrigue) have lower diversity as short-lived species are excluded by rapidly growing shrub dominants. Open woodlands in California and the Mediterranean Basin, however, may have very high local diversity of annuals

and short-lived perennials where grazing maintains open habitat for species establishment.

Regional topographic and climatic heterogeneity *per se*, which might be thought to be logical correlates of comparative diversity between the five MTEs, is a poor predictor of diversity (Table I). Instead, natural selection has operated in a predictable manner to allow for a fine-scale discrimination of habitats and niches under the selective pressures of stable climates, predictably frequent fires, and periodic drought that promote community turnover and diversification. Thus, fynbos in the southwestern Cape Region and kwongan communities in southwestern Australia have evolved species-rich landscapes in topographically homogeneous areas through rapid speciation coupled with low extinction rates.

### IV. BIODIVERSITY IN CALIFORNIA

The political boundaries of the state of California cover an area of  $411 \times 10^3$  km<sup>2</sup>, but the area includes more than the core area of Mediterranean-type climate. These political boundaries include winter rainfall portions of the Sonoran Desert and Mojave Desert, as well as areas of cold desert habitats east of the Sierra Nevada. The California floristic province, as generally defined, excludes these desert areas and adds northwestern Baja California and southern Oregon to the floristic province. Under their definition, the California floristic province covers  $324 \times 10^3$  km<sup>2</sup>. Because of the differences between the political and floristic province boundaries of California, some caution must be used in assessing figures on California biodiversity in the literature.

The geomorphic structure of California is complex and the topographic diversity within the floristic region is very high. Thus, this region covers the Coast Ranges extending north and south along the state, the broad Central Valley, the Sierra Nevada range, and the Transverse Ranges of Southern California. The Coast Ranges reach elevations as high as 2700 m, while Mount Whitney in the Sierra Nevada is the highest point in the continental United States at 4420 m elevation. The Transverse Ranges in Southern California have a number of peaks reaching above 3000 m. The dynamic geologic history of uplift, faulting, and tectonics has produced complex mosaics of soil structure and parent material and sharp climate shifts over the Quaternary with associated glaciation in the high mountains.

The foothill regions throughout most of California are typically dominated by mosaics of evergreen chaparral

ral shrublands and both evergreen and deciduous woodlands with oak species as the typical dominants. These areas commonly receive 400 to 800 mm annual rainfall. Rainfall is strongly centered on the winter months, and 6 months without rain is common. Drier areas along the coast and inland at the transition to desert environments support coastal sage scrub dominated by drought deciduous shrubs and a few species of deeply rooted evergreen sclerophylls. Mountain areas above 1500 m in Northern California and 1800 m in Southern California show a transition to montane conifer forests, subalpine forests, and alpine communities with increasing elevation. Higher rainfall areas along the central and northern coast of California support mixes of conifer and hardwood forests, extending into massive coast redwood forests along the northwestern coast. Mean annual rainfall reaches its highest levels above 2500 mm in this region.

### A. Vascular Plant Species Diversity

California as a political unit contains 4839 vascular plants species. This total includes 99 ferns and fern relatives, 60 gymnosperms (53 conifers), 823 monocots, and 3862 dicots. The largest family in this flora is the Asteraceae with 627 native species, followed by the Fabaceae with 297 species, and the Poaceae with 251 species. The largest five genera make up more than 10% of this total and include *Carex* (131 species, Cyperaceae), *Eriogonum* (112 species, Polygonaceae), *Astragalus* (94 species, Fabaceae), *Phacelia* (93 species, Hydrophyllaceae), and *Lupinus* (71 species, Fabaceae). All of these genera are composed largely of herbaceous perennial and annual species. Notable speciation has also occurred in two shrub lineages, *Arctostaphylos* (Ericaceae) and *Ceanothus* (Rhamnaceae), in response to adaptations to post-fire regeneration. Older published data give a figure of 4452 species for the California floristic province alone.

Endemism is relatively high within the California floristic province. Fifty genera (6.3%) are strictly endemic to this province. If another 14 genera that extend only slightly outside of the region into Arizona or Baja California are included, then 8.1% of the genera are endemic. At the species level, 47.7% of taxa are endemic. This high level of endemism is heavily influenced by the diversity of annual plants that comprise 27.4% of the vascular plants of the California floristic province. For annual species alone, endemism is 65.3%.

The highest species richness of Mediterranean-climate ecosystems in California appears to occur in lightly disturbed grasslands and oak woodlands where

47 to 64 species have been reported in 0.1 ha sites. These levels of diversity at this 0.1 ha scale are also matched in post-fire stands of chaparral where annual plant diversity is very high. Mature chaparral, however, exhibits very low levels of species diversity.

### B. Vertebrate Diversity

The present terrestrial mammal fauna of California includes 160 species, with rodents making up more than half of this total. Among these, however, are 30 species restricted to the desert regions of the state and thus not part of the strict Mediterranean-climate region. That leaves a total of 130 terrestrial mammals native to the shrubland, grassland, woodland, and forest regions of California. An additional five species that once occurred in the state have been extirpated in historical times. These include the grizzly bear, wolf, bison, jaguar (only an occasional visitor in the past), and giant deer mouse. Turnover between habitats (beta diversity) accounts for most of the diversity of mammal faunas, with alpha and gamma diversity relatively low.

Breeding bird species diversity in California is about 357 species. Shore and marine birds make up 39% of this total (139 species). Passerines form the largest group of birds with 145 species (41% of the total). There are 21 species of hawks, vultures, and eagles, 13 species of owls, and 12 species of woodpeckers and flickers. Only two bird species are endemic to California. These are the yellow-billed magpie and endangered California condor. Focusing on passerine birds, the alpha diversity of bird species across landscape gradients peaked in closed woodland and forest habitats, while species turnover between habitats (beta diversity) is greatest in midelevation chaparral.

There are approximately 35 species of amphibians and 68 species of reptiles within the political boundaries of California. For the amphibians, these include 31 newts and salamanders and 14 frogs. The reptiles include 2 turtles and tortoises, 33 lizards, and 33 snake species. However, the desert areas of the state are the habitats for 38 reptiles and amphibian species, thereby reducing the amphibian diversity of the California floristic region to 30 amphibians and 35 reptiles.

## V. BIODIVERSITY IN CHILE

The political boundaries of Chile provide relatively natural biogeographic boundaries because of the topography of desert, mountain, and ocean boundaries. To the north is the hyper-arid Atacama Desert, which reaches

its extreme of conditions in the Tacna-Arica region along the Peruvian border. Here a virtual absence of rainfall separates the Peruvian floristic elements of the desert from the Chilean elements. To the east, the Mediterranean-climate region of central Chile is strongly delineated by the high Cordillera de los Andes. Many peaks in this range reach well above 6000 m and effectively shield Chile from weather fronts moving westward across Argentina. Although major uplift of the Andes began in the mid-Tertiary, at least 14 million years ago, the range continues to be tectonically active today. The elevation of mountain passes along the Cordillera de los Andes in northern and central Chile is too high to allow easy migration of either plants or animals and thus has helped to isolate the flora and fauna of Chile. Only in southern Chile where the Andes are lower has there been easy migration across this range. However, the severe climatic conditions of cold that characterize Patagonia in southern Chile strongly reduce the biological diversity in this area.

Comparisons of species diversity between Chilean organisms and those of other Mediterranean-climate regions deserve some caution in terms of the area included. As with California, the political boundaries of Chile include desert and wet forest ecosystems that are not comparable to core Mediterranean-climate habitats. Most figures for Chilean diversity in the literature are based on political boundaries. An additional issue of political boundaries comes in assessing levels of endemism for the Chilean flora or fauna. Levels of endemism for all groups are much lower if strictly adhering to the political boundaries of Chile rather than to the more natural boundaries of the Chilean/Patagonian biogeographic province.

Much of central Chile, which covers an area of about  $140 \times 10^3 \text{ km}^2$ , shares a physiographic structure parallel to that of California. Moving inland from the Pacific Ocean, there is a coastal range of mountains, a broad central valley, and a high mountain range to the east. The geologically recent Cordillera de la Costa in Chile is relatively tall. West of Santiago at about  $33^\circ\text{S}$  latitude, the major peaks are Cerro Campana (1910 m), Campanita (1510 m), and El Roble (2220 m). These peaks are high enough to intercept moisture from humid southwestern winds, producing woodland areas with significant fog interception and thus improved water relations. Further north in the winter-rainfall desert regions, the coast ranges reach up to 3000 m in the Cordillera Vicuña MacKenna.

The dominant vegetation in central Chile is matorral, an evergreen shrubland similar in general form to chaparral. Along the coast and to the north, this community

grades into a coastal matorral with a greater dominance by drought deciduous shrubs. At higher elevations and on sites with greater water availability, matorral grades into a sclerophyll woodland community, and on the higher parts of the coast range into hygrophilous woodland with species characteristic of the Valdivian forest region of south-central Chile. Much of the central valley of Chile today is dominated by a savanna community termed espinal, with *Acacia caven* as the sole dominant. This community is almost certainly the result of human intervention on landscape processes over the past four centuries. Sclerophyll woodland and matorral would once have covered much of this area.

Unlike California and the Mediterranean Basin, the high mountains of the Andean Cordillera in central Chile do not have a forest zone. The young age of these mountains and lack of soil weathering have produced unstable geological conditions on the west-facing slope of the Andes. Matorral communities on the lower foothills of the Andes give way to a low and scrubby montane matorral community at about 2000 m elevation.

The biological diversity of Chile has an ancient origin that dates back to Gondwanaland. Southern Chile in particular exhibits many broad biogeographical linkages with New Zealand and Australia. Central Chile shows other biogeographical connections with southeastern Brazil, a linkage dating back to mid-Tertiary times before the uplift of the Cordillera de los Andes. Since the Andean uplift, however, the biota of Chile has evolved largely in isolation from other biogeographical regions.

Biosystematic and biogeographical knowledge of the flora and fauna of Chile have increased greatly in recent decades and thus the biodiversity of most groups of vascular plants and vertebrates is relatively well known. These syntheses have been applied more generally to the country as a whole, however, rather than to discrete regional areas. Thus, there is a strong need for more regional studies that evaluate patterns of alpha, beta, and gamma diversity in relation to environmental gradients.

### A. Vascular Plant Diversity

The age and evolutionary isolation of the Chilean flora is clearly indicated by the great number of families that are largely endemic to the Chilean floristic region, which includes adjacent areas of Austral forest in southern Argentina. One family of ferns, the Thyrsopteridaceae, and 9 families of Angiosperms are endemic to Chile. Endemic families entirely restricted to Chile are

the Aextoxicaceae, Gomortegaceae, and Lactoridaceae (restricted to the Juan Fernandez Islands). Other families that are largely Chilean in distribution but cross political boundaries into desert areas of Peru or Austral forests of southern Argentina are the Malesherbiaceae, Nolanaceae, Mizodendraceae, Vivianiaceae, Francoaceae (a segregate from the Saxifragaceae), Heterostylaceae, and Arachnitaceae (sometimes placed in Corsiaceae). Of these endemic families, the Aextoxicaceae and Gomortegaceae, have distributions centered in the Mediterranean-climate regions of central Chile, while the Malesherbiaceae and Nolanaceae are centered in the coastal deserts and adjacent arid montane regions of northern Chile and southern Peru, although one species of the latter reaches the Galapagos Islands.

For continental Chile as a political unit, the vascular plant flora consists of approximately 4600 native species, divided into about 850 native genera and 180 families. This flora includes 124 species of ferns and fern relatives, 13 species of gymnosperms, and about 4500 species of Angiosperms. The flora of central Chile, excluding the desert areas north of La Serena and the forest and moorland areas south of Concepción, is estimated to include about half of this total.

Endemism is high at the generic level in the Chilean flora. Extrapolating from literature, 16% of Chilean genera are strict endemics restricted to the political boundaries of the country. Another 17% of the genera are endemic to Austral regions of Chile and adjacent areas of Argentina. Together, then, a third of the genera are endemic to the Chilean-Patagonian floristic province. There are 22% of the genera with broad South American patterns of distribution and 10% of the genera have Gondwanaland origins with extant species in New Zealand or Australia.

The levels of species-level endemism within groups of the Chilean flora are high, but the values reported vary somewhat depending on whether or not the author is quoting distributions within the strict political boundaries of Chile. A more natural view of endemism would include Andean species or Austral forest and moorland species, which occur within communities of the Chilean-Patagonian biogeographic province that may extend into Argentina. Using the political boundaries of Chile, one authority estimated that 62% of the vascular plant species were endemic, while others estimate about 50 to 55%.

A rich flora high in endemism occurs on the oceanic Juan Fernandez Islands, which lie about 500 km off of the central Chilean coast and contain floristic elements from both the mainland of Chile and Polynesia. The flora includes 361 species of vascular plants, 60% of

which are endemic. In addition, there is an endemic family, the Lactoridaceae.

The 20 largest genera of Chilean vascular plants make up about 30% of the flora. This group is led by *Senecio* (Asteraceae) with 218 species, *Adesmia* (Fabaceae) with 140 species, *Oxalis* (Oxalidaceae) with 111 species, *Calceolaria* (Scrophulariaceae) with 86 species, and *Calandrinia* sensu lato (Portulacaceae) with 70 species. As might be expected, the degree of endemism within the largest families is higher than that for the flora overall. Eleven of the 20 largest families have 80% or more of their species endemic to Chile. These numbers on species diversity and endemism will no doubt change somewhat as these large and difficult genera become better studied. One of the most charismatic of endemic species is the Chilean wine palm, *Jubaea chilensis*. This large palm was once widespread through central Chile but has a more restricted distribution today.

No manual or complete checklist of the flora of central Chile has been produced. For the latitudes of 32° to 36°S, it has been estimated that there are about 2100 species of native vascular plants. Regional floras exist for several areas of central Chile, but these are generally older and incomplete.

Vascular plant species diversity at small scales of habitat in central Chile were studied during the International Biological Program (IBP) in the 1970s. Data were prepared on species diversity in 1 m<sup>2</sup>, 10 m<sup>2</sup>, 100 m<sup>2</sup>, and 0.1 ha plots located at a single site in each of four habitats—coastal matorral, matorral, sclerophyll forest, and espinal. The number of species in the 0.1 ha plots ranged from a high of 114 at the matorral site to only 25 at the espinal site. The sclerophyll forest site (102 species) and coastal matorral site (85 species) were also high in diversity. Excepting the espinal site, all of these values are remarkably high and almost certainly not typical of species diversity in central Chile. More studies of alpha species diversity at these scales are needed.

## B. Mammal Diversity

Mammal diversity in Chile, as in other parts of temperate South America, is low. It has been suggested that severe climatic conditions associated with Pleistocene glacial movements in the Andes may have had a strong impact in reducing the diversity of temperate mammal faunas in South America. Major faunal extinctions were also present in South America at the end of the Pleistocene, as in North America, and these extinctions have been associated with the arrival of early humans. Many ecological niches in temperate South America appear



to be incompletely occupied by mammals in comparison to North America.

Chile has a mammal fauna of 99 native terrestrial species. The largest single group is the Rodentia, which comprise 60% of this total. Next in abundance are the Carnivora with 14%, and the Chiroptera (bats) with 10%. Large mammal species are particularly low in number. These include five species of felids (e.g. puma, Geoffrey's cat, and colo colo), three species of canids (all fox species of *Pseudolopex*), four camelids (guanaco, vicuna, and the domesticated llama and alpaca), and three cervids (huemul, northern huemul, and pudu). Mammal faunas of Chile have been placed into six biogeographical groupings: the summer rainfall Altiplano region of northeastern Chile, the Atacama Desert and adjacent winter rainfall Andes of northern Chile, the Andes of central Chile, the Mediterranean-climate region of central Chile, the Austral forests of southern Chile, and the Patagonian region. In addition to these native mammal faunas, 15 species of terrestrial mammals have become naturalized in Chile. Five of these occur only on the Juan Fernandez Islands where they have had an extreme impact on the structure and composition of the native flora.

Considering Chile in the broad sense to include communities that extend across political boundaries, there are functionally 37 endemic species. Endemism is highest among Chilean mammals, however, in the summer rainfall Altiplano region of northeastern Chile (8 of 27 species) and in Patagonia (10 of 34 species). The Mediterranean-climate region of central Chile is the primary biogeographical range of only 12 endemic species. Observing the strict political boundaries of Chile, only 15 of the 99 species of terrestrial mammals are endemic. These include 13 species of rodents, 1 marsupial, and the Chiloe fox (*Pseudolopex fulvipes*).

On a site basis of comparison for Mediterranean-type shrublands, the small mammal fauna of Chile is also relatively low. IBP studies at matorral, coastal matorral, and dry coastal scrub communities in central Chile found only a total of 10 small mammals. These were nine rodents and a marsupial mouse. A similar gradient of ecological sites in California yielded 17 species.

### C. Bird Diversity

The total bird diversity of Chile is only moderate. Including the oceanic islands of Pascua and Juan Fernandez and the Antarctic territory claimed by Chile, there are reports of 451 native species and 5 introduced species for Chile. These are distributed in 56 families and 222 genera. Of the total native species, 285 are residents,

90 are visitors, and 76 are accidentals. Among the administrative regions of central Chile, the diversity of native bird species is remarkably consistent with a range of 179 to 194 species per region (excluding the smaller metropolitan region of Santiago). There are 10 endemic bird species in Chile, all of them terrestrial. Seven of these occur on the mainland of central Chile, while the remaining three are restricted to the Juan Fernandez Islands.

At a regional scale, the structure and diversity of bird faunas in the Mediterranean-climate region of central Chile have been studied in considerable detail, using structurally parallel communities from the coast into montane habitats of the coast ranges in Chile and California as a comparison. Excluding such guilds as aerial feeders, most parasites, raptors, and nocturnal species, there were only 39 species in the Chilean gradient compared to 69 species in California. Looking more specifically at individual sites, sclerophyll woodlands in Chile supported only 18 bird species compared to 31 species in the structurally similar oak woodlands of southern California. The changes in  $\beta$ -diversity observed along the Chilean and Californian gradients showed a similar pattern, with much higher levels of diversity in California. Chile differed in the opposite direction, however, with much higher  $\gamma$ -diversity than California. When these differences were all merged into assessing the total diversity of resident bird species, Chile and California prove to be almost identical.

The biogeographical isolation of the Chilean bird fauna shows parallels with that of the mammal fauna in the manner in which niches have been filled in unusual manners. Corvids, for example, are absent from Chile. Their role as scavengers has been filled by caracaras.

### D. Reptile and Amphibian Diversity

There are 87 native species of terrestrial reptiles in Chile, divided among 7 families and 18 genera. These include 6 snakes (four genera) and 81 lizards (14 genera). Two of these lizard species (two genera) occur only on Isla Pascua. Seven turtle species (six genera), all marine, are known from the coasts of Chile and its oceanic islands. The diverse lizard fauna of Chile is strongly dominated by the family Tropiuridae, in particular the iguanid genus *Liolaemus* with 53 species. This diverse genus and its evolution have been the focus of a large number of ecological and evolutionary studies. The occurrence of distinctive populations within individual species has led to the designation of a large number of subspecies. Only *Sceloporus* and *Anolis* among New World lizards have a species diversity of this magnitude. Studies in central Chile during the IBP

program found 58 lizard species, a significantly higher diversity than that found along similar gradients in Southern California.

Two lizard genera, *Phrynosaura* with two species and *Velosaura* with two species, are endemic to Chile. However, these genera are only known from the high mountain areas of northern Chile. Endemism at the species level for reptiles in Chile is high, with more than 60% endemic. This endemism is largely concentrated in *Liolaemus* with nearly two-thirds of the species restricted to Chile. There are two endemic species of snakes.

The highest diversity of reptiles in Chile occurs in the northern desert and Andean areas. This region is also the site of the greatest degree of endemism. Within individual administrative regions of central Chile, there are commonly only 10 to 14 species of lizards and 1 to 2 species of snakes. Reptile diversity is low and endemism is totally absent in southern Chile.

The amphibian fauna of Chile exhibits an unusually high level of endemism, in comparison to other vertebrate groups. There are 42 native species in Chile, divided among three families and 13 genera. The highest diversity of these amphibians is located in the forested regions of south-central Chile. There is one additional species of introduced frog (*Xenopus*) that has become naturalized in the area of Santiago. Notable for their absence are salamanders, with only frogs and toad species present.

More than three-quarters (33 species) of these native frogs and toads are endemic to Chile. Many of these endemic species are quite rare and localized in distribution. Ten species, for example, are known only from their type locality. One family, the Rhinodermatidae, is endemic to Chile. It has a single genus with two species whose range is centered in the Austral forests of southern Chile. The range of one species, however, extends into central Chile. Five genera, 12% of the total number, are endemic.

## VI. BIODIVERSITY IN THE MEDITERRANEAN BASIN

The Mediterranean Basin represents the largest area of MTEs in the world, covering a complex landscape with a large amount of topographic and climatic heterogeneity. Covering an area of approximately  $2.3 \times 10^6$  km<sup>2</sup>, it is nearly 10 times greater in size than any other MTE. This area includes more than 20 nations arrayed on both sides of the Mediterranean Sea. While coastal areas are extensive because of the large archipelagos and is-

lands within the Mediterranean, much of this area consists of mountainous terrain with many areas above 2000 m elevation and peaks reaching above 4000 m.

The geographic position of the Mediterranean Basin is also an important factor in understanding the biodiversity of this region. Lying at the juncture of three continental landmasses, it holds a geologic history with dynamic changes associated with plate tectonics, mountain uplift, and active volcanism. In contrast to other MTEs, the great majority of the Mediterranean Basin is underlain by limestone. Local areas of volcanic or siliceous parent material are present, however. Strong climatic shifts that took place during the Plio-Pleistocene period, most notably major glacial episodes, resulted in a telescoping of many communities into the Mediterranean Basin and provided opportunities for geographic isolation and speciation.

The climatic features of the Mediterranean Basin are often used to define this region, but the range of dominant and widespread woody species such as *Quercus ilex* and *Olea europea* also are used as bioindicators of the region. To the north, the Mediterranean-climate region grades into more mesic regions with summer or year-round patterns of rainfall. To the south, the Mediterranean region intergrades with the winter rainfall desert of the northern Sahara. Climates of the Mediterranean Basin are notable for their high interannual variation in both rainfall and temperature extremes.

The large area of the Mediterranean Basin, coupled with its topographic and climatic heterogeneity, makes for complex assemblages of vegetation types. There are extensive woodlands dominated by both evergreen and deciduous species of oak, and evergreen sclerophyllous shrublands of many forms. These shrublands are often differentiated into types depending on the height of the vegetation. Tall sclerophyllous shrublands that may include small evergreen trees are termed maquis. Several species of Mediterranean pine may be present in this community. A middle height shrubland, generally occurring on calcareous substrates, is termed garrigue. Finally, low semiarid evergreen shrublands in the eastern Mediterranean Basin are commonly termed phrygana in Greece and batha in Israel. A long history of human impacts on the natural landscape has strongly impacted community structure and diversity.

### A. Vascular Plant Diversity

The vascular plant flora of the Mediterranean Basin is estimated to include about 25,000 species, making this region the richest among MTEs in total plant diversity. By comparison, the remaining portions of Europe with non-Mediterranean-type climate regimes cover four

times as much area but have only about 6000 vascular plant species. This large flora in the Mediterranean Basin is a broad mixture of species with disparate evolutionary histories and biogeographical origins.

One group of species evolved under subtropical conditions that existed in this region prior to the Quaternary. This group includes such woody plant genera as *Arbutus* and *Calluna* (Ericaceae), *Ceratonia* (Fabaceae), *Chamaerops* (Arecaceae), and *Laurus* (Lauraceae). Another group of taxa represents neo-Mediterranean elements that migrated into the Mediterranean basin after the establishment of a Mediterranean-type climate. Examples of woody genera in this group include *Amelanchier* (Rosaceae), *Clematis* (Ranunculaceae), and *Cistus*, *Halimium*, and *Helianthemum* (Cistaceae). Three groups of nontropical woody elements that evolved after the onset of Mediterranean-type climates have been identified. These groups are a Mediterranean element evolved *in situ* in mountain areas and high in endemism, a desert and cold steppe group of species entering from Africa and the Middle East, and a Holarctic element of species with Eurasian temperate affinities. Endemism is high at the species level in the Mediterranean Basin, with a level of about 50%. No family of vascular plants is strictly endemic to the Mediterranean Basin.

Species richness at the scale of 0.1 ha is often remarkably high in lightly grazed or disturbed Mediterranean woodlands and shrub grasslands, with as many as to 119 to 179 species in such stands. In Israel, with 21 to 29 species in a single 1 m<sup>2</sup>, annual species contributed from 47 to 72% of these 0.1 ha diversities. Protected areas without disturbance and highly disturbed sites support fewer species. This suggests significant coevolution between plants and herbivores.

## B. Vertebrates

Vertebrate faunas of the Mediterranean Basin share with vascular plants the characteristic of multiple biogeographical origins. Dramatic climatic oscillations during the Pleistocene led to a periodic turnovers of Eurasian and African faunal elements and a resulting high beta variety in species diversity.

The present fauna of land mammal species for the Mediterranean Basin numbers about 197 species, of which 25% are endemic. Because of the biogeographical barriers of the Mediterranean Sea and the Saharan Desert, mammal faunas of Mediterranean Europe, the Middle East, and North Africa are somewhat distinct. For North Africa the mammal fauna of the Mediterranean Basin shows its strongest affinities with tropical Africa. A decline in species richness occurred at the end of the

Pleistocene with a combination of sharp climate shifts and human pressures through hunting.

Bird diversity of the Mediterranean Basin includes about 366 breeding species. This number compares with a total of 419 breeding bird species reported for all of Europe. In contrast to mammals, the affinities of bird faunas are more strongly linked to the Asiatic steppes than to tropical Africa. The evolution of these elements of bird faunas can be linked to Eurasian (153 species) and Eremian semiarid habitats (85 species), where Plio-Pleistocene conditions led to ongoing isolation and speciation. Forest birds of boreal origin are widespread and dominant throughout both middle Europe and the Mediterranean Basin. Shrubland bird species characteristic of the region represents only about 12% of the total. Overall, there are 62 endemic bird species in the Mediterranean Basin, 17% of the total.

Reptiles and amphibians of the Mediterranean Basin include 165 and 63 species, respectively, and show distinct holarctic affinities. Much of the endemism within these groups appears to represent archaic lineages that differentiated during the middle Tertiary. Reptile diversity is highest in the eastern Mediterranean Basin and drops steadily moving westward. Species diversity on Mediterranean islands is relatively low. Important reptile groups include lizards of the Lacertidae with 60 species (30% of the world total), snakes of the Viviperidae with 14 species (7.4% of the world total), and tortoises of the Testudinidae with 4 species. Overall, 68% of Mediterranean Basin reptiles are endemic to this region. For amphibian diversity, the pattern is reversed compared with reptiles as the highest levels of diversity are found in the Euro-Mediterranean areas compared to the North African and Middle Eastern portions of the region. Notable groups of amphibians include the Discoglossidae with 10 species (71% of the world total) and the Salamandridae with 19 species (36% of the world total). Endemism for amphibians in the Mediterranean Basin is 59%.

## VII. BIODIVERSITY IN THE CAPE REGION OF SOUTH AFRICA

The Cape Region forms a small area on the southwestern tip of the African continent. It is renowned for its showy and diverse flora that is unlike that of any other area of the world. The landscape of the Cape Region is dominated by steep and strongly folded mountain ranges of quartzitic sandstone that predate the separation of the African continent from Gondwanaland.

These once high mountains have been eroded over the past 200 million years to form low ranges capped by resistant Table Mountain sandstone. Separating the mountains are gentle valleys and undulating plains that are largely underlain by shales with greater nutrient availability. Relatively young Tertiary and Quaternary limestones and sands mantle extensive areas of the coast.

The characteristic vegetation of the Cape Region, particularly on the nutrient-poor quartzite soils, is fynbos. Fynbos is an evergreen shrubland dominated by four major plant morphological groups. These include two shrub groups (the proteoids and ericoids), a sedge-like group (restioids), and geophytes. The proteoids, formed by woody Proteaceae, make up the tallest matrix of the fynbos community and commonly reach to 2 to 4 m in height. The ericoid group gains its name from the Ericaceae but includes more than 3000 species of small-leaved shrubs representing many families. The restioids are primarily members of the Restionaceae, a family with origins in Gondwanaland but its primary diversification in the fynbos. Finally, the Cape Region contains the highest diversity of bulbs and other geophytes in the world, with more than 1500 species. Many types of fynbos have been described, but a simple classification scheme includes proteoid fynbos, ericaceous fynbos, restioid fynbos, asteraceous fynbos, and grassy fynbos.

Another important shrubland vegetation type of the Cape Region is renosterveld (or rhinoceros veld, referring to black rhino that once grazed here). Renosterveld occurs on richer soils with shale parent material and is differentiated by the absence of restioids and minor importance of proteoids. This community once covered more than a quarter of the Cape Region but has now largely been cleared for agriculture and urban expansion.

Woodland and forest communities are surprising rare in the Cape Region. True forests occupy only about 3850 km<sup>2</sup> of moist sites (800–1,200 mm annual rainfall) protected from fire along the southern coast. These forests are low in diversity and represent depauperate outliers of Afro-montane forests of tropical East Africa.

### A. Vascular Plants

The Cape Region contains what is arguably the most unique and diverse flora of any temperate area of the world. Despite the relatively tiny area of this region, phytogeographers have uniformly separated it out from the other parts of Africa and designated it as the Cape Floral Kingdom, one of only six floral

kingdoms that comprise the entire world. This status as a distinct floral kingdom is reinforced by the presence of six endemic families—Geissolomataceae, Grubbiaceae, Penaeaceae, Retziaceae, Roridulaceae, and Stilbaceae. Moreover, there are 193 endemic genera, 19.5% of the total.

Covering an area of only 71,000 km<sup>2</sup>, the Cape Region contains approximately 8550 species of vascular plants. About 7000 of these species grow in fynbos communities. More recently it has been suggested that the winter-rainfall area of the arid Succulent Karoo should be added to the Cape floristic region. Such a change would almost double the number of species present. For the purposes of this review, however, only the Mediterranean-climate shrublands of the Cape Region are considered. The 10 largest genera of the Cape Region account for 20% of the flora. These are *Erica* (Ericaceae, 566 species), *Aspalathus* (Fabaceae, 245 species), *Phyllica* (Rhamnaceae, 133 species), *Agathosma* (Rutaceae, 130 species), *Oxalis* (Oxalidaceae, 129 species), *Pelargonium* (Geraniaceae, 125 species), *Senecio* (Asteraceae, 113 species), *Cliffortia* (Rosaceae, 106 species), *Muraltia* (Polygonaceae, 106 species), and *Ruschia* (Aizoaceae-Mesembryanthema, about 100 species).

Species richness is greatest in the southwestern Cape Region centered around Cape Town. The Cape Peninsula, for example, supports 2256 species (including 90 endemics) in an area of 471 km<sup>2</sup>. Cape Hangklip on the eastern shore of False Bay near Cape Town has 1383 species in 240 km<sup>2</sup>.

Levels of species endemism in the Cape floristic region are among the highest in the world. For the entire region, endemism at the species level is about 68%. The high levels of endemism present in the Cape Region are largely due to the presence of neoendemics rather than paleoendemic species. This dominance of neoendemics is indicated by the predominance of endemic diversity in large, species-rich genera, the widespread presence of sympatric congeners, and the edaphic specialization of many endemics on geologically young substrates. Rather than being a random ecological or phylogenetic assemblage of species, the great majority of endemics are low shrubs killed by fire and dependent on closely dispersed seeds for regeneration. Four families are notably rich in endemics: the Proteaceae, Ericaceae, Rutaceae, and Polygalaceae.

Smaller regional centers of endemism exist within the Cape Region. Dividing the Cape Region up into five floristic zones on the basis of species distributions within seven large families, regional levels of endemism are highest in the southwestern and northwestern Cape (about 50%) and lowest in the eastern Cape and inland

mountain regions with nonseasonal rainfall (18–28%). These patterns of regional endemism have been further demonstrated in studies of distribution of the Proteaceae in the Cape Region. For the entire Cape Region, 99.4% of the 330 species of Proteaceae are endemic. At a regional level, 63% of the Proteaceae in the southwestern region are endemic to that region, compared to only 19% endemism for the coastal mountain and southeastern regions. Point endemism is also widespread involving species that are restricted to highly specific edaphic habitats.

Fynbos species diversity is also extremely high at the alpha-diversity level of small stands. Typical fynbos communities support a mean of about 65 vascular plant species in 0.1 ha, with a range of 31 to 126 species. Renosterveld shrublands have even higher diversities with a mean of 84 species per 0.1 ha and a range 28 to 143 species. Fynbos and renosterveld average about 14 to 17 species in 1 m<sup>2</sup> areas, with a range of 5 to 33 species.

## B. Mammal Diversity

The Cape Region lacks a distinctive mammal fauna. This region contains 127 species of native mammals, with 90 of these being present in the Southwest Cape area. The regional total is less than half of the mammal species occurring within all of South Africa. The largest orders present are the Rodentia and the Carnivora. The rodents are represented by 2 species of mole rats (Bathyergidae), a porcupine (Hystricidae), 2 dormice (Muscardinidae), and at least 21 species of Muridae and Cricetidae. There are 27 species of the Carnivora, ranging from mustelids and civets to larger hyaenas, jackals, and cats. Large browsers and grazers play an important role in this ecosystem in comparison to other MTEs. There are 20 species of Artiodactyla and 5 species of Perissodactyla. Very few of these depend on grazing, however, because of the paucity and poor nutritive value of Cape Region grasses. The Chiroptera is a large group with one fruit bat and 14 species of Microchiroptera in the Southwest Cape. Orders and families present in South Africa but not reaching the Cape Region are hedgehogs (Erinaceidae), pangolins (Pholidota), giraffe (Giraffidae), squirrels (Sciuridae), rock rats (Petromyidae), spring hare (Pedetidae), and cane rats (Thryonomidae).

The abundance and diversity of native mammals was likely always relatively low in fynbos shrublands, unlike the teeming savanna regions of the continent. At the time of European colonization, the highest numbers and diversity of large mammals was present in renoster-

veld or other open communities with better browse. Original large animals that were once common on the renosterveld plains included the bontebok, eland, buffalo, Cape mountain zebra, red hartebeest, and lion.

As fynbos and renosterveld stands have become more limited in extent and more fragmented, it has been increasingly difficult to maintain viable populations of most large mammals in the Cape Region. Widely distributed grazers no longer present in the Southwest Cape Region include eland, buffalo, and hartebeest. Elephants, essential absent from this region today, once grazed commonly along the south coast. Two other grazers, the bluebuck and quagga, are extinct. The absence of large carnivores from many areas today has led to increased populations of small mammals such as rock hyraxes, which may exert a significant impact on plant demography through selective grazing and granivory.

Endemism, as might be expected, is quite low among mammals in the Cape Floristic Region as most of this fauna extends northward or westward into arid or savanna ecosystems. Only 7% of the mammal fauna (nine species) are endemic. These endemic mammals include two bats, three rodents, an insectivore, and three antelope (Cape grysback, bontebok, and the extinct bluebok). The fossorial Cape dune mole rat and burrowing gerbil among the endemic rodents are associated with sandy soil substrates rather than with any specific vegetation type. The colonial behavior and feeding specialization of mole rats on bulbs may be linked to the remarkably high diversity of geophytes in the Cape Floristic Region.

## C. Bird Diversity

Ecological analyses of avian diversity in relation to community structure and function in the Cape Region have shown that species richness in fynbos communities is comparable to that in other MTEs. Thus, the high species richness for vascular plants is a poor predictor of bird species diversity. Fynbos birds generally occupy narrow feeding niches and replace themselves rapidly across gradients of changing fynbos communities. In comparison to bird faunas in California chaparral and Chilean matorral, fynbos birds are more stereotyped to a narrow habitat range and less behaviorally plastic. The nature of low nutrient soils that have led to the evolution of highly sclerophyllous leaves low in nutrient content and high in lignin have made these communities difficult resources for insect herbivores. As a result, the fynbos is low in both the abundance and diversity of insectivorous birds. For the Cape Region avian fauna overall, however, the average density of

birds is quite comparable to that found in California and central Chile in ecologically comparable communities.

Excluding seabirds, there are 288 species of resident birds in the Cape Region, with a notable diversity of Falconiiformes with 22 species. For fynbos communities specifically, there are 101 reported species. The richest habitat for birds is found to be renosterveld with about 77 species.

As with mammals, endemism is low among Cape Region birds. Only 2% (six species) are endemic. Endemic species are largely dietary specialists such as the Cape sugarbird, orange-breasted sunbird, and Protea seedeater that are tied to specific plant resources in the fynbos. The originally direct communication of fynbos habitats with semiarid and savanna shrublands has probably been a factor in limiting the number of endemic fynbos birds. The savanna region of South Africa is far richer in endemic species. At least four of the endemic fynbos birds are characteristic of montane areas or are allied to montane species of East Africa that live in ericaceous shrublands.

#### D. Reptile and Amphibian Diversity

The Cape Region is moderately diverse in reptiles, with 109 species known to occur in this area. Fynbos communities may contain more than 50 species of lizards. The Gekkonidae are the most important group with 18 species. There are 32 species of snakes reported from the Western Cape area. Among snakes, the Colubridae have the highest diversity with 25 species. There are 19 endemic species among the reptiles, 17% of the total. One notable endemic to the southwestern Cape Region is *Psammobates geometricus*, one of the rarest tortoises in the world. The life history of this species seems to have evolved to adapt to fynbos fire cycles, with hatchlings appearing in late autumn after the danger of summer fires is past.

Amphibians are relatively low in diversity in the Cape Region with 38 native species. The largest single group of amphibians is the Ranidae with 13 species. Compared to other vertebrate groups in the Cape Region, endemism among amphibians is relatively high at 50% (19 species), and much of this endemism is centered in the Southwest Cape area. One of the most interest endemics among amphibians is the arum lily frog, *Hyperolius horstocki*, which lives in the flowers of the common arum lily *Zantedeschia aethiopica*.

#### E. Invertebrate Diversity

Relatively little is known about the diversity of invertebrates in the Cape Region, but what data exists suggest

that very high levels of endemism exist in many groups. Among 234 species of butterflies known from the region, 31% are endemic. A regional study of invertebrates on the Cape Peninsula reported a large number of endemic species. The high diversity and rates among insects in the Cape Region is not surprising in view of the important ecological links between insects and plants. Specific pollination mechanisms have been involved in maintaining many sympatric species populations. Ants perform an important role as seed dispersers. About 20% of fynbos species have seeds dispersed by ants, a pattern seen only in western Australia among other MTEs.

### VIII. BIODIVERSITY IN SOUTHWESTERN AUSTRALIA

Core Mediterranean-climate conditions are present in southwestern Western Australia and in South Australia around Adelaide. The core southwestern floristic province of western Australia is approximately  $310 \times 10^3$  km<sup>2</sup> in area, similar in size to the California floristic province. A broader region with transitional rainfall patterns with biseasonal distribution is often included as well, almost doubling the core MTE area. While mean annual rainfall is commonly 350 to 800 mm over much of southwestern Australia, it reaches as high as 1500 mm at the extreme southwestern corner of western Australia and drops to a low of about 250 mm at the eastern edge of the Mediterranean-climate region in a transition to arid communities. Topographic heterogeneity is very low over this region, with elevations almost entirely below 1000 m. Much of the region is a low, laterized plateau dissected into broad valleys with deep *in situ* weathering. Soils of southwestern Australia, as in the Cape Region of South Africa, are generally very old, highly weathered, acidic, and low in nutrient availability.

For the true Mediterranean-climate regions, the highest rainfall zones with 800 to 1200 mm of annual rainfall support evergreen forests and woodlands dominated by *Eucalyptus marginata* (jarrah), *E. calophylla* (marri), *E. diversicolor* (karri), and *E. gomphocephala* (tuart). Low *Banksia* woodlands and coastal heath are also present. At intermediate rainfall regimes of 300 to 800 mm, the dominant vegetation is a mosaic of low woodlands (mallees) and heathland communities termed kwongan. The vegetation and dominant species are finely tuned to small changes in edaphic conditions that influence nutrient and water availability. Human impacts on these ecosystems have been severe over the past century.

High levels of species diversity and endemism characterize the vascular plant flora, with only moderate diversity and endemism present in most vertebrate groups. The high levels of vascular plant diversity has been related to (a) the development of a complex mosaic of landforms and soils during the Tertiary and Quaternary, (b) the geologic history of oscillating moisture regimes through the Quaternary in the absence of glaciation, (c) isolation of southwestern Australia from the east by the arid Nullarbor Plain, and (d) interactions of gene pools from both paleotropical and temperate assemblages.

### A. Vascular Plant Diversity

The South West Botanical Province, which includes the broadly defined Mediterranean-climate regions of western Australia, is estimated to include approximately 8000 species of vascular plants. Thus, this region overall is similar in plant species diversity to the Cape Region of South Africa. The core Mediterranean-climate area of southwestern Australia includes about 3611 species, with 2540 species in kwongan habitats. Woody perennials in four families—the Myrtaceae, Proteaceae, Fabaceae, and Epacridaceae—dominate the flora. Much of the high species diversity within these families is due to extensive adaptive radiation within a few large genera. Large genera for the region include *Acacia* (400+ species), *Eucalyptus* (300+ species), *Grevillea* (200+ species), *Stylidium* and *Melaleuca* (150+ species each), and *Hakea* and *Caldenia* (100+ species each).

Nodes of unusual species diversity are present along the south coast of western Australia (Stirling Range, Fitzgerald River area) and in the sandplains north of Perth (Mount Lesueur area). As in the fynbos of South Africa, large numbers of endemics with highly local patterns of distribution also characterize kwongan. Local scale plant diversity reaches almost as high as areas in the western Cape Region of South Africa. Vascular plant diversity in a sample of 0.1 ha stands of heathland in southwest Australia exhibited a range from 43 to 103 species, while jarrah forests and mallee stands had a lower range of 17 to 55 species. As many as 48 species in 1 m<sup>2</sup>, and 110 species in 100 m<sup>2</sup> quadrats, have been found at Mount Lesueur, with a high turnover as species move to comparable nearby areas.

At the regional level, southwestern Australia exhibits major differences in centers of highest diversity among the most important woody genera. Some genera—as, for example, *Banksia* (Proteaceae), *Adenanthos* (Proteaceae), *Leucopogon* (Epacridaceae), and *Eucalyptus* (Myrtaceae)—are most speciose near the south coast

or in southern kwongan and mallee communities. Other large genera have their highest diversity in northern kwongan—*Grevillea* (Proteaceae), *Conostylis* (Haemodraceae) and *Lechenaultia* (Goodeniaceae). Finally, a large group of genera show bimodal patterns of diversity reflecting the both nodes of high species diversity in northern and southern kwongan—*Calothamnus* (Myrtaceae), *Melaleuca* (Myrtaceae), *Hakea* (Proteaceae), *Darwinia* (Myrtaceae), and *Dryandra* (Proteaceae). Two other genera, *Acacia* (Fabaceae) and *Verticordia* (Myrtaceae), are most diverse in the inland transition area of rainfall.

Three families are endemic to southwestern Australia: Cephalotaceae, Phylodraceae and Emblingiaceae. The latter of these is restricted to kwongan habitats. Endemism is notably high at the generic level, 103 endemic genera in southwestern Australia, and 35 of these are restricted to kwongan habitats. At the species level, endemism for the southwestern botanical province has been estimated to be about 48%.

### B. Vertebrate Diversity

The vertebrate fauna of Australian MTEs shows no unusual level of high diversity of endemism. The majority of the fauna in these MTEs is made up of populations of more typically arid or mesic habitat species whose ranges extend into southwestern or southern Australia. Thus, the majority of the vertebrate appears to have relatively broad ecological niches rather than specialized requirements unique to the MTEs.

For mammals, the Mediterranean-climate Australian fauna of 82 species (which represents one-third of the Australian mammal fauna) has only 16 species (20%) endemic to MTEs. Interpreting the patterns of distribution of large mammals is made difficult, however, by the strong impact that both aboriginal and European populations have had on this fauna. The extinction of a large megafauna in the late Quaternary left the Australian continent without large grazers or predators.

The low level of endemism among southwestern and southern Australian birds is particularly evident. Of the 190 species in western Australia, only nine are endemic to Mediterranean-climate regions.

Endemism is moderate among reptiles. For snakes, 25% of the 43 species native to Mediterranean-climate Australia are endemic. The level of endemism is slightly higher for lizards, with 29% of the 144 resident species endemic. The level of endemism for terrestrial vertebrates in southwestern Australia is highest for amphibians. Among the 36 frog species in this region, 64% are endemic.

## IX. CONCLUSIONS

Mediterranean-type ecosystems provide many opportunities for comparative studies of the controlling factors in the evolution of biodiversity. These five regions share common characteristics of climatic regime, with an independent evolution of plant and animal species adapted to these conditions within each region. Thus, they provide a natural ecosystem experiment with five independent replications. The value of comparative studies then between these regions lies not only with their similarities but with subtle differences in climatic conditions, topographic diversity, evolutionary history, and human impacts that have led to the patterns of diversity that we see today.

There has been a long history of comparative ecosystem studies between California and Chile, and between South Africa and southwestern Australia, and these have traditionally devoted some attention to comparing and contrasting how natural processes operate. The remarkable biodiversity of MTEs, together with the large numbers of rare and endangered species in these regions, gives a special significance to expanding studies of these regions to better understand the evolution of diversity, particularly by vascular plants. Serious threats of habitat transformation and degradation today make it critical that there be a better understanding of the conservation biology and sustainable resource management in all five MTEs.

### See Also the Following Articles

AUSTRALIA, ECOSYSTEMS OF • FIRES, ECOLOGICAL EFFECTS OF • NEAR EAST ECOSYSTEMS, ANIMAL DIVERSITY • NEAR EAST ECOSYSTEMS, PLANT DIVERSITY

### Bibliography

- Arianoutsou, M., and Groves, R. H. (Eds.) (1994). *Plant-Animal Interactions in Mediterranean-Type Ecosystems*. Kluwer, Dordrecht.
- Blondel, J., and Aronson, J. (1999). *Biology and Wildlife of the Mediterranean Region*. Oxford University Press, Oxford.
- Cody, M. L., and Mooney, H. A. (1978). Convergence versus nonconvergence in mediterranean-climate ecosystems. *Annual Review of Ecology and Systematics* 9, 265–321.
- Cowling, R. M. (Ed.) (1992). *The Ecology of Fynbos: Nutrients, Fire and Diversity*. Oxford University Press, Cape Town.
- Cowling, R. M., Rundel, P. W., Lamont, B. B., Arroyo, M. K., and Arianoutsou, M. (1996). Plant diversity in mediterranean-climate regions. *Trends in Ecology and Evolution*, 11, 352–360.
- Dallman, P. R. (1998). *Plant Life in the World's Mediterranean Climates*. University of California Press, Berkeley, CA.
- Davis, G. W., and Richardson, D. M. (Eds.) (1995). *Mediterranean-Type Ecosystems: The Function of Biodiversity*. Springer-Verlag, Berlin.
- di Castri, F., Goodall, D. W., and Specht, R. L. (Eds.) (1981). *Mediterranean-Type Shrublands*. Elsevier, Amsterdam.
- di Castri, F., and Mooney, H. A. (Eds.) (1973). *Mediterranean-Type Ecosystems: Origin and Structure*. Springer-Verlag, New York.
- Gomez-Campo, C. (Ed.) (1985). *Plant Conservation in the Mediterranean Area*. Junk, Dordrecht.
- Groves, R. H., and di Castri, F. (Eds.) (1991). *Biogeography of Mediterranean Invasions*. Cambridge University Press, Cambridge.
- Hobbs, R. J. (Ed.) (1992). *Biodiversity of Mediterranean Ecosystems in Australia*. Surrey Beatty & Sons, Chipping Norton, Australia.
- Keeley, S. (Ed.) (1989). *The California Chaparral: Paradigms Reexamined*. Los Angeles County Museum of Natural History, Los Angeles.
- Kruger, F. J., Mitchell, D. T., and Jarvis, J. U. M. (Eds.) (1983). *Mediterranean-Type Ecosystems: The Role of Nutrients*. Springer-Verlag, Berlin.
- Margaris, N. S., and Mooney, H. A. (Eds.) (1981). *Components of Productivity of Mediterranean-Climate Regions: Basic and Applied Aspects*. Junk, The Hague.
- Marticorena, C., and Quezada, M. (1985). Catálogo de la flora vascular de Chile. *Gayana* 42, 1–157.
- Miller, P. C. (Ed.) (1981). *Resource Use by Chaparral and Matorral: A Comparison of Vegetation Function in Two Mediterranean Type Ecosystems*. Springer-Verlag, Berlin.
- Moreno, J. M., and Oechel, W. C. (Eds.) (1994). *The Role of Fire in Mediterranean-Type Ecosystems*. Springer-Verlag, New York.
- Noble, J. C., and Bradstock, R. A. (Eds.) (1989). *Mediterranean Landscapes in Australia: Mallee Ecosystems and Their Management*. CSIRO, Melbourne.
- Pate, J. S., and Beard, J. S. (Eds.) (1984). *Kwongan: Plant Life of the Sand Plains*. University of Western Australia Press, Nedlands.
- Polunin, O., and Walters, M. (1985). *A Guide to the Vegetation of Britain and Europe*. Oxford University Press, Oxford.
- Raven, P. H., and Axelrod, D. I. (1978). Origin and relationships of the California flora. *University of California Publications in Botany* 72, 1–134.
- Roy, J., Aronson, J., and di Castri, F. (Eds.) (1995). *Time Scales of Biological Responses to Water Constraints: The Case of Mediterranean Biota*. SPB Academic Publishing, Amsterdam.
- Rundel, P. W. (1981). The matorral zone of central Chile. In *Mediterranean-Type Shrublands* (F. di Castri, D. W. Goodall, and R. L. Specht, Eds.). Elsevier, Amsterdam.
- Rundel, P. W. (1999). Disturbance in Mediterranean-climate shrublands and woodlands. In *Ecosystems of Disturbed Ground* (L. Walker, Ed.). Elsevier, Amsterdam.
- Rundel, P. W., Montenegro, G., and Jaksic, F. (Eds.) (1998). *Land-scape Disturbance and Biodiversity in Mediterranean-Type Ecosystems*. Springer-Verlag, Berlin.
- Simonetti, J. A., Arroyo, M. T. K., Sportorno, A. E., and Lozada, E. (Eds.) (1995). *Diversidad Biológica de Chile*. CONICYT, Santiago.
- Specht, R. L. (Ed.) (1988). *Mediterranean-Type Ecosystems: A Data Source Book*. Kluwer, Dordrecht.
- Tenhunen, J. D., Catarino, F. M., Lange, O. L., and Oechel, W. C. (Eds.) (1987). *Plant Response to Stress: Functional Analysis in Mediterranean Ecosystems*. Springer-Verlag, Berlin.
- Trabaud, L., and Prodon, R. (Eds.) (1993). *Fire in Mediterranean Ecosystems*. Commission of European Communities, Brussels.
- van Wilgen, B. W., Richardson, D. M., Kruger, F. J., and van Hensbergen, H. J. (Eds.) (1992). *Fire in South African Mountain Fynbos: Ecosystem, Community and Species Response at Swartsboskloof*. Springer-Verlag, Berlin.







# METAPOPULATIONS

Peter Chesson  
*University of California, Davis*

---

- I. Introduction
  - II. Strict Metapopulations
  - III. Quantitative Effects of Spatial Variation
  - IV. Lessons from Metapopulation Theory
- 

## GLOSSARY

- local population** That part of the population of a species found in a particular habitat patch.
- metapopulation** The collection of local populations in a region.
- patch** An area of suitable habitat for a particular species or particular collection of species, ideally bounded by unsuitable habitat or habitat with different physical properties. Normally, it is one of many such areas in a region.
- spatially structured population** A population whose reproductive and survival rates vary over the region that it inhabits, and whose members stay long enough in a locality to experience the local reproductive and survival rates.
- strict metapopulation** (Also called a classical metapopulation.) A metapopulation satisfying the following conditions: (i) Local populations are partially isolated from one another and are frequently capable of sustaining themselves for several to many generations in the absence of immigration from other local populations, (ii) local population extinction occurs on a timescale of several to many generations, and (iii) migration between local populations leads to

reestablishment of local populations following local extinction.

---

*THE CONCEPT OF A METAPOPULATION* has its beginnings in the suggestion of Andrewartha and Birch and others that some, if not most, local populations of organisms in nature frequently go extinct and are reestablished later by immigration from surrounding areas (Hanski, 1999). The metapopulation—the collection of all local populations—only persists if local extinction is balanced by recolonization. How this balance is achieved is an important focus of metapopulation theory.

## I. INTRODUCTION

Because many human activities fragment natural habitats, humans may artificially create metapopulations or may decrease the density of local populations within existing metapopulations and may put the continued survival of natural populations at risk. Hence, metapopulations are a major topic in conservation biology. Regardless of whether or not a natural population is distributed over easily defined patches of habitat, on which distinct local populations may be recognized, essentially all natural populations are patchily distributed in space. Most ecologists believe that patchiness in space and time has functional roles in population dynamics, i.e., in the manner in which population densities change over time. Most important, population

dynamics at the regional or metapopulation scale are affected by the patchiness of the populations and their physical environment at lower spatial scales.

One very obvious way that patchiness is important is through mate finding. A sparse population evenly distributed over an area may have a low reproductive rate because males and females encounter each other too infrequently for many eggs to be fertilized. Patchiness can be a solution to this problem. High local concentrations or aggregations of a species solve the problem of low encounter rates of males and females consistent with a low average density. Similar effects occur from predation. Individuals may find safety from predators in a group but be vulnerable when isolated. Such effects of sparse populations are called Allee effects and are potentially one reason why patchiness in space is important. However, disadvantages from clumped distributions in space might also occur, for example, due to competition. Clumps of individuals in space deplete resources locally and reduce population growth rates compared to what would be experienced if the population were distributed more evenly. A similar effect results if predators are attracted to clumps of prey. Rather than prey finding safety in numbers, it is possible that they may be more vulnerable in clumps if predators increase their numbers at these locations.

These explanations for the importance of patchy distributions involve interactions between organisms in the same species and therefore involve density dependence—dependence of the probability that an individual survives, or dependence of its reproductive rate, on population density. The most striking theoretical predictions for the effects of patchiness, however, are for interactions between species for which one or more is patchily distributed. Major effects of patchiness on population dynamics also occur from spatial and temporal variation in the physical environment. Naturally, if species are concentrated in better localities, they will have higher growth rates in comparison with random distributions, but such environmental variation has its most striking effects by altering the interactions between organisms, locally in space, and hence by altering the collective outcome of those interactions at the level of regional populations. These ideas are all studied within the context of metapopulations but extend to spatially structured populations generally.

## II. STRICT METAPOPOPULATIONS

The idea of a metapopulation is most commonly invoked when patchiness in space is so extreme that the

regional population splits into local populations, each of which is too small to persist indefinitely (Hanski and Gilpin, 1997; Hanski, 1999). Local populations go extinct but may be recolonized by migration from other local populations. In the strict sense of a metapopulation, local populations are sustained primarily by reproduction of resident individuals but may be subsidized by infrequent immigration from other local populations. The role of immigration, however, is seen mainly as allowing recolonization of a local population after it has become extinct. Local extinction may occur for any of a variety of reasons, as discussed later, but is normally assumed to occur asynchronously in different local populations, or the metapopulation as a whole would be lost. Although nearly all organisms are patchy in space, there is some question regarding how frequently all the previously mentioned characteristics of a strict metapopulation are satisfied (Harrison and Taylor, 1997). Thus, in considering the results of metapopulation theory it is important to distinguish those features applicable only to strict metapopulations and those that apply more generally to spatially patchy populations.

### A. Single Species Considerations

The main interest in metapopulations, strictly defined, is with colonization and extinction dynamics. The variable of prime concern is the proportion,  $p$ , of patches (places where a local population potentially could exist) that are occupied by the species. Two parameters are involved with the dynamics of this occupancy fraction: the extinction rate of local populations ( $\epsilon$ ) and the rate  $c$  at which empty patches are recolonized per fraction occupied. In essence, the fraction of local populations going extinct in one unit of time is  $\epsilon$ . The probability of an extinct population being recolonized in one unit of time is  $cp$  because  $c$  is measured proportional to the fraction of occupied patches and incorporates the idea that the probability of recolonization should depend on the fraction of other patches that are occupied. Thus, the change in the fraction occupied with time may be expressed as the following differential equation, called the Levins equation:

$$\frac{dp}{dt} = cp(1 - p) - \epsilon p.$$

If the rates  $\epsilon$  and  $c$  remain constant with time, it follows that in a metapopulation with a large number of local populations, the fraction occupied will reach an approximate equilibrium at the value  $1 - \epsilon/c$ , provided  $c > \epsilon$ , because at this value extinction is balanced with

recolonization. If  $\varepsilon$  is greater than  $c$ , local extinction exceeds local colonization and the metapopulation as a whole goes extinct. Conversely,  $c$  must exceed  $\varepsilon$  for the metapopulation to persist. This result is often called the threshold condition, and it is of major concern for applications of metapopulation theory in conservation. Because actual metapopulations in nature never have infinitely many local populations, the system is never exactly at the equilibrium  $1 - \varepsilon/c$  but rather fluctuates about this value. In metapopulations with only a few local populations there is a danger that all local populations will simultaneously go extinct, causing the extinction of the metapopulation even though  $c > \varepsilon$ .

On what biological features do these colonization and extinction rates depend? Consider extinction. Local extinction can occur in many ways, some of which stem directly from the activities of humans in habitat destruction or modification and hunting; however, metapopulation models are concerned with the case in which local extinction is a natural and repeated phenomenon, whose rate may be influenced by human activities but whose presence is not. Thus, local populations may become extinct by chance because of small population size. The phenomenon of demographic stochasticity refers to independent chance events in the lives of individual organisms. It is impossible to predict how long any individual will live and how many offspring it will have, and these effects summed over individuals in a population cause it to fluctuate. In a small local population, there is always a definite probability of below replacement reproduction in any year; and in a small population a chance run of such years can lead to extinction. These probabilities decrease dramatically as population size increases, but there are other factors that can also lead to local population crashes, including unstable population dynamics and disturbance.

Two different forms of disturbance are commonly considered. An environmental disturbance is caused by an environmental event such as a storm or fire, and a biological disturbance is caused by the invasion of predators or disease. In the strict metapopulation scenario, the interest is in disturbance events that lead to extinction of the local population either coincident with the disturbance or shortly after, following weakening of the population by disturbance. At any time, disturbances must strike patchily in space to cause extinction of only a proportion of local populations, or the metapopulation as a whole would be lost. It is very common for environmental disturbance agents, such as fires and storms, to distribute mortality very patchily in space even though they may affect a large area in a short period of time.

In a strict metapopulation, to which Levins equation applies, recolonization of a patch is from immigration from occupied patches. Many organisms may have means of recovering from disturbance by regrowth from seeds, roots, resting eggs, or spores within a recently disturbed patch, but with a strict metapopulation such populations have not gone extinct. They have merely died back and have been regenerated from dormant stages in the life cycle or belowground parts. The strict metapopulation idea that colonization occurs from external sources is the reason for the formula  $cp$  for the recolonization probability. Indeed, the rate of regeneration from sources within a patch should not depend on  $p$ . Confusion of regeneration with recolonization could potentially lead to serious errors in calculations and inferences.

Recolonization can be from mobile individuals, seeds, spores, or other propagules that arrive from external sources in sufficient quantity to reestablish a local population. The value of  $c$  will be affected by a variety of factors, such as the spacing of local populations. With large distances between local populations, many propagules may perish after leaving their source population before arrival in suitable habitat. The sizes of local populations will also influence  $c$  because larger local populations should be expected to send out more propagules. The size of the metapopulation, in terms of the number of local populations, is also important. Because the number of habitat patches is always finite, the fraction of patches occupied is not at the equilibrium value  $1 - \varepsilon/c$  but rather fluctuates about this value; the smaller the number of habitat patches, the larger these fluctuations. If the number of habitat patches is too small, there is a danger of simultaneous extinction of all local populations eliminating the metapopulation even if  $c > \varepsilon$ .

Habitat fragmentation by human activities may lead to a metapopulation structure of local populations even if this was not the preexisting situation for the natural population of these organisms. With continuing habitat destruction,  $c$  will decline due to increased spacing and smaller size of suitable patches, and there is the danger that  $c$  will decrease below  $\varepsilon$ , especially because  $\varepsilon$  may increase as suitable patches decrease in size and support smaller local populations. Thus, even though there is habitat capable of sustaining local populations, the danger exists that local populations may be either small or so vulnerable because of other habitat alterations that the rate of local extinction exceeds local recolonization, which would lead to loss of the metapopulation altogether. Alternatively, if the organisms are not able to disperse between habitat patches, habitat fragmentation

is accompanied by zero colonization rates, and the metapopulation decreases to extinction at the rate  $\varepsilon$  the moment it is created.

The simple destruction of habitat reduces natural population sizes, putting species at risk independently of these metapopulation considerations. However, metapopulation theory emphasizes that the viability of populations in the remaining habitat may be seriously affected by its connectedness. A final feature of habitat destruction is that by reducing the areal extent of a population, the probability that extinction events occur simultaneously in different parts of the metapopulation increases. For example, fire might sweep through the whole area occupied by the metapopulation, or climate fluctuations might have severe effects in all parts of the metapopulation, endangering the metapopulation as a whole.

So far, the rates  $c$  and  $\varepsilon$  have been treated as though they are independent of  $p$ , the fraction occupied. This need not be the case. It has been suggested, for example, that extinction rates may decrease as the fraction occupied increases because extant local populations may tend to be larger and less vulnerable to extinction when they are subsidized by immigration. This could be especially important in the recovery of local populations after chance population declines, thus staving off extinction. As such, this phenomenon is called the rescue effect. If the rescue effect is required for  $c$  to exceed  $\varepsilon$ , the metapopulation as a whole may be especially vulnerable to extinction due to chance fluctuations in the fraction occupied; if this fraction becomes too low,  $\varepsilon$  will exceed  $c$ , promoting further decline and total extinction of the metapopulation. There are many other characteristics and complications that one can consider in single-species metapopulations, such as variation in the sizes of local populations, their distances from each other, and variation in habitat quality—features that are discussed by Hanski and Gilpin (1997) and Hanski (1999).

## B. Multispecies Considerations: Predators and Parasitoids

Extinction and colonization rates may depend on other organisms. Indeed, as mentioned previously, as biological disturbance, local extinction may be a consequence of invasion of the local population by predators or disease. In this case, extinction rates vary with the abundance of predators within the metapopulation, which in turn depend on the abundance of prey. It is possible that only the prey has a metapopulation structure be-

cause predators may be more mobile, occupying all of the landscape (Harrison and Taylor, 1997). Their density must be taken into account in determining the extinction rate of the prey, but otherwise their inclusion does not result in major differences in the understanding of metapopulation dynamics, except possibly in terms of the effects of habitat destruction, as discussed later.

If both predator and prey have the same metapopulation structure, a metacommunity results. Simple predator–prey models of isolated local communities are often unstable, causing large fluctuations in their abundances and potentially leading to extinction of the prey or the predator (see Fig. 7). A metacommunity structure has long been suggested as one means by which predator–prey interactions may be stabilized in nature. The interaction between the two locally in space remains unstable: The prey is driven extinct by the predator, but provided the colonization rate  $c$  for the prey exceeds its extinction rate due to predation and other causes, it is able to persist in a metapopulation. Similarly, although the predator depends on the prey, whose local extinction leads to local predator extinction, provided the predator can find patches sufficiently well it can also persist in the system. In this way, it is possible for predator and prey to have a stable equilibrium at the metapopulation level despite unstable dynamics on the local scale (Nee *et al.*, 1997). This metacommunity structure may have important conservation implications. Habitat destruction may more strongly affect the predator than the prey because prey numbers in a region may be buffered at first by reduction in predator numbers, with the effects of habitat destruction on prey being stronger after the predator has been eliminated (Nee *et al.*, 1997).

The potential for metacommunity structure for predator and prey is also of great interest in the study of biological control of insect pests and introduced weeds. Commonly, in biological control, one seeks as a control agent a predator or parasitoid (a parasite that invariably kills parasitized hosts) that is specialized on the pest organism and causes sufficient damage to the pest to reduce its numbers greatly. Such a control agent, however, may have an unstable interaction with its host, leading to large oscillations in agent and pest densities, which is an undesirable outcome. However, there are many successful examples of biological control in which the pest is maintained at relatively low and stable numbers. One possible explanation for such control is that local population dynamics are indeed unstable, but regional dynamics are stable because they are metapopulation dynamics (Briggs *et al.*, 1999).

### C. Multispecies Considerations: Competition

One organism may cause another to go locally extinct not because it is a predator of that organism but because it is a competitor. Many years ago, G. E. Hutchinson and J. G. Skellam suggested that inferior competitors may persist regionally by being good colonizers of newly vacant habitat, where they persist only until their competitors find that habitat and drive them extinct. They termed such good colonizing inferior competitors “fugitive species” (Hanski, 1999). An essential component is a trade-off between dispersal ability and competitive ability: If the better competitors were not poorer colonizers, they would arrive too quickly in vacant habitat, giving insufficient opportunity for exploitation of that habitat by fugitives.

The fugitive species idea has been extended in recent years to consider the possibility that a suite of competitors in a system might have a competitive hierarchy, i.e., be strictly ranked in competitive ability with better competitors eliminating poor competitors whenever they occur in the same patch (Hastings, 1980). Such a competitive hierarchy prevents long-term coexistence in any patch, but if all local communities are ephemeral because disturbance takes place locally in space and removes whatever species happen to be there, then every so often any given local community becomes available for colonization by even inferior competitors. Colonization events in the presence of a competitive hierarchy drive succession (a change in species composition in a particular direction) within a local community toward domination by the best competitor. Assuming that inferior competitors have higher dispersal rates, then they have a higher probability, per unit occupancy (per unit  $p$ ), of arriving in vacant habitat. Any species could actually arrive first, but inferior competitors have an arrival advantage. Any species arriving at a site that is a superior competitor to the current occupant takes that site. The local community shifts progressively in the direction of the best competitor in the system until disturbance occurs, interrupting the process and starting it over again.

Because both extinction and colonization occur patchily in space and time, the metacommunity consists of a mosaic of local communities in different stages of succession (Fig. 1). Each local community may have low species diversity because only the best competitor in the local community persists there for long. However, when considered over space, the metacommunity has high species diversity due to the fact that different local communities are in different successional states and

have different species. It is possible for the metacommunity to be in equilibrium and maintain this high species diversity regionally. This equilibrium is dependent on the inverse ranking of competitive ability and colonizing ability, but it is not sufficient for an inferior competitor simply to be a better colonizer: It has to be better by more than a critical amount that depends on the colonizing abilities of superior competitors and the disturbance frequency (Hastings, 1980).

The disturbance frequency has a particularly important role in the maintenance of regional diversity. A higher disturbance frequency means that fewer patches have the best competitor and there are more patches available for other species. At too low a disturbance frequency, lower ranked species do not have sufficient opportunity for colonization and therefore may not persist regionally: Their  $\epsilon$  values may exceed their  $c$  values. However, at higher disturbance frequencies, lower ranked species may persist and coexist regionally with higher ranked species. If the disturbance frequency is too high, the best competitors, which are by assumption poorer colonizers, have too high an extinction rate,  $\epsilon$ , relative to their colonization rate,  $c$ , and disappear from the system. At extreme levels, only the very best colonizers can persist. It follows that with such competitive hierarchies, both high and low disturbance frequencies lead to low regional diversity. Diversity is maximized at some intermediate value of the disturbance frequency—an idea that is often referred to as the intermediate disturbance hypothesis (Connell, 1978). Disturbances here can be of the physical sort or of the biological sort provided the predators or other agents causing disturbance are generalists and therefore are dependent for their own persistence on all species in the system, not just particular species which they may control separately from other species. Specialist mortality agents may be important in the maintenance of diversity (Chesson and Huntly, 1997), but this idea is not included in the intermediate disturbance hypothesis.

The critical feature of the previously discussed mechanism of diversity maintenance is a metacommunity that exists as a successional mosaic, i.e., the system is diverse because local communities range over a variety of successional stages having different species that provide colonists of other local communities. Colonization moves succession along or reestablishes local communities following local disturbance (Chesson and Huntly, 1997). This successional mosaic model has been suggested to work with local populations consisting simply of single individuals, for example, single herbaceous plants in a meadow (Lehman and Tilman,

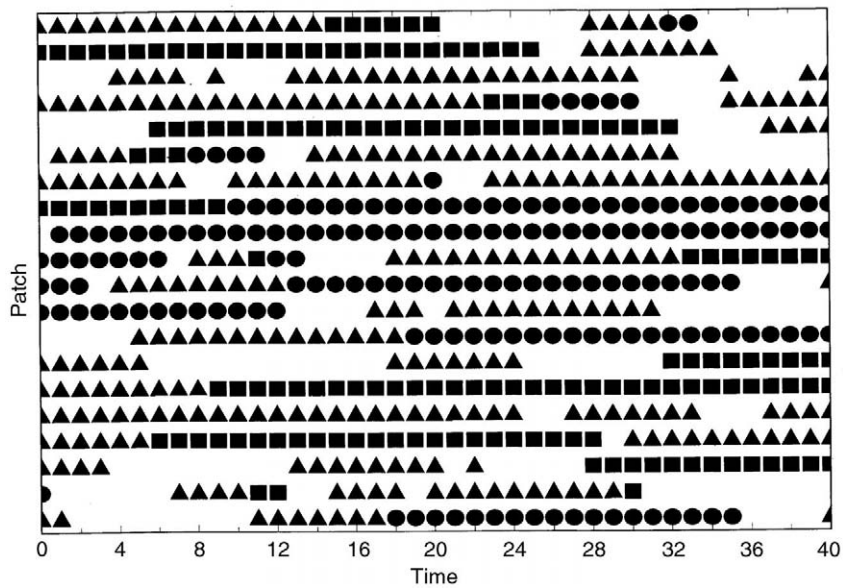


FIGURE 1 Turnover of species in a subset of the patches of a metacommunity with successional mosaic dynamics. Three different species are indicated by the symbols ●, ■, and ▲ in order of decreasing competitive ability and increasing colonizing ability. Blanks mean empty patches.

1997). Although this is not exactly what one might think of as a local population, none of the previous discussion depends on the size of a local population. Death of an individual is then equivalent to local population extinction. Death of an individual may be brought on by the arrival of a superior competitor at the site, disturbance, or simply senescence. Disturbance or senescence open the site to invasion by any species. Otherwise, the site can only be invaded by a superior competitor. The idea that many species may coexist in an area by this mechanism operating on a small spatial scale is referred to in the literature as coexistence by competition–colonization trade-off. This idea is particularly useful in the case of plants or sessile aquatic animals that hold space (Lehman and Tilman, 1997).

These ideas on diversity maintenance in metapopulations have important implications for conservation. If habitat destruction lowers colonization rates, as discussed previously, competitively superior species may be most at risk (Tilman and Lehman, 1997) because their already poor colonizing abilities make them less tolerant of decreases in colonization rates. These models predict that habitat destruction eventually negatively affects all species, however. When the per unit colonization rate,  $c$ , of the best colonizer is reduced below the extinction rate of local populations, all species are doomed to regional extinction even though patches of suitable habitat remain.

### III. QUANTITATIVE EFFECTS OF SPATIAL VARIATION

The discussion so far has made the assumption that the only thing we need to know about a local population is its presence or absence: Is it extinct or not? This is a rather crude accounting because surely the number of colonists or propagules sent out by a local population depends on the size of that local population. Patches may be of different sizes and qualities and therefore support local populations of different sizes and densities, issues that are active areas of research in metapopulation theory (Hanski and Gilpin, 1997). Moreover, local populations vary in size over time in any one patch and in space from patch to patch even when patches of identical size and quality are compared. Ignoring changes in local abundance over time (i.e., ignoring the dynamics of local abundance) would be justified if local population buildup after colonization occurs quickly to some local population equilibrium (a population size at which, on average, reproduction and immigration balance deaths and emigration), around which population fluctuations occur until extinction. A broader variety of behaviors of metapopulations can be examined by taking actual local population sizes into account. At the same time, we can depart from the strict metapopulation assumption that recolonization occurs from ex-

ternal inputs. We shall now assume that there are two spatial scales in the system. The smaller scale is the spatial scale on which interactions between individual organisms occurs (i.e., it is the spatial scale of positive and negative density-dependent effects within species), the scale of competition between species, or the scale of predation, depending on which of these are important in the system of concern. This scale corresponds to the local population scale in the strict metapopulation sense and shall still be referred to in terms of the patch and the local population. The larger scale is the scale of the whole population, which corresponds to the metapopulation or regional scale in the previous discussion.

### A. Single-Species Dynamics

To introduce the fundamental concepts, consider first density-dependent population dynamics applicable to organisms with annual life cycles. The dynamics of a local population in the absence of migration can be defined by plotting local population density (numbers per unit area),  $N_{t+1}$ , at time  $t + 1$  as a function of its density,  $N_t$ , the previous year, as represented in Fig. 2. These curves are dynamical relationships, i.e., by applying them repeatedly, one can plot density as a function of time (Fig. 3). The straight line (relationship I) represents the density-independent case in which individual organisms do not interact with one another, and so an individual's contribution to future genera-

tions is independent of the number of other individuals. Thus,  $N_{t+1}$  is simply proportional to  $N_t$ . Relationships II and III are two cases of density-dependence and this density dependence means that they are nonlinear, i.e., they are curved or humped relationships (certainly not straight lines). For relationship II, the resources sustaining a local population are strictly limited, fixing an upper bound on the local population density, which is achieved if the resources are efficiently utilized. The larger the population at time  $t$ , the closer the population comes to using all resources and the closer the upper bound on population density is approached. Annual plant populations commonly accord at least approximately with this relationship, which is sometimes referred to as contest competition. Some insect populations may have dynamical relationships more like curve III. Above a certain density, the number of insects in year  $t + 1$  is a decreasing function of the number of insects in year  $t$ . One explanation for a relationship such as this is scramble competition: At high densities, a high proportion of the population may starve to death, which leads to loss of the resources that these individuals consumed before death. Thus, a large fraction of the resources that could have been turned into new individuals at time  $t + 1$  is lost, and the population crashes.

The three different dynamical relationships of Fig. 2 give very different local population dynamics (Fig. 3). The linear case (relationship I) gives simply a geometric increase. Relationship II (contest competition) gives highly stable dynamics: The population quickly converges on an equilibrium population size, which is determined by the point at which the diagonal line in Fig. 2 intersects the dynamical relationship. Relationship III (scramble competition), however, gives irregular fluctuations referred to as deterministic chaos which result here from the tendency of scramble competition to cause population crashes following population buildup. What happens when local populations with these various dynamics are connected regionally? If there is no variation in population density in space, then regional population dynamics are the same as local population dynamics. However, if there is a large amount of spatial variation in population densities (a common occurrence in nature), regional population dynamics can be very different from local dynamics (Fig. 4). The density-independent case (relationship I), however, does not show different local and regional dynamics because individuals are not affected by density, and therefore population dynamics cannot be affected by spatial variation in density. On the other hand, the strong density dependence arising from scramble competition leads to corre-

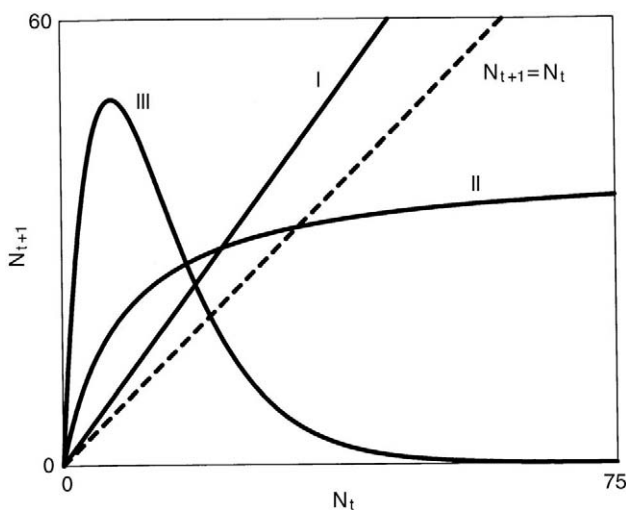


FIGURE 2 Solid curves, dynamical relationships of single-species local populations; I, density-independent dynamics; II, contest competition; and III, scramble competition. The equilibria are the intersections of these curves with the dashed line defining no change in local population size ( $N_{t+1} = N_t$ ).



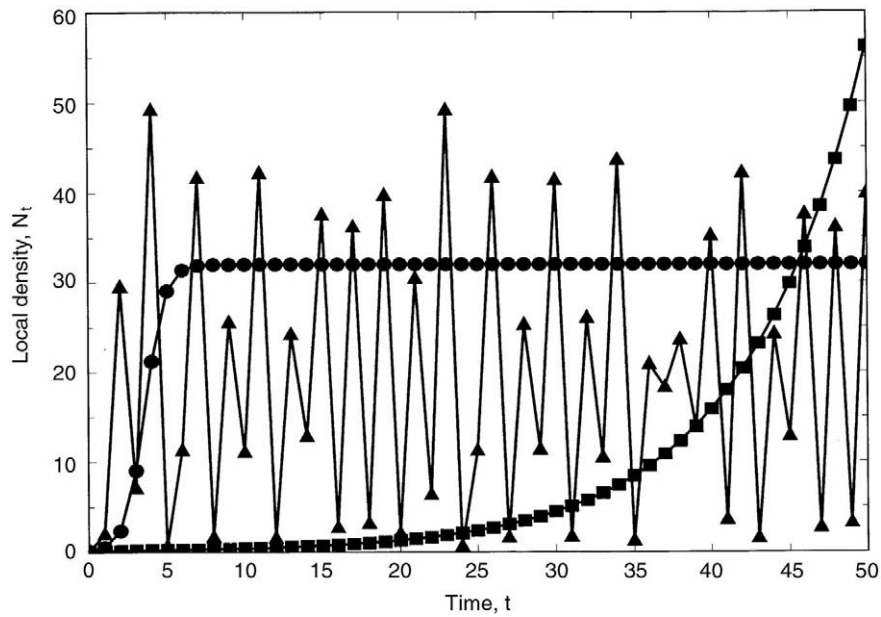


FIGURE 3 Local population dynamics generated by the dynamical relationships of Fig. 2. Relationship I (density independent), ■; relationship II (contest competition), ●; relationship III (scramble competition), ▲.

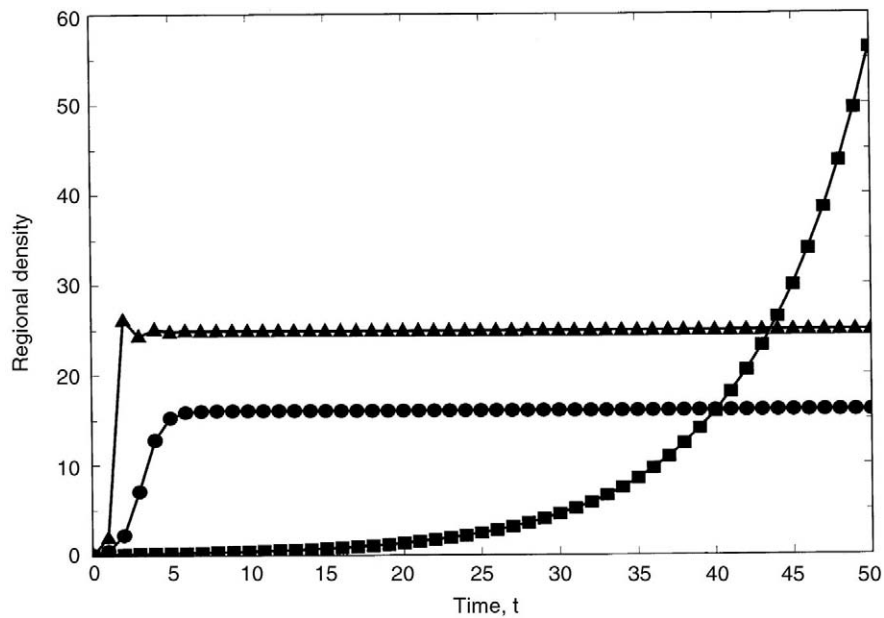


FIGURE 4 Regional dynamics generated by the interaction of the dynamical relationships of Fig. 2 with spatial variation in population densities. Relationship I, ■; relationship II, ●; relationship III, ▲.

spondingly striking differences between regional and local dynamics. For the situation depicted, which is defined in detail later, chaotic fluctuations have been replaced by a stable equilibrium. For contest competition (relationship II), the effect of spatial variation on regional population dynamics is quantitative and not qualitative: The equilibrium density is decreased.

What is the explanation for these changes in dynamics at the regional level in density-dependent situations? To answer this question, we need to work out the dynamical relationship between  $\bar{N}_{t+1}$  and  $\bar{N}_t$  defining the dynamics of population density at the regional level. These regional densities can be defined as the averages of the population densities in local populations, weighted if necessary by patch area. Dynamical relationships such as those in Fig. 2 continue to define dynamical change within patches, but disturbance and fluctuations due to small population size lead to random deviations in these relationships. Thus, these local dynamical relationships are mean relationships converting inputs of population density into outputs, and some of the output population in any patch may then disperse to other patches.

The essence of the difference between local and regional dynamics can be understood by imagining that dispersal occurs early each year and for the rest of the year patches are isolated with mean change governed by dynamical relationships in each patch. Further imagining that after dispersal local populations exist at just two densities (high and low) in equal abundance, then the regional dynamical relationship is easy to derive. This is done in Fig. 5, in which it is assumed that low-density patches have inputs  $1/3$  the regional density,  $\bar{N}_t$ , and high-density patches have inputs  $5/3$  of  $\bar{N}_t$ . The solid curve defines the relationship of local outputs to local inputs (scramble competition) and the dashed curve defines the relationship between  $\bar{N}_{t+1}$  and  $\bar{N}_t$ , i.e., the regional dynamical relationship. This regional relationship is found by connecting pairs of points on the local relationship corresponding to low- and high-density inputs and finding the midpoints of the lines joining these pairs of points. For example, the points A and B in Fig. 5 are one such pair of points, and M is their midpoint. The x coordinate of M is thus the regional input,  $\bar{N}_t$ , which is the average of the x coordinates (local inputs) of A and B. The y coordinate of M is the regional output,  $\bar{N}_{t+1}$ , which is the average of the y coordinates (local outputs) of the points A and B. The complete regional relationship is found by repeating this procedure for every possible value of the input density  $\bar{N}_t$ .

Comparison of the point M on the regional relation-

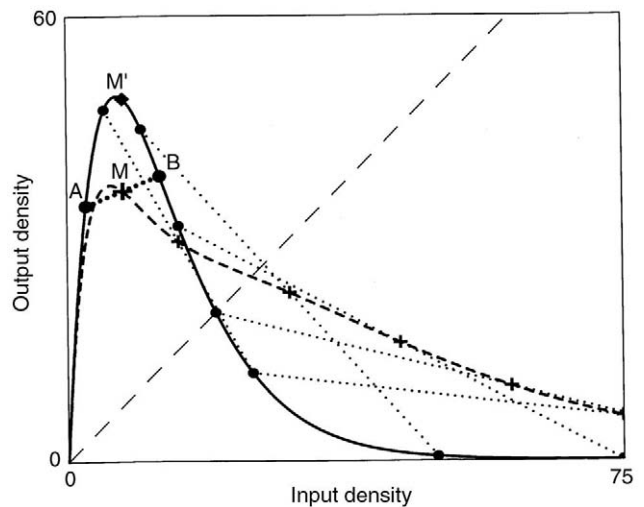


FIGURE 5 Construction of the regional dynamical relationship from the local dynamical relationship III of Fig. 2 and spatial variation between two densities. Solid curve, local dynamical relationship; thick dashed curve, regional relationship.

ship with the point M' on the local relationship reveals the reason for the difference between local and regional relationships. These points both have the same input density, but M' is the output of the local dynamical relationship, whereas M is derived from averaging the outputs of the local dynamical relationship at two different input densities, which both give outputs less than M' due to nonlinearity (curvature) of the local relationship. Thus, nonlinearity in the local dynamical relationship combines with variation in local inputs reducing the hump on the regional dynamical relationship. A similar effect acting on the nonlinearity in the local dynamical relationship after the hump reduces the severity of the decline in the regional dynamical relationship, and it is the combination of these two effects that is responsible for the strong stability of the regional dynamics shown in Fig. 4. Nonlinearity and spatial variation also affect regional dynamics in the case of contest competition, but the effects are quantitative and not qualitative. The regional relationship for contest competition is shown in Fig. 6, in which it is assumed that 50% of local populations are extinct and the others have a density twice the regional density. Not surprisingly, the regional equilibrium is 50% of the local equilibrium, but this reduction, and the regional relationship in total, can be understood using the same averaging approach discussed for scramble competition.

Having local populations take on just two densities for any given average density is clearly unrealistic. However, the qualitative features of the previous results

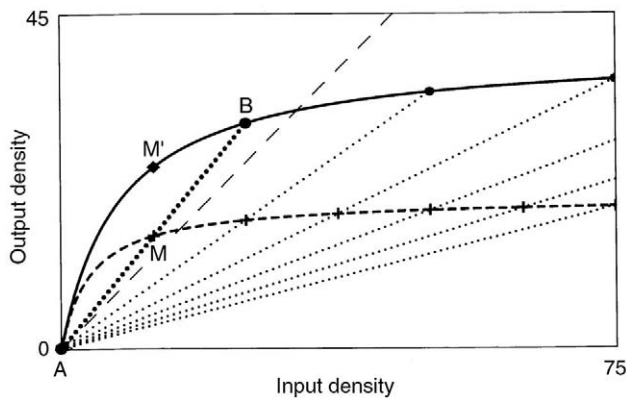


FIGURE 6 Construction of the regional dynamical relationship from the local dynamical relationship II of Fig. 2 in the presence of spatial variation between two densities. Solid curve, local relationship; thick dashed curve, regional relationship.

occur with general sorts of variation commonly observed in natural populations. The most common situation in nature is a negative binomial distribution of possible local population densities, which means that densities are somewhat more clumped than random in space. The striking change in dynamics from local to regional found for scramble competition is enhanced in this case (Chesson, 1998). In the density-independent case (relationship I) the dynamics at the regional level are unaffected by variation from patch to patch:  $M$  always equals  $M'$ , regardless of the pattern of variation. The nonlinearities, i.e., the deviations from a straight line relationship, present in relationships II and III, are necessary for variation in space to have an effect on the dynamics of regional density, and these nonlinearities occur from density dependence.

There are many causes for spatial variation in local densities, of which disturbance and fluctuations due to small population size were mentioned previously. Simple random dispersal processes also lead to some variation in density, but spatially varying environmental factors may have large effects on dispersal and on the ultimate distribution of organisms among patches. For example, in marine organisms such as reef fishes, local currents may transport larvae away from some places and toward others, and these currents may vary with time of year, storms, and other atmospheric conditions, potentially delivering dispersing larvae to reefs in a very patchy fashion. Moreover, organisms commonly actively seek particular places, and the relative attractiveness of different places may vary greatly. Organisms may also seek out their own species, and therefore any existing spatial variation in density may be magnified by

later dispersal. Unstable population dynamics occurring locally in space can be a cause of local variation in density, at least in models. In the illustrations here, spatial variation in density was assumed to have a fixed relationship to regional density [e.g., low-density patches ( $\bar{N}_{t+1}/3$ ) and high-density patches ( $5\bar{N}_{t+1}/3$ )]. More realistically, local densities will not have fixed relationships to the mean, and therefore the regional relationship will in fact change over time as the relationship between spatial variation and regional density changes (Chesson, 1998). Nevertheless, the way in which nonlinearity and spatial variation change the regional relationship, illustrated in Figs. 5 and 6, remains the same despite all these complications. In particular, the stabilizing effect of spatial variation in densities on scramble competition continues to be seen in much more complex circumstances than illustrated here.

These quantitative considerations are very different from the colonization and extinction issues discussed previously for strict metapopulations, but the two approaches can be related. Colonization and extinction dynamics involve implicit density dependence: To count all occupied patches as equivalent, recently colonized patches must quickly increase in density to some sort of steady state, such as a local equilibrium density or fluctuation about a local equilibrium. Local dynamical relationship II would serve in this circumstance, with the mean output taking into account the probability of local disturbance. The example given previously with half the patches extinct corresponds to a  $c$  to  $\epsilon$  ratio of 0.5.

These quantitative considerations also apply to spatially structured populations that are not strict metapopulations, for example, when recovery after disturbance depends to a large extent on regeneration rather than recolonization and also when only part of the life cycle is spatially confined. For example, it is very common for marine organisms that live on a reef, intertidally, or on the ocean floor to have widely dispersing larvae so that only the adult stage has clear spatial structure. The adult may be physically attached to the bottom, as in corals, barnacles, and mussels, or if it remains mobile it may have a territory or at least a home range that is much restricted in extent relative to the distance traveled by dispersing larvae, as in many reef fishes. Such spatial restrictions mean that interactions between individuals are localized because the individuals are localized and are likely most strongly affected by individuals living nearby, for example, on the same coral head or part of the reef. Similarly, in terrestrial plants a spatial unit of major importance is the immediate neighborhood of other plants close enough

to compete with it. Insect populations often have spatially confined larvae and dispersive adults. Although it is the adults that disperse in this case, the fundamental principles are the same in all of these cases, and the effects of variation in density in space are similar, even though in no case does spatial structure of one part of the life cycle qualify these populations as metapopulations in the strict sense. The effects of nonlinear dynamical relationships that we have explored depend on the smaller scale, which corresponds to the local population in a strict metapopulation, being the scale on which these nonlinear relationships apply. Because these nonlinear relationships derive from density dependence, the smaller scale is also the scale of density dependence.

How the scale of density dependence is determined depends on how the individuals in a population affect each other. For example, for territorial coral reef fishes, one source of density dependence may simply be competition for space to set up feeding territories. Other fishes may compete for hiding places from predators. The scale on which fishes may move to secure territories or hiding places defines the scale of density dependence. Predators may cause density dependence in their prey populations by responding to the density of the prey, for example, aggregating in areas of high prey density.

The scale of density dependence is the spatial scale on which predators respond to variation in prey density. Predators may also increase in numbers where prey are dense simply because they have more to eat and therefore reproduce faster. The scale on which this effect occurs is bound to be much larger than the scale of aggregation of predators to variation in prey density. The changes in dynamics with a change in scale demonstrated in Figs. 3–6 should occur in all these instances when one compares the dynamics defined on the scale of density dependence with the dynamics that emerge at the regional or metapopulation scale (Chesson, 1998).

### B. Predators and Parasitoids

Like the dynamics of colonization and extinction, a rich array of phenomena is uncovered when the quantitative effects of spatial structure are examined in terms of interactions between species. In particular, predator–prey and host–parasitoid dynamics (Fig. 7) are often stabilized by local interactions and spatial variation (Hassell, 1997). This phenomenon is best understood in host–parasitoid systems in which a critical issue is spatial variation in the risk of parasitism experienced by a host, which may result from variation in parasitoid

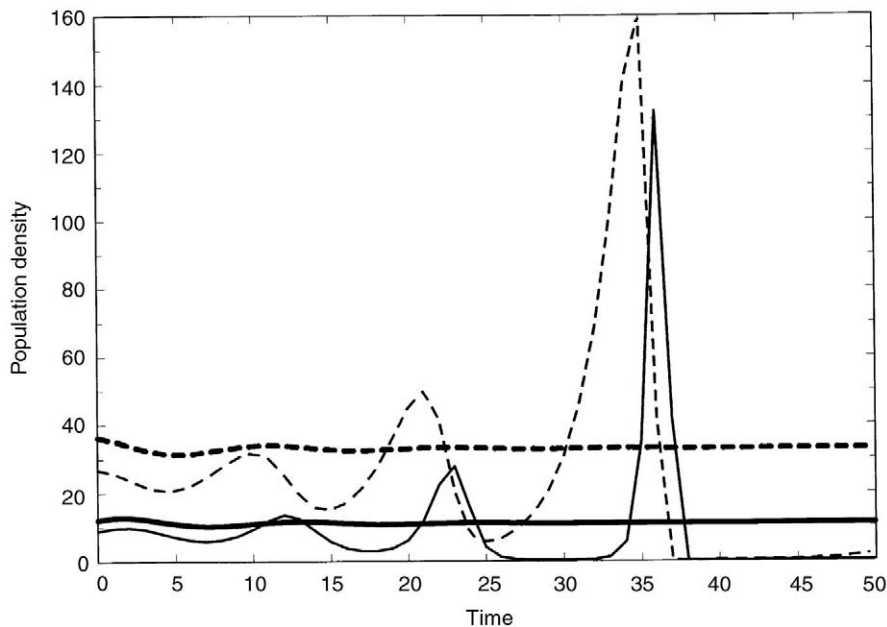


FIGURE 7 Host–parasitoid dynamics. Thin lines, local dynamics of an isolated host–parasitoid community in the absence of migration; thick lines, regional dynamics with parasitoids concentrated in half the patches. Hosts, dashed lines; parasitoids, solid lines.

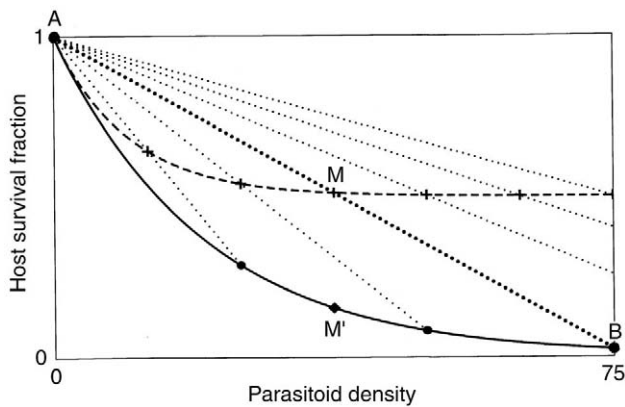


FIGURE 8 Local host survival rates and regional host survival rates as a function of parasitoid density with half the patches having no parasitoids.

density. Figure 8 illustrates how this works. The risk of surviving is thought to be an exponentially decreasing function of parasitoid density, as shown by the solid line.

Very simple two-level variation in parasitoids is assumed for ease of illustration: no parasitoids in half the patches and twice the mean parasitoid density in the other half. The effect of averaging different risk levels, exemplified by Fig. 8, is to moderate the decline in host survival with increasing parasitoid density, which means that host crashes following increases in parasitoid numbers are diminished, tending to stabilize the host–parasitoid interaction regionally (Fig. 7). Although the reasons for spatial variation in parasitoid density are poorly understood in general, an important feature of predator–prey and host–parasitoid interactions is the potential for the unstable nature of the interactions to generate variation in risk, which may then stabilize population dynamics on the larger spatial scale (Hassell, 1997).

### C. Competitive Interactions

Competition–colonization trade-offs are one mechanism allowing coexistence of competing species in a patchy environment. Quantitative approaches allow the consideration of other mechanisms and more general spatial variation. Species fail to coexist (i.e., competitive exclusion occurs) when the species best adapted to the environment depletes resources to levels that are too low for other species to persist. Because the environment varies in space, different species may be the best

adapted under different environmental conditions. Then, many species that depend on the same resources may coexist regionally because each has patches in which it is the best competitor, i.e., each has its own spatial niche. If each of these patches were a closed community (i.e., not connected to other places by migration), then only a single species would dominate in each locality. However, in the presence of migration, each patch may have many species because a species that is not the best competitor in that local community has its reproduction supplemented by immigration from places where it is the best competitor.

If environmental characteristics vary over time also, favoring some species some times and others at other times, local diversity is even more likely to be maintained at high levels because trends in local species abundances will be reversed with reverses in the ranking of competitive superiority in a particular locality: Species tending toward low relative abundance will be partially restored to higher values. A phenomenon with the opposite effect has been discussed by Rosenzweig (1995) and may be particularly important in animals, especially those with complex behavior. Animals detecting a competitively superior species may leave or keep away from a patch, an effect that tends to divide a landscape into exclusive areas containing only the best species for those areas (Rosenzweig, 1995).

The various effects of spatial environmental variation do not require strict metapopulations but simply spatial structure. This more general theory can be applied, for example, to the coexistence of insects, especially various communities of flies such as *Drosophila* and carrion flies, which lay eggs in ephemeral patches of food such as fruit, mushrooms, or dead animals. In general, although the larvae of several species are commonly found developing in the same patch of food, it is also a common finding that there is segregation between species in choice of food patches. Although the underlying cause of the choice is not well understood, the effect is that it promotes stability very strongly (Ives, 1988). A second example of the potential importance of spatial environmental variation is marine communities, especially coral reef fishes but also various sedentary invertebrates that compete for space, with similar effects applicable to plant species. In marine communities, as discussed previously, dispersing larvae characteristically arrive at potential settling sites in a highly patchy manner. Such systems are said to be strongly affected by recruitment variation, where recruitment refers to the process of a larva settling at a particular in patch. Differences in the spatiotemporal patterns of recruitment have the potential to contribute to species coexis-

tence. In this case, the environmental advantage that a species has in a particular patch is proportional to its arrival rate in that patch.

There is a very simple graphical technique for understanding the effects of variation in recruitment rates in space, which serves to illustrate how spatial environmental variation and its interaction with population density may have important effects. Figure 9 assumes a community with two competing species and plots the proportion of new inputs to a local community as a function of the local environment and the regional abundance of the species. The solid curves correspond to two environment types. The  $x$ -axis is the proportion,  $p_t$ , of species 1 in the system as a whole. The  $y$ -axis is the proportion,  $s_{t+1}$ , of the available space at the site taken by new settlers of species 1 during the time interval from  $t$  to  $t + 1$ . The top curve refers to patches favoring settlement of species 1, whereas the bottom curve refers to patches favoring settlement of species 2. The difference between the different sorts of patches is in the relative arrival rates of larvae of the two species: More members of species 1 arrive at the top curve sites, and more members of species 2 arrive at the bottom curve sites. The species compete for available space at

a site, which is assumed to be limiting and thus filled by the larvae arriving each year. Note that if there were only one sort of patch in the system, eventually one species would eliminate the other because it would have higher settlement success everywhere and would increase in abundance relative to the other species. For example, if all sites favored species 1 (Fig. 9, top curve), then the fraction of species 1 settling would always be higher than  $p_t$ . It follows that species 1 would steadily increase until it had taken over altogether, assuming that settled individuals of both species have the same mortality rates (a simple adjustment, however, extends this diagram to different mortality rates; Chesson, 1985).

In Fig. 9, it is assumed that two-thirds of patches favor species 1 and one-third favor species 2. This means that the average proportion of settlers,  $s_{t+1}$ , that belong to species 1 regionally is  $(2/3)$  the top solid curve +  $(1/3)$  the bottom solid curve, which is given by the points dividing the vertical lines between the two solid curves in the ratio 1:2. The dynamics of regional settlement are then given by the wavy line between the two curves. Depending on the actual adult mortality rates of settled organisms, the regional proportion,  $p_{t+1}$ , of species 1 at time  $t + 1$  will lie somewhere between  $p_t$  and the regional  $s_{t+1}$ . Thus, the point where the wavy line crosses the diagonal is the equilibrium point for the system, i.e., the point where  $p_t$  does not change over time. It is a stable equilibrium point: If the value of  $p_t$  is less than the equilibrium, then the regional  $s_{t+1}$  is always greater than  $p_t$ , and therefore species 1 increases in relative abundance ( $p_{t+1} > p_t$ ). However, if  $p_t$  is above the equilibrium, the regional  $s_{t+1}$  is less than  $p_t$ , and therefore species 1 decreases in the system. Hence, whenever away from equilibrium, the system moves back toward it over time.

These results apply no matter whether a patch continues to favor the same species forever or favors different species at different times, for example, if environmental conditions responsible for larval transport change over time. These results have to be modified, however, if instead all patches favor a particular species at a particular time but the favored species changes over time, for example, if the environmental conditions applicable to a particular year favor one species over the other everywhere. In this case, the environment is varying purely temporally. The way temporal variation affects the dynamics of an organism is greatly influenced by life history parameters such as the adult death rate because this determines how long past effects of fluctuations in settlement are reflected by the age structure of the population. A low adult death rate means that a

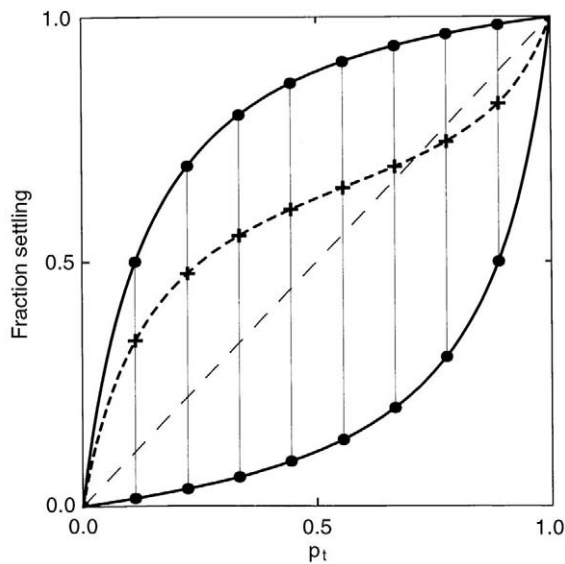


FIGURE 9 Local and regional settling fractions (proportion species 1) as a function of regional fractional abundance for a marine space-holding community with two species. Top solid curve, local environment favoring species 1; bottom solid curve, local environment favoring species 2; wavy dashed curve, regional settling fraction with two-thirds of patches favoring species 1. The regional equilibrium is the intersection of the wavy dashed curve with the diagonal dashed line.

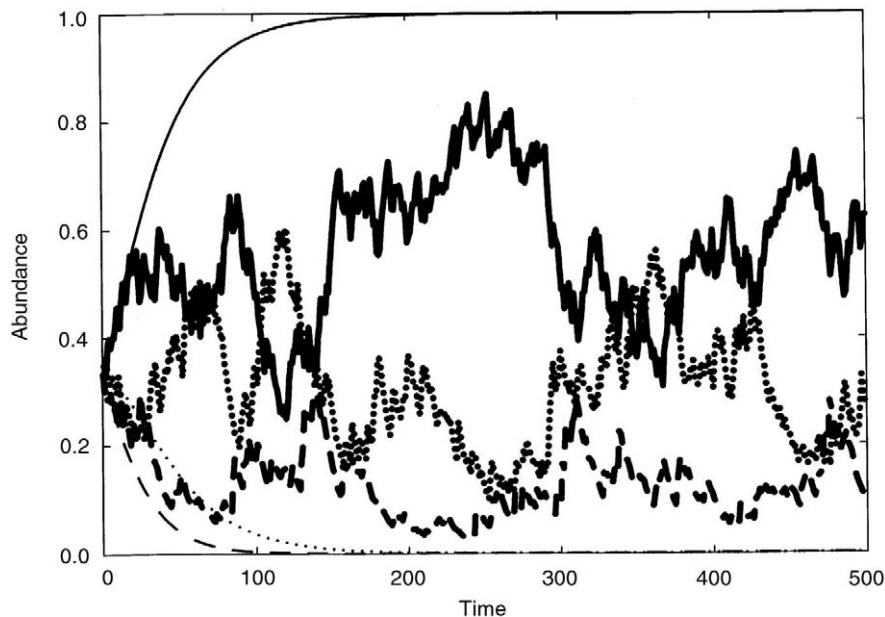


FIGURE 10 Fluctuations in abundance (proportion of sites held) for three competing space-holding species experiencing independent temporal fluctuations in recruitment rates (thick wiggly curves) and constant recruitment rates (thin smooth curves).

species can persist over periods when it is at a competitive disadvantage without its population density declining too greatly. It follows that the ability of species to coexist as a result of temporal variation varies inversely with the adult death rates (Chesson and Huntly, 1997). With irregular temporal fluctuations, the system does not have a traditional equilibrium but nevertheless can still exhibit stable fluctuations about a mean value as depicted in Fig. 10.

Temporal variation is an important feature of many systems, and a mechanism such as that described here has been suggested for coexistence in tropical trees in which highly variable fruit production occurs, in annual plants in which seed germination may vary greatly between years but seeds survive during unfavorable times in the soil seed bank, and also in annual zooplankton in lakes in which a resting egg bank takes the place of the seed bank.

#### IV. LESSONS FROM METAPOPOPULATION THEORY

##### A. Conservation

Few populations in nature exist without important spatial structure. Although there is debate regarding how

often this structure conforms to a strict metapopulation, metapopulations are an important tool in conservation biology theory because habitat fragmentation creates distinct local populations even if they could not be defined before human interference. There is no question that the inhabitants of small isolated habitat fragments are at risk of local extinction. Although a collection of such habitat fragments may genuinely be thought of as a metapopulation, recolonization rates may be so low that each local population will go extinct before any are recolonized. Metapopulation theory emphasizes the need, at a very minimum, for the recolonization rates  $c$  to exceed the extinction rates  $e$ . Habitat destruction not only puts populations at risk by reducing their total size and spatial extent but also, by altering the connectedness of the remaining habitat, it reduces the value of the remaining habitat.

Metapopulation theory also implies that even uniform reduction in recolonization rates with habitat destruction may have different impacts on different species in the system. Therefore, the system may change in character with habitat destruction even under conditions that may not appear to be biased in favor of one species or another. Such effects have the potential to disrupt the functioning of ecosystems. This is not specific to strict metapopulations. The metapopulation perspective also indicates that not just inhabited habitat is

suitable habitat. A species may be only temporally absent or in low abundance in an area, and this habitat may in fact be playing an important role in the system. Empty habitat must be judged in terms of its potential role as part of a larger connected system. Short-term environmental fluctuations may play a role in changing the suitability of habitat patches. Patches that are currently unsuitable may have the potential to become suitable in the future. With the prospect of global climate change, the maintenance of a diversity of patch types and a landscape structure suitable for migration of organisms between patches are of major importance in reducing the loss of biodiversity.

### B. Nonequilibrium Dynamics

The metapopulation perspective highlights the role of nonequilibrium conditions in space, i.e., situations in which local populations do not remain near an equilibrium population size. Simple predator–prey and host–parasitoid models commonly do not have both species coexisting stably in isolated patches. Nonequilibrium locally in space that occurs asynchronously in different patches has an essential role in stabilizing the spatial extensions of these models. Similarly, in successional mosaic systems, perturbation from equilibrium (domination by one species), locally in space, is critical to the maintenance of the system. Such essential roles for nonequilibrium conditions have sometimes been considered as meaning that the system is particularly robust to perturbations, including habitat degradation, by human activities. There is no reason to expect this to be true. Indeed, the message here is that equilibrium occurs on the larger spatial or temporal scales. Interfering with the system through habitat destruction, alteration in the disturbance frequency, alteration in mean environmental states, and alteration of the variance in environmental states all have major consequences and are therefore of serious concern even though on small scales the system may not be at an equilibrium.

### C. Population Regulation

These metapopulation and structured population perspectives also have the potential to resolve the long-standing debate regarding the relative importance of environmental fluctuations versus density-dependent interactions within and between species. Andrewartha and Birch were the first to clearly articulate metapopulation structure as an important feature of nature (Hanski, 1999), but because of the prominent role of stochastic factors (factors with random elements), such as extinc-

tion and colonization, they believed that density dependence was of minor importance. The quantitative analysis here shows that, far from being of minor importance, density dependence within local populations has a major role in shaping the nature of population dynamics at the metapopulation or regional scale. Density-dependent processes lead to nonlinear dynamical relationships, and it has been shown how such relationships interact with fluctuations in time and space to give outcomes at larger temporal and spatial scales that are not predictable on the basis of the separate effects of these factors.

### See Also the Following Articles

COMPETITION, INTERSPECIFIC • DISTURBANCE, MECHANISMS OF • LOSS OF BIODIVERSITY, OVERVIEW • PARASITOIDS • POPULATION DYNAMICS • PREDATORS, ECOLOGICAL ROLE OF • SPECIES-AREA RELATIONSHIPS • SPECIES INTERACTIONS

### Bibliography

- Briggs, C. J., Murdoch, W., and Nisbet, R. M. (1999). Recent developments in theory for biological control of insect pests by parasitoids. In *Theoretical Approaches to Biological Control* (B. A. Hawkins and H. V. Cornell, Eds.), pp. 22–42. Cambridge Univ. Press, Cambridge, UK.
- Chesson, P. (1998). Making sense of spatial models in ecology. In *Modeling Spatiotemporal Dynamics in Ecology* (J. Bascompte and R. V. Sole, Eds.), pp. 151–166. Springer, New York.
- Chesson, P., and Huntly, N. (1997). The roles of harsh and fluctuating conditions in the dynamics of ecological communities. *Am. Nat.* 150(5), 519–553.
- Chesson, P. L. (1985). Coexistence of competitors in spatially and temporally varying environments: A look at the combined effects of different sorts of variability. *Theor. Population Biol.* 28, 263–287.
- Connell, J. H. (1978). Diversity in tropical rain forests and coral reefs. *Science* 199, 1302–1310.
- Hanski, I. (1999). *Metapopulation Ecology*. Oxford Univ. Press, Oxford.
- Hanski, I. A., and Gilpin, M. E. (1997). *Metapopulation Biology: Ecology, Genetics, and Evolution*. Academic Press, San Diego.
- Harrison, S., and Taylor, A. D. (1997). Empirical evidence for metapopulation dynamics. In *Metapopulation Biology: Ecology, Genetics, and Evolution* (I. A. Hanski and M. E. Gilpin, Eds.), pp. 27–42. Academic Press, San Diego.
- Hassell, M. P., and Wilson, H. B. (1997). The dynamics of spatially distributed host–parasitoid systems. In *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions* (D. Tilman and P. Kareiva, Eds.), pp. 75–110. Princeton Univ. Press, Princeton, NJ.
- Hastings, A. (1980). Disturbance, coexistence, history, and competition for space. *Theor. Population Biol.* 18, 363–373.
- Ives, A. R. (1988). Covariance, coexistence and the population dy-



- namics of two competitors using a patchy resource. *J. Theor. Biol.* **133**, 345–361.
- Lehman, C. L., and Tilman, D. (1997). Competition in spatial habitats. In *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions* (D. Tilman and P. Kareiva, Eds.), pp. 185–203. Princeton Univ. Press, Princeton, NJ.
- Nee, S., May, R. M., and Hassell, M. P. (1997). Two-species metapopulation models. In *Metapopulation Biology: Ecology, Genetics, and Evolution* (I. A. Hanski and M. E. Gilpin, Eds.), pp. 123–147. Academic Press, San Diego.
- Rosenzweig, M. L. (1995). *Species Diversity in Space and Time*. Cambridge Univ. Press, Cambridge, UK.
- Tilman, D., and Lehman, C. L. (1997). Habitat destruction and species extinctions. In *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions* (D. Tilman and P. Kareiva, Eds.), pp. 233–249. Princeton Univ. Press, Princeton, NJ.



# MICROBIAL BIODIVERSITY, MEASUREMENT OF

Kate M. Scow, Egbert Schwartz, Mara J. Johnson,  
and Jennifer L. Macalady  
*University of California, Davis*

---

- I. Introduction
  - II. Methods for Measuring Microbial Diversity
  - III. Future Directions
- 

## GLOSSARY

- clone** A population of cells all descended from a single cell; a number of copies of a DNA fragment obtained by allowing the DNA fragment to be replicated by a phage or plasmid.
- fingerprint** A pattern produced by DNA fragments or lipids that represents a community or species.
- function** A process carried out by an organism or group of organisms.
- genotypic** Relating to the genetic composition of an organism.
- oligonucleotide** A short, single-stranded nucleic acid molecule either obtained from an organism or synthesized chemically.
- phenotypic** Relating to observable characteristics of an organism.
- phylogeny** The ordering of species into taxonomic groups based on evolutionary relationships.
- polymerase chain reaction** A method for amplifying DNA *in vitro* involving the use of oligonucleotide primers complementary to nucleotide sequences in a target gene and replication of the target sequences by the action of DNA polymerases.

- pure culture** An organism growing in the absence of all other organisms.
- species** A collection of closely related strains.
- strain** A population of cells all descended from a single cell.
- 

**MICROBIAL DIVERSITY** is the measure of the number or relative abundance of microbial species in a local area or region. Measuring microbial diversity may involve counting individual species, numbers of functional groups, or units operationally defined by the particular method being used.

## I. INTRODUCTION

Microorganisms encompass an extraordinary diversity, in both their taxonomy and their ecological functions. As a group, microorganisms range in size over several orders of magnitude, span the three taxonomic domains (Archaea, Bacteria, and Eukarya), and are uniquely capable of performing all known biochemical transformations. The existence of such an immense variety of organisms, combined with the fact that microorganisms are too small to see without magnification, makes the task of measuring microbial biodiversity challenging and even daunting.

Measuring the biodiversity of microbial communities

is important for immediately practical and more fundamental considerations. Understanding microbial biodiversity may translate into benefits for biotechnology; management of agricultural, forest, and natural ecosystems; biodegradation of pollutants; reclamation of damaged lands; waste treatment systems; and biological control systems. On a more basic level, microbial processes underlie many essential processes shared by all ecosystems. However, because of our current minimal understanding of these groups of organisms, microbial communities are treated as black boxes in most studies of community, ecosystem, and landscape ecology. Increased knowledge of microbial biodiversity would clarify the details of many ecosystem processes and help generate the baseline information needed to predict and perhaps ameliorate the effects of global climate change.

This article discusses major issues involved in measuring microbial biodiversity and provides an overview of methods for measuring microbial diversity in natural, managed, and engineered ecosystems. The content is intended for biologists and ecologists who desire to incorporate the often-neglected microbial taxa into surveys of biodiversity.

### A. Biodiversity of Microorganisms

The unit of measurement used in biodiversity studies of macroscopic organisms is usually a taxonomic one: the species. Microbial taxonomy has undergone enormous changes in the past two decades. Traditionally in microbiology, a species was phenotypically characterized with organisms classified into taxonomic groups based on morphology, physiology, and metabolism. Recently, however, phenotypic classification has been replaced or supplemented by the use of genotypic analysis. Nucleic acid sequence information derived from the small subunit of the ribosome [16S ribosomal RNA (rRNA) for prokaryotes or 18S rRNA for eukaryotes] is used to determine the degree of similarity among groups of organisms and the evolutionary relationships of microorganisms and all other life-forms. Certain analytical methods use the DNA sequence that codes for the rRNA rather than the rRNA sequence itself. The term used to describe this DNA is ribosomal DNA (rDNA). Although genetic sequencing for individual microbes is labor-intensive, particularly when characterizing complex microbial systems, some microbial ecologists maintain that an in-depth understanding of recently described systems is not possible until this information is gathered.

Far more species of bacteria, fungi, and other microorganisms exist than have been described to date. For

example, between 300,000 and 1 million species of bacteria are estimated to exist based on initial in-depth phylogenetic surveys of several specific ecosystems and on calculations based on the reannealing kinetics of DNA extracted from natural environments. In contrast, only approximately 3800 species have been isolated and described by culture-based analyses. As genetic sequence analyses progress, libraries of small subunit rRNA sequences from specific ecosystems are created and compared to sequences in the existing database of known species. Often, these new sequences exhibit a low degree of similarity compared with the sequences of known species.

### B. Background on Measuring Microbial Biodiversity

Although measuring the biodiversity of animal and plant communities involves counting and identifying individual species, such an approach is neither practical nor feasible for microorganisms. Though individual microorganisms can be counted using microscopy, their morphologies do not reflect their diversity. Visualization reveals little that is defining about identity beyond the observations that an individual cell is a fungus or bacterium, is a particular shape (rod or coccus), or has certain cell wall properties (gram positive or gram negative). Furthermore, counting is often hindered by interference from the physical environment in which microorganisms live. Some of the traits used as classification criteria for microorganisms are relatively plastic and, given that horizontal gene transfer appears common among microorganisms, using phenotypic information to classify species in microbiology is problematic. Additionally, the end point of many methodologies used to characterize microorganisms is not species identification. Thus, extrapolating the concept of species used in classifying large eukaryotes (plants and animals) to microorganisms is questionable.

Another important issue in exploring microbial diversity is the question of whether taxonomic (e.g., species or family groups present) or functional (e.g., types of processes present) diversity is of interest. Some important microbial processes include the transformation of elements in biogeochemical cycles, decomposition of organic residues, biodegradation of pollutants, fluxes of gases to and from the atmosphere, symbiotic relations with plants, and parasitism of other organisms. In ecological studies, the processes carried out by microorganisms within an ecosystem may be of greater relevance than knowledge of the specific identities of organisms involved. A considerable amount of functional redun-

dancy among microorganisms is thought to exist in most communities based on the fact that there are far more taxa than processes. To date, specific functions have been difficult to measure and interpret, especially by nonmicrobiologists, in studies of ecosystem biodiversity. However, with the development of rapid molecular assays, measuring microbial community compositions has become more feasible, and taxonomic characterization is more widely used. Compiling functional and taxonomic diversity measurements is important given that the ecological roles played by organisms represented by newly discovered *ssrDNA* sequences are often unknown.

Numerous methods are available for characterizing microbial communities. Determining which approaches to use is a critical step and depends on the objectives of the study. Should the diversity of a community be measured by characterizing its individual members, a task perhaps too large to be feasible, or by directly characterizing the community as a whole? If taxonomic diversity is of concern, what level of resolution is needed? Given that microorganisms are so small, what is the appropriate physical scale on which to sample in a study of biodiversity? These points are discussed later.

### C. Enrichment/Isolation versus Direct Measurement of Microbial Communities

Early studies of microbial diversity involved culturing of bacterial isolates on enrichment media supporting their growth ("culture-based") followed by phenotypic characterization of these isolates based on their morphology, physiology, and specific metabolic capabilities. As the tools of molecular biology began to be applied to microbial ecology, it became evident that culture-based methods were failing to detect large portions of microbial communities (no more than 0.1–5% of bacteria in communities are detected using culture-based techniques). Rapidly growing strains of microorganisms were often favored over other members of the community. A culture-based approach therefore probably offers a skewed perspective of microbial diversity in many environments. Consequently, there has been a shift in studies of microbial diversity toward the use of molecular biology and other biochemical and molecular tools that do not require culturing of organisms. Characterization of taxonomic and functional diversity is conducted using information provided by nucleic acids, lipids, and other cellular constituents. If culture conditions of a microorganism are known, data are gathered after individual strains have been separated from a community using enrichment and isolation techniques. In

other cases, especially in soils and sediments, biochemical constituents may be directly extracted from the environmental samples, bypassing the isolation step. Many of these methods are currently under development and new ones are being developed at a rapid pace. Those seriously engaged in studies of microbial diversity have to keep close tabs on emerging trends in the field.

Newer methods may introduce their own biases in the assessment of microbial biodiversity, although these biases are presumably not as severe as the culturing bias. For example, molecular methods often make use of the polymerase chain reaction (PCR), during which original copies of specific DNA sequence are amplified exponentially. Two important considerations are that amplification efficiency varies among DNA sequences and therefore PCR results are rarely considered to be quantitative, and that a DNA sequence with low amplification efficiency could conceivably be prominent in the environment but not appear in PCR-based diversity measurements (see Section II.B.1).

These direct measurement techniques may also be used to amplify sequences of rRNA or messenger RNA (mRNA) in order to shed light on the distributions, relative activities, and gene expression (functions) of specific groups.

### D. Level of Taxonomic Resolution

When measuring microbial taxonomic diversity, a decision must be made about the level of taxonomic resolution (e.g., species, genus, family, order, and domain) needed to sufficiently assess diversity. The very definition of species in microbiology is subject to intense debate. Though counting numbers of individual species can be relatively straightforward in studies of plants and animals, this approach is not feasible for microorganisms. There is an overwhelming amount of diversity among microorganisms and categorizing all types is beyond the scope and resources of most projects. In some cases, determining families, or even kingdoms, within a previously uncharacterized community or ecosystem provides novel and valuable information.

For instance, the DNA from microbial communities as a whole may be presented in the form of a "fingerprint," which provides a means to classify and describe systems. DNA fingerprints are an abstraction of the original genetic information, but they may be an efficient way to characterize complex microbial habitats in which there is little information linking community structure and community function. They also offer the means to generate hypotheses, test responses of systems to environmental changes and variables, and detect

shifts in microbial communities. Efforts may be concentrated on conducting sequence analysis on changing elements of a DNA fingerprint rather than on sequencing all the members of a microbial community.

### E. Environmental Heterogeneity

Obtaining a representative sample of the microbial diversity of a particular environment may be very difficult. It is obvious that most environments are heterogeneous at the spatial and temporal scale of humans. This heterogeneity is even more pronounced at the scale of microorganisms, often in ways that are not immediately evident to humans. For example, microbial communities within the root zone of plants, in which available carbon is high and soil chemistry has been altered, differ substantially from immediately adjacent communities in unrooted soil. These different communities are often only millimeters apart from one another. Also, a community that temporarily forms in a pocket of decaying plant or animal matter may be strikingly different from a community only several centimeters away. Within sediments, communities change substantially with depth as a function of oxygen availability and carbon distribution. This concept is also applicable on the scale of individual soil aggregates. In assessing biodiversity, one approach is to identify important microenvironments within the ecosystem of interest and measure diversity within each major type of microenvironment. With this approach, the number of samples obtained may soon become overwhelming. Another approach is to combine samples from different microenvironments in an attempt to understand the heterogeneity within the entire ecosystem, in which case it may be difficult to interpret factors governing the diversity. An understanding of the underlying physical structure of an environment is a prerequisite to a good sampling design.

## II. METHODS FOR MEASURING MICROBIAL DIVERSITY

Far more methods for measuring microbial diversity exist than can be presented here. The methods described here are commonly used and, in composite, represent the breadth of approaches available. Methods can be divided into those that involve the counting of cells, measurement of cellular constituents, or determination of activity. Most methods provide either taxonomic or functional information about communities, whereas only a few (e.g., nucleic acid-based methods)

have the potential to integrate both types of information. The following methods were originally developed in pioneering studies of bacteria, but most are also suitable for studies of fungi and other eukaryotes. Commonly, studies of fungal diversity involve isolation and characterization of strains rather than being community-based. Increasingly, community-based molecular methods are being used for eukaryotic microorganisms.

### A. Counting Microorganisms

Since the development of the microscope, microbial population numbers, and to some degree diversity, have been quantified by examining cells under high magnification. Selective growth media, provided either in petri plates or in most probable number (MPN) assays, permit enumeration of culturable organisms that can metabolize specific compounds. Plates and MPN assays thus give potential functional information about a community. Counting organisms by the latter method requires that organisms grow sufficiently to be detectable.

#### 1. Microscopy

Major microbial groups, such as bacteria and fungi, can be distinguished by their morphologies as observed under a light or fluorescence microscope after staining the cells. The microscope provides a quick but relatively superficial survey of microbial diversity based on the sizes, shapes, and staining properties of microorganisms. The relatively low number of morphotypes among bacteria limits descriptions of diversity. To some extent, fungal diversity can be investigated by examining hyphal and fruiting body structures and spore morphology, if present. Taxonomic resolution can be improved by using dyes tagged with fluorescent monoclonal antibodies specific to groups or strains, which permit selective visualization of members of the group to which the antibody binds. Recently, development of fluorescent *in situ* hybridization (FISH) with specific DNA probes (described later) has made the microscope a powerful and highly quantitative tool for use in studies of diversity. The antibody and DNA tagging methods require initial characterization of the organism or group of interest in order to develop the antibody or probe.

#### 2. Plate Counts

Plate counts exploit the fact that individual microscopic cells quickly grow into a colony visible to the unaided eye if provided with suitable growth conditions. Microbial diversity has been assessed by culturing microorganisms on solid nutrient media to which substances are added to specifically promote the growth of target

organisms and/or inhibit the growth of unwanted groups. For example, media containing complex organic substances combined with a low pH tend to select for fungi instead of bacteria. Colonies that grow on the plates are counted and may be further differentiated based on their color or other morphological properties. The diversity of culturable microorganisms is estimated by counting the number of different colonies present on agar media inoculated with a dilution series of the microbial community. Diversity measured based on this technique is unlikely to yield comprehensive, ecologically relevant information about the microbial community as a whole. An advantage of using plating techniques is that strains that grow on plates can often be easily isolated, characterized, and identified by traditional methods of microbiology.

### 3. MPN Methods

MPNs make use of specifically designed culturing conditions to estimate the number of microorganisms in a community able to carry out specific functions. The MPN medium is designed to select a specific trophic group by providing carbon, energy, nutrients, and environmental conditions needed to support the growth of that group. Thus, iron reducers are selected by providing anaerobic conditions, a carbon source, ferric iron, and other nutrients. A series of dilutions of an environmental sample are inoculated into the MPN medium. Measurements of turbidity, substrate utilization, or product formation confirm cell growth or activity. The number of organisms capable of carrying out the particular function being investigated is estimated from cell counts from the series of dilutions and with standard statistical calculations. MPN estimation of numbers is obviously a culture-dependent method, biased toward those organisms able to grow and compete under laboratory conditions.

## B. Analysis of Cellular Constituents

### 1. Nucleic Acids

Methods using information contained in nucleic acids, DNA and RNA, provide the most specific information about an organism and thus potentially the most detailed information about the biodiversity of a microbial community. Figure 1 provides an overview of nucleic acid-based methods used in studies of microbial biodiversity. Extraction of DNA or RNA from environmental samples is the first challenge in using nucleic acid-based methods. One approach is to first extract cells from environmental samples (usually from aquatic environ-

ments) and then extract cellular DNA or perform analyses on cells collected on filters (e.g., using *in situ* hybridization). *In situ* approaches are quantitative because taxonomic information can be correlated with actual cell numbers within the community. The other approach, direct extraction of DNA without first extracting cells, is more commonly employed in complex environmental media such as soil and sediment. Numerous approaches have been developed to overcome extraction problems associated with specific types of environments, such as soils with high clay or humic acid contents. Other methods have been optimized for rapid extraction to accommodate large numbers of samples. There is usually a trade-off between speed of extraction and quality of extracted DNA (e.g., with respect to purity of the DNA). The efficiency of DNA extraction associated with either the direct or indirect method is difficult to estimate and may range from 10 to 99%. Extraction efficiencies associated with different groups of microorganisms differ, which could skew conclusions about a community's diversity.

Nucleic acid-based approaches can be divided into those that employ PCR and those that do not. The PCR makes multiple copies of a specific fragment of RNA or DNA from a mixed pool of nucleic acid fragments and permits detection of sequences originally present at very low densities (Fig. 2). After PCR amplification, fragments are separated in a polymer gel on the basis of their length or nucleotide composition. Concentrations of the fragments form "bands" in the gel which are visualized by staining. The pattern made by the bands constitutes a "fingerprint" of a community. PCR is also useful for detecting specific strains, species, or phylogenetic groups in environmental samples and, if used this way, requires a priori knowledge about sequence variability/conservation in the species or group of interest. The use of primer sets that target evolutionarily conserved regions, such as universal bacterial or fungal primers, circumvents this problem. However, in highly diverse communities, universal primers may amplify so many different nucleic acid sequences that it is not possible to separate them on a gel (e.g., a "smear" may result). Primer sets that target evolutionarily variable regions are useful for differentiating closely related organisms.

PCR amplification efficiencies differ among individual DNA or RNA sequences, introducing PCR bias. This is of particular concern because it is difficult to distinguish artifacts of PCR from true contributions from members of the microbial community. For example, PCR-based diversity measurements may be biased toward populations with sequences most complemen-

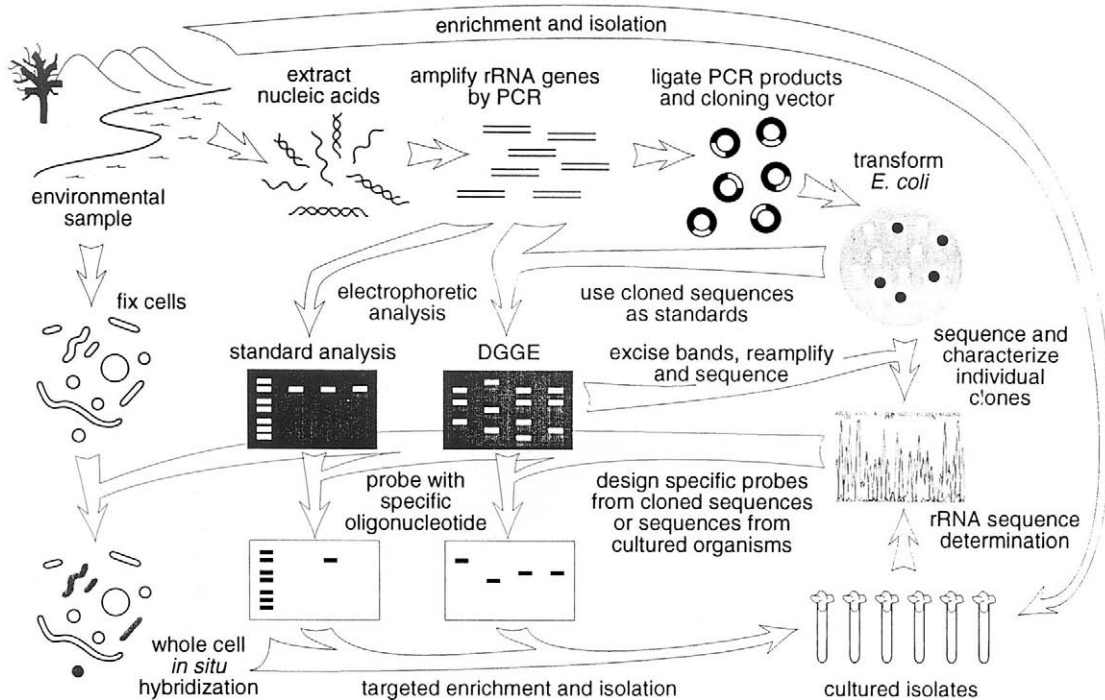


FIGURE 1 A schematic of commonly used approaches in nucleic acid-based microbial ecology studies (reproduced with permission from Head *et al.*, 1998).

tary to primers in the PCR reaction mixture. Consequently, the relative intensity of a PCR product band should only be used as a qualitative measure of the relative abundance of that sequence within the community in the absence of carefully constructed studies that address this issue. Recently developed quantitative competitive PCR methods address this issue. Thermocyclers with detectors that quantify PCR products in real time are also now available.

To obtain information about gene expression (protein or enzyme production), microbial ecologists may utilize a variation of PCR called reverse transcriptase-PCR (RT-PCR). Gene expression is associated with enzyme production directed by another type of cellular RNA called mRNA. Messenger RNA constitutes a small percentage of total RNA but contains the genetic code for individual cellular enzymes. With carefully controlled procedures to eliminate the ubiquitous enzymes called ribonucleases that degrade it, mRNA may be isolated from microbes in environmental samples. With the enzyme reverse transcriptase, mRNA is converted to single-stranded DNA. Once the code is in the form of single-stranded DNA, PCR is performed to produce multiple copies of the code for the enzyme for sequence analysis. Future applications of mRNA analysis will provide insight about the functional diversity of microbial communities.

Finally, PCR may be used as a preliminary step in random cloning and sequencing DNA or RNA from environmental samples from which a "library" of DNA sequences (clones) is generated to represent the diversity present within a community.

PCR-independent methods provide the option of obtaining taxonomic or functional information from nucleic acid extracts or even from intact microbial cells. Hybridization methods entail combining complementary strands of nucleic acids. The complementary strands are genomic DNA in the case of solution hybridization. Other hybridization methods utilize short, single strands of DNA or RNA (oligonucleotide probes) complementary to conserved or variable portions of rRNA or genomic DNA. If the sample contains small amounts of target rRNA or DNA, some of these techniques can be used in conjunction with PCR to reach the threshold of detection. Table I provides a summary of nucleic acid methods commonly used in studies of microbial biodiversity.

#### a. PCR-Dependent Methods

*i. Denaturing or Thermal Gradient Gel Electrophoresis* Gradient gel electrophoresis resolves double-stranded DNA fragments of the same size based on properties defined by their genetic sequence. First, the small subunit rDNA is amplified from microbial com-

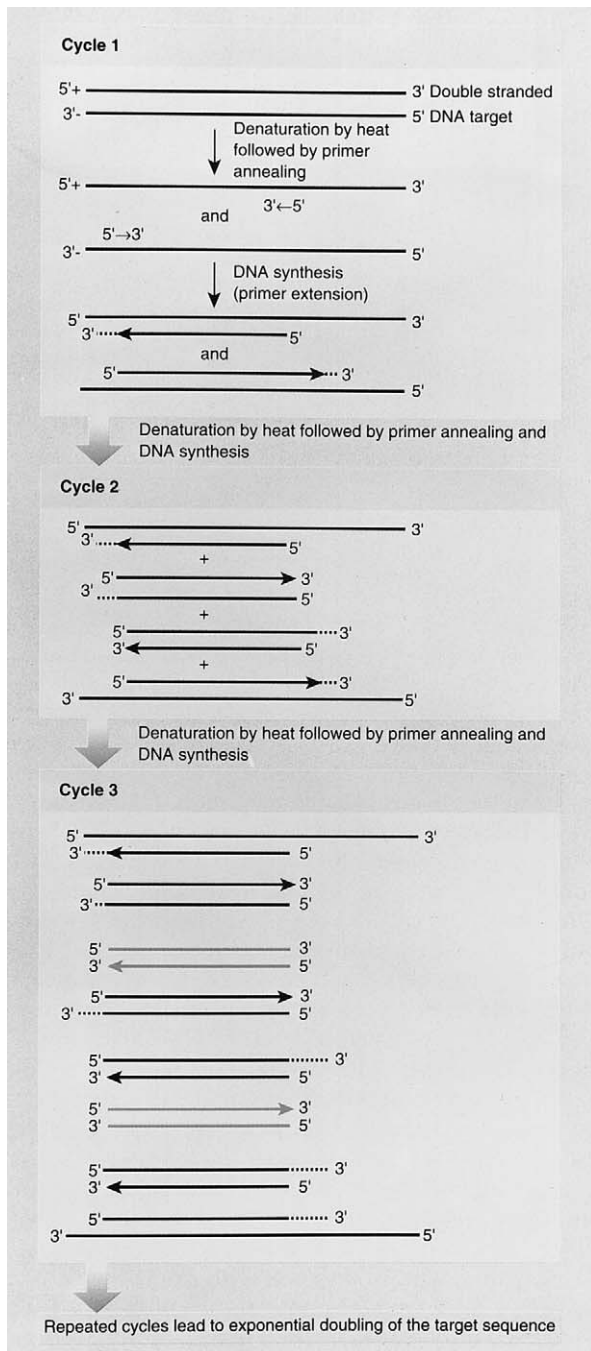


FIGURE 2 The polymerase chain reaction is used to amplify portions of double-stranded DNA from mixtures of DNA from environmental samples. Typically, 30 cycles are completed to generate thousands of copies of the sequences of interest (from Old and Primrose, 1994. Reprinted by permission of Blackwell Science, Inc.).

munity DNA using universal (genetically conserved) or specific (genetically variable) primer sets. PCR products are separated on a gel that contains an increasing chemical gradient of formamide and urea, which differentially

TABLE I Summary of Nucleic Acid Methods	
PCR-Dependent Methods	
<p>The polymerase chain reaction amplifies the signal of microbial DNA from environmental samples. PCR enables detection of microbial populations from very small sample sizes and from extreme environments containing unculturable microorganisms.</p> <p><b>Denaturing/Thermal Gradient Gel Electrophoresis (D/TGGE)</b></p> <p>Small subunit rDNA fragments of the same length are separated based on nucleotide sequence to produce fingerprint patterns. Isolated fragments may be excised and cloned for sequence analysis.</p> <p><b>Intergenic Transcribed Spacer (ITS) Analysis</b></p> <p>Fragments between the small and large rDNA genes are separated based on length producing fingerprint patterns. Isolated fragments may be excised to differentiate strains by sequence analysis.</p> <p><b>Restriction Fragment Length Polymorphism (RFLP)</b></p> <p>Small subunit rDNA PCR products are cut into smaller fragments with restriction enzymes and separated based on length to produce patterns that differentiate simple communities or strains.</p> <p><b>Cloning and Sequencing</b></p> <p>DNA PCR products are inserted in plasmid vectors which are taken up by bacterial cells. The cells are then cloned to generate many copies for sequence analysis and phylogenetic classification.</p>	
PCR-Independent Methods	
<p>Hybridization techniques have all the specificity of PCR-based techniques, are without PCR bias, and have the additional advantage of a means to quantify microorganisms.</p> <p><b>Solution Hybridization</b></p> <p>This method takes community double stranded DNA through thermal dissociation and reassociation process to estimate community DNA complexity.</p> <p><b>Membrane Hybridization</b></p> <p>Group or species-specific probes hybridize with community DNA or RNA immobilized on a membrane to produce estimates of relative abundance.</p> <p><b>Fluorescent In-Situ Hybridization (FISH)</b></p> <p>Fluorescently labeled group or species-specific probes hybridize with RNA in intact cells. Individual cells may be counted and types of organisms quantified by microscopy.</p> <p><b>Oligonucleotide array</b></p> <p>Group, species-specific, or functional gene probes are immobilized on a solid support. Hybridization between probes and community DNA is detected by laser technology.</p>	

denatures or separates the double-stranded DNA fragments. A variant on the method (thermal gradient gel electrophoresis) uses a linearly increasing temperature gradient in the gel, which has an analogous effect on



denaturing double-stranded DNA fragments. PCR product fragments of the same size migrate down the gel at a similar speed until a portion of a particular fragment denatures. At that point the migration of the fragment virtually ceases. Individual bands often correspond to individual microbial strains; however, some strains may produce more than one band (if the strain contains more than one *ssrDNA* sequence).

The result of denaturing or thermal gradient gel electrophoresis (D/TGGE) analysis is a banding pattern that reflects the composition of a microbial community. Statistical analysis involves scoring the bands in the sample as present or absent for a particular migration location on the gel, performing a similarity index based on band information, and then evaluating the relationships of samples with cluster analysis algorithms that create dendrograms. Although D/TGGE patterns have been used directly to determine Shannon–Weaver diversity indexes, the potential for PCR bias necessitates that extreme care be taken in using these banding patterns as the basis for traditional diversity measurements. Often, up to 60 dominant bands can be discerned in a DGGE profile, whereas it is known that thousands of different species may be present in a community. Thus, D/TGGE does not reveal information about the entire community but rather provides a survey of the dominant members in the community (at least those sequences that amplify by PCR). D/TGGE patterns are particularly rich in information since bands in the gel can be analyzed by hybridization methods or excised for cloning and sequencing.

Figure 3 depicts a polyacrylamide gel of a DGGE fingerprint obtained from a bacterial mixed culture grown on the pollutant MTBE (lane 6) or on trypticase soy broth (a rich medium) (lane 7). Also shown on this gel are bands obtained from individual bacterial strains isolated from the mixed culture (lanes 2–5). The organism depicted by the bands in lane 5 is able to biodegrade MTBE.

**ii. Intergenic Transcribed Spacer Analysis** The final portion of the *ssrRNA* gene and the beginning of the large subunit ribosomal RNA (*lsrRNA*) gene contain highly conserved sequences, and primers can be designed to amplify the intergenic transcribed spacer (ITS) region (also called ribosomal intergenic spacer analysis). The *lsrRNA* gene is located downstream from the small subunit gene. The intergenic spacer region consists of both noncoding and coding regions; coding regions include transfer RNA genes in the case of bacteria (Fig. 4) and 5.8S rRNA genes in the case of eukaryotes such as fungi. The noncoding regions are poorly

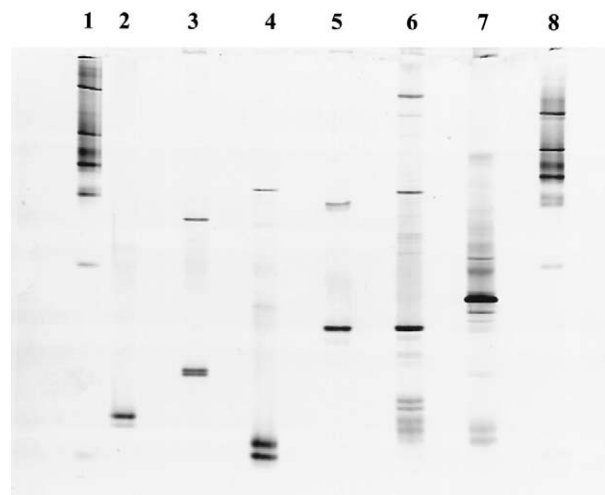
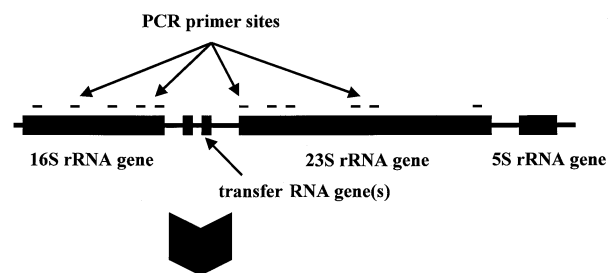


FIGURE 3 Inverted image of fluorescently stained polyacrylamide gel showing DNA bands in DGGE fingerprints. Lanes 2–4, various bacterial isolates from mixed culture; lane 5, MTBE-degrading isolate from mixed culture; lane 6, mixed culture grown on MTBE; lane 7, mixed culture grown on trypticase soy medium. Lanes 1 and 8 are DNA markers.

conserved and contain insertions, deletions, and mutations that do not have a direct effect on the fitness of the organism. Using ITS primers that are universal to bacteria or eukaryotes in a PCR reaction with microbial community DNA generates fragments of different lengths and sequences. These complex banding patterns, like D/TGGE patterns, can be interpreted as a measure of the diversity of the microbial community. As with D/TGGE patterns, a single species may be represented by more than one band. In addition, a single species may have more than one length of ITS region, and one band may represent more than one species.



**16S/23S spacer region is variable in length and sequence**

FIGURE 4 The intergenic transcribed spacer (ITS) region of bacteria contains none, one, two, or more transfer RNA genes between the 16S and 23S ribosomal DNA genes; ITS-PCR products are separated in polyacrylamide gels on the basis of their length.

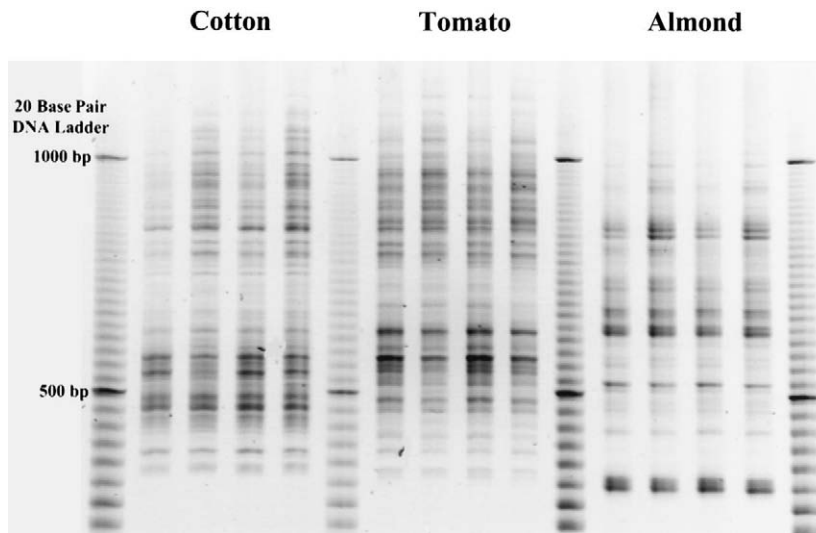


FIGURE 5 Bacterial ITS analysis of whole soil DNA extracts from the crops cotton, tomato, or almond. Each bacterial DNA fingerprint is represented by a pattern that may be categorized by band location and relative intensity and then compared to other patterns by clustering algorithms or other statistical methods.

Because closely related strains may have similar 16S rDNA sequences but dissimilar ITS sequences, ITS analysis is an effective tool to detect diversity in situations in which D/TGGE cannot.

Figure 5 shows microbial communities from agricultural soils planted with three different crops discriminated by ITS analysis. The fingerprints were generated from PCR amplification of total soil DNA using universal bacterial ITS primers. Different lanes within each crop category represent different replicates or PCR reactions.

### iii. Restriction Fragment Length Polymorphism

Small subunit rDNA may also be analyzed by restriction enzymes that recognize and cut double-stranded DNA at precise sequence locations. The fragments derived from the procedure are of different sizes and indicate restriction fragment length polymorphisms (RFLPs). In this method, microbial community *ssrDNA* is either amplified and then digested directly or individual strains are first cloned and RFLP analysis is performed on each clone separately. Restriction enzyme digestion of the *ssrDNA* for a single strain typically results in 2 to 20 DNA fragment lengths, which are separated by gel electrophoresis. The resulting banding pattern is unique and can be used as a fingerprint to distinguish strains or mixtures of organisms from each other. Whole community RFLP fingerprints are extremely complex, resulting in sometimes indecipherable pat-

terns, making this method useful only in systems of low diversity. Cloning the gene first has the advantage of avoiding overly complex community patterns but the disadvantage of limiting investigations due to the additional cost and workload. In terminally labeled RFLP, one of the primers used in amplifying *ssrDNA* is fluorescently labeled for detection in an automated DNA sequencer. As a result, the fingerprint is less complicated than standard RFLP fingerprints because only one of the fragments formed by the restriction enzyme is detected in the analysis.

### iv. Sequencing of *ssrDNA* Clone Libraries

The most direct method for analysis of *ssrDNA* sequences obtained from environmental samples is to clone and sequence them. In cloning, individual PCR products are integrated into a vector (vehicle for reproduction), such as a bacterial plasmid, so that individual *ssrDNA* sequences can be reproduced and analyzed. Although this has traditionally been a time-consuming approach, the increasing availability of automated DNA sequencers is making direct sequence analysis more rapid and widespread. The sequence data are used to construct phylogenetic trees, which represent the evolutionary relatedness of microbial species present in a given habitat. Figure 6 shows phylogenetic affiliations of 16S rDNA sequences of bacterial clones from a hot spring cyanobacterial mat community.

To date, it has been impossible to exhaustively sam-

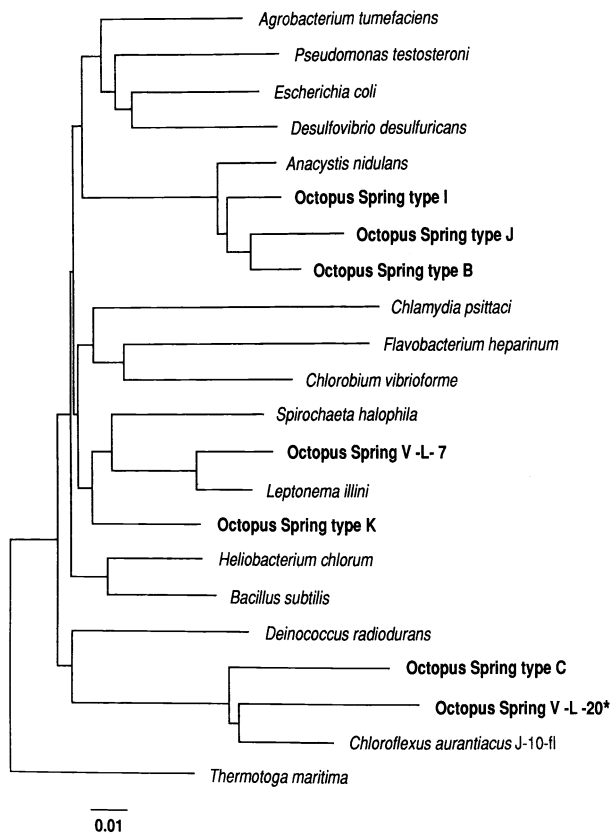


FIGURE 6 A distance matrix phylogenetic tree constructed to include cloned and sequenced members of a cyanobacterial mat community. The sequences of previously undescribed clones are analyzed and fit with their most closely related relatives in the cluster diagram (reproduced with permission from Weller *et al.*, 1992).

ple a microbial community using this approach. For instance, in diversity studies of soil bacterial communities, if 100 clones are picked for analysis, each sequence may be unique, suggesting that species richness is far greater than 100. The clones are chosen randomly; thus, the fraction of clones represented by a common sequence can be used to estimate the relative predominance of that population in the community. In addition to PCR amplification bias, artifacts produced by selective cloning of some fragments over others must be considered.

## b. PCR-Independent Methods

*i. Solution Hybridization* Microbial diversity of an environmental sample may be estimated using solution hybridization (DNA–DNA reannealing kinetics) approaches similar to those used to determine the genetic similarity of two bacterial strains. In this approach, double-stranded DNA is extracted directly from an envi-

ronmental sample, denatured, and then allowed to reanneal (return to double-stranded form). Reannealing of DNA can be monitored using spectroscopic techniques. The rate at which the environmental DNA strands reanneal can be compared with the rate at which perfectly matched DNA strands reanneal, and the difference gives a measure of the number of different genomes present in the sample. Results of these assays should be interpreted with two potential artifacts in mind. First, contaminants in the samples (i.e., humic acids in soil DNA extracts) can affect DNA reannealing. Second, since the size of a genome affects reannealing kinetics, necessary assumptions about the “average” size of a genome may skew absolute estimates of the number of genomes present. The genetic similarity between two communities is best estimated by labeling extracted DNA from one community and measuring the amount of label that hybridizes to DNA from the other community and then vice versa.

*ii. Membrane Hybridization* Specific taxonomic groups can be detected and quantified by Southern blotting with labeled oligonucleotide probes to complementary *ssrDNA* sequences. With this method, the sample is fixed to a membrane, and then through a series of steps the probes are allowed to hybridize to complementary portions of microbial nucleic acids. The probes are either radiolabeled or enzyme linked with fluorescent dyes for quantification by densitometry or fluorimetry. The DNA on a single membrane may be stripped of one probe and then hybridized with others to provide estimates of relative abundance of certain taxa or other phylogenetic information. This technique is more rapid than cloning and sequencing, but the appropriate quantification standards must be used. Precise quantification of the number of individuals per species, however, is not possible by this method because the naturally occurring number of copies of the *ssrDNA* varies among microorganisms. The membrane hybridization technique can also be used with ribosomal RNA, in which case it is called Northern blotting. Northern blots provide quantitative estimates of active taxa because the rRNA concentration is higher in those members of a community that are physiologically active.

*iii. Fluorescent In Situ Hybridization* FISH is a modification of oligonucleotide probing that allows spatial location of specific populations in their native habitat. With this technique, fluorescently labeled oligonucleotides hybridize to *ssrD*/RNA inside cells and are detected with an epifluorescent or laser confocal microscope. Microorganisms are first pretreated to create gaps

in their membranes that permit the labeled oligonucleotides to enter the cells. Group- or species-specific oligonucleotides may be labeled with differently colored fluorescent molecules such that multiple groups of organisms can be observed in a microscopic field. Microscope images can be analyzed automatically to reduce operator errors in identifying and counting fluorescent cells. In soil systems, in which the majority of cells are presumably dormant, FISH results may lead to an underestimation of the microbial diversity due to detection limits. Because this technique allows visualization of individual cells, results are not confounded by variation in the number of *ssrDNA* gene copies among different microorganisms. This method is problematic in soil systems because soil organic matter fluorescence interferes with the probes' fluorescent signals and because probes may bind to soil constituents.

**iv. Oligonucleotide Arrays** A promising new variation of hybridization analysis involves attaching arrays of oligonucleotides to a solid surface rather than attaching sample DNA to traditional blotting membranes. In this technique, multiple oligonucleotide sequences are immobilized on a glass or plastic slide, and the sample DNA is radioactively or fluorescently labeled. After labeled PCR fragments or rRNA hybridize to complementary oligonucleotides on the slide, the signal is quantified using laser technology. As many as 10,000 different sequences can be deposited on a small microscope slide. Due to variation in hybridization kinetics, the results are not strictly quantitative but can show trends in population dynamics. An advantage of this technology is that, given sufficient DNA or RNA, it can be used with nucleic acids extracted directly from the environment without an intermediate PCR step. Interpretation of the wealth of data that results from this method remains a formidable challenge. However, it is currently the only technique that can detect thousands of different populations rapidly, and it likely will have a large impact on microbial diversity studies. As more environmental isolates are characterized and sequenced, oligonucleotide arrays will become more useful for studying biodiversity.

## 2. Fatty Acid Analysis

In addition to DNA-based identification methods, microorganisms can be classified based on their cellular lipid composition. Lipids can be used to identify individual strains or to characterize whole communities. The most commonly analyzed lipids are the phospholipid fatty acids (PLFAs). PLFAs are the primary component of cell membranes and are present in all living

organisms. Fatty acids are distinguished on the basis of chain length, number and location of unsaturations, and location of substituents (e.g., methyls, hydroxyls, and cyclopropane rings).

In lipid analysis of individual strains, cells are grown according to a standard protocol and their cellular fatty acids are extracted. Chromatographic analysis of the fatty acids provides a fingerprint, which is compared to a database to determine identify species. In a modification of the method, PLFAs derived primarily from the cell membrane are analyzed instead of whole cell fatty acids. Analysis of PLFAs in environmental samples is based on an extraction procedure in which polar lipids are extracted and purified using silicic acid chromatography. The ester-linked fatty acids of bacteria, or ether-linked fatty alcohols of Archaea, are analyzed by capillary gas chromatography or gas chromatography with mass spectrometry (White *et al.*, 1979). Figure 7 shows a correspondence analysis plot of PLFA abundance data for 100 sulfate-reducing bacterial cultures together with uncharacterized sulfate-reducing cultures enriched from sediments in Clear Lake, California. Bacterial strains (represented by points) which lie close together have similar fatty acid compositions.

PLFA analysis is also used to characterize entire microbial communities in environmental samples. There

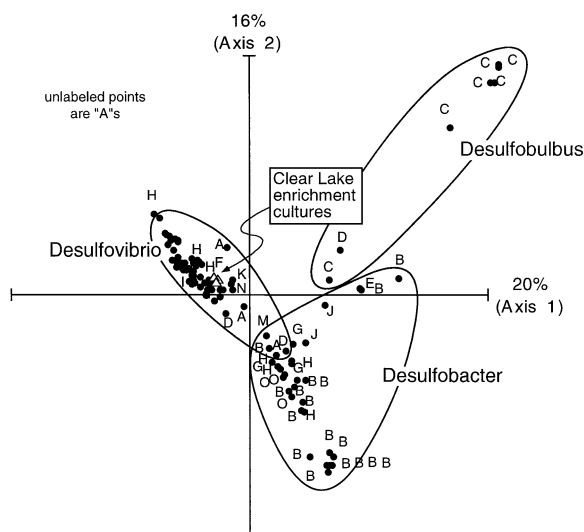


FIGURE 7 A diagram summarizing PLFA abundance data for pure cultures of sulfate-reducing bacteria and uncharacterized strains enriched from polluted sediments in Clear Lake, California. A, *Desulfovibrio*; B, *Desulfobacter*; C, *Desulfobulbus*; D, *Desulfotomaculum*; E, *Desulfobacterium*; F, *Desulfomicrobium*; G, *Desulfuromonas*; H, other (see text); I, *Desulfomonas*; J, *Desulfotobacterium*; K, *Desulfococcus*; M, *Desulfobotulus*; N, *Desulfosarcina*; O, *Desulfomonile*.

are three ways in which PLFAs can provide information about diversity: (i) The entire PLFA profile can be used as a fingerprint which reflects the composition of the total soil community; (ii) signature fatty acids can be used to detect specific subgroups within the community, e.g., sulfate reducing bacteria, methane oxidizing bacteria, fungi, and actinomycetes; and (iii) certain lipids are physiological indicators of environmental stresses, e.g., the ratio of saturated to unsaturated fatty acids, the ratio of *trans*- to *cis*-monoenoic unsaturated fatty acids, and the proportion of cyclopropyl fatty acids.

### C. Measurement of Process or Function

#### 1. Substrate Utilization Patterns

The measurement of substrate utilization patterns is commonly used to evaluate functional diversity of bacterial populations. The commercially available Biolog system, which contains 95 carbon and energy sources on one microtiter plate, is widely used because of its convenience. The plate is inoculated directly with a sample of a microbial community (e.g., a dilution of a soil or sediment suspension) and then incubated under aerobic conditions. Substrate utilization leads to the development of a blue color originating from a redox-sensitive dye. Comparison of well color development patterns among different microbial communities coupled with multivariate statistical analysis may yield information on the functional relatedness of the bacterial populations in the sample.

A major drawback of this method is reliance on the activity of culturable organisms; thus, organisms unable to grow under the incubation conditions are not represented. There is also debate regarding the environmental significance of the carbon sources used in the microtiter plates. These carbon substrates were originally chosen for their ability to distinguish pathogenic bacterial isolates from one another, and therefore bear little relation to substrates encountered by microorganisms in their natural environments. In response to this criticism, plates that contain substrates more common in aquatic and terrestrial environments have been developed and used in some recent applications.

#### 2. Enzyme Assays

A variety of enzymes have evolved to carry out the diverse metabolic processes in which microorganisms are engaged. Six major classes of enzymes have been defined that, in turn, are divided into subclasses on the

basis of the type of reactions they catalyze. Table II provides an overview of common enzymes found in soil. Enzymes found in the environment may be associated with cellular cytoplasm or may be attached to the lipid membranes or surfaces of microbial cells. Other enzymes may be present outside of cells, either in solution or attached to particles in the environment. Enzymes are usually measured based on the type and amount of activity they catalyze under controlled conditions in response to added substrates. "Potential" rather than actual activity is commonly measured.

Some enzymes are present in all organisms (protease and hydrogenase) and give general information about the intensity of biological activity in a sample. Other enzymes catalyze more specific reactions and provide information about the diversity of processes that potentially can be carried out by microbial populations within a given environment. These include enzymes associated with decomposition of polysaccharides (e.g., cellulase, chitinase, amidase, and xylanase), the transformation of phosphorus (e.g., specific kinds of phosphatase) and sulfur (e.g., arylsulfatase), lipid metabolism (e.g., lipase), lignin degradation, and the metabolism of specific environmental pollutants.

### D. Computational Needs

Within the past decade, many large data sets describing microbial communities have been generated using the new biochemical approaches described previously. Much of the information potentially available in these data sets is overlooked because biologists often do not employ the statistical approaches necessary for handling such large data sets. Interesting information may be buried in improperly or unanalyzed data; therefore, serious attention must be given to this problem. One challenge is reducing these large data sets to smaller sets of variables or components. In this way, classification tools can be used and models can be constructed to predict the response of microbial processes to changes brought on by human activity or environmental changes.

Standard multivariate techniques are commonly used to explore the "natural" relationships within a data set (principal component analysis and correspondence analysis) and to determine the correlation of environmental variables to specific components (constrained ordination techniques such as canonical correspondence analysis). In addition, regularized discriminant analysis (RDA) and partial least squares (PLS) are used to test the significance of a priori knowledge about the data sets (e.g., the significance of environmental

TABLE II  
Some Enzymes Extracted from Soils and the Reaction They Catalyze<sup>a</sup>

Recommended name	Reaction	Substrate
<i>Oxyreductases</i>		
Catalase	$2\text{H}_2\text{O}_2 \rightarrow \text{O}_2 + 2\text{H}_2\text{O}$	$\text{H}_2\text{O}_2$
Peroxidase	Donor + $\text{H}_2\text{O}_2 \rightarrow$ oxidized donor + $2\text{H}_2\text{O}$	$\text{H}_2\text{O}_2$ , pyrogallol, chloroanilines, <i>o</i> -dianisidine
Monophenol monooxygenase (polyphenol oxidase)	Tyrosine + dihydroxyphenylalanine + $\text{O}_2 \rightarrow$ dihydroxyphenylalanine + dioxophenylalanine + $\text{H}_2\text{O}$	<i>d</i> -Catechol, <i>p</i> -quinol, <i>p</i> -cresol, 3,4-dihydroxyphenylalanine, <i>p</i> -phenylenediamine
<i>Hydrolases</i>		
Carboxylesterase	A carboxylic ester + $\text{H}_2\text{O} \rightarrow$ an alcohol + a carboxylic acid anion	Malathion, hydroxymethylcoumarin butyrate
Arylesterase	A phenyl acetate + $\text{H}_2\text{O} \rightarrow$ a phenol + acetate	Phenyl acetate, phenyl butyrate, naphthyl acetate
Alkaline phosphatase	Orthophosphoric monoester + $\text{H}_2\text{O} \rightarrow$ an alcohol + orthophosphate	<i>p</i> -Nitrophenyl phosphate
Acid phosphatase	Orthophosphoric monoester + $\text{H}_2\text{O} \rightarrow$ an alcohol + orthophosphate	<i>p</i> -Nitrophenyl phosphate
Arylsulfatase	A phenol sulfate + $\text{H}_2\text{O} \rightarrow$ a phenol + sulfate	<i>p</i> -Nitrophenyl sulfate
Cellulase	Endohydrolysis of 1,4- $\beta$ -glucosidic linkages in cellulose, lichenin, and cereal $\beta$ -glucose	Cellulose, carboxymethylcellulose
$\beta$ -Glucosidase	Hydrolysis of terminal, nonreducing $\beta$ -D-glucose residues with release of $\beta$ -D-glucose	<i>p</i> -Nitrophenyl- $\beta$ -D-glucoside, cellobiose, <i>p</i> -nitrophenyl- $\beta$ -D-glucopyranoside
$\beta$ -Fructofuranosidase (Invertase)	Hydrolysis of terminal nonreducing $\beta$ -D-fructofuranoside residues in $\beta$ -fructofuranosides	Sucrose
<i>Proteinases</i>		
Urease	Hydrolysis of proteins to peptides and amino acids $\text{Urea} + \text{H}_2\text{O} \rightarrow \text{CO}_2 + 2\text{NH}_3$	Casein, gelatin, albumin Urea
<i>Lyases</i>		
Aromatic-L-amino-acid decarboxylase	L-Tryptophan $\rightarrow$ tryptamine + $\text{CO}_2$	<i>dl</i> -3,4-Di-hydroxyphenylalanine, <i>dl</i> -tyrosine, <i>dl</i> -tryptophan, <i>dl</i> -phenylalanine, tryptophan

<sup>a</sup> From Paul, E. A. and Clark, F. E. (1996). Soil Microbiology and Biochemistry. Academic Press, New York. Reproduced with permission.

variables or sample category). RDA uses a regularized covariance matrix estimate for the conventional statistical discriminant analysis methods. PLS is a multivariate calibration tool and can also be applied as a modeling method to solve pattern classification problems. PLS finds a set of latent variables in the measurement variable space (e.g., lipids or DNA banding patterns or sequences) that have a maximum covariance with the dependent variable space. Variable selection can then be used to choose which variables are most important for modeling. Artificial neural net analysis can be used to explore nonlinear relationships within the data sets, (e.g., relating community composition to process rates). Using these approaches, it may be possible to address such questions as which traits are shared by all members of certain classes (e.g., communities on a particular

land use and vegetation type) and which components of a community are linked to specific processes such as pollutant degradation and nitrogen cycling.

### III. FUTURE DIRECTIONS

Great care must be taken to not overstate the reliability of microbial diversity measurements currently available. None of the diversity measurements currently available have the ability to (i) exhaustively sample a community, (ii) quantify population densities, and (iii) resolve individual populations. Ecologists who study macroorganisms have the ability to satisfy all three of these criteria and therefore it is difficult to compare the diversity of large organisms with that of microorgan-

isms. It is likely that a large subset of microorganisms exist at low population densities and are dormant. When this fraction of the community encounters environmental conditions suitable for growth, the dormant populations expand, and only then do we have the ability to measure them. Thus, current techniques tend to be biased toward dominant members of a microbial community. Despite this limitation, ecological studies of microorganisms and this limited perspective have already revealed overwhelming biodiversity.

Once the tools are developed to characterize and analyze data describing microbial communities, numerous questions can be considered. Questions of interest may include the following:

1. What environmental gradients (e.g., vegetation, time, soil texture, and topography) are most strongly correlated with the composition of specific communities or patterns of communities in general?
2. What other (e.g., nongradient) factors also regulate community diversity?
3. Do communities undergo succession within a year? How much do communities vary year to year? Are there seasonal patterns common to all communities across biomes?
4. Are there associations between microbial communities (e.g., facilitation, competition, and interaction)?
5. Does human disturbance alter communities to a greater degree and in different ways than do sea-

sonal and vegetation changes, and are there general patterns that are indicative of disturbance?

## See Also the Following Articles

MEASUREMENT AND ANALYSIS OF BIODIVERSITY •  
MICROBIAL BIODIVERSITY • NUCLEIC ACID BIODIVERSITY

## Bibliography

- Akkermans, A. D. L., van Elsas, J. D., and De Bruijn, F. J. (eds.) (1995). *Molecular Microbial Ecology Manual*. Kluwer, Dordrecht.
- Burlage, R. S., Atlas, R., Stahl, D., Geesey, G., and Sayler, G. (eds.) (1998). *Techniques in Microbial Ecology*. Oxford Univ. Press, New York.
- Head, I. M., Saunders, J. R., and Pickup, R. W. (1998). Microbial evolution, diversity, and ecology: A decade of ribosomal RNA analysis of uncultivated microorganisms. *Microbial Ecol.* 35, 1.
- Old, R. W., and Primrose, S. B. (1994). *Principles of Gene Manipulation*, 5th ed. Blackwell, Oxford.
- Paul, E. A., and Clark, F. E. (1996). *Soil Microbiology and Biochemistry*, 2nd ed. Academic Press, New York.
- Pickup, R. W., and Saunders, J. R. (eds.) (1996). *Molecular Approaches to Environmental Microbiology*. Horwood, New York.
- Van Elsas, J. D., Trevors, J. T., and Wellington, E. M. H. (eds.) (1997). *Modern Soil Microbiology*. Dekker, New York.
- Weller, R., Bateson, M. M., Heimbuch, B. K., Kopczyński, E. D., and Ward, D. M. (1992). Uncultivated cyanobacteria, *Chloroflexus*-like inhabitants, and Spirochete-like inhabitants of a hot spring microbial mat. *Appl. Environ. Microbiol.* 58, 3964–3969.
- White, D. C., Davis, W. M., Nickels, J. S., King, J. D., and Bobbie, R. J. (1979). Determination of the sedimentary microbial biomass by extractable lipid phosphate. *Oecologia* 40, 51.



# MICROBIAL DIVERSITY

Paul V. Dunlap

*University of Maryland Biotechnology Institute*

---

- I. Introduction
  - II. The Scope of Microbial Diversity
  - III. The Biological Significance of Microbial Diversity
  - IV. A New Era in Biological Sciences
  - V. The “Delft School” of General Microbiology
  - VI. The “Woesean Reformation” of Microbiology
  - VII. Major Groups of Microbes
  - VIII. Concluding Comments
- 

## GLOSSARY

- aerobe** An organism that utilizes or requires the presence of oxygen for growth.
- anaerobe** An organism able to grow in the absence of oxygen.
- autotroph** An organism able to utilize carbon dioxide as its source of carbon.
- barotolerant and barophilic** Able to tolerate high pressures and growing better under high pressure.
- bioluminescence** Light production by living organisms.
- chemotroph** Organisms that utilize chemicals as sources of energy.
- cryptoendolithic** Living within the surface of rocks.
- elective culture** The provision of appropriate physical and chemical conditions that elicit the growth of specific metabolic types of microbes.
- extremophile** An organism that grows better at, or re-

- quires for growth, extremes of temperature, pressure, salinity, or other environmental factors.
- halophile** An organism requiring high levels of salts for growth.
- heterotroph** An organism that obtains its carbon from organic carbon compounds.
- microbe** Single-celled organisms, such as bacteria, archaea, protists, and unicellular fungi.
- phototroph** An organism utilizing the energy of light, as in sunlight, for growth.
- psychrophile** An organism that grows better at low temperature or requires low temperature for growth.
- 

**MICROBIAL DIVERSITY** can be defined as the range of different kinds of unicellular organisms, bacteria, archaea, protists, and fungi. Various different microbes thrive throughout the biosphere, defining the limits of life and creating conditions conducive for the survival and evolution of other living beings. The different kinds of microbes are distinguished by their differing characteristics of cellular metabolism, physiology, and morphology, by their various ecological distributions and activities, and by their distinct genomic structure, expression, and evolution. The diversity of microbes presently living on earth is known to be high and is thought to be enormous, but the true extent of microbial diversity is largely unknown. New molecular tools are now permitting the diversity of microbes to be explored rapidly and their evolutionary relationships and history



to be defined. The purpose of this article is to define the scope of, and highlight major themes in, our current understanding of microbial life and to describe recent progress in expanding knowledge of the evolution and biological significance of these organisms.

*“The key to taking the measure of biodiversity lies in a downward adjustment of scale. The smaller the organism, the broader the frontier and the deeper the unmapped terrain.”* (Wilson, 1994)

## I. INTRODUCTION

Rapidly accumulating evidence indicates that microbes most likely account for the vast majority of kinds of organisms on earth. Microbes carry out a stunningly diverse array of metabolic activities, several of which were instrumental in creating conditions for the evolution of other life forms. Through their colonization of diverse and extreme environments, their geochemical cycling of matter, and their biological interactions among themselves and with all other organisms, microbes define the limits of the biosphere and perform functions essential for ecosystem development and health. However, because microorganisms are predominantly unicellular life forms that generally are smaller than can be seen with the unaided eye, they historically have received disproportionately little scientific attention compared to that given to animals and plants. This lack of attention has begun to shift recently as awareness of the diversity of microbes and their biological importance has grown. Of the three presently recognized domains of life, two, the Bacteria and the Archaea, are entirely microbial, and the third, the Eucarya, through its vast array of protists and fungi, is primarily microbial. The essence and full scope of the diversity of microbes is revealed in the dramatic differences among these microorganisms in their phenotypic characteristics of cellular metabolism, physiology, and morphology, in their ecological distributions and activities, and in their genomic structure, expression, and evolution. Appreciation for the true extent of microbial diversity is growing rapidly through the development and use of molecular phylogenetic approaches, which are enabling rigorous analysis of the origins and evolution of microbial life. In combination with classical methods of elective culture, isolation, and phenotypic analysis, the approaches of molecular phylogeny are stimulating the discovery of multitudes of new microorganisms, opening up their biology for study, and providing a clear understanding of the importance of microorganisms as

the functional and evolutionary foundation of the biosphere.

## II. THE SCOPE OF MICROBIAL DIVERSITY

We live on “a microbial planet” (Woese, 1999) in the “Age of Bacteria” (Gould, 1996). Microorganisms, the first cellular life forms, were active on earth for more than 3.0 billion years before the development of multicellular, macroscopic life forms. During that time and continuing into the present, through the invention of a spectacular array of different metabolic and physiological capabilities, microbes evolved to exploit the multitude of environments and microhabitats presented by the abiotic world. They thereby obtained the cellular building materials and energy necessary for growth and reproduction. In so doing, however, they progressively altered the geochemical conditions of the planet, leading to a continual development of new conditions and habitats, abiotic and biotic. Those new conditions and habitats presented both challenges to survival and opportunities to exploit, leading to continuing evolution of distinct microbial types able to endure or take advantage of the biogeochemical changes taking place on earth. Once cellular life began, it is likely that no place on earth containing the molecules and energy conducive to life remained abiotic for long.

The evolutionary trend toward greater complexity, seen in the relatively recent appearance of multicellular life forms (e.g., plants and animals), however, did not cause microbes to be displaced. The appearance of plants and animals did not shunt the unicellular microbes to forgotten corners of the biosphere to hang on and eke out a marginal existence. Instead, multicellular organisms, which themselves can be viewed as highly evolved, complex assemblages of microorganisms, have provided unicellular microbes with a wide variety of new habitats to colonize and exploit. Consider the various microbes whose growth is favored by the different and changing habitats provided by the growth and senescence of roots, stems, leaves, flowers, and fruits during the life of plants. Consider the multitude of physicochemically distinct habitats of the human skin, of our mucous membranes, and the changing environments of our complex intestinal system. Along with these habitats, colonized often by assemblages of several different kinds of microbes, consider the species-specific developmental and metabolic symbioses certain bacteria have established with plants, such as nitrogen-fixing *Rhizo-*

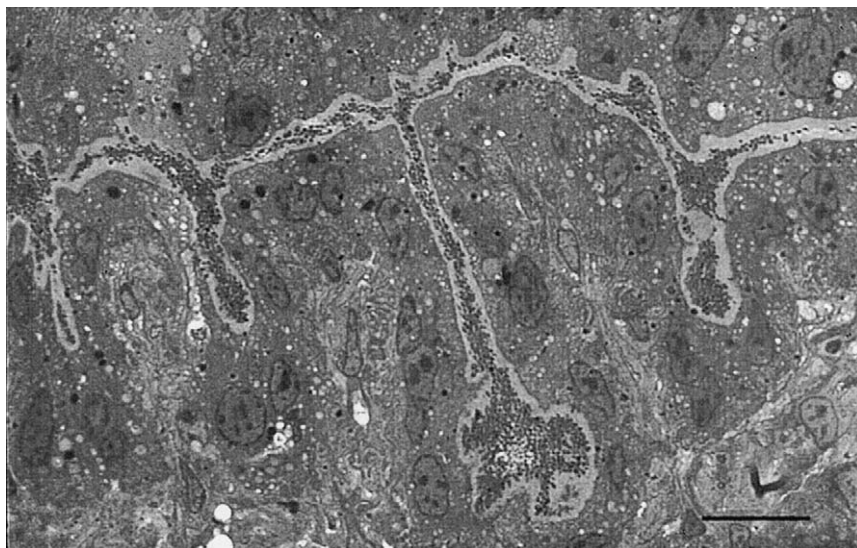


FIGURE 1 Light-micrograph of a section of the light organ of the sepiolid squid *Euprymna scolopes*. The animal harbors a dense population of the luminous marine bacterium *Vibrio fischeri* extracellularly within a ventral tissue complex, the light organ, and uses the light produced by the bacteria in predator avoidance. Reprinted from Claes and Dunlap (1999). Copyright © 1999 Wiley-Liss, Inc.

*bium* with legumes, and with animals, such as bioluminescent *Vibrio* with marine cephalopods and fishes (Fig. 1). Instead of passing on the torch of preeminence in life to the developing multicellular organisms and then politely withdrawing from the life's center stage, microbes were full participants and driving forces in that development, they have continued to diversify, and they remain fully dominant. Plants and animals provide a highly visible but thin multicellular skin over a biologically rich and complex microbial world. A sense of the biological dominance of microbes is given by estimates of the total number of living bacteria, roughly  $5 \times 10^{30}$  cells, with a collective biomass, despite their small size, possibly equal to that of all other life forms (Whitman *et al.*, 1998). We live in the once-and-always age of microbes.

Microbes of one kind or another, especially bacteria, survive and grow almost everywhere on earth. Whether widely distributed, having been spread globally by winds, water currents, and animals, or occurring only in localized areas where they have adapted to grow under specific environmental conditions, microbes dramatically extend our perception of the limits of the biosphere. Previously, that perception was shaped to a large extent by our notion of the conditions under which plants and animals can live. As aerobic organisms, we require the high levels of oxygen present in air for survival. While many bacteria utilize oxygen as

we do, most, however, grow best at lower oxygen levels, and anaerobic microbes of many different types require the strict absence of oxygen to survive, as found, for example, in sediment and gut tracts. Temperatures that from a human perspective are extreme—from just below the freezing point of water in some oceanic waters and sea ice, to several degrees above its boiling point in waters near hydrothermal vents and in hot springs—are not extreme at all to the bacteria that colonize these habitats. Indeed, water temperatures considered cold for humans (i.e.,  $15^{\circ}\text{C}$ ) can be lethally hot to true cold-loving, psychrophilic bacteria. Barotolerant and barophilic bacteria, active at and requiring the extremely high pressures of the deep sea, exist at pressures that would crush the human body, while other microbes, the halophiles, found in salt lakes and solar evaporation ponds, require salt concentrations that would quickly pickle our tissues. Add to this the cryptoendolithic microbes living just below the surfaces of sandstone in the Antarctic, the acid-tolerant and acid-requiring bacteria of acid mine drainage and sulfur springs, and the alkalinity-requiring bacteria of desert soils and alkaline lakes that live at pH levels that would be caustic to our skin. This array of microbial attributes gives a sense of the physical and chemical extremes at which certain bacteria survive and grow (Madigan *et al.*, 1999). These organisms have no “protection” from the environment other than their inherent metabolism and physiology,

which, instead of protecting them, exquisitely suit them to living in these extremes, on which they are dependent. Major research initiatives around the world seek to expand knowledge of “extremophile” biology identifying these “exotic” microorganisms and bringing into study the metabolic and physiological attributes that adapt them to life at the physical and chemical limits of the biosphere.

What about microbial diversity in less extreme environments, away from the limits of the biosphere? Many environments, such as garden soil, coastal seawater, and lake sediments, do not exhibit such dramatic extremes of temperature, acidity, pressure, or other factors. Microbial life in these environments is strikingly diverse. While many different types of microbes can be isolated and grown in laboratory culture from these environments, most so far cannot. A typical expectation is that much less than 1 percent to a few percent of the microbial types seen in an environmental sample will grow in culture (e.g., Amann *et al.*, 1995). That means the vast majority of microbes, even from commonly studied environments such as coastal seawater, have not been brought into study or identified. Thus, discoveries of new types of microbes are waiting to be made, and they are being made. Examples of the unexpected include the discovery of magnetotactic bacteria, which synthesize intracellular chains of magnetic granules that orient the cells to magnetic fields (e.g., Amann *et al.*, 1995) and the occurrence of new members of presumably anaerobic Archaea in oxygenated seawater at shallow depth (e.g., DeLong, 1998). These reports demonstrate that microbial life in more accessible and commonly encountered environments is still poorly understood and far more diverse than presently known or expected. Therefore, most habitats that are known to be rich in microbial life and that have been sampled, such as seawater, soil, and animal gut tracts, have not yet yielded for study anywhere near their full complement of microbes. And what about easily accessible habitats likely to be rich in microbial life but that have not yet been examined at all or at best have been sampled only minimally? One type of habitat is the gut tracts of the several hundred thousand known insect species, only very few of which, such as the termite, have been examined microbiologically. Yet the gut tract of each species—because of the different foods the animal eats, the gut’s specific morphology and physiology for digesting that food, and the environmental conditions under which the animal lives—is likely to host its own very different kinds of microbes. Our “microbial planet” is largely unexplored.

### III. THE BIOLOGICAL SIGNIFICANCE OF MICROBIAL DIVERSITY

Microorganisms are “the foundation of the biosphere” (Staley *et al.*, 1997), providing its “essential, stable underpinnings” (Woese, 1999). Microorganisms have played and continue to play fundamental roles in the evolution of higher life forms on earth. They have done so and continue to do so through the essential ecological processes they carry out in obtaining the materials and energy needed for growth and reproduction. A primary example of the evolutionary role microbes have played is oxygenic photosynthesis, invented by phototrophic bacteria more than 2 billion years ago and which releases oxygen as a by-product of energy generation. Over time, that release of oxygen led to a gradual change in the earth’s atmosphere from reducing to oxidizing. The oxidizing atmosphere, as it developed, allowed energetically more efficient aerobic organisms to evolve and provided a protective shield of ozone against ultraviolet radiation for terrestrial and aquatic organisms. Equally striking examples are the bacterial endosymbiotic origins of chloroplasts, light-harvesting organelles in plants, and of mitochondria, energy-generating organelles, major events in the evolution of plant and animal lineages in the Eucarya. Furthermore, the fixation of atmospheric nitrogen, reducing nitrogen gas to ammonium and converting it into organic nitrogen forms, is an entirely bacterial activity, carried out by various symbiotic and free-living microbes (Madigan *et al.*, 1999).

The ecological processes carried out by microorganisms are equally fundamental. For example, global biogeochemical cycles of major elements, carbon, nitrogen, sulfur, and iron, essential components of all living cells, operate through microbial activity. Specifically, degradation of complex carbohydrates such as chitin, forming the exoskeleton of arthropods, and cellulose, hemicellulose, and lignin, structural polymers in plants, is essential. Without microbial conversion, these polymers would accumulate, removing huge amounts of carbon from the biosphere and blocking a multitude of biological processes that allow micro- and macroorganisms to live. In the absence of these microbial degradative processes, life on earth would soon falter. Besides the microbial degradation of complex organic compounds, consider the range of metabolic diversity in microbes, from oxygenic and anoxygenic photosynthesis, sulfate reduction, methanogenesis, denitrification, iron oxidation, nitrite oxidation and nitrate reduction,

hydrogen and methane oxidation, and so on, all ways by which microbes obtain the energy necessary for growth and reproduction. These considerations form in part the basis for a commonly held view that bacteria and other microbes, in carrying out these processes, serve humans and other higher organisms as environmental recyclers and bioremediators. That view, while essentially correct, overlooks an essential point—these activities and processes are the fundamental biology of this planet. Microbes are the biosphere; their activities create and provide the foundation for all other life.

To gain a perspective on the significance of microbial diversity, imagine a biological survey crew tasked with discovering and documenting life forms on a newly encountered planet. Consider that upon landing, the crew found, remarkably, no macroscopic life. However, suppose that an initial sampling of a cubic centimeter of the planet's surface revealed the presence of millions of discrete microscopic cellular entities. Imagine that with much additional analysis these entities were found by the crew to represent thousands of different kinds of organisms, distinguishable by their morphology or their dramatically different ways of obtaining the energy and nutrients necessary for metabolism and reproduction. Would the survey crew be surprised? What if upon analysis of the genetic material from the different types present, these microbial life forms were confirmed to be different and were found in many cases to be dramatically more distinct from each other evolutionarily than are seaweeds and humans, would the crew be impressed? What if the crew continued sampling, spreading out and choosing other locations of the planet's surface, and found similar "species richness" wherever they looked, but often with little or no overlap in the types of entities present from one environment to the next. Would that start the crew thinking about, to paraphrase E. O. Wilson (1994), "a strange and vastly complex living world virtually without end"? It would, of course. However, there is no need to invoke new planets. This imaginary scenario describes the reality of microbial diversity on earth. The only difference is the microbially driven evolution of macroscopic life forms on earth, giving rise to the plants, animals, and macroscopic fungi.

#### IV. A NEW ERA IN BIOLOGICAL SCIENCES

Despite the complexity and richness of microbial life and the essential evolutionary and ecological roles

played by microbes, awareness of microorganisms remains limited. The diversity and scientific importance of microbes have been largely passed over in human society, in science, in biology, and even in discussions of biodiversity (Hawksworth, 1991), overshadowed by attention to macroscopic forms. The prevailing and erroneous view for many biologists is that bacteria are "primitive, simple and relatively uniform" (Pace, 1996). This view developed naturally from early technical and scientific limitations for discovering and studying bacteria and other microbes in the 18th and 19th centuries, such as the need for high-resolution microscopy and the need to understand cellular structure, biochemistry, polypeptides, and nucleic acids. Such limitations did not hinder as starkly the beginnings of macrobiology. Later, as limitations to the study of microbial life were overcome in the first three-quarters of the 20th century, the bias against microbes as scientifically important biological systems was nonetheless maintained and reinforced through the lack of a comprehensive phylogeny of microbes. This situation was especially true for bacteria, which generally lack distinctive morphological characters and for which sexual reproduction like that of plants and animals is absent. That bias remains largely extant today. For example, most college and university departments of biology and biological sciences are staffed predominantly with animal and plant biologists, with relatively few if any microbial biologists. Yet "the incongruity [between the scientific perception of microbiology and the preeminence of microorganisms in the real world] is astounding; it is worrisome; it cannot be scientifically justified or tolerated" (Woese, 1999). Fortunately, the bias and incongruity are beginning to be eliminated.

#### V. THE "DELFT SCHOOL" OF GENERAL MICROBIOLOGY

The confluence of two distinct but complementary approaches in biological science, one classical and one more recent, is leading to a shift in awareness about microbial diversity and its scientific importance. The classical approach is that of elective culture, also referred to as enrichment culture, by which new microbial types are brought into culture and isolated for phenotypic analysis. The elective culture approach, through the careful design of growth media and conditions, seeks to provide an appropriate physical and chemical environment that, when inoculated with an environ-

mental sample (mud, for example), will elicit the growth of specific metabolic types of bacteria postulated to be active or present in the sample. For success, the approach requires a thorough knowledge of biochemistry, good observational skills, and sensitivity to potential novelty in microbial metabolism and physiology. A recent example of this approach, demonstrating its central importance and value in microbial research, is the isolation of acetogenic spirochaete bacteria from the hindguts of termites (Leadbetter *et al.*, 1999). Spirochetes are major members of the diverse microbial consortium resident in the termite hindgut (Fig. 2), and

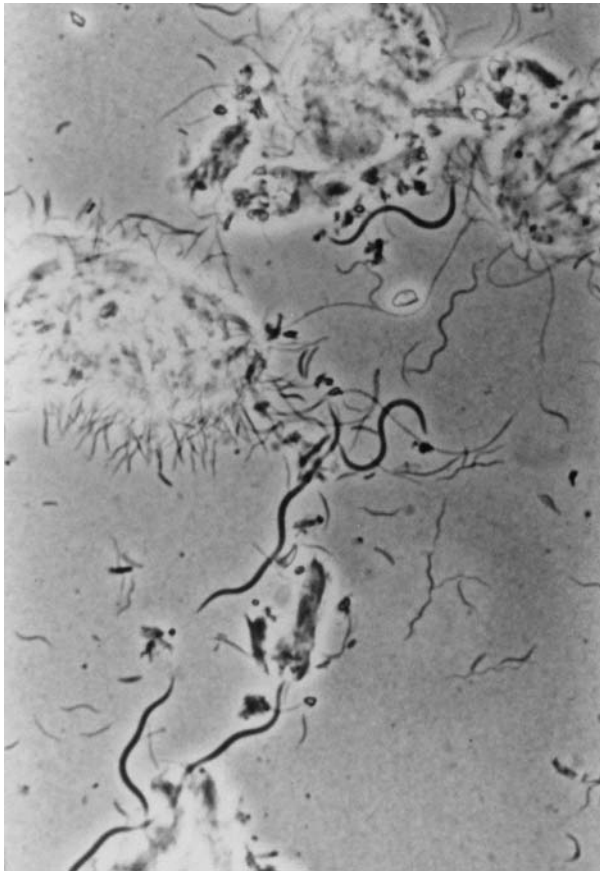


FIGURE 2 Phase contrast micrograph of the contents (diluted) of the hindgut of the termite *Reticulitermes flavipes* (Kollar). The microbial assemblage of this common eastern subterranean termite includes several types of microbes, including a variety of different kinds of spirochetes, both free and attached to cellulolytic protists.  $H_2$  and  $CO_2$ , formed during microbial cellulose digestion, is converted by certain of the spirochetes to acetate (Leadbetter *et al.*, 1999), a major source of nutrition for the animal. The different spirochaetal cells range in length from approximately  $5 \mu m$  to over  $30 \mu m$ . Courtesy of J. Breznak, Michigan State University.

they have been thought to play key roles in the insect's nutrition, which is based on microbial degradation of cellulose and conversion to acetate, a major carbon and energy source for the insect. However, no spirochetes from the termite gut previously had been isolated in pure culture. That inability limited knowledge of the contribution these morphologically distinct and numerically significant bacteria make to host animal nutrition. Leadbetter and coworkers, however, successfully established culture conditions that favored the growth of spirochetes over other bacteria and that simultaneously encouraged the growth of bacteria able to form acetate, from  $H_2$  and  $CO_2$ , breakdown products of cellulose degradation. In this way, acetogenic spirochetes from the termite hindgut were brought into pure culture for the first time.  $H_2/CO_2$  acetogenesis, a type of metabolism previously unknown among spirochetes, reveals an important way, formation of acetate, in which these bacteria contribute to termite nutrition. The ability to design culture conditions that elicit the growth of specific bacterial types known or suspected to be present in an environment is central to development of our understanding of microbial diversity and microbial ecology. Once a novel microbial type has been brought into culture, that organism's special or unique cellular metabolism, physiology, and genetics can then be studied in detail.

This now classical elective culture approach developed through the insights of Sergei Winogradsky, a Russian soil microbiologist, and Martinus Beijerinck and later Albert Kluyver, Beijerinck's successor, in Delft, Netherlands. Their work was instrumental in forming the foundations of general microbiology and microbial ecology. When Cornelis van Niel, a student of Kluyver, moved in 1928 from Delft to the Hopkins Marine Laboratory in Pacific Grove in California, he brought with him the "Delft School" tradition in microbiology, which he continued and further developed through his research and through a course he taught in general microbiology and comparative biochemistry (van Niel, 1949). Through his course van Niel trained a generation of microbiologists. Those individuals have gone on to use the Delft School, van Niel approach, in their research, and they in turn have taught other generations of microbiologists, further disseminating the Delft School tradition. Most notably today, the elective culture and isolation approach to the study of metabolically and ecologically diverse bacteria is fostered in young microbiologists and other scientists through the Microbial Diversity summer course of the Marine Biological Laboratory in Woods Hole.

## VI. THE "WOESEAN REFORMATION" OF MICROBIOLOGY

Despite the progress arising from the Delft School tradition in exploring and defining the diversity of bacteria and their metabolic capabilities, a major problem limited the development of general microbiology in the first three-quarters of this century: the lack of a unifying phylogeny of microorganisms. Various efforts at classifying bacteria and systematizing relationships among them, based on phenotypic characters of morphology and biochemical growth attributes, were attempted and abandoned. The frequent lack of characters, the instability or the shared nature of many characters, and the awareness that phenotypic characterization was not grounded necessarily at the genetic level left these efforts with major flaws. The inability to place microbes, especially bacteria, in an evolutionary context caused general microbiology and microbial diversity largely to languish as mainstream sciences at a time when other areas of biology and microbiology (e.g., clinical microbiology and biotechnology) were developing rapidly (Woese, 1999). What was needed was a unifying phylogenetic framework, founded at the genome level, which would allow the true evolutionary relationships among microbes to be analyzed critically and defined.

The use of informational macromolecules, begun more than 30 years ago, is now fulfilling that need for a unifying phylogeny of microbes (Woese, 1987; 1999) and is reforming our view of the evolution and diversity of microbial life. Analysis primarily of ribosomal RNA (rRNA) sequences, especially for the small subunit 16S and 16S-like rRNAs, has created "the first valid microbial phylogenetic systematics" (Jannasch, 1997). The functionally constant 16S and 16S-like molecules, common to all organisms, contain evolutionarily highly conserved regions, suitable for comparing less closely related organisms, and more variable regions, suitable for assessing evolutionary relationships in more closely related organisms. The universality of ribosomal RNA extends the value of this sequence comparison approach to all life forms, but importantly for bacteria it has established a unified phylogenetically based system with which to begin defining bacterial evolutionary relationships, a "first step in microbiology's reformation" (Woese, 1999). Along with 16S rRNA, other molecules, such as elongation factor Tu, 23S rRNA, and  $F_1F_0$  ATPase, also provide substantial phylogenetic information and can serve as alternative markers for inferring relationships (Ludwig and Schleifer, 1999). A variety of opportunities exist at various institutions for learning

the principles and application of molecular phylogenetic analysis in microbes and higher organisms. Notable among them is the Marine Biological Laboratory's course in Molecular Evolution, an intensive 3-week course dedicated to these topics.

A second step in the Woesean reformation of microbiology is the application of rRNA-based molecular phylogeny to microbial ecology (e.g., Hugenholtz *et al.*, 1998; Pace, 1996). Sequence analysis of 16S rRNAs, for example, extracted from natural environments provides direct access to the diversity of bacteria in that environment, bypassing the need for culturing microbes and giving a rapid and potentially comprehensive assessment of microbial community composition. rRNA-based approaches have opened up for study many microbial activities and associations. The result is a rapidly expanding awareness of the diversity and ecology of microbes, both culturable and not-yet-cultured (Amann *et al.*, 1995; Hugenholtz *et al.*, 1998; Pace, 1996; Pace, 1999).

The strengths of the elective culture and molecular phylogeny approaches make them naturally complementary. The cultivation of a new microbe leads to acquisition of the organism's 16S or 16S-like rRNA sequence in the context of data on its metabolism, physiology and habitat. That sequence provides a defined point of phylogenetic reference and a highly specific tool with which to examine the organism's distribution in the environment. Equally exciting is the opportunity to examine and explore the environment for the full, natural microbial diversity present, without concerns about the bias inherent in and finesse required for culturing. Those explorations identify and define more deeply within a unified phylogenetic framework the diversity of microbes present while also offering potential insights into their metabolism and physiology. That information, in turn, can motivate more refined or novel attempts at cultivation of the sequence-identified microbes. Each approach nurtures and magnifies the strengths of the other. For microbiology in the next century truly to "emerge as the primary biological discipline" (Woese, 1999) the continuing confluence of these approaches must be encouraged.

## VII. MAJOR GROUPS OF MICROBES

Through analysis primarily of 16S and 16S-like rRNA genes, microorganisms can now be placed in a potentially comprehensive phylogenetic framework, one that includes all living organisms and therefore is universal. Examination of the 16S and 16S-like rRNA-based uni-

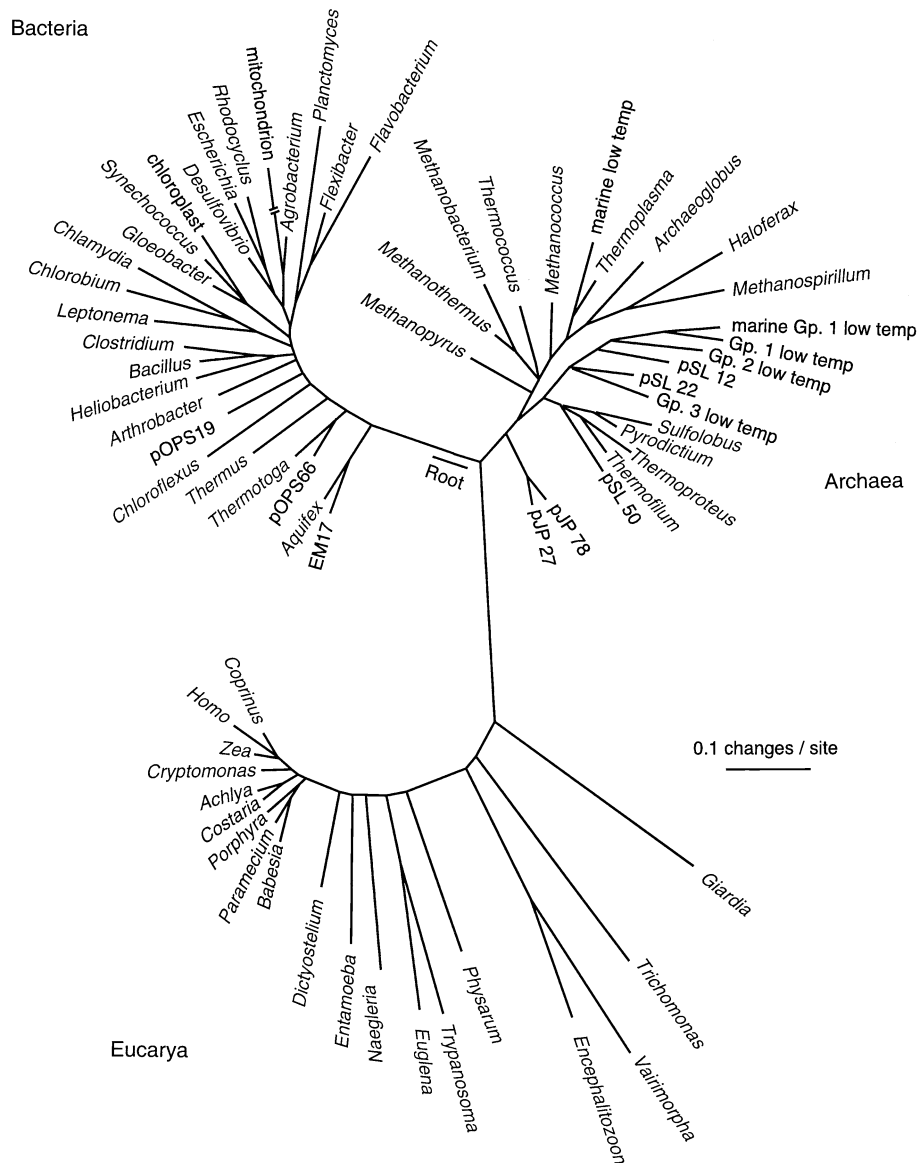


FIGURE 3 The universal tree of life. The three domains of life, Bacteria, Archaea, and Eucarya, represented by small-subunit rRNA sequences of various organisms within each domain, are shown. The domain Bacteria (Eubacteria) and the domain Archaea (previously the archaeobacteria) are entirely microbial, and the domain Eucarya is predominantly microbial. From Pace (1999), with permission.

versal phylogenetic tree (Fig. 3) shows the three currently proposed domains of life, Bacteria, Archaea, and Eucarya. Future analyses, using expanding sequence data sets and markers of phylogenetic relationships other than 16S and 16S-like rRNA (Ludwig and Schleifer, 1999) will serve to test and refine the validity of this evolutionary grouping. Regardless, the current universal tree reveals in a simple, compelling way the

dominance of microbial life forms. The true diversity of life is microbial.

### A. Domain Bacteria

The unit of biological diversity is the species. The classical concept of a biological species, a reproductively isolated interbreeding or potentially interbreeding pop-

ulation of individuals, however, does not work for bacteria and other microbes that do not interbreed, that have undefined or cosmopolitan distributions, and that generally lack distinguishing morphological characters. So, in discussing bacteria, members of the domains Bacteria and Archaea, how is a species defined? The phylogenetic species concept, a group of individuals for which phylogenetic analysis has demonstrated a shared genealogical relationship (Alexopoulos *et al.*, 1996), seems more workable for bacteria and other asexually reproducing microbes. The question then becomes the extent of genealogical relationship between two individuals necessary to designate them as members of the same species. Operationally for bacteria, if the genomic DNA of two strains is 70% or more similar, as determined by DNA-DNA hybridization analysis, they are considered the same species. With respect to rRNA, 16S rRNA sequence, similarity values below 97% are a strong indication that the two bacteria are different species (Amann *et al.*, 1995; Madigan *et al.*, 1999). The combination of these two approaches, both based at the genomic level, is powerful. When further combined with the identification of distinguishing phenotypic characters, such as cell morphology, motility, and flagellation, response to oxygen, requirement for organic growth substrates, growth temperature range, special metabolic attributes, characteristics of the habitat, and so on, a character-rich, biologically meaningful description of a bacterial species can be obtained.

Conservative estimates place the total number of species of bacteria at 50,000 to 3,000,000 (e.g., Staley *et al.*, 1997). As of 1999, approximately 5000 bacterial species had been described (Pace, 1999), a very small portion of the estimated total number, though many of these descriptions do not yet include rRNA sequence information. Most easily accessible environments, even those commonly sampled for microbial life, are likely to contain a multitude of as-yet-uncultured bacteria. Another important consideration in estimates of the number of bacterial species is the extent to which microhabitats, habitats relevant from the microbe's viewpoint, have been sampled. A reasonable assumption is that most biotic and many abiotic surfaces are colonized by bacteria. Each surface presents a different microhabitat and therefore is likely to be colonized by a different individual bacterial type or assemblages of types, with the "many different microenvironments creating an almost infinite variety of selective conditions" (Palleroni, 1994), that is, conditions selecting for the specific metabolic and physiological types. An estimated total number of plant and animal species is approximately 9 million, with insects making up the majority of those

species (Staley *et al.*, 1997). Presumably, the external surfaces of these organisms, and also the mouths, gut tracts, and internal tissues of the animals, provide myriad types of microhabitats for bacterial colonization by assemblages of different microbes (Amann *et al.*, 1995) and in many cases by specific bacterial symbionts, as in the various nutritional endosymbioses of between bacteria and insects. Consequently, these estimates of bacterial species must be too low. When the approximately 2 million species of fungi and protists, many of which undoubtedly also have largely different assemblages of microbes as well as specific bacterial symbionts, also are factored in, then the total number of bacterial species easily could exceed tens of millions. Future generations of microbiologists will likely find even this estimate to be conservative. Regardless, however, of what the actual total number of bacterial species turns out to be, bacteria are stunningly diverse.

Previously, all bacteria were grouped together taxonomically as prokaryotes. That grouping was based primarily on the common lack of a membrane-bounded nucleus in bacteria. Prokaryotic organisms, however, are now separated phylogenetically into two domains, the Bacteria (or Eubacteria) and the Archaea (formerly the archaeobacteria). The Archaea exhibit many similarities to the Eucarya, demonstrating that the prokaryotic body plan is not a phylogenetically definitive character. The bacteria exhibit a wide variety of ways of obtaining for growth the necessary carbon, as in the various kinds of heterotrophic and autotrophic microbes, and energy, as in phototrophic and chemotrophic microbes.

The domain Bacteria presently contains well over two dozen divisions, or kingdoms, of organisms (Fig. 4). Madigan *et al.* (1999) suggest, however, that the true number of kingdoms in the Bacteria may be 50 or more. As the number of these groups indicates, physiological diversity within the Bacteria is profound. The diversity is especially striking in two of the kingdoms, the Proteobacteria and the Gram-positive bacteria (Madigan *et al.*, 1999), members of which exhibit a wide range of different metabolisms, physiologies, morphologies, and habitats. Examples among the Proteobacteria include the human enteric bacterium *Escherichia coli*, pathogenic pseudomonads, light-producing marine photobacteria, the fever-causing rickettsias, gliding bacteria and fruiting-body formers, stalked and budding bacteria, nitrogen-fixers, sulfate reducers, and so on. Within the Gram-positive kingdom are staphylococcal parasites of humans, the milk-sugar fermenting lactobacilli, and spore-forming soil bacteria, among many others. This diversity, however, probably reflects more the relative ease of cultivation and long-



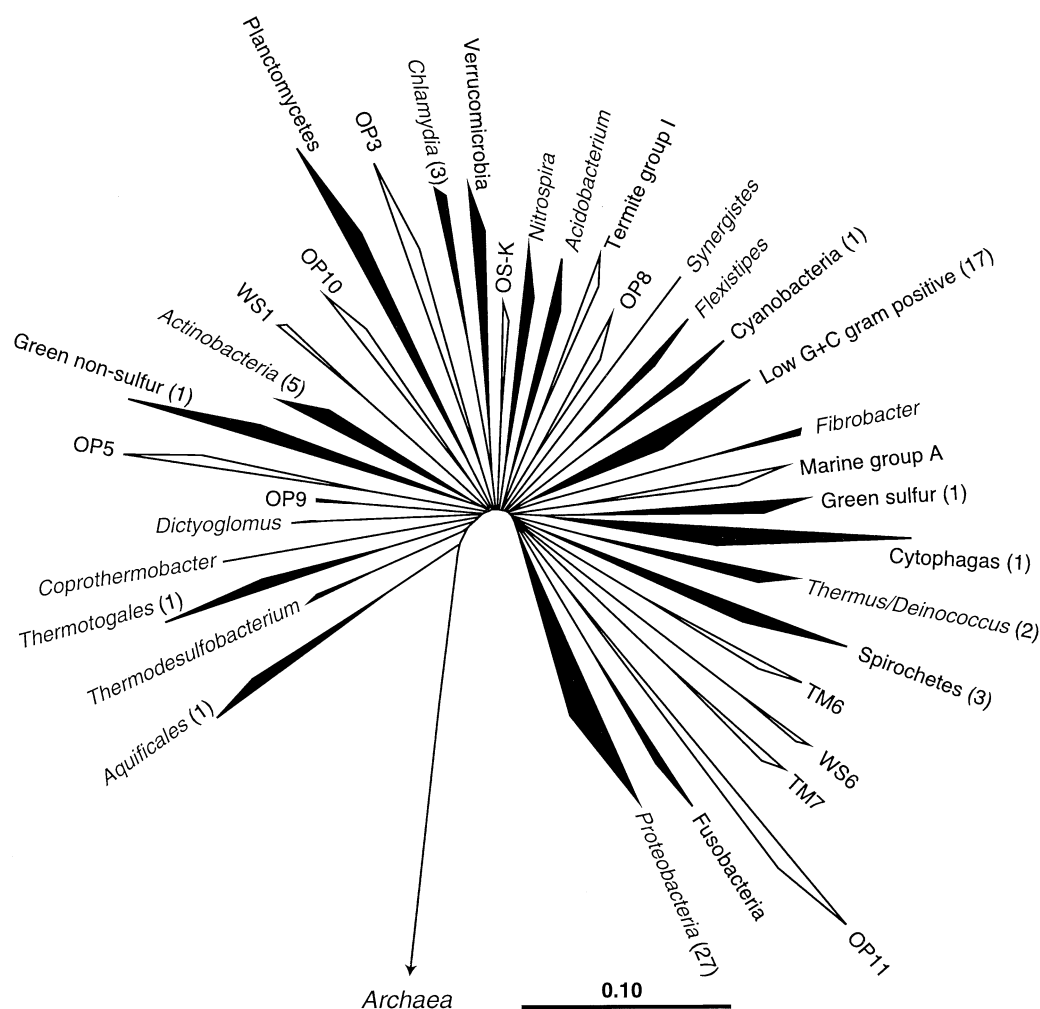


FIGURE 4 Diversity and phylogenetic relationships among members of the domain Bacteria. The division level grouping are of two types: currently recognized divisions, represented by cultivated bacteria (black), and provisional divisions, represented to date by environmental sequences (outline). The scale bar represents 0.1 changes per nucleotide. From Pace (1999), with permission.

term scientific attention members of these two kingdoms have received than the true biological richness of the groups. One can anticipate that much more diversity in other kingdoms of the Bacteria will be revealed with time as bacteria are studied more intensively. Furthermore, as the fusion between elective culture and molecular phylogeny develops, many other types of bacteria, some entirely unexpected, will be discovered and will be found to warrant kingdom status. A recent text (Madigan *et al.*, 1999) highlights the different groups of bacteria, providing information on individual types within a phylogenetic context.

## B. Domain Archaea

The domain Archaea includes the majority of presently known "extremophiles," organisms that live at physical or chemical extremes. Archaea increasingly are being discovered, however, in less extreme types of environments, including the marine plankton, lakes, and sediments (e.g., DeLong, 1998; Vetriani *et al.*, 1999). Diversity within the Archaea is presently less well understood than in the Bacteria and Eucarya because the Archaea often require particular care to culture. Knowledge of how to culture the Archaea has expanded in recent

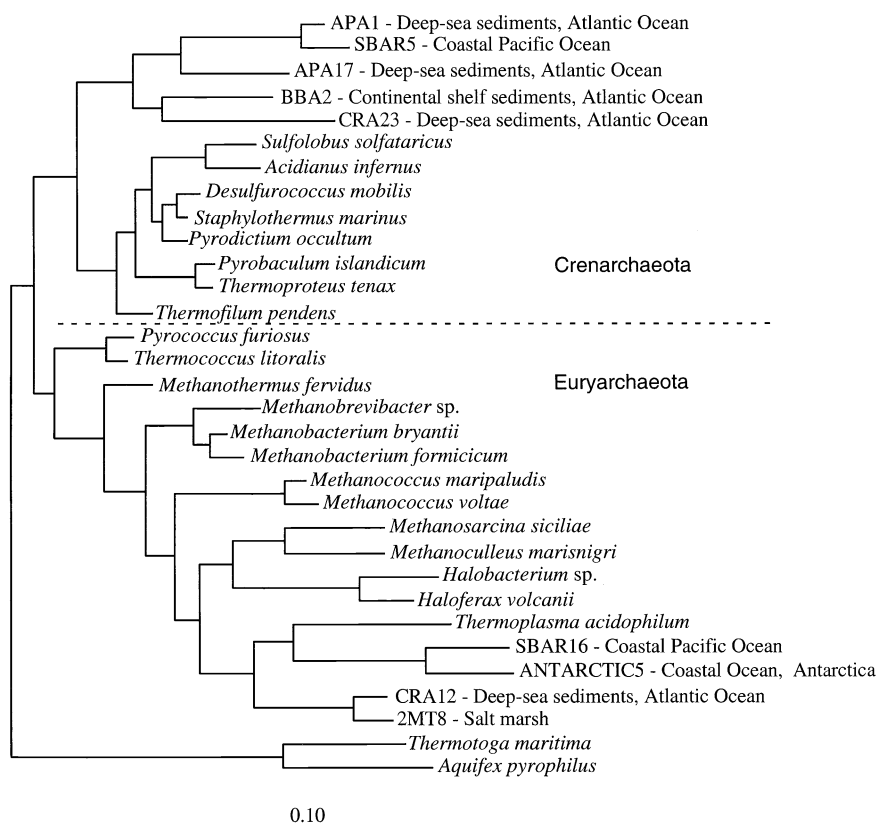


FIGURE 5 Diversity and phylogenetic relationships among members of the domain Archaea. This maximum likelihood tree, based on small-subunit rRNA, shows representatives of the Crenarchaeota (some extreme thermophiles and some marine environmental phylotypes) and of the Euryarchaeota (some extreme thermophiles, methanogens, halophiles, and marine environmental phylotypes). The environmental phylotypes are indicated with acronyms and numbers, with their environmental origins designated. *T. maritima* and *A. pyrophilus* form the bacterial out-group. The scale represents the expected number of changes per sequence position. See Vetriani *et al.*, 1999 for details and references. Prepared and provided by C. Vetriani, Rutgers University.

years, and additional habitats supporting their growth are being actively studied, so the disparity in the numbers of different kinds of Archaea compared Bacteria should progressively diminish. The Archaea consists at this time of three kingdoms, the Euryarchaeota, the Crenarchaeota, and a provisional kingdom, the Korarchaeota (Fig. 5).

Two major types of bacteria, the methanogens and the extreme halophiles, are included within the Euryarchaeota. Methanogens are bacteria that produce methane as an end-product of energy conversion reactions; they occur in a variety of strictly anaerobic habitats, such as sediments, sewage sludge digestors, the rumen of cattle, and the termite hindgut. CO<sub>2</sub> and similar compounds, methanol and other methyl-containing compounds, and acetate are used by different members of

this group as substrates for methanogenesis, often with H<sub>2</sub> as an electron donor. Methane, primarily from methanogenic bacteria, is an important greenhouse gas, accounting for possibly as much as a few percent of total primary production. Some methanogens, such as *Methanococcus jannaschii*, are hyperthermophiles. The extreme halophiles, represented by *Halobacterium salinarum*, are bacteria that require for survival and growth the exceptionally high salt concentrations found in salt lakes and solar evaporation ponds. Other groups of Euryarchaeota include a lineage of extremely acidophilic bacteria containing *Thermoplasma* and *Picrophilus*, and two lineages of hyperthermophiles, represented by *Pyrococcus* and *Archaeoglobus* (Fig. 4).

The Crenarchaeota contains a large and physiologically diverse group of hyperthermophilic, sulfur-metab-

olizing bacteria from terrestrial and marine hot springs and hydrothermal vents. Recently, several Crenarchaeota (and Euryarchaeota) have been identified at the 16S rRNA level as members of the plankton in cold oceanic surface and deep waters, coastal sediments, lakes, and in association with animals, indicating that the Archaea are more cosmopolitan in their distribution than believed earlier (e.g., DeLong, 1998; Vetriani *et al.*, 1999).

A third kingdom of Archaea, the Korarchaeota, was established provisionally based on rRNA sequences obtained from samples of the Obsidian Pool hot spring in Yellowstone National Park and distinct from those of other Archaea (Pace, 1996). Attempts to bring these hyperthermophiles into pure culture are underway. The discovery of this third Archaeal kingdom and the recent discoveries of Archaea in cold, oxygenated habitats clearly indicate that the true diversity of Archaea is likely to far exceed that based on presently identified species and sequences obtained from environmental samples. Implicit in this newly evolving view of Archaeal phylogenetic diversity are substantially broader metabolic capabilities and wider ecological roles for this group of bacteria (DeLong, 1998; Vetriani *et al.*, 1999).

### C. Domain Eucarya

Microbial groups in the Eucarya are the Protista and the Fungi, organisms that, in contrast to the Bacteria and the Archaea, have a membrane-bounded nucleus. Endosymbiotic events are likely to have been major driving forces in the evolution of eukaryotes. Bacteria are thought to have diverged early from a universal prokaryotic ancestor, followed by the Archaea, both of which retained the prokaryotic body plan. Fusion of an archaean and a bacterium may have led to the nuclear line, which through symbiotic acquisition of phototrophic and nonphototrophic bacteria resulted in chloroplast- and mitochondria-bearing eukaryotic cells. Loss of one or both of these organelles, or failure to acquire them initially, along with secondary symbiotic events, can be seen to account for much of the diversity of modern eukaryotes (Madigan *et al.*, 1999).

Previous groupings placed eukaryotes into four kingdoms: animals, plants, fungi, and protists, with a fifth kingdom, Monera, to contain all the bacteria. Current understanding of phylogeny indicates that this five-kingdom system greatly underemphasized the diversity of bacteria while overemphasizing animals and plants, as described earlier. The five-kingdom system also underrepresented the diversity within the protists. The current, developing view, based on a rapidly increasing

database of 16S-like rRNA sequence information, is that the Eucarya consists predominantly of unicellular microorganisms and that many phylogenetically deep, kingdom-level divisions exist within the protists (Sogin *et al.*, 1996). Diversity within the protists dominates that of the other eukaryotic lineages (Fig. 6).

#### 1. Protista

The Protista is a large complex grouping of mostly unicellular eukaryotic organisms. They are morphologically diverse and can be found in most terrestrial, aquatic, and marine habitats as free-living forms and as parasites of other protists, of fungi, and of plants and animals. With their nutritional modes restricted primarily to osmo- and phago-heterotrophy and phototrophy, protists are metabolically much less diverse than Bacteria and Archaea. Along with various independent amoeboid groups, major groupings include the Alveolates, composed of ciliates (e.g., *Paramecium*), dinoflagellates (e.g., *Alexandrium*), and apicomplexans (e.g., *Plasmodium*), and the Stramenopiles, composed of the brown and golden-brown algae, diatoms, chrysophytes, oomycetes, and distinct groups of slime molds, among other groups. Cryptophytes, Rhodophytes, and Haptophytes are other major groupings of protists. Along with these groups are the diplomonads, trichomonads, microsporidia, amoeba-flagellates, and euglenoids (Fig. 6).

#### 2. Fungi

The fungi, *sensu strictu*, are commonly filamentous, multicellular heterotrophic organisms. Though previously thought to be similar to plants but lacking chlorophyll, fungi phylogenetically are not closely related to plants. Instead, fungi are seen now to have diverged from the animal lineage (Alexopoulos *et al.*, 1996). Many fungi are saprobic, feeding osmotrophically on dead organic matter, and many are also parasites or symbionts of animals. As such, they share the limited range of metabolic capability of animals. Approximately 69,000 species of fungi have been identified, and more than 1,500,000 species are estimated to exist (Hawksworth, 1991).

Four major groups (phyla) of true fungi have been defined (Alexopoulos *et al.*, 1996) (Fig. 7). The Chytridiomycota contains a single class, Chytridiomycetes, which uniquely among fungi produces motile cells during its life cycle. The motility organelle, a typical eukaryotic flagellum, probably was retained from ancestral protists (Berbee and Taylor, 1999). Chytrids play important ecological roles in decomposing organic materials. The Zygomycota contains two classes, the Zygomyc-

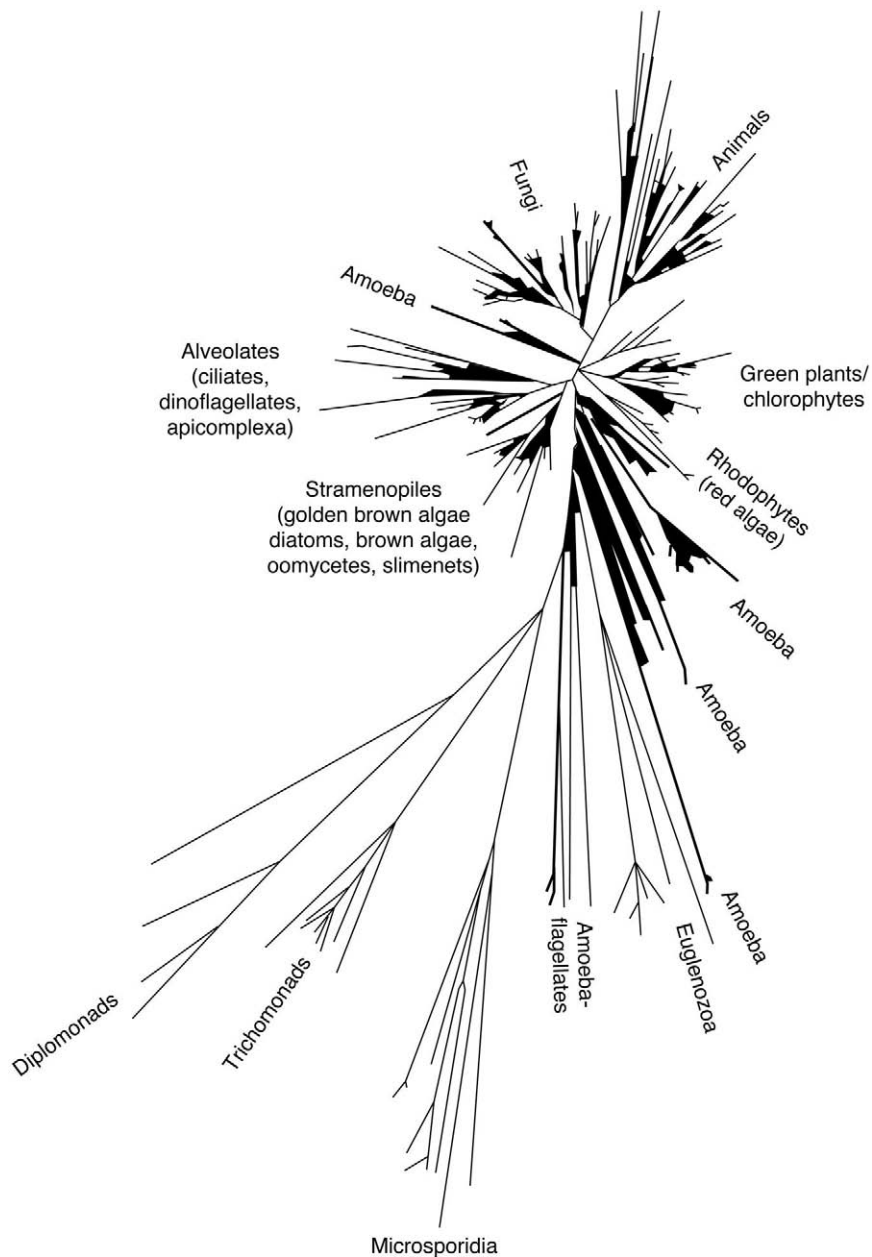


FIGURE 6 Diversity phylogenetic relationships among members of the domain Eucarya. This evolutionary tree reveals the dominance of the protists over the fungi, animals, and plants. From Sogin *et al.* (1996), with permission.

cetes, which form thick-walled resting zygospores, and the Trichomycetes, obligate symbionts of arthropods. The Ascomycota, which form ascospore-carrying asci, contains most of the lichen-forming fungi. The Basidiomycota, which produces sexual basidiospores on specialized basidia, contains many of the commonly recognized fungi, such as mushrooms, puffballs, and bracket

fungi. Plant pathogens in this phylum include the rust and smut fungi. Other fungi-like microbes commonly studied by mycologists now are grouped with the protists. These include the Stramenopile groups of oomycetes, hyphochytrids, and labyrinthulids, the plasmodiophorids, and the dictyostelid, plasmodial, and acrasid slime molds. The phylogenetic diversity and evolution

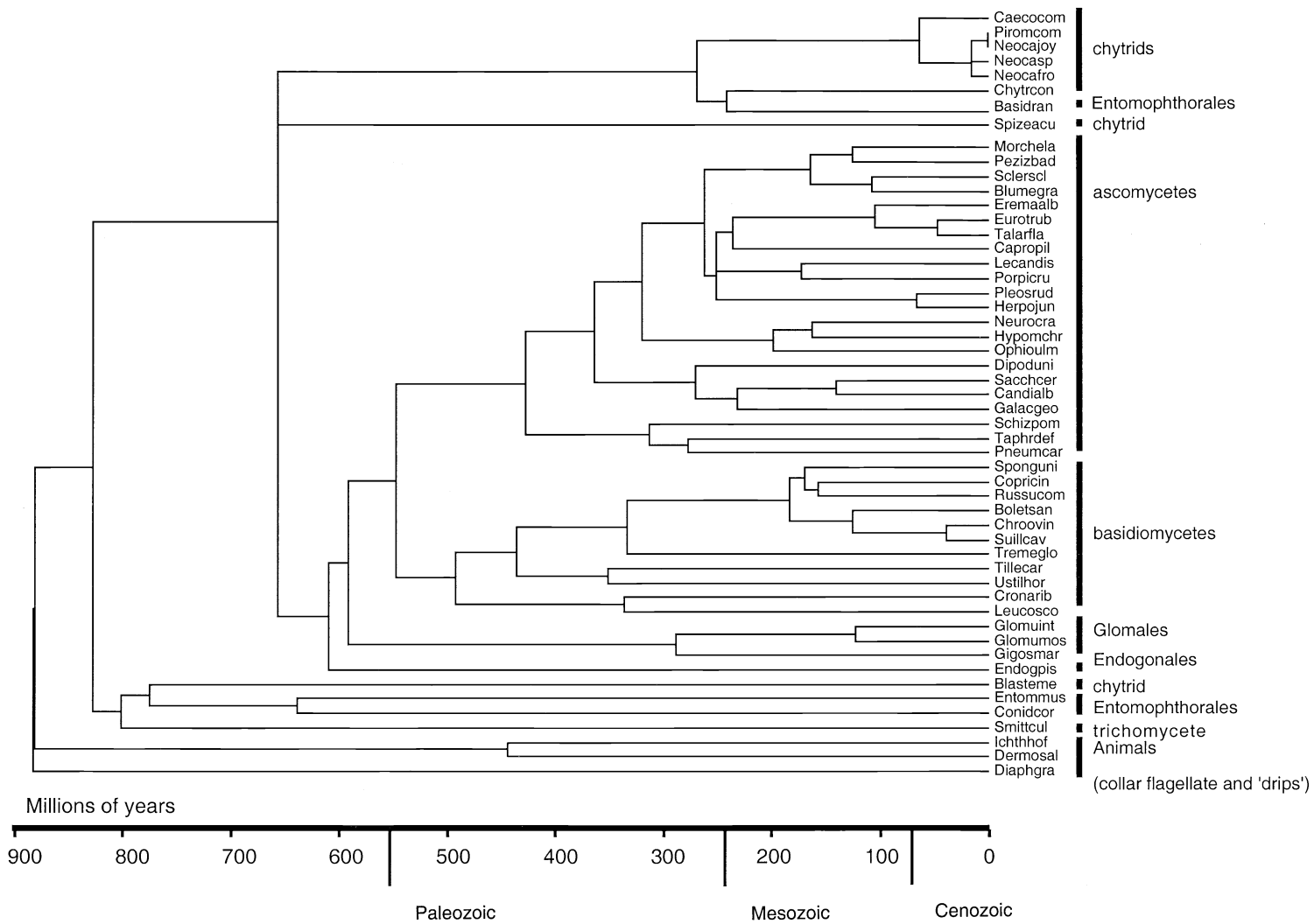


FIGURE 7 Diversity of fungi and the timing of their divergences. See Berbee and Taylor (1999) for details. Species names are abbreviated with the first five letters of the genus followed by the first three letters of the specific epithet (e.g., *Dermocystidium salmonis* = Dermosal). From Berbee and Taylor (1999), with permission.

of the fungi is an area of active research (e.g., Berbee and Taylor, 1999).

## VIII. CONCLUDING COMMENTS

Understanding the extent and significance of microbial diversity “is the primary goal, the necessary objective, of biology in the next century” (Woese, 1999). Where progress toward that goal will take biology, however, is difficult to even guess, since the unexplored, unrecognized diversity of microbes is immense. However, the fusion of elective culture and molecular phylogeny is creating a revolution in our understanding of that diversity and of the ecological and evolutionary significance of microbes, especially bacteria. Molecular phylogeny also is overturning the previously limited view of protist diversity and is enhancing awareness of fungal phylogeny and diversity. Add to this microbial revolution, just underway in the beginning years of the 21st century, the developing influences of genomic sequencing, differential display and proteomics, and one sees a turbulent, radical, and exciting restructuring of biology underway, one that is being spearheaded by new awareness of the extent and significance of microbial diversity. Important beyond measure in the success of that restructuring will be attracting talented students to professions in microbial research. While biotechnological aspects of microbiology will continue to appeal to capable young researchers, many of the best students will be attracted not by the usefulness of microbes but by their inherent biological beauty and by the potential their study offers for insights into the evolution of life on earth.

## Acknowledgments

The author is indebted to M. Sogin, N. Pace, J. Taylor, M. Berbee, K. Suberkropp, J. Breznak, T. Schmidt, J. Fuhrman, C. Vetriani, and E. Leadbetter for providing or suggesting sources of information; to J. Breznak for providing the photograph of termite hindgut microbes; to C. Vetriani for providing the figure of Archaea phylogeny; and to N. Pace, M. Sogin, M. Berbee, and J. Taylor for permission to use published figures. Research in the author's laboratory is supported by a grant from the National Science Foundation.

## See Also the Following Articles

ARCHAEA, ORIGIN OF • BACTERIAL BIODIVERSITY •  
EUKARYOTES, ORIGIN OF • MICROBIAL BIODIVERSITY,  
MEASUREMENT OF • MICROORGANISMS, ROLE OF

## Bibliography

- Alexopoulos, C. J., Mims, C. W., and Blackwell, M. (1996). *Introductory Mycology*, 4th ed. John Wiley & Sons, New York.
- Amann, R. I., Ludwig, W., and Schleifer, K.-H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.* **59**, 143–169.
- Berbee, M. L., and Taylor, J. W. (1999). Fungal molecular evolution: Gene trees and geologic time. In *Mycota* (D. J. McLaughlin, Ed.), vol. 7, pp. 00–00. Springer, New York, N.Y.
- Claes, M. F., and Dunlap, P. V. (2000). Aposymbiotic culture of the sepiolid squid *Euprymna scolopes*: Role of the symbiotic bacterium *Vibrio fischeri* in host animal growth, development and light organ morphogenesis. *J. Exp. Zool.* **286**, 280–296.
- DeLong, E. F. (1998). Everything in moderation: archaea as “non-extremophiles.” *Curr. Opin. Genet. Dev.* **8**, 649–654.
- Gould, S. J. (1996). *Full House*. Three Rivers Press, New York.
- Hawksworth, D. L. (1991). The fungal dimension of biodiversity: magnitude, significance, and conservation. *Mycolog. Res.* **95**, 641–655.
- Hugenholtz, P., Goebel, B. M., and Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180**, 4765–4774.
- Jannasch, H. W. (1997). Small is powerful: Recollections of a microbiologist and oceanographer. *Annu. Rev. Microbiol.* **51**, 1–45.
- Leadbetter, J. R., Schmidt, T. M., Graber, J. R., and Breznak, J. A. (1999). Acetogenesis from H<sub>2</sub> plus CO<sub>2</sub> by spirochetes from termite guts. *Science* **283**, 686–689.
- Ludwig, W., and Schleifer, K.-H. (1999). Phylogeny of *Bacteria* beyond the 16S rRNA standard. *ASM News* **65**, 752–757.
- Madigan, M. T., Martinko, J., and Parker, J. (1999). *Brock Biology of Microorganisms*, 9th ed. Prentice Hall, Upper Saddle River, NJ.
- Pace, N. R. (1996). New perspective on the natural microbial world: Molecular microbial ecology. *ASM News* **62**, 463–470.
- Pace, N. R. (1999). Microbial ecology and diversity. *ASM News* **65**, 328–333.
- Palleroni, N. J. (1994). Some reflections on bacterial diversity. *ASM News* **60**, 537–540.
- Sogin, M. L., Morrison, H. G., Hinkle, G., and Silberman, J. D. (1996). Ancestral relationships of the major eukaryotic lineages. *Microbiology Sem.* **12**, 17–28.
- Staley, J. T., Castenholz, R. W., Colwell, R. R., Holt, J. G., Kane, M. D., Pace, N. R., Salyers, A., and Tiedje, J. M. (1997). *The Microbial World: Foundation of the Biosphere: A Report of the American Academy of Microbiology*. American Society for Microbiology, Washington, D.C.
- van Niel, C. B. (1949). The “Delft School” and the rise of general microbiology. *Bacteriol. Rev.* **13**, 161–174.
- Vetriani, C., Jannasch, H. W., MacGregor, B. J., Stahl, D. A., and Reysenbach, A.-L., (1999). Population structure and phylogenetic characterization of marine benthic Archaea in deep-sea sediments. *Appl. Environ. Microbiol.* **65**, 4375–4384.
- Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci. USA* **95**, 6578–6583.
- Wilson, E. O. (1994). *Naturalist*. Island Press, Washington, D.C.
- Woese, C. R. (1987). Bacterial evolution. *Microbiol. Rev.* **51**, 221–272.
- Woese, C. (1999). The quest for Darwin's grail. *ASM News* **65**, 260–263.





# MICROORGANISMS (MICROBES), ROLE OF

Tom Fenchel  
*University of Copenhagen*

---

- I. Definition of Microbes
  - II. Prokaryotes (Bacteria)
  - III. Unicellular Eukaryotes (Protists)
  - IV. Roles of Microbes over Geological Time and in the Contemporary Biosphere
  - V. Microbes and Man
- 

## GLOSSARY

**archaeobacteria (Archaea)** One of two major groups of prokaryotes, including methane-producing bacteria, extreme halophiles, and extreme thermophiles.

**chemoautotrophy** Metabolism exclusively based on the oxidation of inorganic compounds.

**cyanobacteria** A large group of eubacteria with chlorophyll *a* and oxygenic photosynthesis; formerly known as “blue-green algae.”

**eukaryotes** Organisms with a compartmentalized nucleus in their cells; include animals, fungi, and plants as well as a large number of microbial groups collectively referred to as protists.

**phagotrophy** Feeding on particulate matter.

**phototrophy** Energy metabolism based on light energy.

**prokaryotes** Bacteria; cells without a compartmentalized nucleus.

**protists** Protozoa + protophytes; all microbial eukaryotes.

**protophytes** A diverse assemblage of partly unrelated groups of eukaryotic, phototrophic microorganisms.

**protozoa** A diverse assemblage of partly unrelated groups of eukaryotic, mostly motile and phagotrophic microorganisms; some phototrophic flagellates and fungi-like protists are traditionally included.

**sulfur bacteria** Bacteria that depend on the phototrophic or chemotrophic oxidation of reduced sulfur compounds.

---

**MICROORGANISMS** (microbes) are those life forms too small to be seen by the naked eye; i.e., that require a microscope or other form of magnification in order to be observed. The term microorganism is thus a functional description rather than a taxonomic one, and the grouping includes a wide variety of organisms.

## I. DEFINITION OF MICROBES

Microbes are defined mainly by their size; most are unicellular, but some form colonies which may be macroscopic. Some “microorganisms” even exceed the size of the tiniest animals (metazoa). Taxonomically, microbes include the prokaryotes or bacteria (the eubacteria and the archaeobacteria) and members of various unrelated eukaryotic groups, together referred to as protists or protozoans (including some fungi-like organisms) + protophytes. Contemporary understanding



of phylogeny (Fig. 1) reveals that biological diversity is largely microbial: In this representation, multicellular organisms (plants, animals, fungi, and macroscopic algae) appear as minor and evolutionarily recent additions to the “tree of life.” This is also reflected perhaps by the fact that except for vascular plants (which are largely responsible for primary production on land) biogeochemical cycling of elements is mainly the result of microbial activity.

Although we have defined microbes in terms of size, it should be emphasized that the size range of microbes is immense. The smallest, free-living bacteria measure less than  $0.5 \mu\text{m}$  and the largest protozoan cells measure  $>1 \text{ mm}$  (some foraminiferan species measure  $>1 \text{ cm}$ ). The size range of microbes,  $1 \mu\text{m}$  to  $1 \text{ mm}$  (a factor of  $10^3$  or a factor of  $10^9$  in terms of volume), approximately equals that of all vertebrates (guppies to whales).

Even so, some general properties of microbes that relate to their relatively small size can be identified. Among these, the most important is a high “rate of living”—that is, high weight-specific metabolic rates and high growth and reproduction rates. When organisms spanning a wide size range are compared, weight-specific metabolic rates tend to decrease proportionately to the  $1/4$  power of body weight, and generation times tend to increase proportionately to the  $3/4$  power of body weight. Under optimal conditions the generation (doubling) time of bacteria may be as short as 20 min and unicellular eukaryotes typically have genera-

tion times between 4 and 24 hr. Consequently, a relatively small microbial biomass may be responsible for a relatively large part of the flow of energy and materials in ecosystems. Certain microorganisms may in some circumstances form conspicuous accumulations of biomass in nature, but usually they are macroscopically invisible constituents of, for example, soils and seawater. However, due to their high weight-specific metabolic rates they dominate as agents of chemical transformations in the biosphere.

There are other characteristics of being small. Uptake of solutes from the surroundings (e.g., dissolved oxygen or organic substrates) is by molecular diffusion, a fact that is important in shaping microbial communities and which imposes constraints on physical transport rates. Due to their small size and low swimming velocities, microorganisms live at “low Reynolds numbers”; this implies that motility depends on viscous rather than on inertial forces. Many properties relating to motility and motile behavior of microorganisms therefore appear counterintuitive.

However, functional diversity of microorganisms is (as suggested by Fig. 1) more evident than their similarities. It is therefore expedient to discuss prokaryotic and eukaryotic microorganisms separately.

## II. PROKARYOTES (BACTERIA)

### A. Principle Properties of Bacteria

The prokaryotes include two major groups of life: the archaeobacteria (sometimes referred to as Archaea) and the eubacteria, which are differentiated by numerous genetic and biochemical traits. Structurally and functionally, however, they show so many similarities that it is appropriate to discuss them together. In contrast to eukaryotic cells, bacteria do not have a cytoskeleton and almost all bacteria are enclosed by a rigid cell wall. These features result in certain general properties. The limitation of diffusional solute transport from the surrounding water and within the cell typically constrains bacterial size to 1 or  $2 \mu\text{m}$ . Certain giant bacteria (mainly among cyanobacteria and sulfur bacteria) measure  $5\text{--}10 \mu\text{m}$  or more, but they usually include a large internal vacuole. Bacteria take up only low-molecular-weight solutes from their surroundings. Bacteria that depend on high-molecular-weight polymeric compounds as a source of energy and organic carbon must first hydrolyze their substrates extracellularly, using membrane-bound enzymes, before the resulting monomers can be transported into the cell. This transport

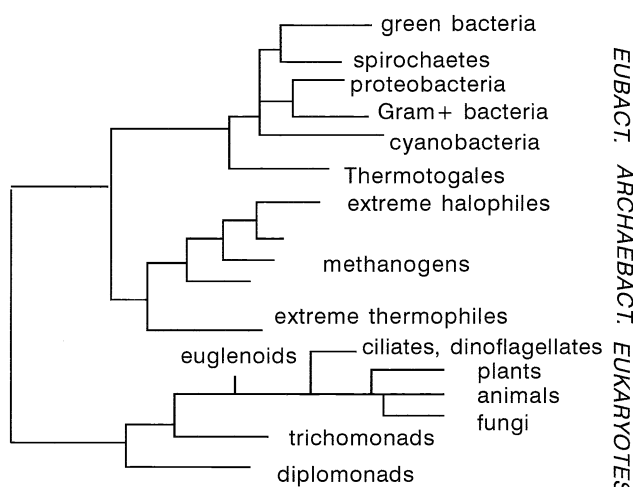


FIGURE 1 The “universal tree” based on rRNA sequencing. Microbes include, in addition to the prokaryotes (Archaeobacteria and Eubacteria), most major eukaryotic groups. For clarity, only a few major groups are included in the tree. Only animals, fungi, plants, and a few (omitted) algal groups represent macroscopic, multicellular organisms; all other organisms are “microbes.”

may be “passive” (facilitated diffusion) or “active” (ATP dependent). The small size of bacteria means that they can exploit and grow in extremely dilute substrate solutions (e.g., approximately 50 nmol per liter of amino acids or glucose).

Two additional properties explain the important role of bacteria in the biosphere; these apply to bacteria as a group rather than to individual species. One is that many bacteria are “extremophiles”: Extreme thermophilic bacteria live at temperatures  $>80^{\circ}\text{C}$  (or  $>100^{\circ}\text{C}$  under hyperbaric pressure); extreme halophiles may grow in saturated brine; acidophilic and alkalophilic bacteria occur at a pH  $<2$  and  $>10$ , respectively; and some bacteria tolerate toxic metal ions (e.g., copper and zinc) in millimolar concentrations. A variety of habitats (hot springs, brines, subsurface, etc.) are inhabited exclusively by prokaryotes. Many natural habitats lack oxygen (e.g., aquatic sediments, stratified water bodies, water saturated soils, or even the interior of soil aggregates and detrital particles) because diffusive supply of  $\text{O}_2$  cannot meet the demand for aerobic microbial respiration. Many bacteria have an anaerobic energy metabolism and they are practically the only inhabitants of anoxic environments. It seems that bacteria have only one absolute requirement for metabolism and growth: liquid water. Active “terrestrial” bacteria (and other microorganisms) are therefore confined to water films surrounding soil particles, leaf litter, plant surfaces, and the like.

Another collective property of bacteria is their metabolic diversity, which far exceeds that found among eukaryotes. Metabolism of bacteria includes processes such as the oxidation and reduction of inorganic sulfur, nitrogen, iron, and manganese compounds and the production and oxidation of methane. Some bacteria can grow using hydrocarbons as a substrate. Some phototrophic bacteria use  $\text{H}_2$ ,  $\text{H}_2\text{S}$ , or  $\text{Fe}^{2+}$  as electron donors (rather than  $\text{H}_2\text{O}$  as in algae, plants, and cyanobacteria) and many bacteria can assimilate atmospheric nitrogen. These are all fundamental (but exclusively bacterial) processes in the biosphere. In addition, all natural and many xenobiotic polymers can be hydrolyzed by at least one type of bacterium.

## B. Metabolic Diversity of Bacteria

It is convenient to distinguish between dissimilatory (energy and catabolic) metabolism and assimilatory (anabolic) metabolism, although the distinction is not always sharp. Heterotrophic bacteria typically use a given organic substrate for dissimilation and a carbon source for growth, whereas autotrophic bacteria must

use energy gained in dissimilatory metabolism for the reduction of  $\text{CO}_2$ . Energy generation and growth are coupled since most energy generated is used for the synthesis of macromolecules (DNA, RNA, and proteins), for active transport of substrates into the cell, and sometimes for assimilatory reductions. In microorganisms there is therefore typically an almost linear proportionality between energy generation and reproductive rate. Here, we concentrate on energy metabolism because it provides an overview of the functional diversity and roles of bacteria in the biosphere, notwithstanding that the synthesis of bacterial biomass is also an important process in terms of providing a basis for phagotrophic food chains.

Table I presents a (simplified) overview of major types of bacterial energy metabolism. Different types of microbial processes predominate in different habitats according to the chemical environment (availability of substrates and electron acceptors), to thermodynamics (energy yields of the different processes), and to certain physiological constraints. Many of the listed processes are interdependent in nature in that one functional type of bacteria requires the presence of other types of bacteria.

Under oxic conditions, aerobic bacteria completely dominate in accordance with the fact that oxygen respiration yields more energy than any other metabolic process. Bacterial diversity in aerobic habitats is to a large extent due to differential abilities to hydrolyze different polymers, but most species are capable of complete mineralization of organic monomers to  $\text{CO}_2$ ,  $\text{H}_2\text{O}$ , and mineral N or of the complete oxidation of reduced inorganic compounds.

Anaerobic bacteria, on the other hand, are specialists and a complete mineralization under anaerobic conditions requires several physiological types of bacteria. Fermenting bacteria play a key role in anaerobic habitats because they are (nearly) the only anaerobes capable of hydrolyzing polymers. The complete fermentation of, for example, carbohydrates (to acetate +  $\text{CO}_2$  +  $\text{H}_2$ ) is thermodynamically possible only if the ambient  $\text{H}_2$  tension is low ( $<$  about  $10^{-4}$  atm). Low  $\text{H}_2$  tension results from the activities of  $\text{H}_2$ -consuming bacteria, notably sulfate reducers and methanogens. This syntrophic interspecies  $\text{H}_2$  transfer is an essential feature of anaerobic habitats. In general, metabolic end products of fermenting bacteria serve as substrates for anaerobic respirers. The type of respiration that will dominate depends on the availability of external energy acceptors and on the energy yield of the different types of oxidation. After aerobic respiration, nitrate respiration is the energetically most favorable process, but its

TABLE I  
Principal Types of Energy Metabolism in Bacteria

Fermentation	Anaerobic processes in which (organic) molecules are dismutated. No external electron acceptor, redox equilibrium; low energy yield. Principal end products (of carbohydrate fermentation): acetate and $\text{CO}_2 + \text{H}_2$ , but many fermentation types are incomplete and metabolites then include other low-molecular-weight fatty acids and alcohols.
Respiration	Use of an external electron acceptor, electron transport phosphorylation.
Aerobic respiration	$\text{O}_2$ as terminal electron acceptor. A variety of organic substrates are used by different <i>heterotrophs</i> . <i>Chemoautotrophs</i> use $\text{O}_2$ to oxidize various inorganic substrates (reduced S, N, Fe, or Mn compounds and $\text{H}_2$ and $\text{CH}_4$ ). Aerobic respiration provides the highest energy yield of any known metabolic process.
Anaerobic respiration	Use of external electron acceptors other than $\text{O}_2$ . <i>Denitrifiers</i> use $\text{NO}_3^-$ (reduced mainly to $\text{N}_2$ or $\text{N}_2\text{O}$ ) to oxidize a variety of organic compounds and reduced S compounds (but not $\text{CH}_4$ ). <i>Sulfate reducers</i> produce $\text{H}_2\text{S}$ while oxidizing especially $\text{H}_2$ or low-molecular-weight fatty acids. <i>Iron and manganese reducers</i> use $\text{Fe}^{3+}$ and $\text{Mn}^{4+}$ for the oxidation of substrates.
Methanogenesis	Methanogenesis is found only among certain anaerobic archaeobacteria. <i>Acetoclastic methanogens</i> dismutate acetate into $\text{CO}_2 + \text{CH}_4$ ; <i><math>\text{CO}_2/\text{H}_2</math> methanogens</i> produce $\text{CH}_4$ from $\text{H}_2 + \text{CO}_2$ . Some methanogens can also produce methane from reduced C-1 compounds (e.g., methanol).
Phototrophy	Phototrophs use light energy for generation of ATP and (with an external electron donor) for the reduction of $\text{CO}_2$ to organic compounds. <i>Oxygenic phototrophs</i> (cyanobacteria) use $\text{H}_2\text{O}$ as electron donor and produce $\text{O}_2$ as metabolite. <i>Anoxygenic phototrophs</i> (several groups of bacteria) use $\text{H}_2\text{S}$ , $\text{H}_2$ , or $\text{Fe}^{2+}$ as electron donors, thus producing $\text{S}^0/\text{SO}_4^{2-}$ , $\text{H}_2\text{O}$ , or $\text{Fe}^{3+}$ .

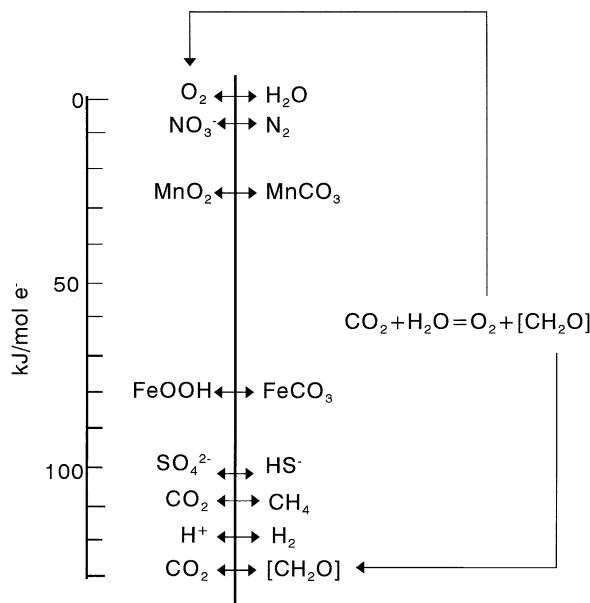


FIGURE 2 A biosphere model incorporating different microbial respiration processes. Oxygenic phototrophs use light energy to produce  $\text{O}_2$  and organic matter ( $[\text{CH}_2\text{O}]$ ); this represents the chemical potential that maintains the biosphere. Microorganisms derive energy by redox processes involving oxidants (left) and reductants (substrates) (right). The standard potential of the half-cell processes decreases from above so that the oxidant has to be situated higher than the reductant if the process is to be thermodynamically possible. Oxidations of organic matter may take place stepwise (e.g., first by sulfate reduction and the resulting sulfide can subsequently be oxidized by  $\text{O}_2$ ). A few processes (e.g.,  $\text{N}_2$  oxidation) are not realized due to kinetic constraints. Data on free energy changes are not precise under physiological conditions because the processes do not necessarily occur under standard conditions and because only a part of the potential energy is conserved as ATP; however, the energetic hierarchy of the processes holds.

quantitative importance is often limited by the availability of  $\text{NO}_3^-$ . When nitrate becomes exhausted,  $\text{Mn}^{4+}$  followed by  $\text{Fe}^{3+}$  are the electron acceptors of choice; when these are exhausted sulfate respiration prevails (Fig. 2). In marine sediments, sulfate reduction is typically the dominant process because of the high concentration of  $\text{SO}_4^{2-}$  in seawater (about 25 mM compared to about 250  $\mu\text{M}$   $\text{O}_2$  at atmospheric saturation; the capacity for oxidation of organic matter through sulfate reduction is about 200 times more than through  $\text{O}_2$  reduction and so, in contrast to  $\text{O}_2$ ,  $\text{SO}_4^{2-}$  supply is not rapidly exhausted). Only when sulfate is depleted will methanogenesis be quantitatively important; this rarely happens in marine environments, but it is important, for example, in lake sediments. In anaerobic habitats, the microbial processes are therefore spatially or temporally

structured; for example, when moving downwards from the surface of sediments,  $O_2$ ,  $Mn^{4+}$ ,  $Fe^{3+}$ , and  $SO_4^{2-}$  are depleted sequentially. The reduced end products of anaerobic metabolism diffuse upwards and are eventually reoxidized by  $O_2$  (or  $NO_3^-$ ) by chemoautotrophic bacteria residing at the aerobic–anaerobic interface. The complex interactions between different types of bacteria in anaerobic habitats and at the anaerobic–aerobic interface are often referred to as a “food chain” by microbiologists. It differs fundamentally from phagotrophic food chains, however, in that the organisms do not feed on each other but rather one functional group of bacteria utilizes the metabolites of another functional group.

Hydrolysis of polymers is often the rate-limiting step in mineralization. Some substrates are inherently difficult to hydrolyze, notably the ligno–cellulose complexes of woody tissue and many other natural plant compounds such as phenols, waxes, tannins, and cork substances (whereas pure carbohydrates, such as cellulose and hemicelluloses, are relatively rapidly degraded under aerobic and under anaerobic conditions). Hydrolysis may also be limited by the availability of mineral nutrients (N and P) that are assimilated (“immobilized”) by bacteria during decomposition of mineral-poor substrates (such as most plant tissue and litter). Crude oils (which contain a mixture of normal and branched paraffins, aromatic hydrocarbons, and other compounds) are degraded by a variety of aerobic bacteria and anaerobically by nitrate or sulfate reducers. Otherwise, easily degradable monomers (e.g., amino acids) may bind to clay minerals or humic substances to become temporarily or permanently inaccessible to bacterial attack. Humic substances are complexes consisting mainly of aromatic rings deriving from lignin, quinones, and phenols. Humic substances are very resistant to microbial attack and their rate of turnover in nature must be measured in centuries.

Some polymers are not, or are only very slowly, degraded under anaerobic conditions. This applies to, for example, lignin and some other plant compounds. Anoxia in conjunction with a low pH, such as found in moors and swamps, strongly limits mineralization, and peat, in which the original structure of plant material is often preserved, accumulates. Even in marine sediments, a certain fraction of the organic input is not mineralized but fossilized as kerogen in sedimentary rocks. Peat is eventually (over geological time through abiologic processes) transformed into lignite and coal. Large accumulations of organic matter in marine deposits may eventually transform into crude oils and natural gas.

## C. Roles of Bacteria for Element Cycling in the Biosphere

The single most important role of bacteria in the biosphere is the degradation and mineralization of organic matter produced by (mainly eukaryotic) phototrophs. In some habitats, however, primary production by prokaryotic phototrophs is also significant.

Mineralization processes are often complex due to spatial and temporal heterogeneity, limiting physical transport rates, and kinetic and physiological constraints. This complexity is especially evident where anaerobic processes are important as in sediments. There are considerable differences between major habitat types and these will be discussed separately.

### 1. The Water Column of Oceans and Lakes

Prokaryotic oxygenic phototrophs play a specific role in the water column. Thus, the unicellular, 1- $\mu$ m-large *Synechococcus* cells are ubiquitous in lakes and in seawater, typically occurring at densities of approximately  $10^5$  ml<sup>-1</sup>. Their role in primary production (relative to that of eukaryotic phototrophs) is especially important in oligotrophic oceanic waters where it may account for 30–70% of the total production. The tiny *Prochlorococcus* has recently been shown to be important in the deepest parts of the photic zone (>100 m) in oceanic waters. Blooms of large colonial filamentous cyanobacteria (*Trichodesmium*) occur periodically in some marine waters; the organism is capable of  $N_2$  fixation and is thus favored by P-rich but N-depleted water masses. Macroscopically conspicuous blooms of various species of colonial cyanobacteria are common phenomena in eutrophic fresh waters.

Aerobic heterotrophic bacteria, however, are the most important group of prokaryotes in the water column. They typically occur in numbers ranging between  $5 \times 10^5$  and  $5 \times 10^6$  cells ml<sup>-1</sup> constituting a volume fraction in seawater ( $\sim 10^{-6}$ ) comparable to that of other important biotic components (e.g., phytoplankton). Their relatively constant numbers imply that they are controlled by grazing (mainly by protozoa) but they are also subject to viral attack, and that their turnover is relatively rapid (varying from <1 to several days).

Bacteria depend on dissolved organic compounds which again derive from several sources. The most important one seems to be the passive excretion of photosynthate from phytoplankton cells; it is estimated that between 5 and 40% of primary production is lost in this way from algal cells to become mineralized by bacteria. Other sources include allochthonous material (e.g., runoff from land), degradation of macroalgae,

and “sloppy feeding” by zooplankton. Dissolved organic matter in natural waters covers a wide range of molecular size from monomers (amino acids and monosaccharides) to colloidal matter. Small monomers have a rapid turnover (sometimes <1 hr) and occur at very low concentrations (nM range) due to efficient bacterial utilization. Larger molecules are utilized much more slowly due to their lower diffusion coefficient and because their utilization requires extracellular hydrolysis. Some macromolecular compounds (humic substances) are very recalcitrant; they have a very low turnover rate but constitute the bulk of dissolved organic matter.

The fact that a relatively large part of primary production is channeled via dissolved organic matter to bacteria, which then enter phagotrophic food chains, has been termed the microbial loop (Fig. 3). The relative amount of the primary production which is channeled through the microbial loop compared to the “classical plankton food chain” (large phytoplankters → copepods → planktivorous fish → carnivorous fish) depends on circumstances. The classical food chain is particularly favored by large or periodic influx of mineral nutrients (such as in upwelling zones); otherwise, the microbial loop seems to dominate in terms of carbon flow.

Special microbial biota are associated with aggregates (“marine snow”) which consist of different types of particles (diatom frustules, fecal material from plankton organisms, etc.) held together by colloidal material. They form in the water column where they constitute approximately 10% of nonliving organic matter; they eventually sink to the bottom and become the basis for benthic life. They are, however, rapidly colonized by bacteria (and other microorganisms) and so a substantial part of the organic fraction is mineralized while still

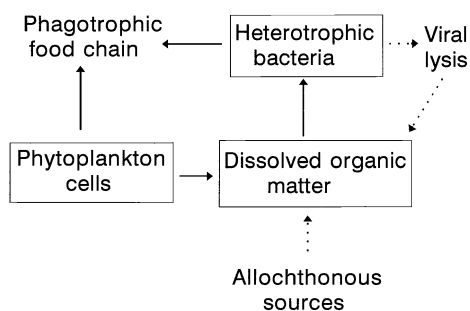


FIGURE 3 The “microbial loop” in the water column. A fraction of the photosynthate is lost from phytoplankton cells to become incorporated into bacteria, which they enter phagotrophic food chains.

suspended in the water column; it has been estimated that 75–80% of the organic matter is lost during a passage of 2000-m depth.

## 2. Aquatic Sediments

The most important property of sediments is that vertical solute transport depends on molecular diffusion. Organic material—produced in the water column—sediments and is in part mixed into the sediment through the activity of burrowing animals. The aerobic metabolism of microbes and animals deplete  $O_2$  at a certain depth, depending primarily on the input of organic material. In shallow-water sediments, anoxia typically prevails a few millimeters beneath the surface; at greater water depths the oxic zone typically measures several centimeters. Beneath the oxic–anoxic interface (the “chemocline”) mineralization is anaerobic. As previously explained, various other electron acceptors are used by bacteria for the oxidation of their substrates (Fig. 2). In marine sediments sulfate reduction dominates and, in shallow offshore sediments, may be responsible for more than 50% of the terminal mineralization. The resulting metabolite, hydrogen sulfide, is thus produced in copious amounts. Some of it combines reversibly with iron to form black  $FeS$ , but most diffuses upwards to become reoxidized by chemoautotrophic sulfur bacteria at the chemocline and therefore a major portion of the  $O_2$  uptake of marine sediments is due to the reoxidation of microbially produced sulfide. Some of the ferrous sulfide reacts (abiotically) with sulfide to form the somewhat more stable pyrite ( $FeS_2$ ), which appears as a type of fossil fuel in sedimentary rocks.

In sulfate-poor freshwater systems methanogenesis plays a major role as the terminal anaerobic mineralization process. Some of the  $CH_4$  formed is reoxidized aerobically in the sediment surface or in the water column; since  $CH_4$  (in contrast to sulfide) has a low solubility, much escapes to the atmosphere via ebullition.

In productive shallow waters the anaerobic zone may reach the surface of the sediments. This is macroscopically visible as a white cover of chemoautotrophic sulfur bacteria (they owe their color to intracellular elemental sulfur). If the sediments are exposed to light, the surface may instead be dominated by a purple layer of phototrophic sulfur bacteria. In stratified water bodies (such as thermally stratified lakes or where there is a salinity difference between the surface and deep waters), vertical, turbulent transport of dissolved  $O_2$  from the surface layers to the deep layers is prevented. The chemocline may then extend above the sediment surface and into the water column. The phenomenon is

common (at least seasonally) in deep, productive lakes and in fjords with a sill; the largest permanent anoxic water body on Earth is the Black Sea, which is everywhere anoxic and sulfidic below 150- to 200-m depth.

### 3. Terrestrial Soils

The microbial life of soils differs in many ways from that of aquatic sediments. The chemical composition of soil organic matter differs because it mainly derives from vascular plants and therefore includes large amounts of structural polymers (especially lignin and cellulose). Vascular plant tissue is very poor in mineral nutrients; C:N ratios typically exceed 60, which affects the microbial nitrogen cycle of soils. Another important feature is that soils represent a complex matrix that includes air-filled spaces. Since diffusion coefficients are approximately  $10^4$  times higher in air than in water, soils are not anaerobic throughout, although anaerobic microniches may occur within individual waterlogged soil particles, and therefore anaerobic microbial processes do occur. In general, the biological activity of soils exceeds that of aquatic sediments reflecting the higher input of dead organic matter. Microbial activity in soils affects soil fertility and some processes (e.g.,  $\text{CH}_4$  oxidation and denitrification) affect element cycling on a global scale.

The overall important, and variable, factor controlling soil microbial activity—qualitatively and quantitatively—is water content. Fungi (ascomycetes and basidiomycetes) are better adapted to water stress and their hyphae can penetrate air-filled spaces; in soils and litter they are therefore important rivals to bacteria as primary decomposers. Certain bacterial groups (including the spore-forming gram-positive bacteria and the fungi-like actinomycetes and myxobacteria) are better adapted to water stress than other types of bacteria and thus play a relatively larger role in soils. For a given type of soil there is an optimum water content for microbial activity: A very low water content prevents microbial activity; when a large fraction of pores are water filled anoxic conditions are more widespread, thus inhibiting aerobic processes. Water contents affect, for example, soil nitrogen cycling; low water contents favor ammonia oxidation (nitrification, an aerobic process), whereas high water content favors denitrification (an anaerobic process leading to loss of reactive N from the soil).

The highest microbial activities are found within the root zone (rhizosphere). Microbes (in part directly attached to plant roots) exploit dissolved organic material which is excreted from the roots. They also form the basis for protozoan grazing. In addition to direct

microbe–plant interactions (such as symbiosis with  $\text{N}_2$  fixing bacteria), microbial activity has a profound indirect effect on soil fertility through its effects on mineral nutrient cycling.

### 4. Extreme Environments

The microbiology of extreme environments (here including hyperthermal and subsurface habitats and brines) has recently drawn considerable interest. Studies have expanded our knowledge of microbial diversity (especially of thermophilic archaeobacteria) and led to the discovery of novel microbial communities; the potential for isolating microbes producing industrially useful enzymes has also spurred interest in extreme environments.

Microbial mats are communities that develop in environments with no or little animal grazing and disturbance. These are complex biota that include a great variety of interacting physiological types of prokaryotes and in some cases also include eukaryotic microbes. Cyanobacterial mats develop in illuminated thermal springs and brines (salterns); when undisturbed, they develop 1-m-thick laminated deposits (stromatolitic mats) over centuries, although almost all biological activity is concentrated in the upper few millimeters. They are a modern analog to Precambrian stromatolites, which seem to have represented the dominating shallow-water biota for approximately 2.8 billion years of Earth's history; they almost disappeared from the geological record in the past 600 to 700 million years, when metazoans made their appearance. Cold or hot seeps of sulfidic water support mats of chemoautotrophic sulfur bacteria. These biota—well-known from deep-sea hydrothermal vents but also occurring elsewhere—are interesting in that their energy support (sulfide, methane, and reduced Fe) derives from geothermal processes rather than from solar energy via photosynthesis (however, the oxygen, which is necessary for exploitation of these energy sources, derives from oxygenic photosynthesis).

Subsurface microbial biota are a recent discovery. Thus, bacterial life has now been found hundreds of meters underground in shales, limestone, and other rocks, and groundwater ages and other evidence indicate that there has been no contact with the surface for at least  $10^4$  and up to  $10^6$  years. The microbial activity seems mainly to be hydrogen based (methanogenesis, acetogenesis, and sulfate reduction), but organotrophs also occur. Oil reservoirs in deep sediments have also been shown to harbor biota of hydrocarbon-degrading thermophilic sulfate reducers.

## D. Symbiotic Bacteria

There are numerous examples of bacteria living in or on other organisms; in many cases the adaptive significance of these associations is not understood. Some cases of symbiosis between prokaryotes and eukaryotes, however, have been studied in detail; they are of scientific interest and sometimes even attract substantial economic interests.

Symbiotic polymer degradation is one such example. With few exceptions, animals are not capable of hydrolyzing and utilizing structural plant polymers directly. Animals do not produce the necessary enzymes (cellulase, xylanase, etc.). Also, plant tissue generally contains insufficient amounts of N and P to sustain growth and various secondary plant compounds are toxic to animals. Some herbivores therefore eat copious amounts of plant tissue with a low digestion efficiency. They are typically confined to a few (taxonomically related) food plants because each grazer has only developed detoxification mechanisms against specific plant toxins; this applies to many insects. Many animals, however, have solved the problems of herbivory by maintaining consortia of anaerobic microbes in their gut system. Such systems have evolved independently in several groups of mammals and in a few groups of birds (ostriches, ptarmigans, and hoatzins), reptiles, and fish; they are also found in termites and cockroaches, in some echinids, and in shipworms.

The best known example is the rumen of cows and sheep. Ruminants exemplify pregastric fermentation. The rumen, which constitutes 10–15% of the volume of the animal, is anatomically a part of the esophagus. Its anaerobic content is a mixture of rumen fluid, ingested plant fragments, and microbes (approximately  $10^{11}$  bacteria and  $10^6$  protists  $\text{ml}^{-1}$ ). Ruminant saliva does not contain hydrolytic enzymes, but it is strongly buffered. Urea (the principal nitrogenous excretion product of mammals) is also excreted into the saliva, supplying the microbial community with N; it is a mechanism for preserving this essential element for the ruminant through internal recycling. Within the rumen a consortium of different bacteria (and some eukaryotic microbes) ferment carbohydrates (including cellulose and xylan) principally into acetate, butyrate, propionate, and  $\text{CO}_2 + \text{H}_2$ . The fatty acids are absorbed in the intestinal tract and they constitute the energy and carbon source of the ruminant. Protein is exclusively provided by the digestion of microbial cells in the true stomach; thus, a cow does not profit from eating proteins because these are microbially hydrolyzed, deaminated and fermented in the rumen prior to the stomach.

In the rumen, methanogens convert  $\text{H}_2$  into  $\text{CH}_4$ ; the methane represents a (necessary) loss to the cow (approximately 15% of the food intake) and it rids itself of the gas through belching. Ruminants are thus totally dependent on their microbial symbionts; ruminants that have artificially been deprived of their rumen biota cannot survive on a normal diet. In addition to ruminants, pregastric fermentation is found, for example, in kangaroos, sloths, and colubine monkeys.

Postgastric fermentation is more widespread (e.g., horses, rhinos, elephants, lemurs, koalas, lagomorphs, and several others; even humans probably profit slightly from their intestinal microbiota in this way). Animals with postgastric fermentation possess one or more caeca containing a consortium of fermenting bacteria. Ingested food is first subject to acid digestion in the stomach, with subsequent absorption of monomers, whereas undigestible plant polymers are eventually fermented in the caecum. In most termites and cockroaches, the principal cellulose decomposers (harbored in the hindgut) are protists (various types of specialized flagellates), although bacteria also play a substantial role in these microbiota.

Nitrogen fixation (the reduction of  $\text{N}_2$  to  $\text{NH}_4^+$ ) is a very energy-requiring, anaerobic process which is known only from bacteria. It is widespread especially among free-living anaerobes and microaerobes, and some aerobes have special adaptations to protect the nitrogenase complex from  $\text{O}_2$  (e.g., some cyanobacteria and *Azotobacter*). Nevertheless, it is estimated that in terrestrial habitats more than 90% of  $\text{N}_2$  fixation is symbiotic.

Symbiotic nitrogen fixation is of immense economic importance and since antiquity it has been known that legumes improve soil fertility. The great majority of plants belonging to the Leguminosae form root nodules that harbor the  $\text{N}_2$ -fixing bacteria (*Rhizobium*). These bacteria normally live in soils; they especially thrive in the rhizosphere of legumes and often adhere to root hairs. Appropriate strains enter root hairs and induce the formation of root nodules. The bacteria transform into nondividing “bacteroids” that fix  $\text{N}_2$  on the basis of carbohydrates provided by the plant. A hemoglobin (leghemoglobin) secures a sufficiently low  $\text{O}_2$  tension while simultaneously allowing for aerobic metabolism, and the plant assimilates the formed ammonia.

Another type of symbiotic  $\text{N}_2$  fixation is found in many unrelated trees (e.g., *Alnus*, *Myrica*, and *Hippophae*); it is due to the bacterium *Frankia* (an actinomycete), which also forms root nodules. The adaptive significance of different examples of symbiotic cyanobacteria in plants seems also to be that of  $\text{N}_2$  fixation

(e.g., in cycads). The water fern *Azolla* harbors  $N_2$ -fixing *Anabaena*; the fern plays a role as green manure in rice cultivation.

There are many other interesting examples of symbiotic relations between bacteria and eukaryotes, including relatively exotic cases such as symbiotic bacteria that confer bioluminescence on certain species of fish and squids. Special mention should be made of various marine invertebrates that farm chemoautotrophic bacteria (especially sulfide oxidizers and, in a few cases, methane oxidizers). First thoroughly studied in the deep-sea hydrothermal vent fauna, the phenomenon has subsequently been found to be widespread in shallow waters as well. Pogonophorans (and the related vestimentiferans) and some bivalves and oligochaetes are gutless and entirely dependent on their symbionts for food; the hosts in turn provide the bacteria with necessary oxygen and sulfide via their circulatory system. Other bivalves maintain sulfur bacteria in their gills, but they are simultaneously capable of filter feeding. Some anaerobic protozoa maintain symbiotic methanogenic or sulfate-reducing bacteria. The protozoa (mainly certain ciliates) have a fermentative metabolism with  $H_2$  production; the symbionts in turn are  $H_2$  scavengers, thus maintaining a low intracellular  $H_2$  tension which again enhances the energy efficiency of the host metabolism.

### III. UNICELLULAR EUKARYOTES (PROTISTS)

Protists are known to include many unrelated eukaryotic groups. Eukaryote cells differ from prokaryotes in possessing a cytoskeleton and membrane-covered organelles, among which mitochondria and chloroplasts are recognized as being descendants of endosymbiotic aerobic bacteria belonging to the  $\alpha$  group of proteobacteria and to cyanobacteria, respectively. Protists have a much greater size range than bacteria; the smallest free-living species measure approximately  $3 \mu\text{m}$ .

Phagocytosis is probably a primary property of eukaryotic cells and most extant species depend on particulate food (mainly bacteria or other protists). Many groups have acquired chloroplasts; this has apparently happened independently in different taxa and examples of "chloroplasts" which represent intermediate stages between an endosymbiont and an organelle are known. Many species have secondarily lost chloroplasts and thus the ability of photosynthesis. Within many groups of phototrophs the ability of phagocytosis has been

retained (mixotrophs, found for example among chrysomonads and dinoflagellates); in other groups (e.g., diatoms and green algae) phagocytosis has been irreversibly lost. A few species (some ciliates and foraminiferans) are capable of retaining chloroplasts from their prey cells; the chloroplasts remain functional for some days and this is exploited by the "host." Many heterotrophic protists harbor endosymbiotic phototrophs. Very few protists subsist on dissolved organic matter; in most habitats they would be inferior competitors to bacteria simply due to their larger size. The majority of protists are aerobes; a few specialized protists, however, are obligate anaerobes depending on a fermentative metabolism. Tolerance to extreme conditions is limited relative to that of some bacteria, but otherwise protists are omnipresent; primarily they are the principal consumers of bacteria in all types of environments and phototroph protists are largely responsible for primary production in aquatic habitats.

#### A. Phototrophs

Macrophytes (macroalgae and vascular plants) are responsible for much of the primary production along shallow coastal waters and lakes. In offshore waters and in the oceans, unicellular phototrophs are solely responsible for light-driven  $CO_2$  reduction and they thus largely constitute the basis for aquatic life. Characteristic annual succession patterns (with respect to species composition) occur in temperate seas and lakes. During winter, mineral nutrients are brought to the photic zone from deeper waters. In early spring, when light conditions allow, a bloom of diatoms develops; the bloom is brought to an end by zooplankton grazing and by nutrient limitation. During summer, phytoplankton is dominated by small forms (chrysomonads and prymnesids) which perform better at low nutrient concentrations; autumn is often characterized by a bloom of large dinoflagellates. In tropical waters which are thermally stratified throughout the year, competition for mineral nutrients is intense and small species prevail. Upwelling areas, such as off the coast of South America, support production of large phytoplankton species and thus represent an exception to this generalization.

Phytoplankton biomass is primarily limited by mineral nutrients—in most marine areas by inorganic N, although P limitation is also known and may be the rule in limnic systems. Diatoms may be limited by available Si and it has been inferred that Fe limits primary production in the Southern Ocean. Increased nutrient additions, especially from fertilizers, via river runoff



and groundwater seepage have caused increased phytoplankton production and biomass in offshore waters, bays, and estuaries in addition to lakes in many parts of the world. Such eutrophication may be detrimental to rooted vegetation (due to competition for light) and in some cases may result in widespread anoxia or hypoxia of bottom waters due to an increased input of organic material.

Large phytoplankton cells are consumed by various zooplankters (copepods, etc.) which again serve as food for fish. Small phytoplankters are to a large extent consumed by protozoa which then enter planktonic food chains. Many benthic filter feeders (e.g., mussels) also depend mainly on phytoplankton cells.

In shallow aquatic sediments phototrophic protists are important primary producers even though the photic zone extends only 2 or 3 mm beneath the sediment surface. In these biota, diatoms, euglenoids, and dinoflagellates are especially important (together with unicellular and filamentous cyanobacteria).

## B. Phagotrophic Protists

Phagotrophic protists seem to be the principal consumers of bacteria in almost all habitats. In plankton, it is particularly the small (4–10  $\mu\text{m}$ ), heterotrophic flagellates that play a role as bacterial consumers, thus closing the microbial loop (Fig. 3). They occur at densities of approximately  $10^3$  cells  $\text{ml}^{-1}$  in most natural waters. Other groups of protists, especially ciliates, mainly consume larger prey such as heterotrophic flagellates and small phytoplankters, whereas heterotrophic or mixotrophic dinoflagellates feed on prey which often exceeds their own size. In oceanic plankton, the large acantharians, radiolaria, and the planktonic foraminiferan *Globigerina* are mainly phagotrophs feeding on large prey; most of them also harbor endosymbiotic phototrophs. The three (unrelated) groups all produce skeletons (made of strontium sulfate, silica, and calcium carbonate, respectively) and the sedimented skeletal remains of the two latter groups characterize many oceanic sediments.

Sediments, especially from shallow waters, harbor high densities of a wide variety of phagotrophic protozoa, including amoebae, flagellates, and ciliates, filling a variety of niches especially with respect to food requirements and oxygen tension. Foraminiferans are characteristic of marine sediments; their niches are in part taken over by testate amoebae and by heliozoans in lake sediments. Deeper marine sediments harbor large foraminifera and in deep-sea sediments the peculiar

xenophyophoreans grow to a size of several centimeters and possess a skeleton made of barite.

In soils, small amoeboid protozoa are most important as bacterial consumers; testate amoebae, heterotrophic flagellates, and ciliates are also present. Soil protozoa have various adaptations to thin water films (small size) and desiccation (formation of desiccation resistant cysts and life cycles which allow exploitation of short periods with high moisture).

## C. Symbiotic Protists in Animals

Like prokaryotes, many protists occur as symbionts in animals; probably all animal species (including humans) harbor several protozoan symbionts. Symbiotic polymer degradation by flagellates in termites has already been mentioned. The most important type of symbiosis involving protists is that between animals and intracellular phototrophs. Many aquatic invertebrates harbor such symbionts. In fresh waters the symbionts are usually the green alga *Chlorella* (e.g., in freshwater sponges *Chlorohydra* and in the ciliate *Paramecium bursaria*). Most marine cases are based on dinoflagellates (*Symbiodinium*), but other groups (e.g., diatoms, chlamydomonads, and prymnesids) are also represented. In most cases the host combines the nutrition derived from the symbiont (usually in the form of carbohydrates) with particulate food; in some cases, it has been shown that the hosts can subsist entirely on the basis of the symbionts, and in a few cases the ability of phagotrophy has been lost. The most important marine example is that of phototrophic symbionts in reef-building corals; the symbionts are not only responsible for a significant share of primary production of coral reefs but also facilitate carbonate deposition of the host during active photosynthesis. A similar situation applies to giant tropical shallow-water foraminiferans; like corals, they are responsible for the formation of limestone deposits (the Cheops Pyramid is built from limestone consisting of the calcareous remains of the Eocene foraminiferan *Nummulites*). Other marine invertebrates harboring phototrophic symbionts include the giant clam *Tridacna* and various coelenterates.

## IV. ROLES OF MICROBES OVER GEOLOGICAL TIME AND IN THE CONTEMPORARY BIOSPHERE

This section briefly explores the role of microbes in the evolution of biogeochemical element cycling on a global

scale. Earth (and the solar system) arose approximately  $4.6 \times 10^9$  years ago; the earliest unambiguous sign of life dates to approximately  $3.5 \times 10^9$  years ago (mid-Archean). It is generally accepted that life arose on Earth, perhaps  $4 \times 10^9$  years ago, under anoxic (and presumably chemically reducing) conditions, and that the atmosphere eventually became oxic as a result of oxygenic (cyanobacterial) photosynthesis. There is convincing evidence that atmospheric  $O_2$  slowly increased from very low levels during the Precambrian. The earliest evidence of life (from Warrawoona in Australia and the Fig Tree Formation from southern Africa) is in the form of stromatolites, which are fossil remains of microbial mats. Some fossils are sufficiently well preserved to reveal organisms that were very similar to present-day cyanobacteria. Stromatolites (sometimes with well-preserved fossil bacteria) are known from throughout the remaining Precambrian. Evidence of an increasing  $O_2$  tension of the atmosphere derives from banded iron formations—laminated ( $>3$  to  $2 \times 10^9$  years old) deposits containing partly oxidized Fe. The general interpretation is that as  $O_2$  (resulting from photosynthesis) appeared in the atmosphere, dissolved reduced iron, in the otherwise anoxic oceans, was oxidized at the sea surface to become insoluble oxidized Fe that again gave rise to the deposits. Atmospheric  $O_2$  levels thus initially remained low due to oxidation of reduced minerals on the surface of Earth (and presumably due to the early origin of biological  $O_2$  respiration). Later, the  $O_2$  level increased, reflecting the accumulation of fossil organic material in sedimentary rocks. There is evidence that the  $O_2$  level had increased to  $>1\%$  of the current atmospheric level approximately  $2 \times 10^9$  years ago (early Proterozoic). When metazoa arose 600 to 700 million years ago, atmospheric  $O_2$  must have reached at least 10% of the current level (according to requirements of extant invertebrates), but it could also have been higher.

The conclusion is that oxygenic phototrophs, quite similar to modern cyanobacteria, must have evolved  $3.5 \times 10^9$  years ago. It seems likely that all basic types of bacterial metabolism had also evolved by then, although this is speculative.

The time for the origin of eukaryotic cells is unknown, although molecular data (Fig. 1) suggests “deep roots”; there is (molecular) evidence suggesting that mitochondria (and thus “modern” eukaryotic microbes) arose approximately  $2 \times 10^9$  years ago (or slightly earlier than fossils that have been interpreted as remains of protists). In all circumstances, when metazoans and the first macroalgae arose 600 to 700 million years ago, almost all major biochemistry and metabolic pathways

had evolved in microorganisms and with them the basic biogeochemical cycles as we know them today. Colonization of land approximately 100 million years later and the evolution of vascular plants and fungi, however, must have had a profound effect on the biosphere and on mineral cycling.

There is not enough space here to describe the major biogeochemical element cycles (notably the C, N, and S cycles) in detail. A view of the nitrogen cycle (Fig. 4), however, further illustrates the immense importance of microbes in the biosphere. Atmospheric  $N_2$  is fixed by bacteria, but reactive N is also added to the biota as N oxides formed during electric discharges in the atmosphere and by industrial  $N_2$  fixation. In cells, N occurs in a reduced form in organic matter and it is released as  $NH_4^+$  when organic matter becomes mineralized. Through nitrification (a bacterial process) it is oxidized to  $NO_3^-$  which can be utilized (through assimilative reduction) by plants. The only pathway back to  $N_2$ , however, is denitrification, an anaerobic bacterial process that is necessary for the completion of the global N cycle and ultimately regulates the amount of reactive N available to the biota.

The important mechanisms that control particular microbial processes are largely understood. However, the complexities of the cycles and the many positive and negative feedbacks render it difficult, or perhaps impossible, to make predictions of the effect of environmental changes on an ecosystem or on a global scale. An example is atmospheric  $CH_4$ , which is a greenhouse gas. The basic sinks (abiotic atmospheric oxidation and microbial oxidation in soils and other biota) and sources

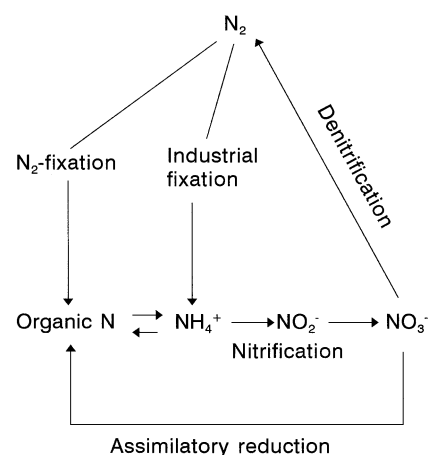


FIGURE 4 The global N cycle (simplified). Crucial processes (biological  $N_2$  fixation, nitrification, and denitrification) are exclusively bacterial. Denitrification and nitrification also imply production of  $N_2O$  and other oxidation levels of N.

(mainly microbial methanogenesis primarily in wetlands including rice paddies, in lakes and in the sea, and in ruminants and termites; to a lesser extent through biomass burning and natural gas production) have been identified. It is also known that atmospheric levels (currently approximately 1.8 ppm) have increased (by approximately 1% year<sup>-1</sup>) during the past century. However, the effect on CH<sub>4</sub> levels as a function of climate change is in practice impossible to predict. Both CH<sub>4</sub> production and oxidation increase with temperature, and melting of tundras will enhance methanogenesis, but other temperature-correlated factors may affect both processes either way. Determinants of methanogenesis also include input of organic C, competitive interactions with denitrifiers, and sulfate- and iron-reducing bacteria; anthropogenic factors such as the extent of rice paddies and ruminant husbandry should also be included.

## V. MICROBES AND MAN

There are many relationships between humans and microbes and they affect us on many levels of our existence. We are hosts to many prokaryotic and eukaryotic symbionts in addition to being victims of bacterial and protozoan pathogens. Humans have also—long before the existence of microbes was recognized—utilized microbial processes and learned to prevent some adverse effects of microbial activities.

Thus, microbiological technologies which have been in use since prehistoric times include a variety of fermented foods involving lactic acid bacteria, propionic acid fermenters, and the production of vinegar from ethanol. In addition to their gastronomic qualifications, these techniques have served to prevent or control undesirable or even dangerous microbial spoilage of food, which is also the purpose of salting, smoking, or acidifying food. It has also been suggested that the use of spices and drinking of wine, rather than water, also served to control pathogenic microbes.

A quite different use of microbes is acid mine leaching, which was also in use long before the underlying mechanism was understood. Circulating water through crushed copper ore leads to acid conditions and dissolution of the ore; metallic Cu can then subsequently be recovered from the leachate by chemical methods. The underlying mechanism is that a consortium of acidophilic, chemoautotrophic bacteria (including *Thiobacillus ferrooxidans*) oxidize both the reduced S and the reduced Fe of pyrite (FeS<sub>2</sub>), which is omnipresent in

many ores. The resulting sulfuric acid in turn dissolves the ore.

Biological sewage treatment serves primarily to mineralize organic material. Various types of sewage treatment are in use, depending, among other factors, on the scale of the plant; most systems involve aerobic and anaerobic microbial processes. An important aspect is the flocculation of bacteria, a process which is enhanced by the presence of protozoa. Removal of nitrate by microbial denitrification is another important function of biological sewage treatment, whereas phosphate is primarily removed by chemical precipitation. Mainly in smaller plants, anaerobic mineralization can be exploited to produce methane, which can be collected and subsequently used for heating.

In recent times, mass production of certain species of bacteria for the production of enzymes and antibiotics has played an important industrial role. Recently, genetically engineered bacteria that express human protein genes (e.g., insulin and other hormones) have been used in the pharmaceutical industry.

Microbial diseases, of which there are many, represent the most direct encounter between humans and microbes. Through recorded history such diseases have played an important role for human populations, most dramatically illustrated, perhaps, by the recurrent plague epidemics in Europe from medieval times to approximately 1700; however, many other bacterial diseases, such as cholera, tuberculosis, leprosy, and typhoid fever, and protozoal diseases such as malaria were also important. In North America, Europe, and in some other parts of the world serious bacterial and protozoal diseases have, especially after World War II, largely been brought under control due to the combined effects of hygienic measures, vector control (mosquitoes and rats), immunization programs, and antibiotics and other forms of chemotherapy. However, globally, tuberculosis and malaria remain among the most frequent causes of death. Many bacterial and protozoal diseases of livestock also remain economically significant. Evolving resistance to antibiotics and other types of chemotherapy in agents of disease in man and animals may represent an increasing problem and indicate that our interactions with pathogenic microorganisms is not a closed chapter in human history.

### See Also the Following Articles

ARCHAEA, ORIGIN OF • BACTERIAL BIODIVERSITY • ECOSYSTEM SERVICES, CONCEPT OF • EUKARYOTES, ORIGIN OF • MARINE SEDIMENTS • MICROBIAL BIODIVERSITY • NITROGEN AND NITROGEN CYCLE • PROTOZOA • PSYCHROPHILES • THERMOPHILES, ORIGIN OF

## Bibliography

- Balows, A., Trüper, H. G., Harder, W., and Scheifer, K. H. (eds.) (1991). *The Prokaryotes*, Vols. I–IV. Springer-Verlag, New York.
- Coleman, D. C., and Crossley, D. A., Jr. (1996). *Fundamentals of Soil Ecology*. Academic Press, New York.
- Falkowski, P. G., and Woodhead, A. D. (eds.) (1992). *Primary Productivity and Biogeochemical Cycles in the Sea*. Plenum, New York.
- Fenchel, T. (1987). *Ecology of Protozoa. The Biology of Free-Living Phagotrophic Protists*. Springer-Verlag, Berlin.
- Fenchel, T., King, G. M., and Blackburn, T. H. (1998). *Bacterial Biogeochemistry. The Ecophysiology of Mineral Cycling*, 2nd. ed. Academic Press, London.
- Harris, G. P. (1986). *Phytoplankton Ecology: Structure, Function and Fluctuation*. Chapman & Hall, London.
- Hausmann, K., and Hülsmann, N. (1996). *Protozoology*, 2nd ed. Thieme Verlag, Stuttgart.
- Madigan, M. T., Martinko, J. M., and Parker, J. (1997). *Biology of Microorganisms*, 8th ed. Prentice Hall, Upper Saddle River, NJ.
- Schopf, J. W., and Klein, C. (eds.) (1992). *The Proterozoic Biosphere*. Cambridge Univ. Press, Cambridge, UK.
- Zehnder, A. J. B. (ed.) (1988). *Biology of Anaerobic Microorganisms*. Wiley, New York.





# MIGRATION

Mace A. Hack and Daniel I. Rubenstein  
*Princeton University*

---

- I. Introduction
  - II. What Is Migration?
  - III. Migratory Patterns: A Taxonomic Survey
  - IV. Attributes of Migrants Affecting Susceptibility to Human Disturbance
  - V. Ecological Consequences of Human Disturbance on Migrants
  - VI. Evolutionary Consequences of Human Impacts on Migrants
- 

## GLOSSARY

**circannual rhythms** Endogenous, or internal, rhythmic cycles of one year in duration that govern the onset and cessation of migratory behaviors.

**compass orientation** Navigation in a particular direction without reference to landmarks or sites of origin or destination. Migrants are known to use compass information from magnetic fields, chemical gradients, and visual features such as the stars, sun, and planes of light polarization.

**diadromy** Migrations that take individuals between fresh and salt-water habitats, a common phenomenon for many migratory fish species.

**partial migration** The case where intrapopulational variation in migratory behavior leads some individuals to migrate while others within the same population may only migrate locally or remain sedentary.

**zugenruhe** Restlessness exhibited by some migratory

species, especially birds, if not allowed to migrate during their usual migratory period. It reflects an underlying physiological transition to a migratory state.

---

**MIGRATION DESCRIBES** the movement of individuals between spatially separate ecological communities, typically on a seasonal or annual schedule. Several characteristics of migrations distinguish them from other forms of animal ranging behavior, including more persistent movements, of greater duration, that follow a more direct path with fewer turnings. Furthermore, migrants do not respond to resources along their path, but show a heightened response to the same resources near the migratory journey's end. This latter feature distinguishes migrations from typical foraging and dispersal movements. Behavioral specializations may include specific activity patterns particular to departure and arrival, and unique patterns of energy allocation to support long-distance movements. The ecological consequences of migration are that they take a species from one community of organisms to another and they partition life histories so that specific phases or events occur in different ecological communities.

## I. INTRODUCTION

Migrations capture the human imagination like few other animal behaviors. The single-minded struggle of

the salmon fighting its way upstream, the barely visible formation of geese piercing the sky overhead, and the thundering line of wildebeest snaking across the open savanna—all speak of ancient rhythms that drive life on our planet. The often great distances moved and the large numbers of individuals involved make animal migrations a conspicuous and essential aspect of many regions' biodiversities. Animals migrate for many reasons but in general do so to avoid temporarily unfavorable conditions or to locate particularly favorable areas that can meet specific biological needs, such as reproduction. However, it is somewhat ironic that migrations might actually increase a species' risk rather than reduce it. Protecting migratory species requires preservation not only of their final destinations but their migratory routes and stopover points as well. In effect, migration inextricably links the fates of biotas across the length and breadth of the globe. No natural phenomenon makes the point better that biodiversity on local and global scales shares the same actors and the same processes that drive them. In this chapter we explore the nature, scope, and patterning of animal migrations as a prelude to discussing how human-caused changes in the environment are likely to impact migratory species. We not only address direct effects on migratory species behavior, ecology, population dynamics, and evolutionary potential, but we also consider indirect impacts on the ecological communities and regional biodiversities that disruption of migrations can produce.

## II. WHAT IS MIGRATION?

Definitions of "migration" abound in the scientific literature, many of which capture the nature of the behavior for specific taxa but fail to generalize across taxa. Commonly cited elements include long distances traveled and movements from one place to another, then back again. But what is "long"? Dark-eyed juncos (*Junco hyemalis*) and Blackburnian warblers (*Dendroica fusca*) are both small passerine birds that migrate with the seasons; both breed at higher latitudes in northeastern Canada (among other sites) yet juncos may migrate only a few hundred kilometers to their overwintering range while Blackburnian warblers fly several thousand kilometers to overwinter in the Andean forests of Ecuador. The lengths of these respective trips differ by an order of magnitude, yet in examining the ecology and behavior of each species, we might well consider both to constitute migrations. For example, both involve moving between distinct ecological communities on a consistently timed, seasonal basis. Similarly, "typical" migrants, like

most birds, travel the same circuit each year. Others, however, may only complete a single circuit in their lifetime (e.g., salmon) or only *part* of a circuit before they die, such as insects in which successive generations continue the journey their predecessors began. Should we exclude such species from the "migratory" category despite other aspects of behavior, physiology, and life history held in common with typical migrants?

From a biodiversity perspective, migration drives a species' life history and pattern of resource use, and it ties together ecological communities in different regions. Thus, defining migration in ecological terms of a species' use of space over time is imperative to understanding its functional impacts on biodiversity. However, one also needs to ask, does biodiversity refer to just the specific organisms found within a region, or does it encompass the specialized behaviors such organisms display? Ecological consequences aside, are sedentary populations of Canada geese interchangeable with migratory ones when summing up a region's biodiversity? If migratory and nonmigratory races or subpopulations—whether genetically or phenotypically different—each represent unique components of biodiversity, then we must also define "migration" in terms of the behavioral mechanisms distinguishing them.

The combined ecological and behavioral dimensions of migration have plagued attempts to come up with a single definition suitable from both perspectives. Ecological definitions have included "the act of moving from one spatial unit to another" by Baker (1978) and "the persisting change that is left over when all other, minor excursions are removed" by Taylor and Taylor. Both capture the essential idea of migrations extending the ecological space used by individuals, although ambiguity over what constitutes a spatial unit or minor excursion renders each too imprecise to be of much use in distinguishing migratory from nonmigratory species.

Dingle (1996) has taken Kennedy's lead to propose a definition of migration that emphasizes the differences between migrations and other forms of animal movement. He uses clear behavioral terms to distinguish migrants from nonmigrants, yet he uses each species as its own reference for differentiating movements with different ecological consequences. Generally put, nonmigratory (ranging) movements are driven by the immediate need to acquire or safeguard resources, and they cease when suitable resources are encountered or are adequately defended. The length and timing of foraging and territorial defense movements will vary daily, or on even shorter time scales, depending on the frequency of encountering resources or potential threats. Commuting daily between refugia and feeding

locations, such as the diel movements of zooplankton and other aquatic animals, is similarly directed by resources and is responsive to their location and abundance. Even dispersal to establish a new home range or to leave a natal group usually ceases as soon as a suitable new home range is found. In contrast, once they set out, migrants will ignore many of the resources they encounter, or use them briefly to refuel, and only become responsive again to resource abundance and quality after all or a critical part of their journey has been completed. Migratory movements thus entail specific changes in behavior and physiology that distinguish them from nonmigratory movements, even those that are circular in nature. Defining migrations relative to a species' other forms of movement results in a more generally applicable definition and one that encompasses the diversity of the migratory behaviors animals display.

Specifically, five mechanistic attributes distinguish migratory behavior: movements are (a) more persistent and of greater duration than ranging movements and (b) follow a more direct path with fewer turnings. There is (c) an initial suppression of responses to resource-derived stimuli, but often a heightened response to such stimuli near the migratory journey's end. Migrants may also have (d) specific activity patterns particular to departure and arrival and (e) unique patterns of energy allocation to support long-distance movements. Of course, not all migrants will show all five characteristics, but as a group these traits circumscribe the suite of distinct and specialized behaviors entailed in migration. To these behavioral attributes we add two functional hallmarks of migrations: (f) they take a species from one community of organisms to another, and (g) they partition life histories so that specific phases or events occur in different ecological communities.

Two common uses of the term "migration" will not be considered in this chapter: (a) migration in the paleontological sense of species shifting their historical distributions with climate change or geological events and (b) migration as geneticists use the term to refer to gene flow between populations.

### III. MIGRATORY PATTERNS: A TAXONOMIC SURVEY

The diversity of migratory behavior in animals overwhelms attempts to neatly summarize its character and function. Generally, animals migrate to escape unfavorable conditions or to exploit favorable ones, yet defining

"favorable" or "unfavorable" is often specific to the taxon examined and the life-history function at hand. Polar and cold temperate habitats tend to have more migrant species than tropical ones because they vary so strongly in productivity and habitability, although this tendency differs widely among the major taxonomic groups. Migration distance and duration similarly vary to a great degree, even within groups of species that migrate for the same reason. Differences in physiology account in part for this variability, since larger body sizes can store relatively more energy to fuel longer trips and lower the weight specific cost of transport while certain forms of locomotion, such as flying or swimming, are more efficient than others. In addition to physiology, physical forces such as winds and currents act in concert with habitat topography to further shape the migratory route and schedule. Biogeographic history may even play a role in these features of migration since routes may be hard-wired genetically and slow to adapt, or learned and dependent on knowledge within lineages. Indeed, migrations present a truly fascinating mix of evolutionary puzzles in behavior, ecology, physiology, and biogeography (see Dingle, 1996, for an excellent introduction to these). The broad surveys that follow provide an overview of migration prevalence, function, and character for the major taxa with specific emphasis on the biodiversity of migrants per taxonomic group and region.

#### A. Birds

Birds epitomize the act of migration for many people. Whether one lives in the temperate zone or tropics, New World or Old, Southern or Northern Hemisphere, migratory species constitute a significant and conspicuous fraction of the avifauna. At the most northern and southern latitudes, close to 100% of bird species migrate out of the region for part of the year. For example, 135 species breed in the arctic zone, yet all migrate south to spend most of their year elsewhere. In more temperate regions, a majority of species, and nearly all insectivores, migrate to more tropical latitudes after breeding. Approximately 200 species migrate from North America to the West Indies, Central, and South America each year, while many more make shorter distance migrations into the southern United States and Mexico. In Europe, 177 species, or 40% of the region's avifauna, migrate from temperate breeding grounds to overwinter in Africa; 104 species from western, central, and eastern Asia join them there. Moreau has estimated that 5 billion individuals make this migratory journey south to Africa alone; the comparable number for the entire



globe is surely several times this. Seasonal fluctuations in temperature ultimately drive food production cycles at higher latitudes, setting the schedule for migratory movements, but in tropical climates distinct wet and dry seasons may function in an identical manner. Substantial numbers of birds follow regular migratory routes that track resources such as fruiting or flowering trees, seeding grasses, or invertebrate flushes brought about by rains. Indeed, this form of short-distance movement may be the most common type of migration in birds. For example, out of 1450 breeding species in Africa, 532 species have been classified as intra-African migrants.

Bird migration is clearly a ubiquitous feature of our planet's biodiversity, yet migrant species are not distributed equally around the globe. In particular, birds that migrate in excess of 1000 km—moderate and long-distance migrants—are more prevalent as breeders in the northern hemisphere. This is largely due to the much greater land area at higher latitudes in the Northern than in the Southern Hemisphere. For example, only 20 species among the hundreds that breed in the temperate regions of southern Africa migrate as far north as the equator, while a much larger number and proportion of Palearctic species overwinter in the same equatorial region (noted earlier). Similarly, only 8% of the nearly 600 species breeding in Australia and Tasmania migrate in and out of the temperate zone. In the New World, 31 species of shorebird (e.g., plovers, sandpipers, curlews) breed at higher latitudes in the Northern Hemisphere and migrate thousands of kilometers to spend their nonbreeding seasons in the Southern Hemisphere, yet not a single shorebird species that breeds in the south even migrates as far north as the equator during its nonbreeding season. Although the actual bases for these fascinating patterns remain relatively unexplored, they suggest a greater sensitivity of northern avifaunas to factors that threaten moderate and long-distance migrants off their breeding grounds.

Long-distance migrants accomplish truly stunning feats. Blackpoll warblers (*Dendroica striata*), weighing only 10 to 20 gm, embark from Cape Cod on a nonstop, 86 hr, 3500 km flight 2000 m above the waves of the Atlantic to the northeastern coast of South America. Tundra-breeding shorebirds such as the Pacific golden plover (*Pluvialis fulva*) make nonstop 5000 to 7000 km journeys to wintering grounds in the South Pacific. Arctic terns (*Sterna paradisaea*) have the longest migration of any bird, and perhaps any animal, traveling an annual circuit that can exceed 40,000 km. They breed in the boreal summer along the northern edges of the Old and New World continents, then migrate along

the western edges of North America, South America, Europe, and Africa to feed in the rich waters off Antarctica's pack ice during the austral summer. Close to 8 months of each year are spent in transit between these two endpoints.

The Arctic tern's migratory journey seems extraordinary but it typifies why many birds migrate. In doing so, they exploit highly seasonal flushes of food resources, especially to meet the increased demands of breeding. High variability in food supply at higher latitudes, especially in insects and other invertebrates, makes it difficult for species to reside permanently in the habitat. However, the very predictable nature of these fluctuations allows mobile species to rely on them for part of their annual cycle. The absence of resident competitors may make these ephemeral resources even more abundant and accessible to migrants. Thus, the typical migratory movement for birds involves making an annual round trip between seasonally resource-rich breeding sites and nonbreeding areas where resource abundance may be lower but less variable over time.

Seasonal resource variability explains why many birds migrate, but what determines how far they go and where their final destinations lie? Proximal determinants of migration distance and route include weather patterns, history, the distribution of resources along the way, and the character of the landscape. The north-south orientation of mountain ranges in North, Central, and South America tend to funnel migrants along corridors also running north-south, whereas Eurasia mountain ranges run east-west, forcing many Palearctic migrants to make large westerly movements before they are able to fly south. The importance of trade winds or winds generated by weather fronts to migratory journeys can be seen in the often very different routes followed when moving north versus south (e.g., Arctic tern study by Alerstam). The migratory path of the white-rumped sandpiper (*Calidris fuscicollis*) extends the length of the Western Hemisphere and appears carefully choreographed to coincide with seasonal pulses of invertebrate prey along the route. History too seems to shape migratory paths and destinations as suggested by the very indirect route (via the eastern Mediterranean) taken by red-backed shrikes (*Lanius collurio*) migrating from the Iberian Peninsula to their nonbreeding range in Central Africa.

Habitat suitability is a more ultimate determinant of migration distance and destination. Species with very specific habitat requirements may have limited options for suitable nonbreeding areas. For example, the upland sandpiper (*Bartramia longicauda*) breeds in North American grasslands and must travel 10,000 km to over-

winter in the pampas of Argentina—the only similar Southern Hemisphere grassland habitat in the New World. Many tundra-breeding shorebirds migrate to the southern coastlines of South America, Africa, Southeast Asia, and Australia where extensive intertidal mudflats and rocky beaches provide abundant invertebrate prey. Most neotropical migrants—the warblers, flycatchers, vireos, swifts, hummingbirds, swallows, tanager, orioles, and raptors—migrate more moderate distances to forest and scrub habitats in Central America, the West Indies, and northern South America. In addition to the general structure of the habitat, competition with resident forest species and other migrants certainly influences where migrants settle during their nonbreeding seasons.

Often the question of how far migrants move depends on which population one examines. Many migratory birds conduct “partial” migrations where some individuals migrate while others remain as year-round residents on the breeding grounds or move only short distances (e.g., European blackbird, *Turdus merula*). Different age and sex classes may pursue different migratory strategies, as in the dark-eyed junco of the eastern United States. Alternatively, all individuals may migrate but different populations travel different distances. A common pattern in this case is for higher latitude breeding populations to migrate the longest distances, “leap-frogging” beyond the migratory movements of lower latitude populations (e.g., fox sparrow, *Passerella iliaca*). Some sanderling (*Calidris alba*) populations migrate only a short distance from their tundra breeding grounds to overwinter in the northwestern United States while others travel 7500 km to nonbreeding areas in Chile. Apparently, the much greater energetic and exposure costs of migrating to South America are offset by much richer food resources and a more hospitable climate, so that the payoffs for each strategy are equivalent, and both short- and long-distance migrants persist.

The presence of both migratory and nonmigratory strategies in the same species underlines the opportunistic and flexible nature of bird migration. As environmental changes occur or as a species expands its range, migratory behavior can often adapt to fit the new circumstances. The European starling (*Sturnus vulgaris*) is a widespread, permanent resident of Britain today, yet in the 18th century starlings regularly migrated out of Britain to overwinter in warmer regions of Europe. Long-term climate change has presumably made Britain a more hospitable place for starlings to spend their winters. Conversely, several species resident year-round in Europe have extended their breeding ranges north

over the past century. These more northerly populations have developed full migrations to southern Europe for overwintering. Migrants may also establish nonmigratory populations along traditional migratory routes as resources become more abundant or habitats are altered, for example, the barn swallow (*Hirundo rustica*) in Argentina and the Canada goose (*Branta canadensis*) in eastern North America. With industrial parks in North America supporting extensive expanses of grassy lawns some Canada geese cannot only forage sufficiently well at temperate latitudes during the winter, but high levels of vegetative production during the spring and summer enable them to stay and breed.

A fascinating form of bird migration not related directly to food abundance is molt migration. It is particularly common among waterfowl that molt all their flight feathers simultaneously and thereby lose the ability to fly. To escape predation, they may migrate to coastal areas, large lakes, or far off-shore until their new feathers have grown in.

## B. Mammals

Small (<5 kg), nonvolant mammals represent a large fraction of our planet's mammalian biodiversity—rodents alone account for 40% of all mammal species—yet very few species are known to migrate. Rather than leave highly seasonal habitats when resource abundances decline, many small mammals hibernate or reduce activity levels and wait for conditions to improve. Species known to migrate include lemmings in the arctic tundra and a variety of rats, mice, and shrews living on the Kafue River flats of Zambia studied by Sheppe. In both cases, these small mammals move several kilometers to escape flooding conditions as snow melts (lemmings) or rivers flood (Kafue mammals).

Migration occurs more commonly among a second, highly diverse mammal group, the bats (25% of all mammals). The ability to fly and travel large distances more efficiently than via terrestrial forms of locomotion must account in large part for this difference between bats and other small mammals. Fruit-eating bats in West Africa migrate 1500 km annually, following the movement of rains and the consequent pattern of fruit abundance. As in birds, different sex and age classes may migrate different distances; juveniles migrate further, perhaps trading-off reproduction for more abundant food and greater survivorship. Similar migrations tracking fruit, nectar, and insect abundances have also been found in Australia and the New World tropics.

Insectivorous bats breeding in the temperate regions of North America and Europe commonly migrate,

showing two distinct patterns of movement. Like many birds in these regions, species such as the Mexican free-tailed bat (*Tadarida brasiliensis*) and the hoary bat (*Lasiurus cinereus*) travel a maximum of 800 to 1700 km south to overwinter in warmer sites where their insect food can still be found. These individuals do not hibernate, but other individuals of the same species may make only part of the southward journey to hibernate at sites along the migratory route. The second migratory pattern involves movements of shorter distances (10–200 km), and more variable orientation, from summer ranges to particular winter hibernation roosts. For example, as Griffin discovered, the little brown myotis (*Myotis lucifugus*) hibernates in a few select caves and mine tunnels in northern New England, yet migrates usually less than 100 km back to summer ranges both north and south of these wintering sites. Seasonality determines the timing of migration in these species but the very specific requirements of hibernation, such as temperature stability, humidity, and protection from predators, drive migration distance and direction. It is not uncommon for 75 to 95% of the population in a large geographical area, or even an entire species, to migrate and hibernate in only a few caves. Both migratory patterns are particularly common among species that roost in trees during the summer, presumably because trees make very poor, exposed winter hibernation sites.

Among terrestrial mammals, migrations are most common, and most spectacular, in the ungulates, or hoofed mammals. In the Serengeti ecosystem of Kenya and Tanzania, more than 1 million wildebeest (*Connochaetes taurinus*) and several hundred thousand plains zebra (*Equus burchelli*) and Thomson's gazelle (*Gazella thomsonii*) migrate seasonally along a several hundred kilometer annual circuit. A spatially varying seasonal rainfall pattern, and the consequent growth in grass that it generates—the primary food of all three species—determines the migration's timing and route, although drinking water availability and salinity may also play important roles. Plains zebra and wildebeest populations elsewhere in Africa make similar migrations between wet and dry season grazing areas, but in smaller numbers. These species are not obligatory migrants, however, since resident, nonmigratory individuals can be found in most populations. Other ungulate migrations driven in part by seasonal food availability include the movements of caribou and reindeer (*Rangifer tarandus*) in the Holarctic region from tundra in the summer to taiga during winter. The more wooded taiga also provides greater protection from harsh winter conditions; eastern North American populations of “woodland” caribou do not migrate at all.

Ungulates in temperate regions often migrate altitudinally with the seasons, moving to lower altitudes in winter to avoid harsh weather, low food abundance, and low food accessibility due to deep snow. Such species include bighorn sheep (*Ovis canadensis*), elk (*Cervus elaphus*), mule deer (*Odocoileus hemionus*), and feral horses (*Equus caballus*) in the Nearctic and chamois (*Rupicapra rupicapra*) in the Palearctic. Migratory distances in these species are usually short. An interesting exception is the mule deer, where each individual or family has a definite summer and winter range yet the migratory route is often not the most direct link between them. Individuals have been observed to travel 150 km in straight-line distance, crossing six mountain ranges in a winding route to do so. Tradition appears to have a tremendous effect on defining seasonal ranges and migratory routes in this species.

For marine mammals, annual migrations are the rule rather than the exception. The baleen whales, or mysticetes, migrate the longest distances, some moving thousands of kilometers from tropical to polar waters and then back again each year. Northern and Southern Hemisphere populations both migrate in this way, but their opposite schedules prevent overlap in the tropics. Food availability and quality at higher latitudes drive the timing of these movements. During the polar summer, long day-lengths lead to phytoplankton blooms which, in turn, generate huge abundances of the zooplankton, such as krill, that baleen whales specialize on when feeding. The lipid content of krill and small fish prey, or energy per mouthful from the whale's perspective, also rises during the summer months. Indeed, polar waters provide such an amazingly rich food source that many species forego feeding the rest of the year, consuming enough during a 3-month summer binge to not only maintain themselves for the remaining 9 months but to complete a hemispheric round-trip and breed. Species such as the fin whale (*Balaenoptera physalus*) both mate and give birth 11 months later in warm, low latitude waters. Lactation lasts for 6 to 7 months and the calf is weaned on the summer feeding grounds; fat reserves for lactation just after birth may constitute 50 to 75% of a female's body mass. Lower thermoregulatory costs in warm water presumably allow the binge-migrate-and-fast strategies of baleen whales to be successful, especially since warm, tropical waters may also provide a thermal environment more conducive to calf growth.

Odontocetes, or toothed whales such as dolphins, pilot whales and beaked whales, may show regular seasonal movements between breeding and nonbreeding grounds, but long-distance migrations rarely occur. Most toothed whales feed on single food items like fish

and squid and, as a consequence, must feed every day to maintain themselves. Many have large feeding ranges through which they travel 20 to 60 km per day. Short-distance migrations typically occur when food abundance in an area changes seasonally, such as when fish schools move off-shore or concentrate in river mouths. Some Arctic odontocetes, such as the narwhal (*Monodon monoceros*) and beluga (*Delphinapterus leucas*), migrate short distances seasonally as the polar pack ice retreats and opens up calving grounds in warmer-water estuaries. When the ice returns, they move back with its advancing edge to deeper, ice-free waters. Only the sperm whale—the largest odontocete—migrates distances comparable to those seen in the baleen whales. Adult males move two to three times the distances of females, reaching richer feeding areas at higher latitudes in order to achieve and maintain their much larger body sizes.

Among pinnipeds, 44% of phocid (true or hair seals) species migrate seasonally while only 14% of otariids (fur seals, sea lions) do so. This difference may arise from the phocid's greater reliance on seasonally changing pack ice, rather than solid land, as a substrate for giving birth, raising young, and molting. For example, harp (*Phoca groenlandica*) and hooded seals (*Cystophora cristata*) whelp on pack ice in March near Newfoundland, among other sites, then migrate north with the retreating pack ice to feed. Annual round-trips may cover 4000 km, although there is much variation in migratory distance and route among individuals and populations throughout the North Atlantic. Northern elephant seals hold the current record for the longest migration of any mammal: 18 to 21,000 km annually as they move twice between Californian islands used for breeding (January–February) and molting (July–August) to higher-latitude feeding areas rich in cephalopod prey. Northern fur seals, perhaps the only true migrants among the otariids, migrate in the opposite direction of most phocids, breeding at higher latitudes in the summer months and migrating south to temperate waters to overwinter.

Although clearly not mammals, sea turtles closely resemble pinnipeds in their migratory behavior. Adults migrate to specific island or continental beaches where eggs are laid. Hatchlings venture immediately to the sea where they may spend 30 or more years before maturing and returning to the same beach to complete the cycle.

Tropical marine mammals, such as dugongs and manatees, are generally nonmigratory, although the West Indian manatee (*Trichechus manatus*) population in Florida may move several hundred kilometers north in summer to exploit new feeding areas, retreating back

to warmer waters in the winter. Their need for warm water in which to overwinter has in some cases made them dependent on the warm-water discharges of coastal power plants.

### C. Fish and Other Aquatic Species

Fish show a range of migratory patterns, both among different species and different populations of the same species. Migration is relatively common in this group, particularly as a means of linking rich feeding habitats with specialized spawning grounds that provide refuge for eggs and young fish. For diadromous species—the most studied and conspicuous migratory fishes—this journey requires moving from marine to freshwater environments in order to breed (anadromous: e.g., salmonids, shad, sticklebacks, lampreys) or, more rarely, the converse (catadromous: e.g., freshwater eels, southern trout, southern smelt). Anadromous species predominate in cold-temperate and subpolar waters in both hemispheres, while catadromous species more commonly occur in warm-temperate and subtropical waters. The benefits of moving from lower to higher productivity habitats to feed and grow may account for this pattern since marine productivity is higher at high latitudes while freshwater productivity is higher at low latitudes. Anadromous species constitute a higher proportion of total coastal fish diversity in the Pacific than they do in the Atlantic, largely due to the Atlantic's much higher coastal fish diversity overall.

As with marine mammals (e.g., baleen whales), migrant fishes with separate feeding and spawning areas must feed intensively and store sufficient energy reserves for both migration and spawning. Although many species migrate only a few kilometers between feeding and breeding sites, some fish migrations are truly remarkable in length (also discussed later). Upstream distances in diadromous species include 300 to 400 km in lampreys, 500 km in shad, and well over 1000 km in some salmon and sturgeons. Sockeye salmon migrating upstream expend 70% of their available energy in reaching the spawning grounds and the rest on spawning itself; not surprisingly, both sexes die shortly thereafter. The combined toll of fasting, migrating upstream, and spawning results in semelparity for most diadromous species, but some species, such as the Atlantic salmon (*Salmo salar*) and northern populations of American shad (*Alosa sapidissima*) may migrate and spawn more than once. The dramatic physiological transition required to move between freshwater and marine environments may also be a factor in the prevalence of semelparity among diadromous species.

Salmonid migrations typify anadromous life cycles,

but they also illustrate the very flexible migratory strategies of fish. Pacific salmon (*Oncorhynchus* spp.) spawn in streams at cold-temperate and subpolar latitudes on both sides of the Pacific. After hatching, they either (a) travel immediately downstream to the ocean or, (b) before migrating to the ocean, move downstream to lakes where they spend the next 2 to 4 years, or (c) remain in their hatching stream for one or more years. Once in the ocean, some species migrate the breadth of the Pacific to reach feeding areas while others remain closer to their spawning streams; accordingly, the time spent feeding and growing in the marine environment varies among species but is generally 2 or more years. The same species may have several distinct upstream migrations throughout the year. Further variation on this pattern is shown by individuals (e.g., sockeye salmon: *O. nerka*) that never migrate to the ocean but remain in lakes and move upstream to spawn with ocean-returning migrants. In contrast to other anadromous fish, such as shad and lampreys, salmonid ancestors resided year-round in freshwater and evolved the ability to migrate to marine habitats, perhaps accounting for the great variety of migratory strategies now seen in this group.

The best known catadromous migrants are the freshwater eels (*Anguilla* spp.). They occur throughout the world, including North America and Europe where most studies of their life cycles have been conducted. In North America, adult eels live in rivers and brackish estuaries from the Gulf of Mexico to Labrador, migrating into the Atlantic to spawn. The actual spawning sites still remain a mystery, but small larvae (leptocephali—once thought to be a separate species with no affinity to eels) have been found in the Sargasso Sea, suggesting spawning is here or very nearby. Over the first year, currents carry the leptocephali toward inshore habitats where they metamorphose into glass eels and leave the marine environment. After several years in fresh water, they are ready to migrate, spawn, and complete the cycle, dying after they spawn. Interestingly, adult females are found at greater distances from the spawning grounds—further up rivers, in lakes, and at higher latitudes—than adult males and should thus have greater costs of migration. Females also grow larger at maturity suggesting very different life history strategies for the two sexes.

Oceanodromous species, or those that migrate throughout the marine environment alone, often travel complex migratory circuits between sites optimal for different stages of the annual or life cycle. Ocean currents and coastal topography shape these movements, and those of prey items, adding further complexity to

the annual migratory pattern. For example, Atlantic herring (*Clupea harengus*) have numerous northern subpopulations, each with distinct spawning and feeding grounds and each following its own migratory route and schedule without intermingling to any great extent. Immature fish have separate migrations, moving inshore during warmer months and offshore to deeper waters during winter. Upon reaching maturity, they follow the migratory circuit of their parents. Other oceanic long-distance migrants include cod (*Gadhus morhua*) and plaice (*Pleuronectes platessa*) at higher latitudes and several tuna species in temperate and tropical waters. In most cases, these migration circuits are on the order of 1000 to 3000 km in diameter. Littoral migrations also occur commonly among oceanodromous species, but they are generally much shorter. Those species moving to track food resources usually move offshore in winter, while species migrating inshore during winter may do so to spawn in relatively predator-free habitats.

Species living solely in freshwater (i.e., potamodromous) commonly migrate from deep to shallow waters in order to spawn. A large variety of potamodromous fish migrate throughout the world's lakes and rivers, including freshwater rays, sturgeons, suckers, minnows, pikes, sunfishes, darters, perches, and catfishes. Shallower, upstream habitats may exclude certain predators or have water characteristics (e.g., oxygen, temperature, or silt levels) more suitable to the development and growth of eggs and young fish. The South American characin, *Prochilodus mariae*, has both non-migratory, lake-breeding and migratory, stream-breeding individuals within the same population. The observation that lake breeders expend five times more energy on egg production than stream breeders suggests that stream habitats provide sufficient benefits in offspring survival to compensate for the high energetic costs of migration.

Beyond fish and marine mammals, migration in aquatic animals is less common—or perhaps just understudied. Spiny lobsters in the family Palinuridae demonstrate a very curious migratory behavior, queuing up in long lines of up to 50 individuals that snake their way along the ocean bottom towards deeper, more sheltered habitat. These movements, as much as 30 to 50 km in length over several days, usually occur seasonally in response to a greater incidence of polar storm fronts, which bring colder water temperatures and greater wave disturbance to the shallow water habitats the lobsters feed in during warmer months. Migration thus allows spiny lobsters to exploit a resource-rich, yet seasonally stressful, habitat.

Plankton show distinct migrations in the water column driven by the seasonal availability of nutrients. During months of high productivity, plankton migrate to surface waters; as productivity declines, they return to lower depths and often enter into a state of diapause. Since changes in light work in concert with endogenous rhythms to set the schedule for these movements, the maximum depth reached during the nonproductive season will depend on how far light penetrates the water. In the Northeast Atlantic, typical maximum depths are 1200 meters. Seasonal planktonic migrations are most prevalent at higher latitudes and in regions where currents cause seasonal upwellings of nutrients. (Many planktonic species demonstrate diel vertical “migrations” in the water column, but because these constitute daily ranging movements rather than migrations by the definition given earlier, we do not consider them further.)

#### D. Amphibians and Reptiles

Amphibian life cycles often require a return to water to reproduce, resulting in migrations from feeding areas or refugia to seasonal ponds, streams, and other water sources. However, salamanders and anurans are not known for their great mobility and consequently most species do not travel far from where they were born. Red-bellied newts (*Taricha rivularis*) may hold the distance record for their migrations of 1 km from feeding range to breeding site. Similar migrations in anurans may extend several kilometers. Still, migrating anurans are remarkable for their fidelity to specific breeding sites and the precision with which individuals are able to return to the same few meters of shoreline each year.

Lizards very rarely migrate, but those that do also move to seek out suitable nesting sites. Green iguanas and a few related species that nest on islands may move as much as several kilometers to lay eggs at specific sites; the scarcity of appropriate soil in their island habitat presumably drives these migrations. The red-sided garter snake (*Thamnophis sirtalis parietalis*) migrates to winter hibernacula, such as limestone sinkholes, aggregating in the thousands to buffer the harsh winter conditions of Manitoba.

#### E. Insects and Other Terrestrial Invertebrates

Our current state of knowledge regarding insect migrations contains a number of spectacular examples but gives the overall impression that the phenomenon is

not common in this group despite its huge contribution to our planet's biodiversity. Well-known examples include the Eurasian milkweed bug (*Lygaeus equestris*) and the North American ladybird beetle (*Coleomegilla maculata*), which perform seasonal movements from summer feeding and breeding sites to more protected areas a few kilometers away. There they diapause and last out the winter surviving on stored fat. Individuals may coalesce at specific sites and form spectacular aggregations. As in other migratory species we have considered, migration allows both species to exploit rich, but seasonally variable, habitats. Because many insects have highly specialized feeding and breeding requirements, even small seasonal variations in temperature, moisture, and light levels may be sufficient to trigger migrations to sites where individuals can wait for local conditions to improve (i.e., diapause). The general migration pattern illustrated by ladybirds and milkweed bugs may thus be widespread among both temperate and tropical species. However, small insects are likely to migrate only short distances and be overlooked and understudied, especially if migrations occur without individuals aggregating into conspicuous groups.

An alternative migration pattern takes individuals from habitats declining in quality to richer sites that allowing feeding and breeding to continue. Among the noctuid, or armyworm, moths—a group found throughout the world—several species follow annual round-trip migratory circuits of several hundred kilometers that take them to successively higher latitudes or to habitats recently freshened by rain with host vegetation more suitable for breeding. In many cases, these species have become significant agricultural pests having a life history ideally suited to quickly exploiting ephemeral but rich food sources. Interestingly, single generations may complete only part of the migratory circuit, raising fascinating questions regarding the evolutionary genetics of migration timing and navigation. Perhaps the most famous migrant of this type is the monarch butterfly, *Danaus plexippus*, of North America. Monarch larvae feed on milkweeds and the timing of migration coincides with seasonal milkweed growth. In autumn, adult butterflies from as far north as eastern Canada migrate more than 1000 km south to overwinter in huge aggregations at only a few critical sites in the mountains of central Mexico. In spring, these same adults migrate 200 to 300 km to the northern edge of the Gulf of Mexico where they breed and die. Successive generations then move north tracking the milkweed growing season.

When large-scale migrations occur in insects, they are truly astounding behaviors if one considers the size

of the travelers and the distances covered. Because insects have only limited capacity for fuel storage, they depend more greatly than other migrants on winds to propel and direct their movements. North American aphids, for example, fly for 2 to 24 hours and travel from 50 to 1100 km before setting down. Differing wind speeds at different altitudes account for the variation in flight time and distance moved, but aphids probably exert some control over these features by choosing an altitude at which to fly.

In response to deteriorating local conditions, either due to seasonal reductions in food or increased densities, some insects develop long-winged (i.e., macropterous) forms able to travel relatively large distances to find more suitable habitat. Examples include planthoppers and aphids. These irruptive movements also occur in other taxa, such as birds, but they resemble dispersal processes more than the migrations we have so far considered.

Terrestrial crustaceans, such as hermit crabs and ghost crabs, must migrate from nonbreeding, feeding habitats to high-salinity water in order to breed. These movements occur seasonally, especially in response to tidal cycles, and may require movements from 10 to 3000 m each way.

#### IV. ATTRIBUTES OF MIGRANTS AFFECTING SUSCEPTIBILITY TO HUMAN DISTURBANCE

As we have seen, migrations appear to serve a few common functions. In general, they enable species to temporarily avoid harsh conditions or to meet important biological needs that are separated by great distance. Not surprisingly, there are certain attributes of migrants that affect their susceptibility to human disturbance.

First, migratory species often use a variety of habitats, leaving them vulnerable to multiple points of disturbance. Often harm is felt mostly at one destination. Bachman's warbler (*Vermivora bachmanii*), for example, was driven to extinction by the destruction of its overwintering habitats in the tropics. For others, however, such as many Neotropical migrant birds, the impacts of human activities are felt at both endpoints of their migrations. Globe-trotting species like the white-rumped sandpiper typify the precarious dependence of migrants on habitat health across a tremendous geographic scale. But harm need not be limited to the endpoints of a migrant's journey. Any diminution in

quality of refueling sites along the way could winnow a population and limit its ability to replenish its numbers before the next cycle of migration. And given that migratory trajectories for many species are shaped by the vagaries of prevailing winds or currents, conservation strategies entailing the protection of all stopover areas becomes almost impossible. Even where resting points are protected, unintended consequences associated with the normal nonintrusive behavior of naturalists can put migrants at risk. Steady viewing by bird-watchers at refuges along coastal fly routes has been known to force birds to move too far offshore where they can no longer feed on inshore marine invertebrates exposed at low tides.

Sea turtles probably illustrate best the effects of migratory species being vulnerable at many life-history stages to the excesses of human behavior. Not only are their eggs sought after and easily harvested by indigenous peoples, the beaches themselves are often degraded by the activities of affluent humans. Entire breeding populations are eliminated when beach habitats are developed or severely compromised in their abilities to launch young when dune buggies destroy nests or excessive night lighting disrupts the water-seeking behavior of newly hatched young. In addition, for those young that mature to subadults and return to the breeding grounds, nets of fishermen, particularly shrimpers, provide the *coup de gras* to the species by drastically reducing the pool of future breeders. In fact, it has been shown for that the most vulnerable period in the life cycle is not the nestling, but rather the subadult, stage. Although protecting beaches and increasing the number of functioning beaches gives the species a head start by diversifying recruitment sites, sea turtles are most vulnerable to extinction if the number of subadults is reduced. It is at this point in the species' life history where a long life of breeding commences and reproductive value is highest. Only by insisting that shrimpers insert turtle excluder devices (TEDs) in nets so that the turtles can escape before drowning will there be any hope that those migratory species already endangered will survive.

Second, migratory species often aggregate when traveling. While this behavior might reduce each individual's risk of falling prey to nonhuman predators, both aquatic and terrestrial species in groups are easy targets for the advanced harvesting gear employed by commercial fishers and hunters. In prehistoric times, hunters stampeded herd animals off cliffs and into canyons where massive kills occurred. Such actions have been implicated in the extinction of many of the North American megafauna. Today even the smaller, more elusive

and difficult to capture prey are at risk. Many marine fisheries involving migratory schooling fishes, such as those of cod and haddock, have been overfished. And despite moratoria, many are not recovering.

Third, since many migratory species move according to very precise schedules, any delays caused by human alteration of the habitat or disruption of normal movements could lead to a cascade of effects, jeopardizing a species' ability to replenish its population numbers reduced by ordinary mortality. As noted earlier, migrating shorebirds if forced to linger longer at refueling refuges, could have breeding seasons shortened sufficiently to lose the ability to lay one or more clutches. And since selection favors renesting because nest predation rates are already high, any force constraining such efforts could severely limit a species' recruitment potential.

Despite these attributes that threaten species survival there are some beneficial traits exhibited by migrants. By occupying, at least temporarily, many different habitats, migrants can spread risks and thus can escape catastrophes that befall one location. Unless all individuals in a species simultaneously occupy the same habitat, survivors from habitats not impacted by the catastrophe can serve as sources for increasing species numbers. Migratory species are also more likely to find and colonize new habitats as they are opened up by climate change or other human-induced and natural changes in the environment.

## V. ECOLOGICAL CONSEQUENCES OF HUMAN DISTURBANCE ON MIGRANTS

As humans fragment landscapes, many migratory species will find themselves in peril. For species that must move from one region to another in order to meet a variety of biological needs, any barrier disrupting these movements could lead to local extinctions. In southern Africa, fences are being built either to prevent mixing of wildlife and livestock and the concomitant transmission of disease (ensuring that exported beef is disease free) or to prevent wildlife from gaining a competitive edge over livestock when both compete for critical resources such as food near water. Without access to these resources at the time of year most critical for developing juveniles, any migratory population's recruitment will be severely curtailed. Many local populations will either go locally extinct or will be transformed from "sources" of new emigrating individuals to "sinks" where excess individuals from healthy populations find

refuge. Even for populations that still boast numbers in the hundreds of thousands, such as the plains zebra or even African elephants (*Loxodonta africana*), fragmentation of the large populations is well underway. With human populations expanding into habitats only marginally suited for horticulture, the cry for fencing areas to prevent crops being consumed by migrating herbivores is increasing.

When the movements of these populations are disrupted, a cascade of indirect effects on both the recycling of nutrients in ecosystems and the structuring of animal communities are likely to occur. Since migratory species represent some of the largest aggregations and highest densities of individuals seen in the animal world, any disruption to migration is likely to reduce local density. Thus it is quite possible that the important impact on the structure of vegetation, or more generally the recycling of nutrients, will be altered. It has even been proposed that the overharvesting of sperm whales and the removal of their large carcasses from the ecosystem have impacted the community stability of the deep-sea communities associated with deep-sea vents. By preventing carcasses from falling to the ocean floor, an important source of renewing "island" resources has been eliminated, thus making it more difficult for species tightly associated with vents to hop from one vent to another. With respect to structuring communities, if plains zebras are unable to move in large numbers across the landscape from the tops of catenas where they forage during the rains to the wetter valleys where they move when the rains cease, then their facilitative effect on diversifying the herbivore community will be reduced. Typically, plains zebra move to the wetter valleys and graze down the tougher, fibrous forage that dominates these areas making the greener and more nutritious forage available for the ruminants that require such higher quality vegetation. Thus by disrupting the movements of such a "keystone" species, the diversity of large grazers could be reduced. Furthermore, exclosure experiments suggest that the grazing community could shift from one dominated by large ungulates to one populated by small rodents and lagomorphs. Since rodents often come in contact with people and harbor human diseases, the cascade of effects could be quite profound.

Perhaps one of the greatest impacts of disrupting movements will be seen via the effects that elephants have on the landscape. As populations become restricted to reserves if their numbers are allowed to increase, then their ability to transform a mosaic landscape of trees and grasslands into one of mostly grasslands will mean the disappearance of resident for-



est and woodland dwelling species of colobine monkeys and antelopes.

Even for those subdivided populations that can adapt to the nonmigratory habit—and in certain places such as the Ngorongoro Crater in Tanzania, resident and migratory populations of zebra, Thompson's gazelle, and wildebeest coexist—genetic effective population sizes will shrink and the loss of heterozygosity could leave the population vulnerable to the emergence of novel diseases or the devastating effects of inbreeding. The trend to isolate migratory wildlife into reserves that are too small, although serving as an immediate panacea to human population expansion and development, could have adverse long-term consequences for the health and survival of species unless they are designed to be of appropriate size or have appropriate corridors for migration included in their design. The creation of multinational parks in South Africa, Mozambique, and Zimbabwe that include important mixes of habitats required by a wide variety of migratory species is on the drawing board and offers some hope that subdivision and isolation of populations will be reduced.

Migratory species are also likely to be directly affected by global climate change. It has been suggested that the arctic tundra as a habitat might disappear. If the upper 2 to 3 m of permafrost, which helps bind the tundra together, were to disappear, then the tundra as a habitat would vanish. The effects would drastically reduce populations of caribou that traverse large distances across it as they move from summer breeding to more protective wintering feeding sites. Such transformations of animal communities have occurred in the past when climates have dramatically changed and migratory species were excluded from important parts of their ranges. For example, much of the Canadian-Alaskan megafauna vanished 10 to 20 million years ago when the forests closed and sedge and other patchy but nutrient-rich herbaceous species disappeared. But the scope of these impacts are likely to be greater in the near future because both the magnitude and rate of change associated with the climate effect is increasing and is being coupled with widespread land-use changes that are occurring in areas where the migrants will be forced to move.

## VI. EVOLUTIONARY CONSEQUENCES OF HUMAN IMPACTS ON MIGRANTS

While it is clear that human behavior disrupting migrations will have ecological consequences, the effects on

the evolutionary potential, or even the future characteristics, of species are also likely to be large. Nowhere is this impact more evident than in North America where the range of the bighorn sheep has been severely reduced and fragmented. Bighorn migrate seasonally from meadows to mountaintops to avoid harsh climate and in search of grass that changes seasonally in quality at each location. As human populations have expanded, populations have been isolated into refuges, numbers are shrinking, and genetic diversity is being lost. In addition, hunting for trophy males is removing just those individuals with the best and most diverse genotypes. As a result, the ability of the isolated populations to remain genetically diverse enough to forestall problems associated with inbreeding or to avoid the ravages that would occur if a novel disease appeared has been severely compromised.

Migratory Atlantic salmon (*Salmo salar*) provide perhaps the clearest example of where human activities are causing a species to change under our own eyes. Atlantic salmon typically are born in fresh water and develop for 1 to 2 years before smolting and heading to sea to grow and fatten by feasting on a rich supply of marine invertebrates. Once they attain a certain size, they become sexually mature and return to rivers. There they travel long distances upstream until they reach clear and cool breeding grounds. Some individuals, however, never head to the sea. Instead, they remain in their natal streams and mature sexually at young ages and at small sizes. Such individuals are called parr and they never go through the smolting process that adapts them to a marine lifestyle. Under pristine environmental conditions, the fraction of the population that becomes parr is small since the reproductive gains of such a strategy are low. When competing for mates with larger more aggressive males, parr fare poorly. What matings they obtain are derived by "sneaking" among a mating pair and releasing milt at just the right time. Such events are rare and the mixing of milt is poor so such sneak matings result in few young being sired by parr. But the survival prospects of parr are high since they do not incur the risks of going to sea and, moreover, they begin breeding at a very early age. Thus over a lifetime, reproductive success is modest. But as human fishing increases and the netting of the older larger salmon intensifies, the *relative* lifetime reproductive success of parr is improving. Since the costs of migrating to sea and back again have increased dramatically, the long life span of parr gives them a relative advantage. And given that competition with the larger males for mates is also declining, the chances of parr securing matings are also improving. Thus it is not

surprising that the composition of the population is changing as parr increase markedly in abundance. The impact on the long-term stability of the population is unclear. But the long-term impact on the species appearance is clear, just as it is for yet another fishing industry that will go into decline as the large fish disappear.

### See Also the Following Articles

ARCTIC ECOSYSTEMS • BIRDS, BIODIVERSITY • DISPERSAL  
BIOGEOGRAPHY • MOTHS

### Bibliography

Able, K. P. (1999). *Gatherings of Angels: Migrating Birds and Their Ecologies*. Cornell University Press, Ithaca, NY.

Baker, R. R. (1978). *The Evolutionary Ecology of Animal Migration*. Hodder & Stoughton, London.

Berthold, P. (1993). *Bird Migration*. Oxford University Press, New York.

Curry-Lindahl, K. (1981). *Bird Migration in Africa*. Academic Press, New York.

Dingle, H. (1996). *Migration: The Biology of Life on the Move*. Oxford University Press, New York.

McDowall, R. M. (1988). *Diadromy in Fishes*. Timber Press, Portland, OR.

McKeown, B. A. (1984). *Fish Migration*. Timber Press, Portland, OR.

Rankin, M. A. (Ed.) (1985). *Migration: Mechanisms and Adaptive Significance*. University of Texas Marine Science Institute, Port Aransas, TX.

Rappole, J. H. (1995). *The Ecology of Migrant Birds*. Smithsonian Institution Press, Washington, D.C.

Reynolds, J. E. I., and Rommel, S. A. (1999). *Biology of Marine Mammals*. Smithsonian Institution Press, Washington, DC.

Terborgh, J. (1989). *Where Have All the Birds Gone?* Princeton University Press, Princeton, NJ.





# MOLLUSCS

David R. Lindberg  
*University of California at Berkeley*

---

- I. Introduction
  - II. Temporal and Spatial Distributions
  - III. Major Groups
  - IV. Ecology
  - V. Molluscs and Humans
  - VI. Classification
- 

## GLOSSARY

- ctenidia** (*ctenidium sing.*) The molluscan gill, comprising a characteristic muscular axis from which multiple filaments arise. The relative placement of ciliary bands, nerves, muscles, and blood spaces is similar in all extant taxa.
- mantle** An ectodermal tissue layer that secretes calcium carbonate in the form of spicules or shell.
- radula** A ribbon-like structure bearing transverse rows of small “teeth,” typically used to collect food and move it into the mouth.
- torsion** The 180° counterclockwise rotation of the visceral mass and mantle cavity during the larval stage of gastropods.
- 

**THE MOLLUSCA** are a taxon of soft-bodied metazoans, typically with distinct “head” and “foot” regions, and are covered dorsally by calcium carbonate spicules or a shell(s); shells may also be internal.

## I. INTRODUCTION

The molluscan body plan typically consists of four body components: (1) a head with tentacles and eyes, (2) a ventral muscular foot, (3) a dorsal visceral mass, and (4) the enveloping mantle that secretes the shell. A space between the covering mantle and the side of the foot forms a mantle (or pallial groove). In most molluscs the groove deepens in the posterior region to form a cavity that contains a pair of gills, or ctenidia, as well as openings of the rectum, paired renal organs, and gonads.

The mouth typically is equipped with jaws and a radula. Numerous glands and sacs are associated with the mouth (or buccal chamber) and esophagus, which opens into a stomach. A digestive gland also opens into the stomach. The hindgut, or intestine, is typically long and highly looped or coiled. The nervous system is concentrated in the head and consists of four ganglia—cerebral, visceral, pedal, and pleural. They are connected by commissures; the paired pedal nerve cords extend ladder-like through the foot. The circulatory system consists of a heart enclosed within the pericardium. The heart has multiple auricles and a single ventricle. The system is typically open except in cephalopods, which have capillaries. The excretory system is paired and connected to the pericardium as well as the gonads in some taxa. The gonads are also paired, but can be fused into a single structure (Polyplacophora) or reduced to a single organ (Gastropoda). Separate gonoducts are present in some taxa, and in other taxa

the gonads empty into the kidneys. These connected, mesodermal structures (pericardium, kidneys, and gonads) likely represent the reduced coelom of the Mollusca.

The characters that diagnose the Mollusca are a shell-secreting mantle with a tripartite mantle edge divisible into outer, middle, and inner folds; the secretion of an outer proteinaceous layer, or periostracum; the secretion of multiple crystal types and arrangements (microstructures) within the shell; and a pericardium. Plesiomorphic characters include a mantle groove with flow-through water current patterns, bilateral symmetry, and cephalic tentacles.

This body plan has been substantially modified both among and within groups. Diversification appears to have occurred very early in the history of the Mollusca, but there has been surprisingly little change in some groups. For example, the shells of late Cambrian monoplacophorans are almost identical to those of living taxa despite 450 million years of evolution. Other examples of little change include protobranch bivalves, nautiloids, and scaphopods. There have also been some large and notable extinctions, including those of the bellerophonoids, hyoliths, rostroconchs, and ammonites.

## II. TEMPORAL AND SPATIAL DISTRIBUTIONS

The Mollusca include some of the oldest metazoans found in the fossil record. Late Precambrian rocks of southern Australia and the White Sea region in northern Russia contain bilaterally symmetrical, benthic animals with a univalved shell (*Kimberella*) that resembles those of molluscs in many respects. The earliest unequivocal molluscs are found in the early Cambrian, and there is even some support for a Precambrian presence (*Coeloscleritophora*). Most of the familiar groups, including gastropods, bivalves, monoplacophorans, and rostroconchs, all date from the early Cambrian, whereas cephalopods are first found in the middle Cambrian, polyplacophorans in the late Cambrian, and the Scaphopoda in the middle Ordovician. Most of these taxa tend to be small (<10 mm in length). After their initial appearances, taxonomic diversity tends to remain low until the Ordovician, when gastropods, bivalves, and cephalopods show strong increases in diversity. For bivalves and gastropods this diversification increases throughout the Phanerozoic, with relatively small losses at the end-Permian and end-Cretaceous extinction events.

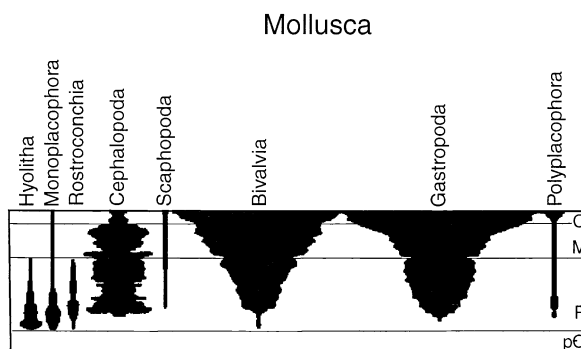


FIGURE 1 Spindle diagrams of the relative diversity of major molluscan groups through time. (Valentine, James; *Phanerozoic Diversity Patterns*, Copyright © 1985 by Princeton University Press. Reprinted by permission of Princeton University Press.)

Cephalopod diversity is much more variable through the Phanerozoic, whereas the remaining groups (monoplacophorans, rostroconchs, polyplacophorans, and scaphopods) maintain low diversity over the entire Phanerozoic or became extinct (Fig. 1).

Molluscs occur in almost every habitat found on Earth. Large concentrations of gastropods and bivalves are found at hydrothermal vents in the deep sea, whereas other gastropods live above tree line in Arctic tundra and on mountain slopes. Molluscs occur in the wettest environments of tropical rain forests and in the driest deserts, where annual activity patterns may be measured in hours. They live below ground in the lightless world of aquifers and caves, and even interstitially in groundwater (stygiobionts), but are also familiar organisms of the seashore, rivers, and lakes and even our gardens. The only thing molluscs do not do is fly, unless the passive aerial dispersal of minute land snails among the Pacific Islands by high winds is considered to constitute flying.

Like many other organisms, marine molluscs reach their highest diversity in the tropical western Pacific and decrease in diversity toward the poles. Marine diversity is also highest nearshore and becomes reduced as depth increases beyond the shelf slope. In terrestrial communities gastropods can achieve great diversity and abundance: as many as 60–70 species may coexist in a single habitat and abundance in leaf litter can exceed more than 500 individuals in 4 liters of litter. Abundance and diversity for some groups can also be higher in temperate communities than in tropical settings. In freshwater communities, species diversity tends to be lower but abundances of sessile molluscs such as the zebra mussel *Dreissena polymorpha* can exceed more than 3000 individuals per square meter.

### III. MAJOR GROUPS

The major groups of living molluscs are clearly dissimilar from one another and have long been recognized as distinct taxa. However, not all were originally recognized as belonging to the Mollusca. For example, the worm-like bodies of the aplousobranchs were perplexing to early biologists and required study of their internal anatomy to ultimately recognize their affinities with the other molluscan groups. This problem becomes especially acute with fossil taxa; the extinct groups (indicated with a dagger) discussed below may or may not be molluscs in our current delimitation of the taxon based on living representatives. However, there is little doubt that at some more inclusive grouping these fossil taxa share common ancestors with living molluscan groups.

The converse problem obtains for living taxa. For example, whereas it is possible to relate living taxa to one another using both morphologic and molecular characters, there exists the real possibility that the living taxa do not share a single most recent common ancestor, but may have had multiple, independent derivations from distantly related molluscs or mollusc-like taxa that are now extinct (see below). These alternatives require that both fossils and living taxa be studied and incorporated into evolutionary scenarios and hypotheses of molluscan relationships, especially when the fossil record provides such a wealth of fossils and putative relatives.

#### A. Coeloscleritophora†

The worldwide presence of small, hollow, calcareous sclerites in numerous Precambrian and Cambrian sediments (collectively referred to as “small shellies”) was an enigmatic component of molluscan evolutionary studies. However, in the early 1990s, an articulated fossil was found in the lower Cambrian of Greenland that was covered with small shellies. It was immediately apparent that what had been thought to be the remains of individual organisms were actually parts of a single larger animal (Fig. 2). Recent work in the Cambrian of Europe, Asia, and Australia has greatly expanded our knowledge of Coeloscleritophora, and although their relationship to the Mollusca remains uncertain, they likely share a common ancestry with the molluscs as well as with annelids and brachiopods.

#### B. Hyolitha†

Hyoliths (Fig. 3) had cone-shaped shells with a (presumably ventrally) flattened side. The aperture was closed by an operculum, and a single pair of anterior

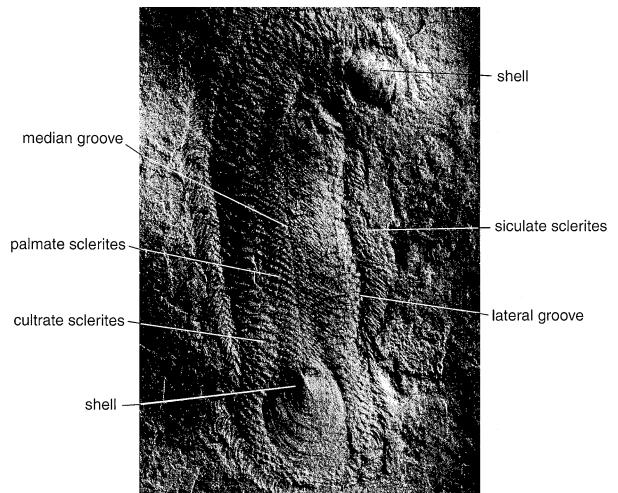


FIGURE 2 Articulated halkieriid from the early Cambrian of Greenland. (Peel, J. S.; Functional morphology, evolution and systematics of Early Paleozoic univalved molluscs. Bulletin Grønlands Geologiske Undersøgelse 161. Copyright © 1991 by Geological Survey of Greenland. Reprinted by permission of Geological Survey of Denmark and Greenland.)

appendages (or Helens) is present in some taxa. The operculum was attached to the shell by pairs of symmetrical muscles. Shells were constructed of crossed lamellar calcium carbonate crystals, and the apical tips bore larval shells. Two forms of larval shells—a smooth globose shell and a point shell with growth lines—perhaps reflected direct and indirect development, respectively. A looped alimentary system connected the ventral mouth to the dorsal anus. Hyoliths were sessile, epifaunal deposit feeders, scraping sediments from the seafloor. They first appear in the early Cambrian and became extinct at the end of the Paleozoic.

#### C. Rostroconchia†

Rostroconchs (Fig. 4) begin life with a single conch (or valve), but the single shell transforms ontogenetically into a nonhinged bivalved shell that gapes at its margins. Based on muscle scar patterns, the animals appear to

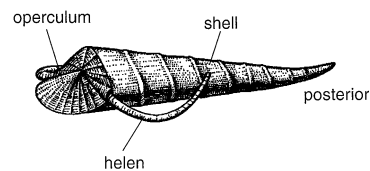


FIGURE 3 The hyolith, *Haplophrentis carinatus*, middle Cambrian. (Briggs, D. E. G., D. H. Erwin, and F. J. Collier; The Fossils of the Burgess Shale, Copyright © 1991 by Smithsonian Institution. Reprinted by permission of Smithsonian Institution Press.)

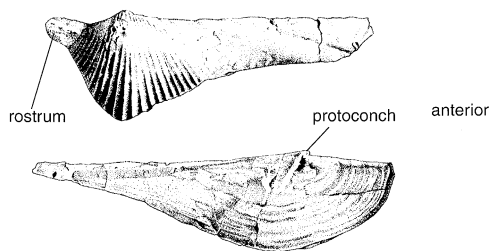


FIGURE 4 Rostroconchia. Top, *Conocardium* sp., Carboniferous. Bottom, *Pinnocaris* sp. Ordovician.

have had anterior feeding tentacles that were likely used for deposit or suspension feeding, and both motile and nonmotile forms have been recognized. Rostroconchs first occur in the lower Cambrian and undergo an extensive late Cambrian and early Ordovician radiation. The last known rostroconchs occur in the Permian. Rostroconchs are thought to share common ancestry with the Bivalvia.

### D. Polyplacophora

Polyplacophoran molluscs, or chitons (Fig. 5), have a dorsal shell that is secreted as eight valves or plates. The plates of individual animals differ in size and shape based on their sequential position; this plate dimorphism is also present in the earliest chitons of the late Cambrian. Although known since the early Paleozoic, chitons do not show a marked increase in diversity until the Cretaceous. Chitons occur worldwide in intertidal habitats and at depths in excess of 7000 m. There are about 850 species and they live on a variety of firm substrates ranging from rocks to algae.

Shell plate morphology has been an important character in determining taxa and relationships. The plates are composed of two distinct shell layers. The outer surface of each plate is high in organic content layer and overlies an inner calcareous layer. Plates may be

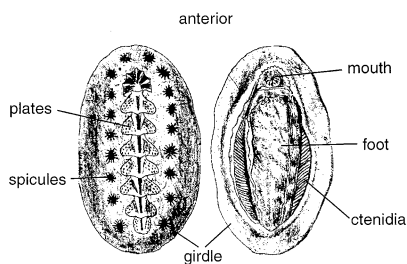


FIGURE 5 External morphology of the Polyplacophora. Left is dorsal view, right is ventral view. Modified from Gray, Maria; Figures of molluscos animals, selected from various authors. 1857–1859.

covered by a thin periostracum. The posterior and anterior edges of each plate have projecting flanges and slits for articulation with other plates, and the outer surfaces are sculptured into distinct regions. The outer plate layer is penetrated by numerous vertical canals that connect to the mantle edge via a canal system that lies between the outer and inner layers. The canals come in two sizes and connect sense organs (including eyes) on the dorsal surface of the valves with the mantle's nervous system.

The plates are surrounded by the girdle, which is formed by a strongly differentiated mantle epidermis. The mantle secretes the plates and surrounds their articulated form. Around the outer plate boundaries the mantle epidermis becomes thicker and secretes a vast array of scales, spines, bristles, and other structures on its dorsal surface. In some species the anterior portion of the mantle epidermis is modified into a large hood that is used to trap food; most species feed by grazing on hard substrates. A poorly defined head and elongated foot that is surrounded by the ventral mantle epidermis of the girdle mark the ventral surface of the animal. The gills are located in the mantle groove that is formed between the girdle and the lateral sides of the foot. The excretory pores and gonoducts also open into this groove.

The visceral mass of the chiton is dominated by a long looped alimentary system. Above the digestive system is a bilobed gonad that appears to result from the fusion of a single pair of gonads. Two separate gonoducts lead from the gonads to opposite sides of the mantle groove. The excretory organs are paired and each connects with the pericardial cavity via a short canal. The center of the pericardial cavity contains an elongated ventricle that is surrounded on either side by enlarged auricles that collect blood from the lateral gills. The polyplacophoran nervous system is ladder-like and almost identical to that of the aplacophorans and monoplacophorans; however, it differs from both by lacking cerebral ganglia. The radular musculature is also similar to that of the monoplacophorans, and the radular tooth configuration is docoglossate like that of the monoplacophorans and patellogastropod limpets. The typical chiton radula has a weak central tooth that is flanked by strong lateral teeth and weaker marginal teeth, some of which are only plates.

Chiton development has been well studied. The larvae come from relatively large eggs and are nonfeeding. The eggs are covered by a thick hull with elaborate spines and knobs. The chiton embryo hatches as a trochophore, a globular cell mass with an apical tuft and central band of cilia, and subsequently elongates and differentiates into a small chiton. Larval eyes are

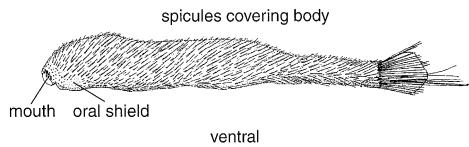


FIGURE 6 External morphology of a chaetodermomorph Aplacophora. (Redrawn from Scheltema, 1985, *Biological Bulletin*, 169:496, Fig. 3L. Reprinted by permission of the author.)

present but disappear after settlement. The most marked external development is the formation of the foot on the ventral surface of the larva and the formation of transverse bands on the dorsal surface of the larva that mark the sites of valve formation. As the valves form, the apical tuft and prototroch are lost.

### E. Aplacophora

The morphological distance between an aplacophoran mollusc (Fig. 6) and, for example, a cephalopod mollusc stretches anyone's concept of the rank of phylum. Aplacophora were the last class to be placed in the Mollusca and have until recently been one of the least studied groups. The aplacophoran molluscs include two different morphological groups: the Chaetodermomorpha or Caudofoveata, and the Neomeniomorpha or solenogastres. The Chaetodermomorpha are footless vermiform molluscs that live in sediments. The Neomeniomorpha are more elongate, have a narrow foot, and typically live in association with cnidarians such as hydroids and alcyonaceans. They are found throughout the world and live at depths between 18 and 6000+ m. It is estimated that there are about 250 species of aplacophorans.

The shell of the aplacophoran mollusc is a myriad of calcareous spicules that are secreted by the mantle epidermis. Spicule morphology varies over the body of the aplacophoran, and in some taxa spicules are modified into scales.

The calcareous spicules that cover the bodies of most aplacophorans give the animals a striking sheen. In the Neomeniomorpha there is a pedal groove on the ventral surface of the animal. The pedal groove expands into an anterior pedal pit and contains the narrow foot. The Chaetodermomorpha lack a pedal groove but do have a posterior mantle cavity that contains a gill. In some Chaetodermomorpha there is a slight constriction and distinct change in spicule morphology at the midline of the body.

It is the internal anatomy that provides evidence of the molluscan identity of the aplacophorans. In both groups the anterior end of the alimentary system in-

cludes a radula and odontophore. In the Chaetodermomorpha the radula and odontophore are strongly developed, and the alimentary system is more differentiated than in the Neomeniomorpha. Both groups have a dorsal gonad that opens into the pericardium, which contains the heart. From the posterior portion of the pericardium there extends a coelomoduct that loops or bends and ultimately opens into the mantle cavity. In the Neomeniomorpha the posterior portion of the coelomoducts are modified for reproductive functions such as sperm storage or brooding young. The nervous system is ladder-like with a well-developed cerebral ganglion. Radular configurations are quite variable and show a wide range of tooth development and modifications that include jaw-like structures, denticles with cones, and sweepers. This is one of the greatest ranges of radular variation found in a Mollusca, and it stands in marked contrast to the lack of variation in the radular configurations found in both the Monoplacophora and Polyplacophora.

Development of the aplacophoran mollusc includes a test cell larval stage in which the three tissue types (mesoderm, ectoderm, and endoderm) align and differentiate within an exterior cell layer constructed of large test cells. Aplacophoran eggs are relatively large and free-spawned in the Chaetodermomorpha and fertilized internally in the Neomeniomorpha; some Neomeniomorpha brood their young to various stages of development. After the formation of the test cell larva with an apical tuft and prototroch, the posterior development of the differentiating larva quickly outgrows the constraints of the exterior test and develops directly into the juvenile aplacophoran.

After many years without students, the Aplacophora are now the subject of extensive studies. These studies and interpretations of aplacophoran phylogeny have focused attention on this small group of molluscs. Overall, the aplacophoran body plan is very similar to that of the chiton. Aplacophorans and polyplacophorans differ from the monoplacophorans by having a dorsal gonad rather than a posterior gonad. The pericardium is similar in all three groups as are many of the other organ systems and positions. Major differences are found in the type of shell secreted by the dorsal mantle epidermis.

### F. Bivalvia

Bivalves (Fig. 7) are laterally compressed molluscs with a hinged, two-part shell. A noncalcified ligament serves as a hinge and connects the two valves along their dorsal surfaces and acts to force the valves apart. The remaining shell surfaces may tightly close or gape with



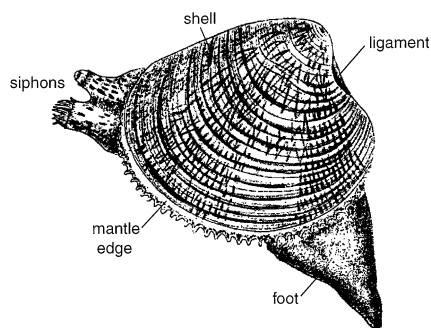


FIGURE 7 External morphology of the Bivalvia. Modified from Gray, Maria; Figures of molluscous animals, selected from various authors. 1857–1859.

contraction of the musculature. Adult valves can vary greatly in shape and size and certain shell forms appear associated with specific habitats. The shell can be internal and reduced (or even absent) and the bivalve animal worm-like such as in “shipworms” (*Teredo*). Although the first occurrences of taxa referred to the Bivalvia are found in lower Cambrian deposits, it is not until the lower Ordovician that bivalve diversification, both taxonomic and ecological, explodes in the fossil record. This diversification continues unabated through the Phanerozoic, with relatively small losses at the end-Permian and end-Cretaceous extinction events. Today over 10,000 species of bivalves are found in most marine, brackish, and freshwater communities. They may be infaunal or epifaunal, and epifaunal taxa may be either sessile or motile. Bivalves typically display bilateral symmetry both in shell and in anatomy, but there are significant departures from this theme in taxa such as scallops and oysters.

Bivalve shells are constructed of different shell fabrics, including crossed lamellar, nacreous, and foliated microstructures. Most of the variability in shell structure sorts along higher taxon divisions. For example, nacreous structures are present primarily in the basal members of the group (Protobranchia) whereas crown taxa have primarily crossed lamellar shells (Heterobranchia). The outer shell surface is often covered by a periostracum, and the free margins of the shell of some taxa may be uncalcified and flexible. In some taxa such as oysters, one valve is cemented to the substrate. The hinge area often contains an array of sockets and pegs, referred to as “teeth” or hinge dentition. These teeth align the valves as they close and prevent shearing of the valves. The inner surfaces of the valves have scars formed by the adductor shell muscles, typically positioned in the anterior and posterior regions and con-

nected by the pallial line that marks the ventralmost attachment point of the mantle.

The mantle that envelops the bivalve animal forms a large ventral chamber or mantle cavity with paired ctenidia. The mantle edge along the posterior portion of the shell often forms a pair of siphons that facilitate inflow and outflow of water through the mantle cavity. Although the pleisomorphic feeding state for bivalves is likely deposit feeding, the ctenidia provide an effective filter-feeding mechanism in most taxa with numerous levels or grades of organization. Bivalves lack a well-defined head and a radula. The foot is often triangular and in some species, such as mussels, the larval byssal gland is retained into the adult and attaches the individual to the substrate with strong fibers.

The visceral mass is primarily situated above the mantle cavity and continues ventrally into the foot. The mouth is often flanked by palps and opens into a stomach. The intestine is irregularly looped and opens dorsally into the excurrent flow near the exhalant siphon. Also opening into this region are the paired kidneys and, when separate from the kidneys, the gonopores of the paired gonads. The heart typically lies below the center of the valves and consists of two auricles and a single ventricle that supplies both anterior and posterior aorta. The nervous system is made up of three pairs of ganglia. These innervate the musculature, mantle, viscera, ctenidia, and siphons. They receive sensory input from statoliths, osphradium, various siphonal sensory structures, and photoreceptors along the mantle margin.

Bivalve development has been well studied, primarily because of the economic importance of bivalves (see below). The eggs are typically small and not very yolk-rich. Fertilization is usually external, but in brooding species occurs in the mantle cavity. Cleavage patterns are spiral and both polar lobes and unequal cleavage patterns are present throughout the group. The first larval stage is typically a trochophore that transforms into a veliger. Although morphologically similar to the gastropod veliger stage, phylogenetic analyses suggest that these larval stages are homoplastic rather than homologous. The initial uncalcified shell grows laterally in two distinct lobes to envelop the body. Larval bivalves have a byssal gland that may assist with flotation while planktic but later attaches the juvenile to the substrate.

## G. Monoplacophora

For more than 50 years the Monoplacophora (Fig. 8) were known only from the fossil record and were recognized by the presence of multiple, symmetrical muscle

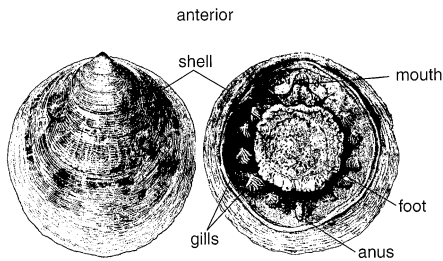


FIGURE 8 External morphology of the Monoplacophora. Left is dorsal view, right is ventral view. (Reprinted by permission from Nature (179:413–416), copyright 1957 Macmillan Magazines Ltd.)

scars in patelliform shells. The discovery of living monoplacophorans, beginning with *Neopilina galathea* in 1957, provided anatomical characters for the group. About 20 living species of monoplacophorans have since been discovered worldwide, living at depths between 174 and 6500 m. They are found both on soft bottoms and on hard substrates on the continental shelf and seamounts. Paleozoic taxa are associated with relatively shallow water faunas (>100 m).

In Recent and fossil patelliform monoplacophoran shells, the apex is typically positioned at the anterior end of the shell, and in some species actually overhangs the anterior edge of the shell. Aperture shapes vary from almost circular to pear-shaped. Shell height is also variable and ranges from relatively flat to tall. Shell sculpture varies from rugose concentric ridges to fine growth lines. In some taxa radial sculpture produces a fine reticulate pattern.

The monoplacophoran animal has a poorly defined head with an elaborate mouth structure on the ventral surface. The mouth is typically surrounded by a V-shaped, thickened anterior lip and postoral tentacles; postoral tentacles come in a variety of morphologies and configurations. Below the head lies the semicircular foot. In the mantle groove, between the lateral sides of the foot and the ventral mantle edge, are found five or six pairs of gills (less in minute taxa).

Internally, the monoplacophoran is organized with a long, looped alimentary system, two pairs of gonads, and multiple paired excretory organs (four of which also serve as gonoducts). A bilobed ventricle lies on either side of the rectum and is connected via a long aorta to a complex plumbing of multiple paired atria that in turn are connected to the excretory organs. The nervous system is ladder-like and has weakly developed anterior ganglia; paired muscle bundles enclose the visceral mass. The most interesting anatomical structures in the original description of a living monoplacophoran were the paired “dorsal coeloms” that are connected to

the anterior excretory organs and are topographically similar to the fused gonads in chitons. Later studies suggest that these structures are more likely extensions of glands associated with the pharynx. The monoplacophoran radula is docoglossate: it has a rachidian tooth, three pairs of lateral teeth, and two pairs of marginal teeth. The lateral teeth are in a stepped configuration: the first and second pairs of lateral teeth are aligned with the rachidian tooth, and the third lateral pair is slightly posterior and lateral to the first two pairs. The cusp of the innermost marginal tooth is frilled.

Developmental studies of monoplacophorans have not been done.

Recent monoplacophorans form a distinct clade, and their similarities and differences with the other extant molluscan groups are easily recognized. There is little question that some Paleozoic taxa also are members of this clade. However, the characters that distinguish some Paleozoic monoplacophorans from torted molluscs (i.e., the Gastropoda) and vice versa are open to alternative interpretations, and the relationships of several major groups of early-shelled molluscs have therefore been the subject of much lively debate.

## H. Scaphopoda

The scaphopod mollusc (Fig. 9) combines gastropod-like shell morphology with bivalve-like development with cephalopod-like anatomy. Members of the class first appear in the early Paleozoic and the taxon has maintained a slow and steady rate of increase in morphological diversification since then. Our knowledge of the biology of recent species comes primarily from members of a single genus, *Dentalium*. Scaphopods are infaunal organisms and feed on foraminiferans and other interstitial prey. Approximately 350 species occur from the intertidal zone to depths in excess of 7000 m and are present in all the major oceans.

The scaphopod shell is a calcium carbonate tube with equal or unequal apertures; the tube may be either inflated or bowed. The shell microstructure includes

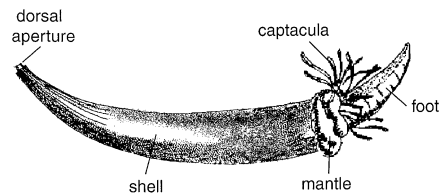


FIGURE 9 External morphology of the Scaphopoda. Modified from Gray, Maria; Figures of molluscous animals, selected from various authors. 1857–1859.

prismatic and crossed-lamellar components; the latter is similar in structure to elements seen in members of the Bivalvia.

The head and foot extend through the ventral aperture of the scaphopod shell. The head bears numerous tentacles (captaculae) that are used to capture and manipulate prey. Foot morphology is variable and has been used as a taxonomic character. The pattern of water circulation through the scaphopod mantle cavity is also unique among the molluscs, because water both enters and exits through the small dorsal aperture. Scaphopods also lack gills.

Unlike the previously discussed groups, scaphopods have a U-shaped gut rather than an anterior–posterior configuration of the mouth and anus. The stomach and digestive gland are in juxtaposition, and the intestine loops before passing through the excretory organ and opening into the mantle cavity. The posterior portion of the digestive gland overlies the gonad that connects with the mantle cavity via the excretory organ. There may be a reduced pericardium, but a heart is absent. The radula consists of a rachidian plate, a single first lateral tooth, and a lateral plate.

The ontogeny of several species has been documented. The trochophore larva has an apical tuft and prototroch and develops mantle folds that divide the larva into equal lateral halves. A velum forms from the prototroch on the anterior end and the dorsal and lateral sides of the larvae begin to secrete a shell that fuses along the ventral surface of the larva but leaves an opening at the posterior end. The trilobate foot forms ventrally under the velum, and the larva attains the tubular shell morphology of the adult scaphopod. The loss of the velum and completion of organogenesis after settling complete metamorphosis.

Scaphopods have an intriguing set of molluscan characters that have been allied to several scenarios of molluscan evolution and relationships. Their shell structure and development suggest bivalvian affinities, but scaphopods also have a radula. The gross morphology of the scaphopod gut is U-shaped, like that of the gastropods and cephalopods, rather than linear as in monoplacophorans, polyplacophorans, and aplacophorans. However, like monoplacophorans, scaphopods have a gonad that lies under the alimentary system rather than on top of it as in aplacophorans and polyplacophorans. Some workers have suggested that aplacophoran larvae and scaphopod larvae share some features; however, others are unconvinced by these arguments and compare the scaphopod velum to that of higher bivalves.

It has been suggested that scaphopods are descended from ribeiriid rostroconchs and they have been grouped

with the Bivalvia. Though there is little doubt that scaphopods share some characters with the Bivalvia, the direct derivation of the scaphopod mollusc from a ribeiriid rostroconch is contradicted by the U-shaped gut present in scaphopods. Rostroconchs are thought to have had a linear gut. Thus, the transition from the ribeiriid condition to a scaphopod morphology would include the construction of a U-shaped alimentary system from a linear one, although the ancestral linear gut would have been an ideal exaptation for a scaphopod shell morphology open at both ends and with flow-through water circulation. Like many other characters, novelties and confusing patterns are common in molluscan morphology but so are convergent grades (especially in molluscan shell form).

## I. Gastropoda

Externally, gastropods (Fig. 10) appear to be bilaterally symmetrical; however, they are one of the most successful clades of asymmetric organisms known. The ancestral state of this group is clearly bilateral symmetry (e.g., chitons, cephalopods, bivalves, see above), but gastropod molluscs twist their organ systems into figure-eights, differentially develop or lose organs on either side of their midline, and generate shells that coil

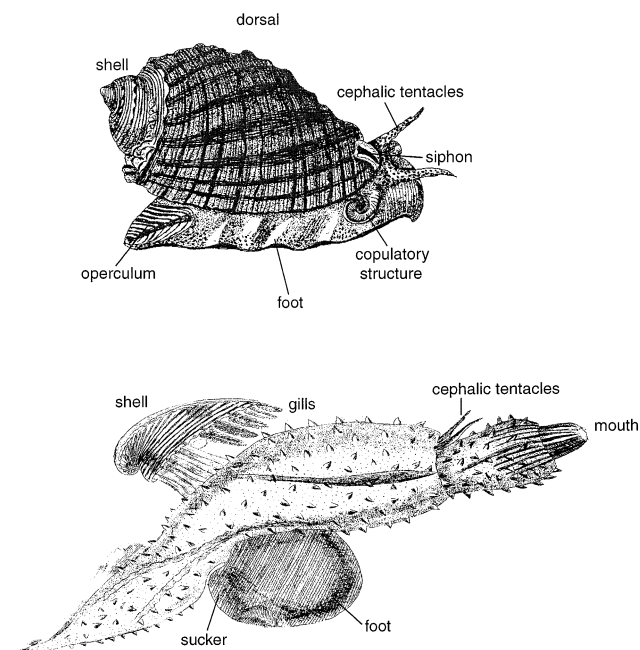


FIGURE 10 External morphology and diversity of the Gastropoda. Top, typical benthic gastropod mollusc. Bottom, a pelagic heteropod. Modified from Gray, Maria; Figures of molluscous animals, selected from various authors. 1857–1859.

to the right or left. And although they have left the more common molluscan state of bilateral symmetry, gastropod molluscs are one of the most diverse groups of animals, both in form and habitat. They occupy habitats ranging from the deepest ocean basins to the highest mountains and from the tropics to high latitudes. Estimates of total extant species range from 40,000 to over 100,000, but may be as high as 150,000, with about 13,000 named genera for both Recent and fossil species. They have figured prominently in paleobiological and biological studies, and have served as study organisms in numerous evolutionary, biomechanical, ecological, physiological, and behavioral investigations. They have a long and rich fossil record that shows periodic extinctions of subclades, followed by diversification of new groups.

The best documented source of gastropod asymmetry is the developmental process known as torsion. Like other molluscs, gastropods pass through a trochophore stage, and then form a characteristic stage of development known as the veliger. During the veliger stage a 180° rotation of the mantle cavity from posterior to anterior places the anus, and renal openings over the head, and twists organ systems that pass through the snail's 'waist' (the area between the foot and visceral mass) into a figure eight. This rotation is accomplished by a combination of differential growth and muscular contraction. In some taxa the contribution of each process is about 50:50, but in other taxa the entire rotation is accomplished by differential growth. Although the results of torsion are the best known asymmetries in gastropods, numerous other asymmetries appear independent of the torsion process. Anopodal flexure, sometimes considered a feature of torsion, is widely distributed in the Mollusca; it is present in the extinct hyoliths as well as in the Scaphopoda and Cephalopoda (and to a lesser extent in the Bivalvia).

Trends in gastropod evolution often feature a reduction in the complexity of many characters. These include reduction of the number of radular teeth, simplification (thought to be due to shell coiling) of the reno-pericardial system (loss of right auricle and renal organ), reduction of ctenidia (loss of the right gill), and associated circulatory and nervous system changes. There is also a reduction of diversity of shell microstructures, simplification of the buccal cartilages and muscles, reduced coiling of hindgut, and simplification of the stomach. Other characters show an increase in complexity, such as life history characters (e.g., internal fertilization with penis and spermatophores and associated reproductive organs). This increase in complexity is correlated with the ability to produce egg capsules and the evolution of planktotrophic larvae and direct

development. There is also an increase in chromosome number, and greater complexity of sensory structures (e.g., eyes, osphradium). In the pulmonates (land snails) the pallial cavity is modified into a pulmonary cavity or lung, while in the opisthobranchs (sea slugs) there are secondary gills and elaborate neurosecretory structures.

Generally, gastropods have separate sexes. One major exception is the heterobranch clade (pulmonates & opisthobranchs), which is exclusively hermaphroditic. Basal gastropods typically spawn gametes directly into the water column, where they are fertilized and develop. Brooding of developing embryos is widely distributed throughout the gastropods, as are sporadic occurrences of hermaphroditism. These basal groups also have non-feeding larvae. Eggs are relatively small in most taxa, although eggs of taxa with non-feeding larvae tend to be a little larger than those taxa that have feeding larvae. Egg size is reflected in the initial size of the juvenile shell or protoconch: when preserved in fossil taxa, this feature has been useful in distinguishing feeding and non-feeding taxa in the fossil record. The first gastropod larval stage is typically a trochophore that transforms into a veliger and then settles and undergoes metamorphosis to form a juvenile snail.

## J. Cephalopoda

Cephalopods (Fig. 11) are dorsoventrally elongated marine molluscs that may or may not have a recognizable external or internal shell. Cephalopods are the most complex and motile of the nonvertebrate metazoans and

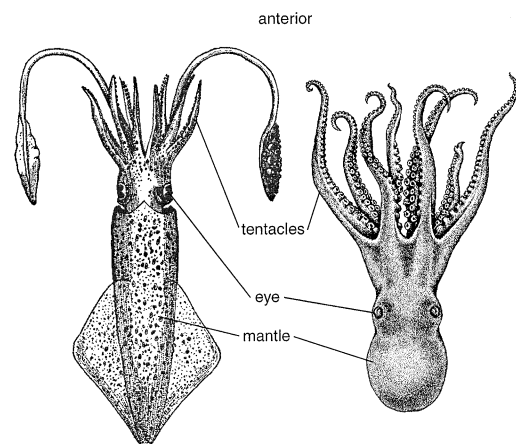


FIGURE 11 External morphology and diversity of the Cephalopoda. Left, squid; right, octopus. (Octopus redrawn from Beesley, P. L., G. J. B. Ross, and A. Wells, *Mollusca: The Southern Synthesis. Part B. Fauna of Australia, Vol. 5*. Copyright © 1998 by Commonwealth of Australia. Reprinted by permission of Australian Biological Resource Study.)

show numerous modifications of the general molluscan body plan. The nautiloids first appear in the late Cambrian and undergo a rapid diversification in the Ordovician. Ammonites (Coleoidea) appear later in the Devonian. Cephalopods are much more variable in their diversity through time than other molluscan groups. They are hit by numerous extinctions (e.g., terminal Permian, Triassic, and Cretaceous events) but typically showed rapid replacement by the survivors. The cephalopods include the largest living as well as the largest extinct molluscs: ammonite shells range to over 2 m across and body sizes of living squid range up to 8 m, with tentacles exceeding 20 m in length. Over 10,000 species are known with about 700 living taxa. All are active carnivores in marine benthic and pelagic habitats from nearshore to abyssal depths.

Cephalopods are thought to have evolved from monoplacophoran-like ancestors. Septa formed at the apex as the animal grew and withdrew into a newly formed body chamber. The old chambers are gas-filled and provide buoyancy for the organism. The foot was modified into a funnel and provides jet propulsion for movement. Prehensile arms with suckers surround the mouth on the head (cephalic in origin) and capture prey. In cephalopods with external shells (ammonites and nautiloids), the shells are composed of an inner nacreous layer and an outer prismatic layer. In other cephalopod taxa the shell is typically internal and reduced to a linear stiffener as in cuttlefish and squid or virtually nonexistent as in the octopus, whose pelagic ancestors reinvaded the benthic realm.

Much of the external anatomy of cephalopods appears associated with their highly motile, pelagic habits. These modifications are so strong that it is difficult to readily identify the typically molluscan body axes in a cephalopod. For example, as a squid jets through the water, the most posterior portion of the body, relative to the direction of movement, is the squid's head with its prehensile arms. The most anterior portion is the dorsal surface of the muscular mantle and the posterior mantle cavity is located ventrally. In addition to jets of water expelled through the funnel, squid and some octopi use undulating movements of paired fins at the distal end of the mantle for swimming as well.

The cephalopod mouth contains strong beak-like jaws that are used to deliver a lethal bite to prey items as well as shred them; a radula is also present. Salivary glands can produce highly toxic venoms in some squids and octopi, and ink sacs are present in most taxa except *Nautilus*. The digestive system differs from that of most other molluscs in having a cecum in juxtaposition with the stomach and a relatively short intestinal tract. There

are two pairs of ctenidia in living *Nautilus*, but all remaining cephalopods (cuttlefish, squid, and octopus) have a single pair of ctenidia. The circulation system also differs markedly from that of other mollusks in being a closed system capable of maintaining a high blood pressure, rather than open and diffuse as in other taxa. The heart consists of two auricles and a single ventricle that supplies multiple arteries. In all cephalopods except *Nautilus*, a pair of branchial hearts situated at the base of the ctenidia pump blood through the ctenidia. The nervous system is highly concentrated and developed in cephalopods. The three major ganglia are fused and organized into lobes, each with its own specific function. *Nautilus* has fewer lobes than other cephalopods, and lobe size may vary among groups with different lifestyles. Coleoid cephalopods also have two large stellate ganglia on the mantle that control both respiratory and locomotory functions of the mantle. Sensory structures include statocysts, olfactory organs, and eyes. Coleoid eyes are surprisingly convergent with vertebrate systems and are capable of resolving brightness, shape, size, and orientation. Cephalopods are also able to change color patterns using an elaborate chromatophore system that is under nervous control; it provides them with the ability to display incredible camouflage.

Cephalopods have separate sexes and spermatophores are transferred between males and females by modified tentacles. Eggs are large and yolk-rich, and embryonic development of cephalopods is different from that of all other mollusks. There is no larval form, just direct development into juveniles. Both the eggs and young may be brooded, benthic, or pelagic.

#### IV. ECOLOGY

Molluscs occur on a large variety of substrates, including rocky shores, coral reefs, mudflats, and sandy beaches. Gastropods and chitons are characteristic of these hard substrates; bivalves are commonly associated with softer substrates, where they burrow into the sediment. However, these patterns are not inflexible. The largest living bivalve, *Tridacna gigas*, occurs on coral reefs, whereas microscopic gastropods live interstitially between sand grains. On hard substrates gastropods are often grazers that feed either selectively or indiscriminately on algal, diatom, or blue-green films and mats or animal aggregations. On soft substrates bivalves typically are suspension or deposit feeders.

The adoption of different feeding habits appears to have had a profound influence on molluscan evolution.

The change from grazing to other forms of food acquisition is one of the major features in the adaptive radiation of the group. Based on our current understanding of relationships, the earliest molluscs were likely carnivores and detritivores that indiscriminately grazed on encrusting animals and detritus. Truly herbivorous grazers are relatively rare in the molluscs and are limited to the polyplacophorans and gastropod groups, whereas living aplacophorans, monoplacophorans, scaphopods, and cephalopods are carnivorous. Some protobranch bivalves are also carnivorous; most bivalves are either suspension or deposit feeders that indiscriminately take in particles but then elaborately sort them based on size.

Cephalopods are typically active carnivores specialized on mobile prey such as fish, crustaceans, and other cephalopods. Because they are so abundant in pelagic systems, cephalopods are often important food sources for larger fishes, marine mammals, and even seabirds. In the gastropods several groups such as Janthinidae are planktic pelagic carnivores whereas the heteropods (Caeongastropoda) and the pteropods (Opisthobranchia), like the cephalopods, are active swimmers in search of prey. These taxa spend their entire lives in the water column where they primarily feed on Cnidaria, other molluscs (including small cephalopods), and even fishes.

Bivalve-like suspension feeding has evolved in some gastropods such as the vetigastropod *Umbonium* and in the pelagic gastropod group Thecosomata. Some groups with carnivorous diets have undergone what appear to be true, explosive, adaptive radiations (e.g., the Neogastropoda). Other carnivorous taxa such as the aplacophorans and scaphopods have low diversity and abundance.

In addition to these more typical trophic strategies and interactions, some molluscs (especially among the gastropods) have also become both endo- and ectoparasitic in and on other invertebrates—most often echinoderms. Galeommatoid bivalves and eulimid gastropods occur as ectoparasites on holothuroid and asteroid echinoderms, respectively. The reduced, worm-like Entoconchidae occur in the internal body cavity of holothuroids, whereas the larvae of freshwater unionid bivalves parasitize fish and amphibians, although the adults are free-living.

Molluscan groups are ubiquitous and diverse in marine habitats, but only the bivalves and gastropods have invaded freshwater habitats, and only gastropods have invaded terrestrial ones. The major terrestrial clade is the pulmonates, which originated at least by the Carboniferous, but other nonmarine groups such as the Neritopsina (marine, freshwater, and terrestrial) are

likely Devonian in origin. Often the terrestrial groups are among the most basal of the extant taxa in the clade. For example, in the Neritopsina terrestrial taxa are thought to be more basal than marine members of the group. These patterns could result from competition among sister taxa and the relegation of one taxon to a unique habitat while the other diversified in the ancestral setting.

Shell morphology is often thought to be correlated with lifestyle and habitat and some substantial changes in body form are clearly associated with major adaptive changes. However, in many cases gross differences in morphology within a group are not readily correlated with habitat. Conversely, similar shell morphologies do not necessarily indicate similar habits or habitats, making paleocological studies with molluscs problematic. For example, limpets occur on wave-swept platforms, on various substrates in the deep sea, in fast-flowing rivers, or in quiet lakes and ponds. It is often suggested that strong wave action selects for limpet morphologies in several groups, but it is obvious from their current distributions that limpets do very well in a wide range of habitats.

## V. MOLLUSCS AND HUMANS

Molluscs and humans are most often associated economically. Molluscs have many important commercial benefits such as fisheries and mariculture but can also be responsible for tremendous economic loss and human suffering. Molluscs are found in some of the earliest human habitation sites in southern Africa over 100,000 years ago, and it is likely that humans have included molluscs in their diet and as material resources for millions of years. More recently, commercial fisheries have focused on bivalve, gastropod, and cephalopod taxa for food (oysters, abalone, scallops, clams, cephalopods, and mussels) as well as pearl and button fisheries.

In freshwater habitats molluscs may serve as intermediate hosts for parasites such as digenetic trematodes and are an integral part of the Schistosomiasis contagion that infects millions in the tropical regions of the world. Nonhuman, molluscan-borne parasites are also responsible for more than \$2 billion worth of losses worldwide to the livestock industry each year. Most people are familiar with “garden snails” and the damage they can do to favorite plants, but this local damage is minimal compared to the role of molluscs as pests. Molluscan pest species cause millions of dollars worth of damage to horticultural and food crops throughout the world every year, and as the global economy expands, the

probability of potentially damaging introductions increases dramatically. For example, over the past 5 years, almost 5000 molluscs from 100 different countries were intercepted entering the United States.

One of the most notable introductions has been that of the freshwater bivalve *Dreissena polymorpha*, or zebra mussel. Zebra mussels were first detected in Lake St. Clair (Great Lakes chain) in 1988. It is thought that zebra mussels entered the Great Lakes in ballast water of commercial ships from eastern Europe, and without natural enemies to control their numbers, they have thrived and spread to 19 states in less than 10 years. In the United States and Canada, they have caused widespread economic and environmental damage. They typically concentrate where water flows rapidly and therefore can quickly clog intake pipes and cut off water flow to hydroelectric and nuclear power plants, industrial facilities, and public water supplies. One mussel removal project in Lake Michigan removed 400 cubic yards of zebra mussels at a cost of \$1.4 million. Zebra mussels can also have tremendous effects on freshwater ecosystems by displacing native species and increasing levels of toxic algae. As zebra mussels have spread across the country, they have devastated native freshwater clam populations (especially the unionid species) that were already endangered and threatened by pollution and habitat modification.

On a more positive note, molluscs are also important in biomedical and biotechnology research. Although not as well known as model organisms such as the fruit fly *Drosophila*, opisthobranch molluscs are commonly used in neurological studies of learning and memory functions, and bivalve molluscs have served as models for environmental and genetic interactions in tumor research. Extracts of numerous other molluscs have been examined for antibacterial and other properties potentially useful in the medical and health sciences. Molluscs also provide important model organisms for environmental health studies. For example, bivalve molluscs are frequently used in studies on the toxicology and environmental monitoring of heavy metals and other pollutants.

Molluscan conservation issues are primarily centered on freshwater and terrestrial taxa. The largest number of documented extinctions has involved nonmarine molluscs, and nonmarine molluscs are second to the arthropods in numbers of currently threatened species. Terrestrial species are impacted primarily by the destruction of native vegetation and cover, and freshwater species by alter flow, sedimentation rates, and pollution. Although marine taxa are thought to be less impacted by human activities, populations of some taxa have

been substantially reduced by overfishing. These include abalone (*Haliotis*), clams (*Tridacna*), and scallops (*Pecten*). Marine taxa, like freshwater and terrestrial taxa, are also threatened by habitat destruction, pollution, changes in sedimentation patterns, and other factors.

As discussed above regarding zebra mussels, the introduction of exotic species can also seriously impact native taxa. This is especially true on islands where the effective population size may be small. One of the best-documented cases was the apparent extinction of the endemic land snail taxon *Partula* from the island of Moorea in French Polynesia. This resulted because of the intentional introduction of the predatory snail *Euglandina rosea* to the island as a biocontrol agent for the Giant African Snail, *Achatina fulica*.

## VI. CLASSIFICATION

Given the state of flux in molluscan and metazoan phylogeny and the wealth of new data that is now appearing from molecular, morphological, and paleontological work, any classification proposed here would be rapidly outdated. Several traditional classifications are available in the references cited in the bibliography. However, few are based on hypotheses of relationships, but are instead based on overall similarity and ad hoc scenarios of evolution.

Classifications based solely on morphology have been especially problematic, and much of this confusion has resulted from problematic taxa such as the aplacophorans, scaphopods, and bivalves where possible reduction and loss of organs or other secondary simplification has produced morphologies that may be argued as either primitive or highly derived. Many classifications have also focused exclusively on the morphology of living taxa and have ignored potential, fossil members of the Mollusca. If extinct fossil taxa are included in evolutionary scenarios, they are typically limited to distinctive clades such as the Rostroconchia and Bellerophonata. Other more problematic extinct taxa (e.g., hyoliths) are systematically ignored, arbitrarily excluded from the Mollusca without analysis, or shoehorned into extant groups.

The sister taxa of Mollusca have included the Platyhelminthes, Annelida, Sipuncula, and the Kamptozoa. Within the Mollusca both Polyplacophora and Aplacophora have been argued as the most primitive taxon, and thus the outgroup to all Conchifera.

Most classifications have assumed a single cladogenetic event in the origin of the Conchifera from the

supposedly more primitive placophoran groups. Alternative hypotheses have derived the conchiferans in an unresolved polytomy from a hypothetical ancestral mollusc, or HAM. Some workers have interpreted the Cambrian Burgess Shale taxon *Wiwaxia* and other less complete halkieriid-like fossils as molluscan while others have argued *Wiwaxia* to have annelid worm affinities. However, the discovery of an articulated halkieriid from the lower Cambrian and the existence of these and other multishelled placophorans necessitate the reexamination of long-held assumptions of molluscan ancestry and monophyly. The rapidly increasing knowledge of Coeloscleritophora diversity suggests that they may harbor independent ancestors for extant molluscan groups.

Molecular phylogenies for the Mollusca have not fared much better than the morphological studies. Nuclear and mitochondrial DNA sequences have had very limited success in resolving a monophyletic molluscan clade or even producing robust or reasonable groupings within the Mollusca (e.g., the bivalves and gastropods). These problems most likely result because of the deep, Paleozoic divergence of many of the molluscan taxa and the variable rates of change across molluscan genomes.

The following rank-free classification is conservative and only denotes the major clades within the Mollusca.

Coeloscleritophora  
Sipuncula  
Hyalitha  
Mollusca  
  Chaetodermomorpha  
  Neomeniomorpha  
  Polyplacophora  
  Conchifera  
    Rostroconchia  
    Bivalvia  
      Protobranchia  
      Pteriomorpha  
      Heterodonta  
      Anomalodesmata  
  Monoplacophora

Scaphopoda  
  Dentaliida  
  Gadilida  
Gastropoda + Cephalopoda  
  Gastropoda  
    Patellogastropoda  
    Vetigastropoda  
    Neritopsina  
    Caenogastropoda  
    Heterobranchia  
  Cephalopoda  
    Nautiloidea  
    Coleoidea  
      Aminonoida  
      Decapoda  
      Octopoda

### See Also the Following Articles

INVERTEBRATES, FRESHWATER, OVERVIEW •  
INVERTEBRATES, MARINE, OVERVIEW •  
INVERTEBRATES, TERRESTRIAL, OVERVIEW

### Bibliography

- Beesley, P. L., Ross, G. J. B., and Wells, A. (Eds.). (1998). *Mollusca: The Southern Synthesis. Fauna of Australia*, Parts A and B, Vol. 5. CSIRO Publishing, Melbourne.
- Broadhead, T. W. (Ed.). (1985). *Mollusks. Notes for a Short Course* (Studies in Geology 13). Department of Geological Sciences, Univ. of Tennessee, Knoxville, TN.
- Davis, G. M. (Ed.). (1999). *Interactions between man and molluscs. Malacologia* 41(2), 319–509.
- Johnson, P. A., and Haggart, J. W. (Eds.). (1998). *Bivalves: An Eon of Evolution—Paleobiological Studies Honoring Norman D. Newell*. Univ. Calgary Press, Calgary, AB.
- Moore, R. C. (Ed.). (1960). *Treatise of Invertebrate Paleontology. The Mollusca. Part I. Mollusca I*. Univ. Press of Kansas, Lawrence, and Geol. Soc. Am., Boulder, CO.
- Ponder, W. F. (Ed.). (1988). *Prosobranch Phylogeny. Malacol. Rev.*, Suppl. 4.
- Ponder, W. F., and Lindberg, D. R. Towards a phylogeny of gastropod molluscs—A preliminary analysis using morphological characters. *Zool. J. Linnean Soc.* 119, 83–265.
- Taylor, J. (Ed.). (1996). *Origin and Evolutionary Radiation of the Mollusca*. Oxford Univ. Press, London.







# MOTHS

David L. Wagner  
*University of Connecticut*

---

- I. Introduction
  - II. Higher Classification
  - III. Fossil History
  - IV. Species Diversity and Biogeography
  - V. Life History
  - VI. Feeding Biology
  - VII. Natural Enemies and Chemical Defense
  - VIII. Dispersal and Migration
  - IX. Key Adaptations
  - X. Pheromones
  - XI. Coloration, Diurnality, Nocturnality, and Attraction to Light
  - XII. Importance
  - XIII. Conservation
- 

## GLOSSARY

**aposematic** Warningly colored; boldly colored, usually involving reds, oranges, or yellows, as well as black and white.

**crochets** The minute hooklets on the fleshy abdominal prolegs of a caterpillar.

**diapause** A period of delayed development that is generally associated with weather conditions that are unfavorable to survival or reproduction.

**detritivory** Feeding on fallen and generally dead organic debris; although the term covers both plant or animal matter, in moths the term is especially apt to apply to leaf litter feeders.

**hypermetamorphic** A life cycle that includes two or

more larval forms, with each often specialized for a different feeding function.

**instar** A larval stage; the first instar hatches from the egg and upon molting enters the second instar. Most moths undergo five or six instars prior to pupation.

**maxillae** The second pair of mouthparts, located between the mandibles and the labium (the third pair of mouthparts).

**monophyletic** A group with a single evolutionary origin, which includes a common ancestor and all of that ancestor's descendants.

**Mullerian mimicry** When two or more distasteful species come to resemble one another. In Batesian mimicry there is a model (unpalatable species) and one or more mimics (palatable species).

**parasitoid** An insect whose larval stage feeds on a second (host) species, killing the host in the process; a parasitoid attack is equivalent to delayed predation and therefore the phenomenon is distinguished from true parasitism.

**pharate** A "cloaked" or hidden stage, for example, the adult moth just prior to its emergence from the pupa.

**pheromone** A chemical released by one individual that elicits a response in a second individual of the same species.

**polyphagous** Eating plants from more than two unrelated plant families.

---

**WORLD MOTH DIVERSITY** is believed to be in excess of 225,000 species, virtually all of which are dependent

on the explosive Cretaceous radiation of the flowering plants. This chapter begins with a synopsis of the higher classification and diversity of the order Lepidoptera, then reviews several aspects of the order's biology. Throughout the chapter, emphasis is placed on the basal lineages on which the group's success is rooted. Special effort is made to discuss the key innovations and circumstances that may have led to the enormous global diversity and importance of moths.

## I. INTRODUCTION

Every night at dusk one and a half million Mexican free-tailed bats flood out from beneath the Congress Avenue Bridge in Austin, Texas. Bats from this single colony harvest more than 10,000 pounds of insects every evening, most of which are moths. Moth caterpillars account for much of the above-ground insect biomass in temperate forests; without them one wonders whether there would be songbirds to usher in each spring. Estimates of species-richness for moths have climbed steadily over the past decade, with the more modest extrapolations falling between 200,000 and 300,000 species. Whether measured in terms of biomass, influence on terrestrial ecosystems, or species-richness, moths must be held as one of the most successful lineages of macroscopic organisms on this planet.

Body sizes span two orders of magnitude: Nepticulidae and Heliozelidae, some with wingspans under 3 mm, are little more than aerial plankton (Fig. 1). The world's largest moth, a neotropical noctuid, has a wingspan that sometimes exceeds 275 mm. All members of the order Lepidoptera can be diagnosed by the presence



FIGURE 1 *Stigmella variella* (Nepticulidae); adults have a wingspan of only 4 mm. Note most of the wing surface area is made up of cilia-like scales.

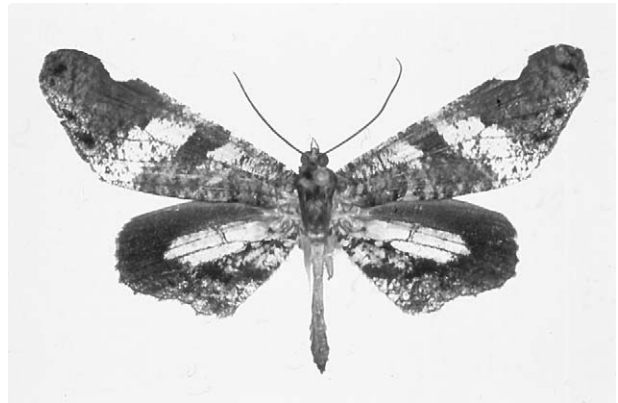


FIGURE 2 *Macrosoma hyacinthina* (Hedylidae); both morphological and molecular evidence suggests these moths are closely related to, if not a subgrouping within, the butterflies (Papilionoidea and Hesperioidea).

of scales, which cover the wings and body. A second readily observable structure that is part of the order's ground plan is the epiphysis, a cuticular flap on the inside of the foreleg, which is used by many moths to clean the antenna. A curious lepidopteran attribute is the production of apyrene (anucleate) sperm—in some species 80% or more of the sperm is apyrene and therefore incapable of fertilization. Kristensen (1984, 1999) identified nearly two dozen other synapomorphies, which, taken as a whole, incontrovertibly established the monophyly of the Lepidoptera and its sister-group relationship with the Trichoptera (caddisflies). Molecular studies of both nuclear and mitochondrial genes have corroborated the major findings of his benchmark studies.

Butterflies and moths make up the order Lepidoptera. Because butterflies derive from within one (or two) moth lineages, it is not possible to identify uniquely evolved features shared by moths that are not also found in butterflies. Recent phylogenetic reconstructions indicate that butterflies have their closest relatives in the Hedylidae, an odd-looking group of neotropical moths that were formerly classified in the Geometroidea (Fig. 2). But regardless of their phylogenetic position within the order, since butterflies are, evolutionarily speaking, nothing more than diurnal moths (and negligible in terms of species richness), “Lepidoptera” and “moths” are used synonymously throughout this chapter.

## II. HIGHER CLASSIFICATION

Phylogenetic relationships among the early lineages of Lepidoptera are among the best known of any insect

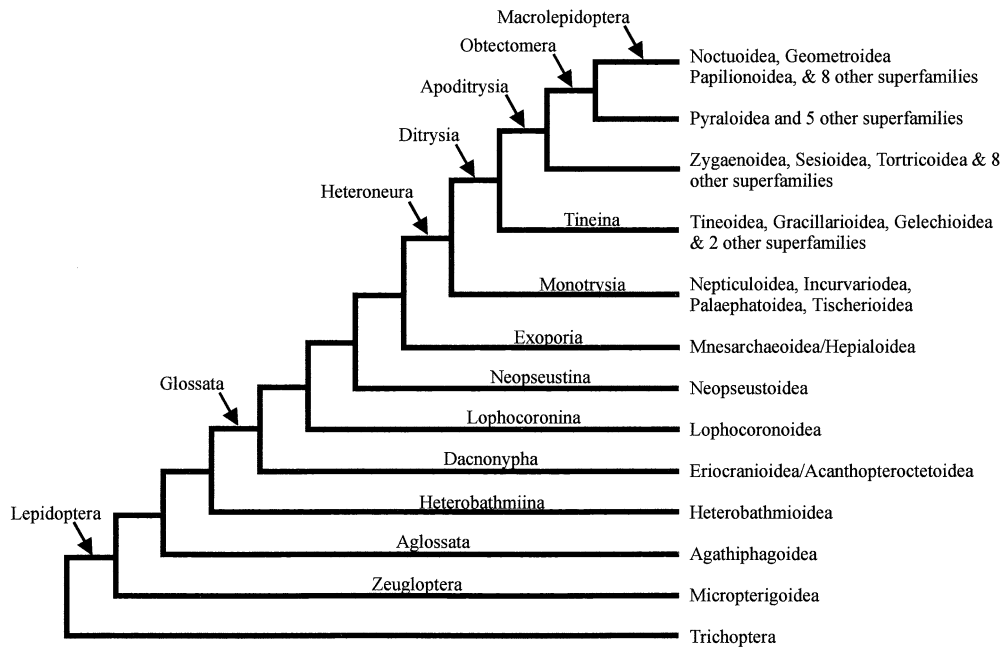


FIGURE 3 Phylogenetic tree for major lineages of moths. Adapted from Kristensen (1999) and Heppner (1998).

order, so well that they have even been used as a “truth table” for testing the phylogenetic signal in several gene sequencing studies. Despite this agreement, there is little consensus as to how to render the phylogeny (Fig. 3) into a classification. Kristensen (1999, p. 28), frustrated by the redundancy and instability of higher classifications for the order, adopted a phyletic sequencing scheme (i.e., he arranged taxa in a hierarchical outline without formal assignation of many higher category names).

### A. The Four Suborders

The most recent higher classification for the Lepidoptera recognizes 46 superfamilies, all but two of which (the Hesperioidea and Papilionoidea) are moths (Kristensen, 1999, and authors therein; Fig. 3 and Table 1). The 120 families have been arranged in four suborders: the Zeugloptera, Aglossata, Heterobathmiina, and Glossata. The first and most basal lineage within the order contains but a single family, the Micropterigidae, with circa 180 known (121 of which are described) species distributed across the major continents. Decidedly primitive, this family shares many morphological traits with the Trichoptera. Noteworthy are the functional mandibles in the adult (and absence of a proboscis). The larvae are bizarre creatures with little morphological affinity to other lepidoteran caterpillars (Fig. 4a), so odd that at one time these moths were placed in their

own order. The Aglossata contains but a single family (Agathiphagidae) with two species, restricted to Australia, New Caledonia and a few Pacific islands. The third suborder, the Heterobathmiina, is represented by nine (three described) species from temperate areas of Chile and Argentina. All are classified in a single genus in the Heterobathmiidae. The remaining diversity occurs in Glossata, which possess, among other traits, a coiled proboscis.

### B. The Glossatan Infraorders

The most basal lineage within Glossata is the Dacnonypha, with a single principal family, the Eriocraniidae. These are small leaf-mining moths of the Northern Hemisphere with 24 recognized species. The remaining glossatan infraorders have been grouped in the Coelolepida, all of which share “normal” wing scale ultrastructure and lack ocelli (light-sensitive but non-image-forming lenses over the dorsum of the head). (Ocelli reappear in some derived moth lineages.) The placement of the Acanthopteroctetidae, another small family, remains in question; some authors place it in the Eriocranioidea (Dacnonypha) and others in the Coelolepida. Australia’s lophocoronids are superficially similar to both of these families but are accorded their own family, superfamily, and infraorder. Although the life history of all six lophocoronids remains unknown, the piercing ovipositor indicates that the larvae are internal

TABLE I  
Diversity of the Lepidoptera

Superfamily	Approx. no. families	Approx. no. described species <sup>1</sup>
Suborder Zeugloptera		
Micropterigoidea	1	121
Suborder Aglossata		
Agathiphagoidea	1	2
Suborder Heterobathmiina		
Heterobathmioidea	1	3
Suborder Glossata		
Infraorder Dacnonypha		
Eriocranioidea	1	24
Acanthopteroctetoidea	1	5
Infraorder Lophocoronina		
Lophocoronoidea	1	6
Infraorder Neopseustina		
Neopseustoidea	1	11
Infraorder Exoporia		
Mnesarchaeoidea	1	8
Hepialoidea	5	609
Infraorder Heteroneura		
Division Monotrysia		
Section Nepticulina		
Nepticuloidea	2	902
Palaephatoidea	1	60
Tischerioidea	1	80
Section Incurvariina		
Incurvarioidea	1	588
Division Ditrysia		
Section Tineina		
Simaethistoidea	1	4
Tineoidea	5	4,350
Gracillarioidea	4	2,300
Yponomeutoidea	8	1,428
Gelechioidea	15	15,540
Elachistidae		3,270
Oecophoridae		3,150
Gelechiidae		4,530
Section Apoditrysia		
Galacticoidea	1	17
Zygaenoidea	11	2,548
Sesioidea	3	1,356
Cossoidea	2	676
Tortricoidea	1	8,000
Choreutoidea	1	405
Urodoidea	1	60
Schreckensteinioida	1	5
Epermenoidea	1	83

continues

Continued

Superfamily	Approx. no. families	Approx. no. described species <sup>1</sup>
Alucitoidea	1	148
Pterophoroidea	1–3	986
Whalleyanoidea	1	2
Immoidea	1	246
Copromorphoidea	2	313
Hyblaeoidea	1	18
Thyridoidea	1	760
Pyraloidea	2	17,763
Pyralidae		6,133
Crambidae		11,630
Mimmallonoidea	1	200
Lasiocampoidea	2	1,575
Bombycoidea	9	3,554
Saturniidae		1,590
Sphingidae		1,269
Axioidea	1	6
Calliduloidea	1	60
Hedyloidea	1	40
Hesperioidea	1	3,500
Papilionoidea	4	13,600
Lycaenidae		6,000
Nymphalidae		6,000
Drepanoidea	2	675
Geometroidea	3	21,740
Noctuoidea	8	41,300
Arctiidae		11,000
Lymantriidae		2,500
Noctuidae		25,000
Notodontidae		2,800
	120	145,677

<sup>1</sup> With but few exceptions numbers were taken from Kristensen 1999. The figure for the Tortricoidea was suggested by John Brown; those for the Bombycoidea and Noctuoidea were suggested by Ian Kitching. The classification is adopted from Kristensen (1999) and Heppner (1998). Species-numbers are given for all 46 superfamilies. Families, other than the nominate, that contain more than 2500 described species are also given. A few additional families that are mentioned repeatedly in the text are also included.

feeders. The neopseustids are peculiar insects that in some ways resemble lacewings. Larval stages are unknown, but morphological evidence suggests that they feed internally; the 11 species are found in South America, China, and Southeast Asia. The Exoporia, with six families and more than 600 species, are the only

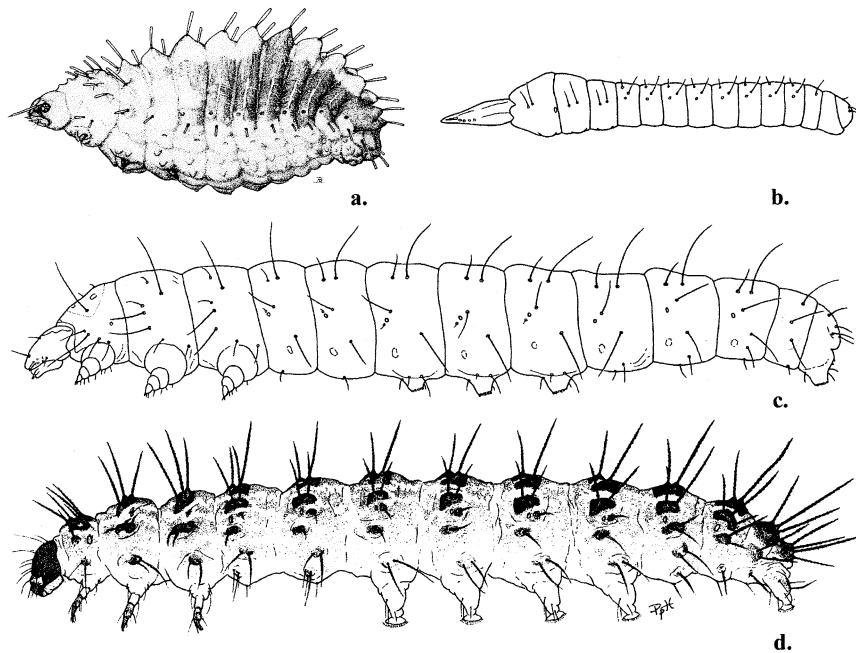


FIGURE 4 Moth larvae (caterpillars): (a) *Sabatinca* sp. (Micropterigidae). From (1999). *Lepidoptera, moths and butterflies*, Vol. 1. In *Handbook of Zoology*, Vol. IV, Part 35 (N. P. Kristensen, Ed.). Used with permission. (b) *Phyllonorycter blancardella* third instar (Gracillariidae). (c) *Phyllonorycter blancardella* fifth instar (Gracillariidae). (d) *Hypoprepia fucosa* (Arctiidae). Figures b, c, and d are from F. W. Stehr (1987). *Immature Insects*. Copyright 1987 by Kendall/Hunt Publishing Company. Used with permission.

basal lineage with appreciable species diversity. Three of the families are Gondwanan: Mnesarchaeidae (New Zealand, 14 species; 8 described), Prototheoridae (South Africa, 12 species), and Anomosetidae (Australia, 1 species). The largest family, the Hepialidae, with 587 species is cosmopolitan; the most generic diversity within the family also is Gondwanan. The remaining 37 superfamilies usually are placed into a single taxon, the Heteroneura—the name referring to the venational differences between the fore- and hindwing—the monophyly of which is still in question.

### C. The Heteroneurans

Phylogenetic relationships among many heteroneuran groups remain equivocal even in the face of detailed morphological and molecular studies. This uncertainty may be because these early evolutionary splits correspond to a period of rapid diversification within the order—the events occurring so closely together in time that it will be difficult to accurately reconstruct their sequence. There is weak morphological evidence sug-

gesting that four superfamilies (Nepticuloidea, Incurvarioidea, Paleaphatoidea, and Tischerioidea) form a monophyletic group, the Monotrysia. All share an unusual configuration of the female reproductive system, whereby there is a single gonopore for copulation and oviposition (and in the Incurvarioidea this is fused with the anus to form a cloaca). But other morphological and molecular data indicate that the characters holding this group together are symplesiomorphic (i.e., that they should be regarded as part of the ground plan for all heteroneurans). In Freidlander *et al.*'s recent molecular analyses employing gene sequence data from dopa decarboxylase, there was little support for the monophyly of the Monotrysia.

### D. The Ditrysia

The remaining 33 heteroneuran superfamilies are placed in the Ditrysia, a group whose monophyly is well supported by the presence of separate copulatory and ovipositional pores, an internal duct that connects the two, and the unique organization of the proboscis

musculature. This morphologically homogeneous clade contains 98.5% of the described species of Lepidoptera. Published phylogenies for ditrysian superfamilies and families are replete with uncertainty. In all likelihood considerable gene sequence data will have to be collected before consensus is reached about phylogenetic relationships within this remarkably successful group of insects.

Many guides and faunal works divide the order into the so-called Microlepidoptera and Macrolepidoptera. Microlepidoptera roughly corresponds to the basal groups, most of which are small to minute, possess a fused CuP vein in the hindwing, have the larval crochets arranged in a circle, and possess a jugum or the remnants of one (all of which are primitive features that were part of the ground plan for the order). It is a grouping of convenience as the "macros" are clearly derived from the "micros," and therefore some micros are more closely related to the macros than they are to other micros. The monophyly of the Macrolepidoptera, which includes the last 11 superfamilies in Table I, is still being debated: the crochets are arranged in a linear series parallel to the body axis, the first axillary sclerite in the wing base is elongate, and, obviously, the wingspans average consistently larger than those of the micros. Scott discussed a number of other characters that argued for the monophyly of the Macrolepidoptera, but exceptions are so numerous that some authors are reluctant to employ the term in their classifications.

### III. FOSSIL HISTORY

While the fossil record of the Trichoptera, the sister group to the Lepidoptera, is adequate and dates into the Permian, the record for lepidopterans is meager, especially for Mesozoic-aged fossils, when the basal lineages were diversifying. There are five putative Lepidopteran fossils from the Triassic; these cannot be assigned to the order with certainty. The oldest unequivocal moth, *Archeolepis mane*, dates to the Lower Jurassic. By definition, sister taxa are of equal age, and thus it is curious that the oldest caddisfly and moth fossils are of such disparate ages. Skalski has raised an interesting point—that some early stem group Lepidoptera may be unrecognized because currently they are assigned to the Trichoptera.

More than a dozen moth fossils from the Upper Jurassic and Cretaceous are unassignable to a family. While most of these clearly belong to preglossatan taxa, this is not true of all. *Protolepis cuprealata* from Upper

Jurassic deposits in Kazakhstan appears to have a coiled proboscis. This finding has special significance because phylogenetic reconstructions for the order have an angiosperm-feeding lineage (the Heterobathmiina) branching off prior to the evolution of the Glossata (Fig. 3). Because the oldest documented angiosperms are Lower Cretaceous (and *Protolepis* is Upper Jurassic), this suggests that either specialized angiosperm-feeding evolved at least twice within the Lepidoptera or that the date for the origin for angiosperms is, in fact, much older.

Micropterigids have been identified from several amber and sedimentary deposits. Examples from the lower Cretaceous represent extinct genera, while those from the Tertiary are assigned to modern genera. There are no fossil records for either the Aglossata or Heterobathmiina. Glossatans, including a putative lophocoronid, are well substantiated from the Upper Cretaceous. By 97 mya, modern genera of ditrysians are recognizable, suggesting that early evolution of the order occurred on nonseed plants and gymnosperms during the late Jurassic (Labandeira *et al.*, 1994). Not surprisingly, most of the modern superfamilial and species diversity appear after the angiosperms began radiating in the early Cretaceous. The oldest unambiguous macrolepidopteran fossils are Paleocene.

The most enigmatic macrolepidopteran fossil is the noctuid egg reported by Gall and Tiffney from the Upper Cretaceous (75 mya). Because noctuids are thought to be among the most derived ditrysian superfamilies (Fig. 3), their presence in the Cretaceous would push back the origin of virtually all lepidopteran higher taxa. Arguing against Gall and Tiffney's identification is the observation that all present-day noctuids have a thoracic "ear" sensitive to the ultrasonic frequencies emitted by bats and that the oldest bat fossils date back only to the Eocene. Perhaps then, it is not surprising that the oldest unequivocal noctuid fossils also date to the Eocene.

Leaf-mining families from the Northern Hemisphere have an extensive fossil record that is believed to reflect their ages of origin as well as their relative abundances. Labandeira *et al.*'s (1994) discoveries of mines of *Ecctoedemia* and *Stigmella* (both Nepticulidae) and a phyllocnistine (Gracillariidae) from Cenomanian deposits (97 mya), which agree in detail with those of modern species from closely related host taxa, are noteworthy in that they demonstrate the antiquity of many moth-plant associations. Another especially well-represented group of fossil moths is made up of the tineoids, which are richly represented in Dominican amber deposits that date from 15 to 40 mya.

#### IV. SPECIES DIVERSITY AND BIOGEOGRAPHY

There is still considerable doubt about the species richness within the order, an alarming observation given that Lepidoptera are among the most well-studied group of invertebrates. Two recent compilations, which are far from independent, give estimates of 145,677 (Table I) and 146,565 (Table II) for the world lepidopteran fauna. Heppner (1998) suggested that half again as many species remain to be described, and that species richness for the order would climb to 255,000 (Table II). Kristensen and Gaston have offered even higher estimates, in the range of 300,000 to 500,000 species.

The world's least speciose biogeographic realm, the Palearctic region, is the only one that can be regarded as well known. Heppner's tallies indicate that 92% of all the Lepidoptera that occur in the Palearctic have received names (Table II). The richest biogeographic realm, the Neotropics, with an estimated 35% of the world's butterfly and moth fauna, has only about 50% of its fauna named. Yet these figures are probably overly optimistic. In recent revisions of neotropical tineids or tortricids, Davis and Powell, respectively, found that 75 to 90% of the species were new. Of the more than 250 species of Gracillariidae that have been collected at the La Selva Biological Station in Costa Rica, 98% is believed to be undescribed.

There can be little doubt that microlepidopterans will be a frontier for descriptive taxonomy for decades if not centuries. In general, small cryptically colored taxa have greater proportions of unnamed species. Also, species-rich taxa that are intractable taxonomically have proportionately greater numbers unstudied. Microlepidopteran superfamilies with especially high numbers of undescribed species include the Gracillarioidea, Tineoidea, Gelechioidea, and Pyraloidea. But even among the Macrolepidoptera, considerable taxonomic work re-

mains, especially for neotropical taxa. The only two appreciably diverse moth families where the descriptive taxonomy is approaching completion are the Saturniidae and Sphingidae, veritable behemoths among insects that have long been favored among collectors.

Present tallies suggest that the Noctuoidea is without parallel in richness, with more than 41,300 described species, most of these falling within the nominate family (i.e., the Noctuidae, Table I). It is the most species rich superfamily on all continents except Australia, where its preeminence is superseded by the Gelechioidea. Current tallies of described species place the Geometroidea second in richness with some 21,740 species worldwide, nearly all of which are in the nominate family. Probably its ranking will fall to at least fourth once world microlepidopteran faunas are studied. Presently, there are approximately 15,500 described species of Gelechioidea and 16,000 Pyraloidea. The former likely contains in excess of 50,000 species and may well prove to be the most species-rich superfamily within the order; the latter contains at least double that of the described number. The butterflies, with 17,500 species, most of which have been given names, will rank no higher than fifth in global importance. In the most well-studied faunas (i.e., those of the north temperate zone and Australia) moths make up 93 to 95% of the lepidopteran species diversity, with butterflies making up the remainder. In terms of biomass or number of individuals, moths and their caterpillars probably have even greater importance, except in open and early successional habitats where butterflies thrive.

Because tropical faunas are very incompletely known, especially for microlepidopterans, a quantitative assessment of global diversity patterns for the entire order is not yet possible. Nevertheless, virtually all lepidopterists are in agreement that greatest species, generic, familial, and superfamilial diversity is found in

TABLE II  
Worldwide Lepidopteran Species Diversity by Biogeographic Region (1758–1990)

	Nearctic	Neotropical	Palaeartic	Ethiopian	Oriental	Australia	Total
Described species	11,532	44,791	22,465	20,491	27,683	19,603	146,565
(% of world fauna)	(7.9)	(30.6)	(15.3)	(14.0)	(18.8)	(13.4)	(100)
Estimated richness	14,000	90,000	25,000	38,000	50,000	38,000	255,000
(% of world fauna)	(5.5)	(35.3)	(9.8)	(14.9)	(19.6)	(14.9)	(100)
% of fauna known	82	51	92	51	53	50	57

From Heppner (1998).



tropical rain forests. Preliminary estimates of the moth fauna of Costa Rica (51,100 sq. km) run between 13,000 and 16,000 species and thus are roughly comparable to the species diversity of the entire North American continent north of Mexico or that of Australia, areas 380 and 150 times greater in size, respectively. Families with exceptional tropical richness include the Arctiidae, Cosmopterigidae, Gelechiidae, Gracillariidae, Oecophoridae, Lymantriidae, Notodontidae, and Sphingidae.

Given that moths are quintessentially phytophagous, their species diversity should parallel that of vascular plants. If this is correct, the highest global richness should occur on the low to mid-elevation slopes around the Amazon basin. The forests of Southeast Asia and Indo-Australia also would be expected to harbor hyperdiverse moth faunas. In temperate regions, the floras of South Africa and southwestern Australia are among the richest in the world. These latter areas are noteworthy because they support comparatively high percentages of nonditrysian moths and hence can be regarded as two of the world's most important "living museums." Obviously, not all taxa are hyperdiverse in the tropics. Familial, generic, and species diversity for the nonditrysian moth lineages is highest in extratropical latitudes. In temperate areas of the Northern Hemisphere, the Coleophoridae are among the most numerically important Lepidoptera in grassland, scrub, and desert ecosystems, yet the family appears to be absent in the rain forests of Central America. The Tortricoidea also appear to be less diverse than would be predicted based on the latitudinal richness gradients seen in other Lepidoptera.

On a local scale, Lawton has shown a correlation between herbivore richness and host plant architecture. Trees have more architectural complexity and therefore more herbivores than shrubs, shrubs more than herbs, and so on. A poignant example of the importance of trees is provided by white oak (*Quercus alba*)—across its range in eastern North America it hosts more than 400 species of moths (this number includes both specialist and generalist feeders). It follows that ecosystems with high tree diversity (e.g., lowland tropical rainforests) would have especially rich moth faunas.

Island faunas, especially those of distant islands, are always unbalanced with respect to those of the nearest continents. Macrolepidopterans that tend to be well represented on islands include the Arctiidae, Geometridae, Noctuidae (especially the Catocalinae, Heliotionae, and Noctuinae), and Sphingidae. Despite their small size, microlepidopterans are present even on remote islands, presumably because they disperse effectively as aerial plankton. Hawaii has been colonized by

dozens of different microlepidopteran lineages, and all can be characterized as weak fliers. Interestingly the richness of all leaf-mining families (minute to small moths) on Santa Cruz Island, off the coast of California, is nearly that of the mainland, while most groups of larger moths and other insects are represented by less than half of the mainland fauna.

## V. LIFE HISTORY

### A. Egg

The ova are variable in shape and especially surface ornamentation. Microlepidopterans often have flat, smooth eggs, many of which are nearly transparent once laid on a leaf surface; those of macrolepidopterans tend to be raised. The latter may be round, square, disk shaped, or spindle shaped; the chorion (outer surface) varies from smooth to richly ornate. Females generally deposit the egg within or on the surface of the host; in the latter case, secretions released from the accessory glands help bond the egg to the substrate. Females may leave behind abdominal scales, which discourage natural enemies or advertise to other females that the resource already has been found. A few polyphagous species (e.g., arctiids and hepialids) broadcast the eggs as they fly. The latter family also includes the world's most fecund lepidopteran—a single female *Abantiades hyalinatus* laid 30,000 eggs (and died with another 20,000 ova remaining in her abdomen).

### B. Larva

The larva (Figs. 4–9) is the principal feeding stage, increasing its mass by four (to five) orders of magnitude



FIGURE 5 Moth caterpillar defense (also see Figs. 6–9). *Isa textula* (Limacodidae). *Isa* caterpillars are armed with hundreds of stinging setae.

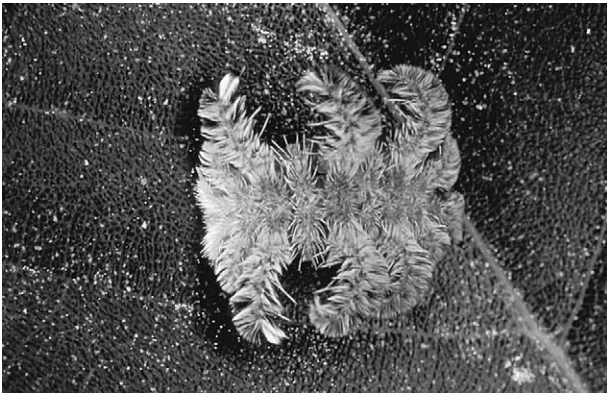


FIGURE 6 *Phobetron pithecium* (Limacodidae). *Phobetron* caterpillars are thought to mimic the cast “skins” of tarantulas, which contain thousands of deciduous urticating hairs—*Phobetron* caterpillars do not sting and are thought to be completely harmless.

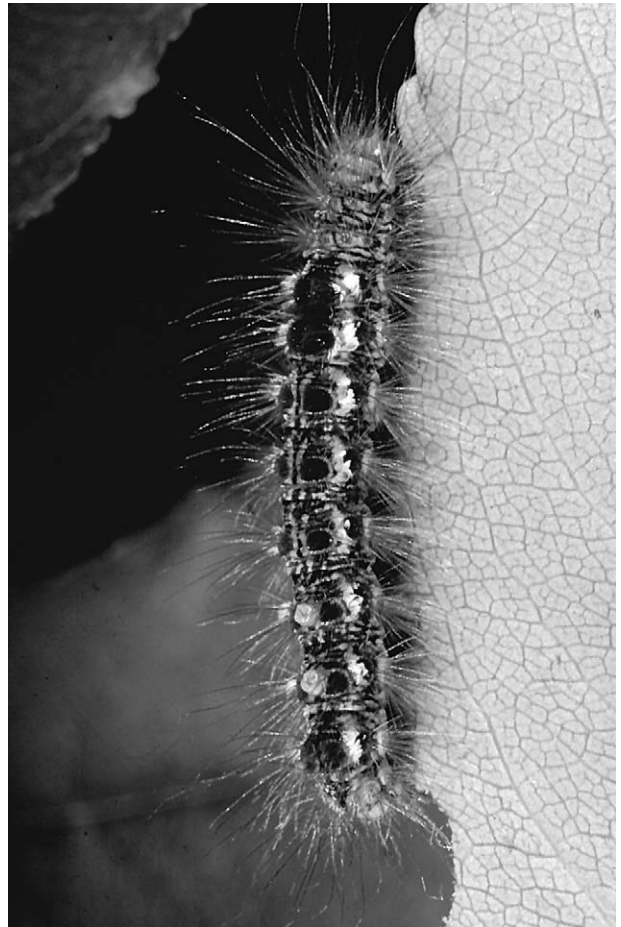


FIGURE 8 *Euproctis chrysorrhoea* (Lymantriidae). *Euproctis* caterpillars have red middorsal glands on the sixth and seventh abdominal segments that yield a *pot pourri* of defensive secretions—this caterpillar is responsible for at least two human deaths.



FIGURE 7 *Automeris* sp. (Saturniidae). *Automeris* caterpillars is fortified with a battery of urticating setae. The effect of which is not unlike that of stinging nettle.

during its development. There are from three to more than a dozen larval instars, with five being the mode. Dyar was the first to note that there is commonly a linear increase of 1.3 to 1.4 in head dimensions between successive instars. Dyar’s rule holds across the Lepidoptera and often is employed to infer the number of larval instars when only incomplete life history data are available. In bombycoids and a few other groups the females may pass through an additional instar. Development is remarkably fast in some species, especially those that feed on ephemeral tissues such as fruits and new leaves. Larger moths of high latitudes may take 2 or 3 years to mature. In *Gynaephora groenlandica*—a lymantriid found well north of the Arctic Circle—caterpillars mature over a 14-year period.

Hypermetamorphosis—where two or more larval



FIGURE 9 *Hemeroplanes* sp. (Sphingidae). Courtesy of Phill De Vries. *Hemeroplanes* caterpillars are superb snake mimics, despite the fact that mature larvae are but several centimeters in total length.

forms occur, each specialized for different functions—is rare within the order. The best-known example occurs in the minute to small leaf-mining moths of the family Gracillariidae. In basal gracillariid genera the first two or three instars are flat, legless, possess anteriorly directed mouthparts, lack a spinneret (silk-spinning organ) (Fig. 4b), and feed in a single cell layer, usually the epidermis. The diet is liquid, the larva filtering plant fluids from damaged cells. In later instars the larva takes on a rounded form that is legged, possesses a spinneret and ventrally directed mouthparts (Fig. 4c), and consumes whole cells and tissues. The spinneret is critical in that it is employed by the larva to lay down silk within the mine drawing it into a bubble, thereby creating a space where a rounded body and legs are appropriate. Metamorphosis in some *Bucculatrix*, epipyropids, cyclo-tornids, opostegids, and doubtless others is also hyper-metamorphic. Nonfeeding instars are rare, generally occurring in the first or last instars. An example of the

former is provided by some zygaenoids. In a number of gracillariids (e.g., *Cameraria* and *Chrysaster*) there may be as many as two nonfeeding prepupal instars. Whereas a general loss of color is common among prepupae, some caterpillars turn almost scarlet prior to pupation (e.g., some Prodoxidae, Gracillariidae, Elachistidae, and Notodontidae).

### C. Pupa

In the most basal lineages the wings, legs, and antennae are weakly fused to the main body of the pupa (Fig. 10). In several of these families, the pupal mandibles, which may be impressively large (Fig. 10a), are used to cut through the cocoon or work the pharate adult free from the pupal crypt. In heteroneurans the appendages are fused to the body and mandibles are very reduced and immobile (Fig. 10c). Another phylogenetic trend involves the degree of fusion between adjacent abdominal segments. In the most primitive moths the first seven abdominal segments are movable. Intermediate degrees of fusion occur among microlepidopterans. In the Obtectomera, a clade which includes 17 of the most derived lepidopteran superfamilies (Fig. 3), the first four abdominal segments are fused and are immobile with respect to one another. Taxa with appreciable abdominal mobility often extrude the distal end of the pupa from the cocoon or pupal crypt prior to emergence (e.g., most microlepidopterans). Those with a high degree of fusion hatch within the cocoon or pupal crypt and must crawl free before expanding their wings (e.g., all Obtectomera). Male pupae have a single set of paired genital marks on the venter of the ninth abdominal segment; females have an additional paired set of genital “scars” on the eighth segment as well (e.g., 10b). The pupal stage may be as short as 10 days or last several years, the latter is not unusual among species that inhabit arid environments.

Most moths spin a cocoon on or within the larval substrate or find their way to the ground and pupate in litter or duff. The cocoon may be a highly elaborate affair that takes days to spin. The most famous example is that of the Oriental Silk moth (*Bombyx mori*); the cocoon is fashioned from a single strand of silk that, when unwound, stretches for more than half a mile (0.8 km). A curiously elaborate cocoon is made by *Marmara* and related genera in the Gracillariidae: the cocoon is ornamented with up to 100 anally extruded bubbles, each of which is nearly the diameter of the caterpillar’s body. Cocoons represent a significant energetic investment in that silk is entirely protein, and the nitrogen that goes into its production is generally in short supply in plants. In addition to silk, the cocoon

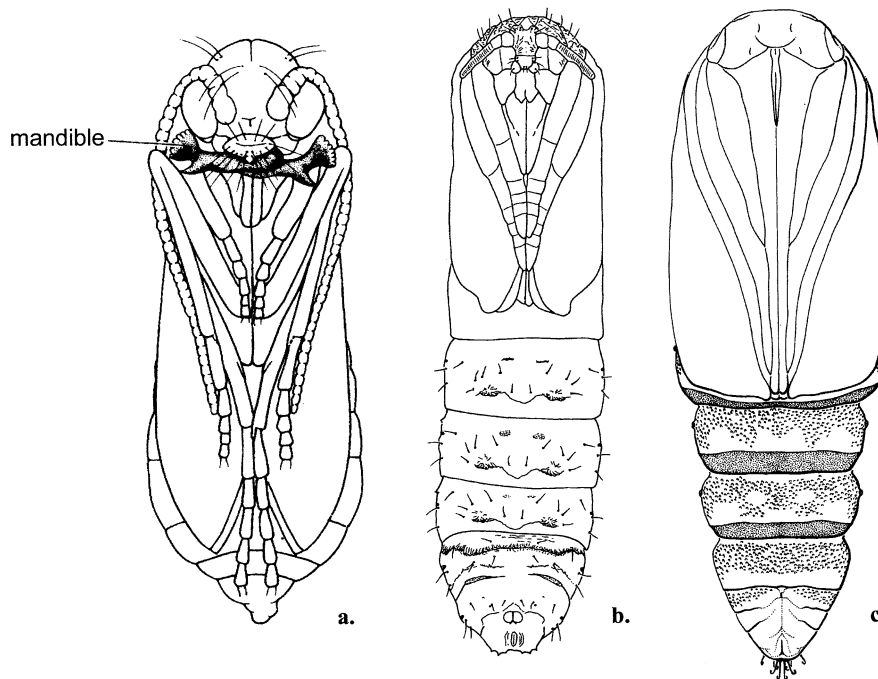


FIGURE 10 Moth pupae. (a) *Agathiphaga* sp. (Agathiphagidae). From Lepidoptera, moths and butterflies, Vol. 1. In *Handbook of Zoology*, Vol. IV, Part 35 (1999). (N. P. Kristensen, Ed.). Used with permission. (b) *Sthenopsis auratus* (Hepialidae). Reprinted with permission from *Journal of the New York Entomological Society* 97, 7. (c) *Asota* (Noctuidae). From W. De Gruyter (1999). Lepidoptera, moths and butterflies, Vol. 1. In *Handbook of Zoology*, Vol. IV, Part 35 (N. P. Kristensen, Ed.). Used with permission. Across this phylogenetic progression there is reduction in the mandibles, increased fusion of appendages, simplification in the external expression of the appendages, and decreased abdominal mobility.

may incorporate excrement, scrapings from the larval substrate, or larval setae (especially if these are stiff, poisonous, or urticating). Prepupal limacodids discharge a calcium oxalate solution over the inner surface of the cocoon that hardens the cocoon as it dries. Open, netlike cocoons are often spun by species with aposematic pupae. Taxa that pupate in concealed sites such as below ground or in wood often forego spinning a cocoon. A few moths (e.g., elachistids, pterophorids, sterrhine geometrids, and hedyliids) pupate naked, exposed on surfaces, like butterflies.

#### D. Diapause and Quiescence

Diapause may occur in any of the four life stages. The largest fraction of temperate moths pass through the winter months in diapause as prepupal larvae or pupae; the second largest fraction overwinter as eggs. Moths that overwinter as adults invariably need to feed over the five to seven months of winter—some of these are among the unwelcome guests in the buckets used to

collect maple syrup. Reproductive diapause is the rule among moths that overwinter as adults, ovarian maturation mating and being delayed until the arrival of spring. Summer diapause and aestivation occurs in many groups that live in habitats with a pronounced dry season. Enormous aggregations involving tens of thousands of moths (usually noctuids) are known to assemble in caves or on (hill) mountain tops. Moths in these aggregations become important sources of protein for vertebrates, even for grizzly bears. The diapause longevity record belongs to the desert-inhabiting prodoxids—single larval cohorts may yield adults over a period of 3 decades.

## VI. FEEDING BIOLOGY

### A. Larvae

Given the enormity of the order it is remarkable how little diversity there is in feeding habits. More than 99% of all moths are fundamentally herbivorous. Within this

realm, however, they have radiated to exploit virtually every corner of niche space. Terrestrial algae (including blue-greens), lichens, fungi, bryophytes, pteridophytes, and seed plants are eaten. All plant tissues and organs are consumed, but especially leaves, meristems, and reproductive structures. The majority require living tissues, sometimes of a very specific age, while others subsist on dead and fallen leaves, fruits, seeds, and wood. Nymphuline pyralids, which number more than 700 species, are aquatic as larvae, with some even adapted for life in rapidly flowing water. Other lineages have members that are subaquatic or bore into aquatic plants.

The most basal lepidopteran lineages have diets that include ancient plant lineages. Among the Micropterigidae, there are many species that include bryophytes in their diet. Larvae of agathiphagids—representing the order's second most ancient lineage—feed within the seeds of *Agathis* (Araucariaceae), a family that traces back to the Jurassic. Heterobathmiid larvae are leaf miners in new foliage of *Nothofagus* (Fagaceae), an archaic Gondwanan angiosperm. Basal families of the Glossata are strictly associated with angiosperms, hence the great species, generic, and even familial radiations within the Lepidoptera were reliant on the diversification of the angiosperms.

Few of the basal lineages account for appreciable species diversity. The first radiation of note, as measured by extant richness, was that of the Hepialidae, with 587 species worldwide. Their success may be due to their diet, which is arguably the most diverse of any moth family. Larvae eat bryophytes, pteridophytes, conifers, and hardwoods; leaf and especially woody root and stem tissues; leaf litter and fungi, especially in early instars. Many are cannibalistic, and some even attack and consume soft-bodied arthropods. Another early lineage with remarkably catholic diets are the tineoid families, a high percentage of which are fungivores and detritivores. Members of the Tineidae (e.g., the clothes moths) are noteworthy for their ability to metabolize keratin, the normally undigestible protein that makes up fur and feathers. Others feed in bird and bat guano, dung, and even the dead insects that accumulate below spider webs.

As the Lepidoptera radiated so did their host associations and feeding habits. The majority of microlepidopterans that followed are concealed feeders, tunneling into plant tissues, forming leaf mines or shelters, or feeding from within silken cases. Galls are made by some members of more than a dozen families of Lepidoptera, but none of the lineages can be regarded as particularly successful. Diets tend to be specialized in

internal feeders because the host is not only the food but also the environment for some or all of the immature stages. The most intimate host associations occur in taxa where the adults and larvae utilize the same host plant, the most famous example being that of the yucca moths in the genera *Tegeticula* and *Parategeticula* (Prodoxidae). The female yucca moth uses special maxillary “tentacles” to collect yucca pollen and then flies to a second flower where she scrapes the pollen over the stigmatic surface. She then deposits one to a few eggs in the developing ovary, and in so doing, leaves behind a marking pheromone, which can be detected by other females; she then collects more pollen and flies to another flower. The larvae are seed predators but do not consume all the seeds in a fruit. The mutualism is obligate, as neither moth nor yucca can survive without the other.

Leaf miners, minute moths whose larvae feed between the two surfaces of a leaf, account for the most significant early radiations. The Nepticulidae, whose fossil history traces back more than 97 million years to the Cenomanian, contains more than 800 species with at least this number remaining to be described. The Gracillariidae, with 2000 recognized species and perhaps three times as many unrecognized, also traces its fossil history back to the Cenomanian. The latter family contains members that are among the first shelter-forming Lepidoptera. Early instar gracillariines mine in leaves as do other gracillariids, but middle instars may exit the mine to form a leaf shelter in which they complete their development.

The bulk of the microlepidopterans, including the three largest superfamilies (the Gelechioidea, Tortricoidea, and Pyraloidea) are concealed feeders, mostly leaf rollers or shelter formers, but others are borers, leaf miners, gall formers, and detritivores. A handful are predators, inquilines in galls or nests of social insects, and so on. Leaf rollers and shelter formers exhibit intermediate levels of host plant specialization, but polyphagy is not common. An obvious exception to this generalization is provided by many tortricids, especially members of the nominate subfamily. Several microlepidopteran families have representatives that are external feeders for one or more instars (e.g., the Bucculatricidae, Schreckensteinidae, Pterophoridae, and others), but none accounts for appreciable present-day diversity. The largest group of externally feeding microlepidopterans are the zygaenoids, which as a group appear to be well protected—their arsenal includes urticating spines; abundant, long, or dense setae and hairs; and cyanoglucosides (Figs. 5 and 6).

Approximately 60% of the described Lepidoptera are

macrolepidopterans, which are principally external feeders. Host plant associations are diverse—some macrolepidopterans are exceptionally specialized while others are widely polyphagous. Polyphagy is far more common among external feeders and is the mode for many macrolepidopteran families (e.g., Lasiocampidae, Saturniidae, ennomine Geometridae, Lymantriidae, and several subfamilies of the Noctuidae). It is worth noting that many species that are host plant generalists across their geographic range specialize on just one or a few plants at any one locality. Regardless of the mode of feeding, internal or external, caterpillars that feed on highly poisonous plants tend to be host plant specialists.

Fungivory appears throughout the order but is especially common in the Hepialidae and Tineoidea. The line between fungivory and feeding on fallen plant material is often an arbitrary one. Not surprisingly, both of these groups are also detritivores. Other sizable lepidopteran clades, with members specializing on fallen leaves, decaying wood, and other nonliving plant tissues, include the herminiine Noctuidae, Blastobasidae, Oecophoridae, and a few pyralid subfamilies. In dry oak forests of the eastern United States, herminiines may account for more than half of all macrolepidopterans in light trap collections during the month of July. Australia has an exceptionally rich fauna of litter-feeding Oecophoridae and Tortricidae. There, numbers are so great in eucalyptus forests that the larvae are believed to play a significant role in litter processing. The sloth moths of the neotropics are an oddity. Adults of three chrysaugine pyralid genera are found on or about sloths, awaiting the once-weekly release of dung, which serves as the substrate for the developing larvae.

The most common type of carnivory in the order is cannibalism, often showing up in taxa where the host substrate is limited (e.g., annual plants and architecturally reduced perennials) or where movement between hosts is either impossible or exposes larvae to considerable risks (e.g., internal borers). Essentially phytophagous species that occasionally eat other caterpillars occur sporadically across the order. The Epipyropidae and Cyclotornidae are obligate ectoparasites of homopterans. The latter is only parasitic in the first instar, as the second to final instars are brood predators in ant nests. Only about 200 lepidopterans are regarded as predaceous, many of which are butterflies (see De Vries this volume). Nearly all predatory caterpillars feed on sessile or nearly sessile prey, such as the brood of hymenopterans or aphids, scale insects, and related homopterans. Hawaii has spawned the most exceptional predaceous caterpillars—members of the genus *Eupithecia* (Geometridae) are sit-and-wait predators that snatch small in-

sects that are so unwise as to crawl over the caterpillar's body.

## B. Adults

Micropterigids are mandibulate and use their jaws to grind up pollen and fern spores. A crude proboscis, formed from the maxillae in adult *Dacnonypha*, is used to collect fluids from bark and other surfaces where the adults are perching. These fluids may include dissolved honeydew, but adults do not visit flowers for nectar. Nectar feeding is first seen in the *Incurvariidae* and is retained by virtually all the larger lineages of moths that follow. In addition to nectar and other sources of dissolved sugars, moths use the proboscis to imbibe water, liquids from injured plants and tree flows, dissolved salts (including human perspiration and lacrymal secretions), pus, animal excreta, the crushed bodies of other insects, rotting fruit, dung, carrion, many other substrates. A few anomalous noctuids use the proboscis to pierce fruits; related species have even taken to blood feeding on large mammals. In both cases the tongue is short, stout, and the two sides of the proboscis are able to slide back and forth against each other as the tongue is worked into the fruit or flesh.

A favorite collecting technique for moths is “sugaring.” A standard bait is made by mixing fruit, beer, and sugar. While overripe bananas seem to be the most universal ingredient, collectors sometimes add watermelon, apricot, grapes, rum, jellies, molasses, and a rather long list of other substances, some of which might best be left unnamed. Once the mixture has been allowed to ferment it is then applied to tree trunks or offered on sponges. Virtually all of the winter-active moths (e.g., Oecophoridae, Tortricidae, and Noctuidae) feed at bait. During summer months, many moths in dry woodlands also come to bait. Presumably, moths that are attracted to bait normally would be feeding at tree wounds, sap flows, rotting fruit, or the honey dew secreted by homopterans.

Plant tissues tend to be low in sodium relative to the metabolic needs of animals. Many lepidopterans reach adulthood with inadequate supplies of this essential element, but they are able to acquire it by imbibing fluids from mud puddles and water courses, urine, sweat, lacrymal secretions, and so on. Although “puddling” is well known among butterflies, it also occurs in a diverse array of diurnal and nocturnal moths. In most moth and butterfly species, 90 to 100% of the individuals that puddle (or feed at other sodium-rich substrates) are males. A remarkable case is provided by *Gluphisia septentrionis* (Notodontidae). Puddling males

will drink (and forcibly excrete) more than 600 times their own body weight in a single evening. *Gluphisia* males transfer sodium to females with their ejaculate; females in turn pass much of the sodium to their offspring.

Many adult moths lack functional mouthparts or digestive systems, including many (or all) Arctiidae, Bombycidae, Hepialidae, Lymantriidae, Notodontidae, Saturniidae, and Zygaenoidea. Adults of late fall and early spring geometrids tend to be non-feeding, whereas those that fly during the summer commonly consume nectar.

## VII. NATURAL ENEMIES AND CHEMICAL DEFENSE

### A. Natural Enemies

Moths are fecund invertebrates that produce 100 to 10,000 times more ova than will ultimately survive. A panoply of biotic and abiotic mortality factors come into play during every life stage to limit population growth. Biotic factors include an enormous range of pathogens, parasites, and predators. Population numbers of many pest species, especially in forests and other stable communities, often are controlled by viruses, bacteria, and fungi. Viruses are often highly specific, and one wonders if there are not nearly as many lepidopteran viruses as there are moths and butterflies. There are comparatively few recorded species of bacteria and fungi that attack insects, but the taxonomy of insect pathogens, especially those that cannot be easily cultured, is grossly understudied; while some are highly specific, others are thought to have broad host ranges. *Bacillus thuringiensis*, the most widely used pathogen for the control of Lepidoptera in gardens, agriculture, and forests, has a host range that includes hundreds of species across many superfamilies. Fungal pathogens, mostly in the Entomophthorales and Fungi Imperfecti, often require special conditions such as a period of high humidity before appreciable infection will occur. A species of special note is *Entomophaga maimaiga*. This fungal pathogen was introduced into North America from Japan in 1910 and 1911 as a biological control agent to stem the ever-increasing populations of the European gypsy moth (*Lymantria dispar*). The introduction was considered unsuccessful until 1989 when the fungus began showing up in caterpillar cadavers throughout the Northeast. It has since spread throughout much of the range of the gypsy moth and continues to be a principal mortality agent for the gypsy moth

caterpillar. A diverse assemblage of nematodes and nematomorphans also attacks Lepidoptera, especially the larvae, but their importance is generally regarded as modest. Microsporidians affect both fertility and longevity in many moths.

Lepidoptera are primary hosts for a multitude of parasitoids. Three of the largest families on this planet owe much of their evolutionary success to the Lepidoptera: the Braconidae (>10,000 species), Ichneumonidae (>10,000 species), and Tachinidae (>20,000 species). Another 20 or so families of insect parasitoids attack Lepidoptera, all but two of which (Diptera: Bombyliidae and Sarcophagidae) are parasitic Hymenoptera, mostly in the Chalcidoidea and Proctotrupeoidea. All four life stages are exploited by parasitoids, although most are specialized to attack the larva. Some of the smallest insects, trichogrammatid and mymarid hymenopterans, with body lengths of only 0.2 mm, are egg parasites. The great majority of insect parasitoids are specialists that utilize but one or a few related species. Large, long-lived caterpillars support rich guilds of parasitoids: across eastern North America, 29 fly and wasp parasitoids are recorded from *Hyalophora cecropia*, at least 3 of which are believed to be specialists on this saturniid and its congeners. Such caterpillars may be at the base of an entire food web, as evidently there are six hymenopteran hyperparasitoids that parasitize the parasitoids of *Hyalophora*, and one wasp hyperparasitoid that principally attacks the hyperparasitoids. Askew and others have noted that leaf miners and gall-forming caterpillars have an exceptionally rich parasitoid fauna. *Phyllonorycter apparella* (Gracillariidae) has a wingspan of only 7 mm, yet larvae and pupae of this leaf miner are known to host 20 species of hymenopteran parasitoids and hyperparasitoids.

Invertebrate predators of major importance include mites, spiders, predaceous stink bugs, beetles, robber flies, ants, and wasps. Many of the physical, chemical, and behavioral defenses of caterpillars are known to be effective in thwarting the pillages of ants. No doubt, these defenses are of special importance in tropical ecosystems where ants make up much of the insect biomass. Mammals, birds, reptiles, and amphibians are the principal vertebrate predators. Birds are such an important selection pressure for externally feeding caterpillars that there is a growing consensus that they have shaped not only what many caterpillars look like, but also what, when, and how they feed. Their influence on adult color patterns and behavior also seems to be preeminent. Bats, too, have been an exceptionally important evolutionary force in the diversification of the order. Hearing organs, sensitive to the frequencies emit-

ted by bats, have evolved independently at least five times (discussed later); within one of these groups, the Arctiidae, some chemically protected taxa have evolved the ability to signal back to bats. Upon detecting ultrasound frequencies, these arctiids begin clicking, presumably advertising their whereabouts (and toxicity).

## B. Defensive Chemistry

The defense strategies of caterpillars and adults are legion and make for fascinating reading and study (see Stamp and Casey, 1993, and Scoble, 1992, for larvae and adults, respectively). Chemical defense against parasitoids and predators likely has played a prominent role in the diversification of the order. The ecological and evolutionary constraints facing unpalatable species differ widely from those faced by palatable taxa. For example, chemically protected caterpillars are often brightly colored, feed day or night, and rarely take measures to conceal their feeding damage; many are gregarious. Conversely, palatable species are apt to be cryptic in color, to be nocturnal, to move slowly, and to only rarely leave conspicuous signs of feeding. Across the order there is a spectrum of palatability, ranging from those species that can be eaten in large numbers (by birds or indigenous peoples) to those where contact or ingestion can be fatal. There is an enormous literature on this subject—this discussion and examples presented here are only meant to be illustrative.

Stinging setae, hollow and filled with poison, occur widely in the Limacodidae, Megalopygidae, and Hemi-leucinae (Saturniidae). In rare instances, human fatality has been reported after encounters with the latter two. The body setae of *Euproctis chrysorrhoea* (Lymantriidae) (Fig. 8) contain histamine, protease, lipase, and phospholipase, and are thought to be responsible for at least two human deaths. Gypsy moth caterpillars and other lymantriids drag the body setae through the dorsal abdominal glands, “arming” them with a histamine and a diverse mixture of noxious substances that cause itching and other allergic responses, which can be severe in some individuals. Other taxa with large numbers of species with urticating setae include the Lasiocampidae, Thaumetopoeinae (Notodontidae), and Arctiidae.

In some taxa, caterpillars and adults synthesize noxious compounds from simple dietary precursors. Zygaenid larvae produce cyanoglucosides from two basic amino acids. Chemical protection in many Notodontidae also appears to be based on simple manufactured compounds (e.g., the late instars of many Heterocampi-

nae eject formic acid and ketones from a cervical gland if molested).

Other moths sequester toxins or precursors from their (poisonous) host plants; toxin concentrations in the hemolymph may greatly exceed those that occur in any of the plant tissues on which the caterpillar has fed. Examples include *Euphorbia*-feeding caterpillars, which retain dietary terpenols, and the milkweed caterpillar (*Euchaetes egle*) (Arctiidae), which, like the monarch butterfly caterpillar, sequesters high concentrations of a cardenolide. Other arctiids actively seek out and consume host plants high in pyrrolizidine alkaloids (PAs). Experimental studies have demonstrated the effectiveness of PAs in repelling both vertebrate and invertebrate predators as well as their importance in courtship, mating, and paternal investment (discussed later). If bright coloration can be used as a proxy for chemical protection, there is still a universe of protected moth species whose natural chemistry remains unstudied and unappreciated.

In many species the chemicals sequestered or manufactured by the caterpillars are retained by the adult. PAs collected by male *Utetheisa bella* (Arctiidae) caterpillars are used to synthesize the male courtship pheromone. Eisner and his students found that female *Utetheisa* (Fig. 11) raised on diets devoid of PAs are readily consumed by spiders, but females that have mated with males laden with PAs are cut out and dropped from webs. The eggs and even young caterpillars, resulting from the latter type of pairing are also chemically protected by the alkaloids collected by the father. Arctiids are not completely dependent on the larval diet because adults of some species are able to locate and ingest

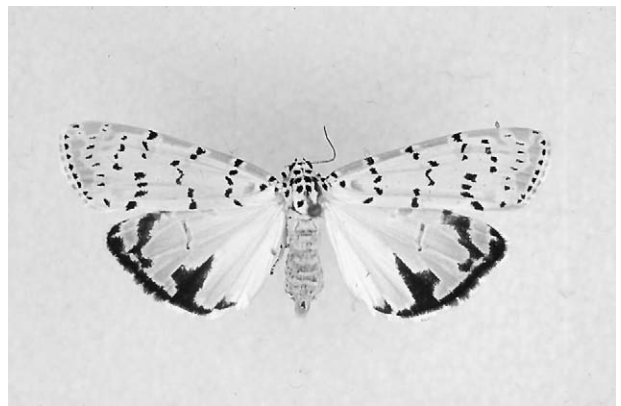


FIGURE 11 Tiger moths (Arctiidae) (also see Figs. 12 and 13). *Utetheisa bella*; in addition to sperm, the male transfers alkaloids during mating that not only protects his mate subsequently, but also her offspring.





FIGURE 12 *Grammia virguncula*; adults of many tiger moths bubble out alkaloid-laden hemolymph from their thoraxes when disturbed.

dissolved PAs. They will imbibe fluids from the surfaces or wounds of PA-containing plants (e.g., Asteraceae, Boraginaceae, and Fabaceae) or even the crushed bodies of alkaloid-rich caterpillars or moths. Many arctiids bubble out a defensive secretion from the prothorax upon disturbance. In *Grammia*, the fluid is mostly clear and green and may represent nothing more than hemolymph (Fig. 12); in others it is more frothy, sometimes milky, and loaded with acetylcholine, histamine, and possibly pyrazines. Adult female lymantriids collect larval hairs from the cocoon and incorporate these into their egg mass. Not all adults are able to retain the protection enjoyed by their larvae (e.g., adult saturniids and limacodids are generally palatable). Both caterpillars and adults of chemically protected species tend to be aposematic (Fig. 13). Adults are often diurnal and may be appreciably slower fliers than their unprotected



FIGURE 13 *Platyrepia virginalis*; chemically protected moths are often boldly colored and, like this species, diurnal.

sister lineage. Not a small number are members of Müllerian mimicry complexes.

## VIII. DISPERSAL AND MIGRATION

Long-distance movements occur in many groups, not an insignificant number of which are pest species. One of the most common patterns is for populations to cycle through the winter in tropical and subtropical latitudes, then to move north (or south) during the spring and summer. Some migrants move on the leading edge of storm fronts, where both high wind and high humidity facilitate long-distance dispersal. Occasionally their numbers may be so great that they show up on radar screens. Swallows, swifts, and other insectivorous birds follow these fronts as well, gleaning insects as they go. A second common pattern is for moths to move between wet and dry forests, with adults trickling out of the latter over the course of the dry season then reinvading with the return of the rainy season. The dazzlingly beautiful sunset moths (Uraniidae) are famous for their mass diurnal migrations in Africa and Latin America.

## IX. KEY ADAPTATIONS

One can only speculate as to why there are so many moths. One morphological feature that stands out as universal is the presence of scales over the wings and body. Those covering the wings are easily abraded, so much so that the wings feel slippery. Regardless of the context in which scales evolved, there can be little doubt that they provide considerable protection from natural enemies: adults can fly into a spider web and, if they are lucky, they will fly out, leaving behind a hundred or a thousand scales still attached to the web. Scales allow the wings to slide past one another and often out of the bill or jaws of a would-be predator.

The coilable siphon or proboscis has allowed Lepidoptera to exploit liquid diets. More than any other feature, this structure has been tied intricately to the coevolutionary history of moths with flowers. Nearly all insect-pollinated flowers with deep corolla tubes (or nectar spurs) are pollinated by Lepidoptera. Charles Darwin knew this, and upon being told of a Madagascar flower with a 25 cm corolla, predicted that a sphingid hawkmoth would one day be found on the island with a tongue of at least this length—and it was. The tongue also allows the collection of dietary constituents that are scarce in plants, such as nitrogen and sodium (noted earlier). No nonfeeding moths are known to migrate,

suggesting that the presence of a tongue is a critical attribute for long-distance movement. Similarly all temperate moths that overwinter as an adult have a tongue.

A tympanum or analogous hearing structure, sensitive to the ultrasonic frequencies used by hunting bats, has evolved independently in no less than five moth lineages (Scoble, 1992, and Fig. 12). It may be no coincidence that three of these (the Geometroidea, Noctuoidea, and Pyraloidea) are among the most species-rich moth clades. Fenton and Fullard estimated that 85% of all macrolepidopterans possesses some type of hearing organ. Noctuids that detect ultrasound first attempt to fly away from the source, but if the hunting frequencies are sensed at close range, these moths either drop to the ground or initiate a highly erratic flight while spiraling downward. Further testimony as to the importance of the ear is provided by the mites that infest the tympanal cavities of certain noctuids (especially *Leucania*) (Fig. 14). Treat found that mites placed onto *Leucania* always move into one ear, leaving the other fully operational.

Silk has played a major role in the evolutionary success of the order. Its most universal use is in the formation of the cocoon or in the attachment of the chrysalis to the pupation substrate. Most microlepidoptera use silk to form a shelter or to line a chamber within the larval feeding substrate. Many caterpillars lay down silk whenever they walk, which presumably allows them to quickly retrace their step, for example, in times of danger. Case making, where the caterpillar constructs a portable silken case, has evolved in no less than a dozen moth groups; two especially successful lineages include the Psychidae and Coleophoridae (both families are appreciably more species rich than their non-case-making sister taxon). Some gregarious caterpillars spin con-



FIGURE 14 “Ear” of *Leucania* (Noctuidae) located immediately below hindwing; courtesy of Asher Treat.

spicuous nests. Not only do these nests offer considerable protection from natural enemies, but they can be important in thermoregulation, heating up well above ambient temperatures on cool days. Many species, especially among those taxa with flightless females (e.g., some Psychidae, Geometridae, and Lymantriidae), disperse by “ballooning.” Early instars drop from silken lines and wait for winds to blow them about. Among microlepidopterans, it is common to see prepupal larvae dropping from trees on silk lines on their way into leaf litter where pupation will occur. At night many geometrid caterpillars drop from their perches, suspended on a silken thread, and dangle for hours before climbing back onto foliage, always before daylight. In a similar fashion many caterpillars, if disturbed, drop from their perch on a silk belay and crawl back up after the danger has evidently past. Processionary caterpillars (Notodontidae: Thaumtopinae) and tent worms (Lasiocampidae) lay down a silk trail, impregnated with pheromone, that siblings follow. Some microlepidoptera (e.g., Bucculatricidae) form a molting cocoon within which the vulnerable transition between instars takes place. Prior to molting, many species lay down a light sheet of silk into which the crochets are engaged; in so doing the new caterpillar can crawl more easily out of its old integument.

One behavioral trait that stands out is the ability of Lepidoptera to exploit the hours of darkness. In many ecosystems moths are the most diverse and numerically important group of nocturnal flying insects. Perhaps 90% of all Lepidoptera are night active. Associated with their nocturnal habits is the order’s remarkable elaboration of morphological, neurophysiological, and behavioral traits that facilitate pheromonal communication (discussed later).

Their success may also just be a matter of being in the right place at the right time. Because the order was essentially phytophagous and beginning to diversify on nonseed plants and gymnosperms prior to the origin of angiosperms, it was well poised to ride along with the explosive radiation of flowering plants that was to follow. Indeed, the moths are the largest monophyletic lineage to have evolved with plants.

## X. PHEROMONES

While visual communication is thought to be of modest importance in nocturnal moths, chemical communication is highly developed and perhaps universally important. The first animal pheromone was identified from the Chinese Silk Moth (*Bombyx mori*) and was appropri-

ately dubbed bombykol. In nearly all nocturnal moths, virgin females release a pheromone that attracts males from distances of up to several hundred meters (reports of up to a mile are likely exaggerated) (Fig. 15). Males move upwind, in and out of the pheromone plume, until they arrive at its source. The male antenna and associated neurophysiology make up one of the most sensitive chemical detection systems known in nature. Not surprisingly, males often have more elaborate antennae than females, sometimes with secondary or tertiary branches that increase the surface area several fold (e.g., many bombycoids, Fig. 16). One interesting exception, and one that remains unexplained, are *Ar-rhenophanes* females (Tineioidea)—the female's antenna is considerably more complex than that of the male.

Female sex pheromones usually are released from glands on the abdomen, especially the intersegmental regions between abdominal segments seven and eight or eight and nine. The complete pheromone contains an isomeric “cocktail” of two to several components, usually with one or two major components that are responsible for most of the male attraction. The individ-



FIGURE 15 *Hemileuca maia* (Saturniidae); unmated females expose an abdominal gland—here, the small structure at the tip of the abdomen, extended nearly perpendicular to the body axis—when “calling” for males.



FIGURE 16 *Antheraea polyphemus* (Saturniidae); male saturniids often have strikingly plumose antennae, which contain thousands of setae that capture and respond to the female sex pheromone. Courtesy of Alexander Klotz.

ual components of the pheromone are often odorless (at the concentrations produced by moths), straight chain hydrocarbon backbone, with one or more double bonds, and an active moiety (e.g., an acetate, aldehyde, epoxide, ketone, or alcohol). Closely related species tend to have different (enantiomeric or isomeric) blends, minor components, or “calling” times.

Once the two sexes are in proximity, male courtship pheromones may come into play. The structures by which males deploy their pheromones are as varied as any structures in the order. In the simplest case they are individual scales peppered over the wings or other parts of the body. Male sex scales (or androconia) that are grouped into tufts, brushes, or “pencils” occur over almost any part of the body including the wings, legs, mouthparts, and antennae. They are especially common about the genitalia, where they provide final signals prior to coupling (Fig. 17). Remarkably complex scent organs occur in a number of arctiids, hepialids, and noctuids (Fig. 18). A hallmark of androconia is that their abundance and development may differ markedly among closely related species, which suggests that they are under strong sexual selection.

Male pheromones are thought to serve in species recognition and isolation and to promote female acceptance. In general they are used at close range in the vicinity of the female. Role reversal, with males calling females from long ranges, is very rare but does occur (e.g., see Fig. 18). Chemically, male pheromones are exceptionally heterogeneous, ranging from simple molecules to long-chain hydrocarbons and ring compounds. Many have a detectable and often pleasant odor. The male pheromone blend of the oriental fruit moth contains both (*Z*)-methyl epijasmionate and (*E*)-



FIGURE 17 Male scent-releasing structures (also see Fig. 18). *Trichoplusia ni* (Noctuidae); end view of abdominal brush that encircles the male genitalia. Courtesy of Ken Haines.

ethyl cinnamate; the former is the active constituent in the odor of jasmine, a common ingredient in women's perfumes. Many male scents are based on secondary plant compounds that are sequestered by the larva or



FIGURE 18 *Estigmene acrea* (Arctiidae); “calling” male with fully extended coremata (eversible sacs) lined with scent scales. Courtesy of Mark Willis. Reversed calling, with males broadcasting the sex pheromone, such as exhibited here by *Estigmene acrea*, is very rare among moths.

collected by the adult. In some arctiids, the female is reluctant to copulate with males that lack appreciable quantities of pyrolizidine alkaloids (PAs) on their androconia.

Beyond mate location, courtship, and mating, pheromones are also known to serve in a number of lesser roles. Some female moths (e.g., yucca moths) leave behind a “marking” pheromone at the time of oviposition that alerts other females that a resource already has been located. The aggregation behavior displayed by some arctiids and noctuids is likely mediated by a pheromonal signal. During mating, some male Lepidoptera transfer an “anti-aphrodisiac” pheromone to the female that, at least temporarily, repels other would-be suitors.

## XI. COLORATION, DIURNALITY, NOCTURNALITY, AND ATTRACTION TO LIGHT

As a general rule, moths that are active principally at night are subtly colored in earth tones, especially browns, grays, black, and white. The patterning in nocturnal taxa is often variable, presumably because details of the coloration are under relaxed selection pressure. But exceptions to both of these statements abound. Most notably, there are dozens of lineages with brightly colored hindwings that are covered by the forewings—the bright coloration being exposed as a flash or warning signal after a moth has been disturbed. Bright coloration is associated with moths that are diurnal or both diurnal and nocturnal. Aposematic and highly reflective (e.g., blue, silver, and gold) coloration are associated with diurnal taxa.

Assuming nocturnality as the ground plan for the order, or at least the Glossata, it is evident that diurnality has evolved repeatedly, perhaps hundreds of times within the order. The most conspicuous example is that of the butterflies, which are, from an evolutionary sense, nothing more than a group (or two) of diurnal moths. Taking the Geometridae as an example, there are dozens of species and genera that are diurnal; yet there are essentially no species-rich clades (genera, tribes, subfamilies) within the family that are active principally during the day. Diurnality is common among moths that live in alpine or arctic habitats or that fly during the coldest months of the year.

From a phylogenetic perspective, diurnality is well distributed across the order. Members of the most basal family, the Micropterigidae, are both diurnal and nocturnal in habit, with some genera exhibiting splendidly

metallic golds and reflective purples. These same bright colors are found in a host of other primitive families (e.g., Heterobathmiidae, Eriocraniidae, and Incurvariidae). Like in the Micropterigidae, adults may be active both day and night. Across the Microlepidoptera diurnality is widespread; essentially diurnal lineages include the Adelidae, Choreutidae, Heliodinidae, Heliozelidae, Scythrididae, and Sesiidae. Many moths are crepuscular, some flying at both dawn and dusk and others limiting their activity to one of these two periods of rapidly changing light intensities.

Nearly all nocturnal moths can be collected at light. Indeed, light traps are routinely employed to conduct moth inventories, to monitor populations of agricultural pests, and to collect gravid females for life history studies. The most commonly asked question about moths is "Why are they attracted to light?" Although it is difficult to answer this question with certainty, the moon-compass explanation has great appeal. If a moth had a need to fly in a straight line at night it could do so by maintaining a constant angle to any celestial light source. When this same behavior (constant orientation to a distant light source) is applied when a light is proximate, a moth will spiral into a nearby light source (Fig. 19). Perhaps the greater mystery is why only a fraction of the moths present in a given area are attracted. Anyone who runs flight interception traps, collects at bait, or knows what caterpillars are feeding in the garden appreciates the fact that the individuals attracted to light represent but a small fraction of the whole. Sex ratios in light-collected samples are often heavily biased toward males. Hepialids in the genus *Sthenopsis* provide a counterexample as females predominate in museum collections. There are many moths known from but one sex—the sex that is attracted to light.

Moths are not entirely dependent on celestial navigation. There is ample evidence that moths also possess a geomagnetic compass that allows them to move in a directed fashion, even on moonless nights and under overcast skies.

## XII. IMPORTANCE

Lepidoptera, in particular moths, are among the most important forest and agricultural pests. There is almost no plant tissue that is not eaten by one or more species of moths. Most of the forest pests are of cyclical occurrence with natural enemies eventually controlling population numbers. Introduced species offer an obvious exception to this generalization—their populations

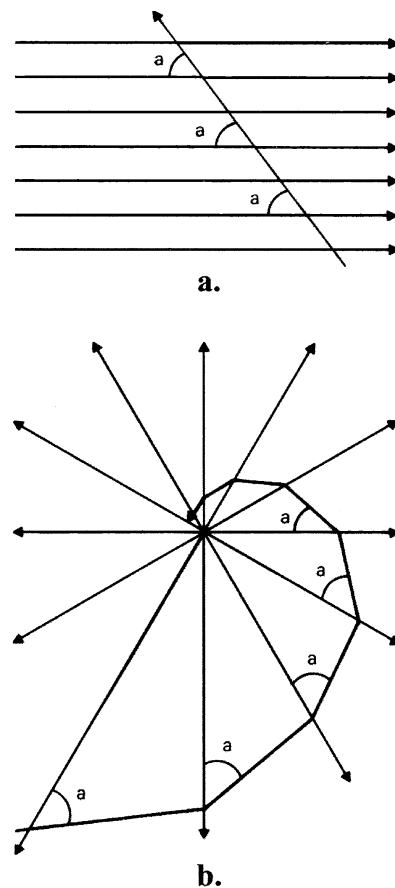


FIGURE 19 A moth can proceed in a straight line by moving at a constant angle ( $a$ ) to a distant light source. However, if the light is proximate, (b) this same orientation behavior will cause the moth to spiral into the source.

may go unchecked for many years before natural enemies can be introduced or recruited from the native fauna. Agricultural pests tend to be more chronic and, as noted earlier, many are migrants preadapted to be crop pests (because they are highly dispersive insects that locate resources quickly). Moreover, crop pests often have short life cycles that allow them to complete one or more generations before natural enemies can build to appreciable numbers. A few, such as the Indian Meal Moth (*Plodia interpunctella*) (Pyralidae), are stored product pests that attack stored grains, cereals, bird seed, dry dog food, and even candy that has sat too long in the pantry. Clothes moths (*Tinea* species) (Tineidae) are well known for their spoils of woolens.

Most moths should be regarded as beneficial principally for their role in terrestrial food webs, especially those that include birds, bats, and rodents. Many songbird nestlings are dependent almost entirely on caterpil-

lars. With the exception of large, bat-pollinated flowers, most white-flowered plants are pollinated by moths under cover of night. Scents from moth pollinated flowers may be exceptionally fragrant (e.g., honeysuckle, jasmine, and some orchids). Sphingidae are the primary pollinators for about 10% of the tree species in Costa Rica's seasonal dry forests. These large, strong fliers are especially important in tropical forests where individuals of the same species can be widely spaced. Many moth caterpillars feed on leaf litter, decaying wood, and fungi and thus play at least a minor role in litter decomposition and nutrient cycling. The most famous product derived from a moth is silk. This fiber, with a tensile strength greater than that of steel, is still made the same way it has been for more than 4600 years by boiling the cocoons of *Bombyx mori* and unraveling them a single thread at a time. Moths have been used successfully as biological control agents of invasive weeds. Australia had lost 50 million acres of pastureland to introduced prickly pear cactus (*Opuntia*) before the pyralid, *Cactoblastis cactorum*, was introduced from southern South America. Within a decade of its release, most large stands of the cactus had been utterly destroyed. The moth also has been effective in controlling invasive *Opuntia* in Hawaii, Mauritius, South Africa and the West Indies. This story recently has taken an odd and unfortunate twist, as this same moth has found its way into the Florida Keys, where it now threatens stands of three rare species of *Opuntia*.

### XIII. CONSERVATION

Across much of Europe and the United States lepidopteran faunas are being increasingly employed as bioindicators by state, provincial, and federal agencies and nongovernmental conservation organizations to evaluate natural areas and management regimes. Moths offer many advantages as bioindicators: many macrolepidopterans are easily identified and their habitat requirements are reasonably well known. By any standard they are hyperdiverse and therefore provide considerable data relative to the amount of effort that goes into their sampling and sorting. Part of their value as bioindicators is that they represent another trophic level—two grasslands that are botanically equivalent would have different moth faunas if one were being managed too intensively (e.g., by fire). One major disadvantage of moths as bioindicators is that samples often include many transients. For example, more than one-third of the species of moths that occur in Denmark have been collected in a light trap atop the Zoological Museum

on the University of Copenhagen campus at the periphery of the city.

The principal threat to moths is the same as for other terrestrial invertebrates—habitat destruction and fragmentation—and nowhere is this more true than for tropical forests where moth diversity is thought to be highest. In other ecosystems the obverse, reforestation and succession, pose threats to moths. Fire suppression in grasslands, chaparral, barrens, scrub forests, and other fire-maintained ecosystems has been detrimental to many local invertebrate populations. And the reverse, ill-timed and too aggressive prescribed burning, can be detrimental. There has been much discussion about the impacts of artificial lighting on moth populations, much of it focusing on the decline of saturniids and other large moths. Up to this point the published commentary has been anecdotal, but there can be little doubt that street lamps and other high-intensity lamps have had some impact on moth populations—within days, bats learn to forage at them by night, and they are among the first places that birds visit at dawn.

Introduced or alien species can be a threat at many levels. Invasive plants can displace native hosts upon which either the larvae or adults of native moths depend. In exceptional cases, introduced plants can provide strong ovipositional cues for species whose larvae either fail entirely or experience high levels of mortality. Because introduced moths often arrive without natural enemies, their numbers can cause widespread overexploitation of host plants. It is hard to imagine that the widespread defoliation by the gypsy moth (*Lymantria dispar*) has not been responsible for at least local impacts. Top-down threats from the release of introduced predators, pathogens, and parasitoids for biological control efforts merit special attention. The tachinid, *Compsilura cocinnata*, was introduced from Europe into eastern North America to control the gypsy moth. The parasitoid's host range includes more than 200 native species. The moth has one generation a year, the fly three to four. It has been estimated that a single acre of a gypsy moth infested forest will yield up to 10,000 adult *Compsilura* in June, which then must seek out alternate (native) hosts.

Although DDT and other pesticides have been implicated in the local extirpation of Lepidoptera, the evidence is invariably circumstantial. While suppression activities (single applications) would rarely be expected to result in local extinctions, eradication efforts (multiple applications, especially in a single season) certainly impact many moths and the insectivorous animals that are dependent on them. The most widely used biological pesticide to control Lepidoptera is *Bacillus thuringiensis*

var. *kurstaki* (Btk)—a bacterium that either kills a caterpillar outright or leads to a systematic infection that results in death. Btk is used by home owners to control garden pests, by farmers in agricultural systems, and by government and industry to combat forest defoliators. Product labels indicate that Btk is effective against a taxonomic array of Lepidoptera (and Coleoptera). Susceptibility to Btk ranges from taxa that are not affected to those that are killed weeks after a single application; saturniids and papilionids (swallowtail butterflies) are among the groups reported to be especially susceptible. From one perspective Btk is a great control agent because it affects few nonlepidopterans and has no mammalian or avian toxicity. Yet from a lepidopterists' perspective it is hard to view Btk in such a positive light. The use of Btk and other pesticides in natural ecosystems is most likely to be a problem over critical habitats that are isolated, restricted, or in some other way imperiled, especially when these occur within a matrix that contains a primary pest. For example, the glades, balds, or barrens communities that occur in an otherwise forested landscape are apt to suffer from the aerial pesticide applications targeted for forest pests.

With the exception of the Chinese silk moth, which no longer exists in the wild, there is no evidence that collectors have ever represented a significant risk to moths. Collecting is likely to represent a serious threat only to highly imperiled and geographically restricted taxa. Where population sizes are very small, the use of light traps that employ a killing agent should be avoided. A far greater threat is undercollecting, as it seems probable that less than half of the world's moth fauna has been described. And even in those regions

where most of the basic taxonomic work has been completed, it is not until we have reliable occurrence data on a broad geographic scale that conservation biologists will know with certainty what species are slipping toward extinction.

### See Also the Following Articles

BUTTERFLIES • INSECTS, OVERVIEW • INVERTEBRATES, TERRESTRIAL, OVERVIEW • PARASITOIDS • PESTICIDES, USE AND EFFECTS OF

### Bibliography

- Common, I. F. B. (1990). *The Moths of Australia*. E. J. Brill and Melbourne University Press, Melbourne.
- Heppner, J. B. (1998). *Classification of the Lepidoptera. Part 1. Introduction*. Holarctic Lepidoptera. Volume 5, Supplement 1.
- Kristensen, N. P. (1984). Studies on the morphology and systematics of primitive Lepidoptera (Insecta). *Steenstrupia* 10, 141–191.
- Kristensen, N. P. (Ed.) (1999). Lepidoptera, moths and butterflies. Volume 1. Evolution, systematics, and biogeography. Part 35. Arthropoda: Insecta. Volume IV. In *Handbook of Zoology* Walter de Gruyter, Berlin.
- Labandeira, C. C., Dilcher, D. L., Davis, D. R., and Wagner, D. L. (1994). Ninety-seven million years of angiosperm-insect association: Paleobiological insights into the meaning of coevolution. *Proceedings of the National Academy of Sciences* 91, 12278–12282.
- Scoble, M. J. (1992). *The Lepidoptera. Form, Function, and Diversity*. Oxford University Press, Oxford.
- Stamp, N. E., and T. M. Casey (Eds.) (1993). *Caterpillars: Ecological and Evolutionary Constraints on Foraging*. Chapman Hall, New York.
- Stehr, F. (Ed.) (1987). *Immature Insects. Volume 1*. Kendall/Hunt, Dubuque, IA.
- Young, M. (1996). *The Natural History of Moths*. Academic Press, New York.



# MUSEUMS AND INSTITUTIONS, ROLE OF

Paul Henderson and Neil Chalmers  
*The Natural History Museum*

---

- I. Collections
  - II. Research and Biodiversity
  - III. Exhibitions
  - IV. Education and Training
  - V. Publications and Outreach
  - VI. Advisory
- 

## GLOSSARY

**bioprospecting** The search for commercially valuable biochemical and genetic resources in plants, animals, and microorganisms.

**descriptive taxonomy** The description of new species usually involving written accounts, with illustrations, of the characteristics of a specimen.

**parataxonomist** A layperson who has received training in practical basic biology, ecology, and taxonomy, as well as in the collection and preparation of biological specimens, so that he or she can undertake a specific part of a biodiversity inventory.

**revisionary taxonomy** The reevaluation of entire groups of organisms based on new and old evidence.

---

**THE COLLECTIONS HOUSED** in natural history museums and similar institutions are essential to research, education, and a better appreciation of the natural world and its diversity. Since it is estimated that only approxi-

mately 15% of the species alive today have been described and named, taxonomic and related research is still essential if we are to understand, conserve, and manage biodiversity properly. Furthermore, increasing awareness of the value of biodiversity has led some countries to establish institutions with the express aim of making inventories of their biodiversity and to become the agents for the information that arises from those surveys. The research and collections of museums are applied to a wide range of issues, including human and animal health, bioprospecting, environmental quality, resource management, and the implementation of legislation. Museums are also active contributors to a wide range of educational objectives, including those incorporating biodiversity. This is achieved through exhibitions, training, field studies, and outreach programs. New technology, such as the Internet, is enabling museums to order and make readily available the vast amounts of information on biodiversity they already hold.

The role of museums and related institutions is to conserve and develop the collections in their charge; to improve public understanding through exhibitions and other means; to undertake research involving the collections; and to make their information, expertise, and objects as widely accessible as possible.

## I. COLLECTIONS

Collections are the *raison d'être* of natural history museums and similar institutions. They form the prime basis



for the museums' research, education, exhibition, and advisory work. For the larger national and equivalent museums, the collections of animals, plants, fossils, and sometimes rocks and minerals have been assembled over more than two centuries. They are therefore a major source of information on biodiversity in both space and time. They also reflect our knowledge of the natural world and something of the processes that we have used—past and present—to describe and understand it.

### A. Content

It is estimated that the world's 6500 natural history museums contain approximately 3 billion specimens covering the animal, plant, and mineral kingdoms, excluding microorganisms. Add to these the untold numbers of specimens (living and dead) in the numerous botanical gardens, arboreta, zoos, and institutions with culture collections and it is clear that the resource is huge. To this must also be added the collections of books, monographs, research papers, and artwork amassed over many years by the libraries of these institutions. They form an invaluable part of the collections and represent, in part, the developing knowledge and perceptions of biodiversity. The particular nature of the collections in each museum or institution will depend on its function and a host of other factors. Two European examples (Paris and London) showing the broad makeup of their collections are given in Fig. 1.

In most museums the collections are maintained by curators or collection managers whose task is to ensure that the specimens are well conserved and available for users. Collection management and conservation is a skilled job often requiring considerable patience and a sound taxonomic knowledge of the organisms. As interest in biodiversity increases, so does the use of related collections, and this brings associated threats principally to the specimens themselves. Conservation practice must therefore be proactive so as to allow increased access to the material while developing improvement in its care. Conservation is thus seen as the employment of best practice to prevent or arrest the long-term physical deterioration of natural history specimens and associated artifacts and documents to preserve their scientific and cultural worth.

The methods used in conservation are the subject of global interest and conferences on the topic are now being held; the first major one was the International Symposium and First World Congress on the Preservation and Conservation of Natural History Collections, which was held in Madrid in 1993. A significant litera-

ture now exists. One topic of particular interest is the preservation of an organism's DNA in museum specimens. The evidence suggests that most current practices are not particularly deleterious but additional research is needed.

The collections of any institution are a manifestation of its research, collecting, and acquisition policies possibly over a considerable period of time. The British Museum, established in 1753 (and as such is one of the earliest museums to house natural history and other objects), developed from the initial collections of Sir Hans Sloane who had acquired specimens from many parts of the world. The natural history part of this collection grew by donation, purchase, and active collection by its staff over the succeeding decades to the extent that accommodating it became a major problem. As a result, the Natural History Museum was opened in 1851 in South Kensington. The collections continued to grow and have become one of the world's most renowned, along with those in Washington, DC, Paris, New York, and Berlin. This long history of collection development involving major periods of exploration in different parts of the world—on land and at sea—has led to this museum having specimens from numerous countries and waters, collected over a wide time span. It is perhaps no surprise to find numerous specimens of extinct species as well as those that have been crucial to the development of ideas on evolution (such as the finches collected by Darwin from the Galapagos Islands) and other concepts. Other museums have acquired their own identity through complementary histories leading to distinct collections. No two museums have very similar holdings of specimens.

Culture collections are of increasing importance and require special storage and conservation approaches. Such collections are usually housed in institutions specially established for this purpose. An example is the American Type Culture Collection (ATCC) comprising mainly cultures of bacteria and fungi. Others may be very specific, such as the Provasoli–Guillard National Center for the Culture of Marine Phytoplankton, also in the United States. A principal function of these centers is to distribute cultures to the research community. The ATCC is also working to set up local culture collections in developing countries.

## B. The Uses and Role of Collections

### 1. Taxonomic Studies

For most specimens in the collections, the question "what exactly is it?" usually has to be answered. The

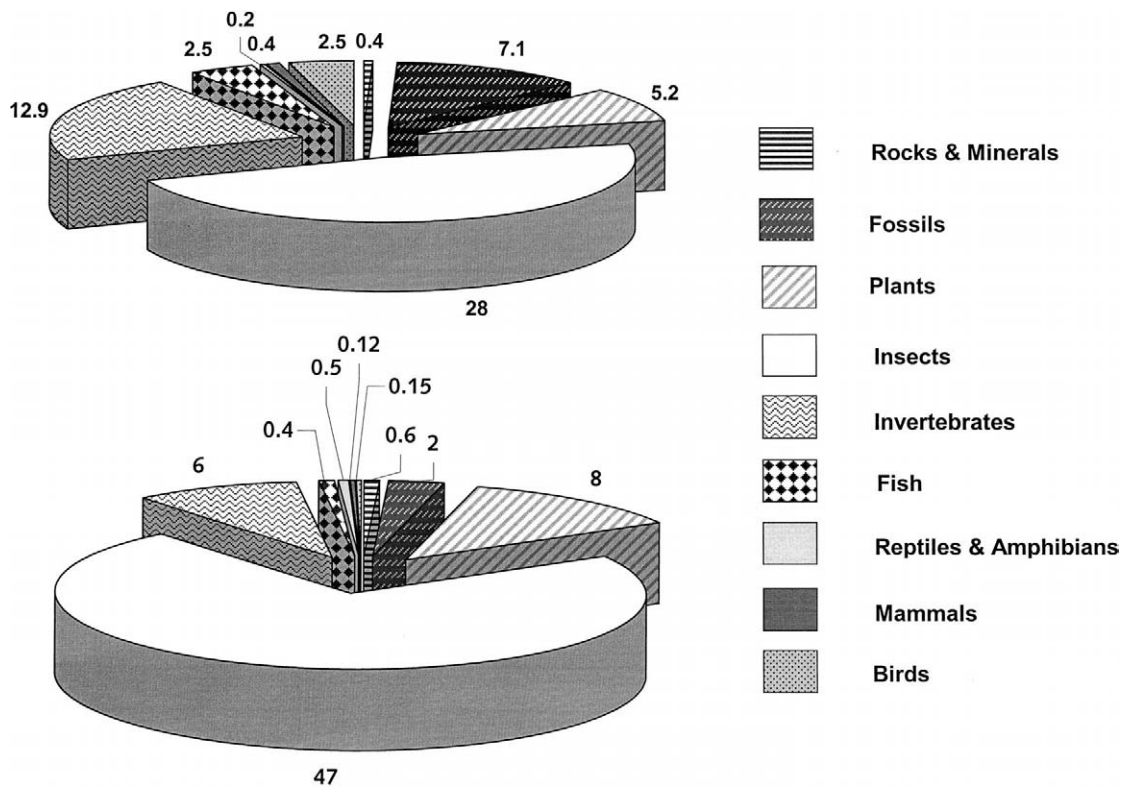


FIGURE 1 Diagram showing the number of specimens (in millions) of various groups of animals, minerals, and plants in the collections of the Natural History Museum, London (top), and the Museum National d'Histoire Naturelle, Paris (bottom).

initial information usually sought is that which is intrinsic to the specimen itself, i.e., properties such as shape (or morphology), color, molecular characteristics (e.g., DNA), and sometimes its behavior. With such information the specimen can be named and placed within a classification. This process is the key to virtually all the other uses of the collections: Without a proper basic knowledge of the material, other studies and applications would probably be worthless.

For accurate taxonomic nomenclature “type” specimens are essential. When formally naming a species, it is normal practice to select a particular preserved specimen (the type) to act as a “name bearer.” It is given a unique and permanent name. If, subsequently, any doubt arises about how the given name should be used, then identification of the type specimen will show which name belongs to which recognized species. As the taxonomic system changes because of new discoveries, reexamination of type specimens is often essential for accurate use of names, and this need may arise repeatedly over decades and centuries. Since type specimens are very important, they are usually deposited in

museums or other public institutions. It follows that there is a major obligation for such institutions to conserve them and make them available for study by the international scientific community.

Systematists, by examining the intrinsic properties of different species, work to establish the relationships between organisms. Multiple specimen samples can provide information on variation within species, on geographical distribution, and on changes over time. If other relevant information is available, such as the habitat from which the specimen was collected, then other findings are clearly possible. The collections are therefore curated and developed with these needs in mind.

## 2. Biological and Related Research

Collections are an essential resource for a range of biological research activities, including evolution, genetics, ecology, and epidemiology. Some of this research is carried out in museums and related institutions (see Section II), but much is also done elsewhere. The museums, while sometimes providing the material, usually

act primarily as a resource for verifying the status of organisms on which research has been done. Without this backing, the findings can lack credibility or repeatability.

Bioprospecting can only be undertaken on a sound basis if the species are accurately identified. Previously collected museum specimens are perhaps rarely used directly in such work, especially when it must be destructive, but the collections and the taxonomists' expertise are often vital for verification and for planning field-collecting programs. Moreover, cladistic and other predictive classifications, usually based in whole or part on comparison of museum specimens, can provide helpful guidance for selecting taxa for bioprospecting. They help in searching for taxa that are likely to have similar properties to those of valuable species already discovered.

### 3. Environmental Conditions

Specimens collected at different times and in different places can be of significant value in documenting changes in environmental conditions, including those resulting from human activity. Examples include documenting the presence of DDT in different areas and times, documenting the presence of radioactivity from weapons testing or environmental pollution from radioactive waste, and assessing the spread of invasive species.

### 4. Inventories and Conservation

Management of biodiversity, in almost any sense of the term, requires knowledge of the species (or at least the genera) in the region of interest. For example, conservation work requires an inventory of those organisms that are of relevance. It is sometimes argued with good reason that a requirement to conserving biodiversity is to discover, identify (and possibly describe), and inventory the species in a region or locality. The museums can and do provide the resources for such work, but these are far too little for them to have a sufficient impact on their own. It is thus necessary for others (often local people) to be trained and given the tools for the task. For this, a local or regional collection of what are often called "voucher specimens" may be established. These are correctly identified specimens that serve as a reference for the fieldworkers in identifying the specimens that they have collected. National and international publications (such as the International Union for the Conservation of Nature's *Red Lists of Threatened Animals and Plants*) giving information on endangered species are also based on information stemming initially from collections.

Since collections have often been developed over extended periods of time (some for two or more centuries), it follows that some of the information associated with a specimen, such as date and place of collecting, is very useful for establishing past distributions and their changes. Unfortunately, although many collections have these data, a significant amount of work may be necessary to extract the data from labels or registers.

### 5. Implementation of Legislation

Many countries have legislation in place to help protect against pests and diseases, whether they originate from or via organisms that are domestic or imported. Legislation prevents the growing or rearing of certain species. Some countries do not allow the export of certain endemic or other species for a variety of reasons, including the concern that others might benefit commercially from them in some way but without any benefit accruing to the country of origin. Implementing this kind of legislation requires people, such as custom officers, to either have or be able to call on the appropriate expertise or identification facility. The tools for such jobs (e.g., guides and charts) are based on the work and collections of museums and other institutions. Some cases are relatively straightforward. One well-known case involved efforts to prevent the introduction of the Colorado potato beetle into the United Kingdom. Posters showing this bright and distinctive beetle and warning of its potential impact were displayed at points of entry into the United Kingdom for many years. Most other cases, however, are much more complex and require regular contact between customs and other authorities and museums scientists or similar specialists.

Some international treaties and conventions require the resources of collections and associated expertise for them to be enforced or implemented. The Convention on International Trade in Endangered Species (CITES) is one example which involves the collaboration and cooperative working of museums, botanic gardens, etc.

### 6. Resource Management

The numerous related commercial activities of farming, cultivation, breeding, and fishing involve the use of collections through the need for pest identification or even occasionally for the identification of the specimens collected (e.g., fishing). The exploitation of resources through activities such as mining or oil exploration and extraction entail impact studies in which the affected land or water is monitored before and after for its biodiversity. Correct identification of organisms is essential for understanding these impacts and any programs for possible remedial work and restitution.

## II. RESEARCH AND BIODIVERSITY

A core research activity for natural history museums, in addition to other institutions such as botanical gardens, zoos, and aquaria where these have research staff, is the pursuit of systematics and taxonomy. This research is effectively aimed at the fundamental units of biodiversity, namely, species, by describing and naming them and understanding their relationships. Much of this research refines what we know and also charts the unknown. Phylogenetic studies, in addition to giving important insight into evolution and its processes, help to provide a sound and strongly predictive framework for inferring unknown properties of taxonomic groups.

The systematic research carried out in museums is done at different levels. The important base level is descriptive taxonomy (sometimes called "alpha-taxonomy"), which involves the description of new species and is heavily reliant on their morphological characteristics. Revisionary taxonomy is a higher level involving studies on the relationships of species and it can involve the establishment of phylogenies. Other levels of systematic study can be concerned with life cycles, ecology, and genetic aspects.

It is variously estimated that we have named and described only approximately 10–15% of the species alive today. The huge task of describing the rest falls for the most part to taxonomists in museums and related institutions. Such institutions, however, are expected to increase the breadth of their research to embrace ecology, conservation, and theoretical studies. These pressures arise because of the demand for information on a wide range of environmental and resource issues at a time when there is a much greater extinction rate of species (recently estimated to be approximately 10,000 times greater) relative to the normal rate over geological time. The research is quite labor-intensive and requires relevant skills, many of which are acquired through years of experience. Modern methodologies and technologies, through approaches such as molecular biology and electronic pattern recognition, are helpful, but the task and its importance remain enormous. It is thus becoming increasingly important to establish priorities for applying the research resource.

Different institutions have responded in different ways to this imperative. Several museums are organizing their research around broad themes or programs that address significant current and often global issues. By focusing different types of expertise onto specific and timely problems, greater impact can be achieved while still providing some of the basic information for

more generic uses. Some authors advocate that priority should be given to groups of organisms that are ecologically and economically important (Raven *et al.*, 1998), including those that we know relatively little about such as bacteria, fungi, and nematodes. However, despite international recognition of the need for good priority setting, there has been little discussion or agreement on how this should be done, primarily because of the current shortage of taxonomists and financial and other constraints placed on the relevant institutions.

Some institutions, however, do have a very specific research agenda, especially those concerned with charting the biodiversity of their own country. Indeed, a few institutions have been set up with this express purpose, as private or government agencies charged with the management of biodiversity information. Principal examples include the Environmental Resources Information Network in Australia (<http://kaos/erin.gov.au/psg/erin/index.html>), the National Commission for the Knowledge and Use of Biodiversity (La Comision Nacional Para El Conocimiento Y Uso de la Biodiversidad) in Mexico (<http://www.conabio.gob.mx/>), and the National Biodiversity Institute (Instituto Nacional de Biodiversidad; INBio) in Costa Rica.

INBio is different from the other examples in that it holds collections and has wider functions. Its mission is to promote a new awareness of the value of biodiversity and thereby achieve its conservation and use to improve the quality of life. Its functions involve making inventories of biodiversity, including those concerned with bioprospecting, and administering and disseminating biodiversity information. It works in collaboration with other research centers in Costa Rica to obtain corroboration of inventory information and to help conduct chemical processing in prospecting. It was established in 1989 and has since developed a significant staff ranging from researchers to trained parataxonomists. Its four principal objectives are

- To assume responsibility for developing and executing a national biodiversity inventory
- Where feasible, to locate national collections within one physical space and under one administration
- To centralize biodiversity information now and in the future
- To put this information into easily accessible and user-friendly formats available to a wide variety of users and to promote its use by Costa Rican society

Further information is available from its web site (<http://www.inbio.ac.cr/>).

Research in the form of inventories can provide sig-

nificant benefits in addition to those accruing from bioprospecting and the development of new drugs. These include a better knowledge of pests, invasive species, and disease vectors and the control of all of these. These in turn lead to a more sustainable agriculture, better resource management, and improvements in human health.

The type of inventory can vary significantly. Commonly, this will involve collecting examples of all members of a particular set of organisms, such as plants, in a particular area using the expertise of both trained taxonomists and parataxonomists. Perhaps more rarely they will be an All Taxa Biological Inventory in which most, if not all, organisms in an area are identified, together with information on ecological relationships. However, inventories in any form usually require considerable resources, especially manpower, if they are to achieve a significant degree of usefulness. In this regard, this activity involves long-term research goals.

Research on systematic biology carried out in museums also has other uses:

1. In medicine through studies involving disease vectors and pathogens
2. In the pharmaceutical industry, especially through bioprospecting (see Section I,B,2)
3. In fisheries and agriculture, including studies of pests and pathogens
4. In environmental quality assessments involving water and air pollution and the effects of food production and human and animal health
5. In appreciation of our environment, including raising awareness and knowledge of biodiversity through good-quality guides and keys

The research agenda sometimes includes work directed toward conservation of biodiversity by modeling and manipulating data on the distribution of species. The aim is to give assessments of species richness (i.e., the number of species in a given area), species turnover, habitat occurrences, and other parameters that are useful in planning programs with strategic conservation objectives.

### III. EXHIBITIONS

For modern museums of natural history and equivalent institutions, biodiversity provides a new and integrating focus to their exhibition program. In many institutions, there is a move away from displays on particular taxonomic groups toward exhibitions encompassing

ecology, environmental change, resource exploitation, conservation, and related topics. Concepts are being emphasized, which is an approach that helps to aid education and understanding. Issues are also being addressed, such as the strongly deleterious effects of certain human activities on the environment and on biodiversity. Increasingly, members of the visiting public are being given the opportunity to find answers to their specific questions and to see how they might help in conserving biodiversity and to understand some of the implications of its loss.

The approach to some exhibitions on biodiversity is to demonstrate the nature and variety of the animal and plant kingdoms. These tend to be based on what we know. Few museums deal with the problems of our limited knowledge of living species and habitats, and how we are endeavoring to overcome them, despite the recognition by the systematics profession and others that both our understanding and our fundamental knowledge base are relatively poor.

The scale and remit of a museum will, of course, determine the broad nature of the exhibits. Local and regional museums tend to show the biodiversity of their area and address particular educational needs for their community. They may be able to stage activities, including field days or courses, on particular aspects. These can provide an important educational and social contribution and help raise awareness of the scope and processes of local natural history and some of the associated threats and opportunities. Larger museums are less likely to focus on particular areas, although some describe the biodiversity, in selected ways, of their country or large area.

Even allowing for the fact that museums are less likely to emphasize taxonomic groupings, the approaches to exhibitions can be very different from one museum to another. Most attempt to engage visitors and will strive to meet their needs. The Grande Galerie de l'Evolution (part of the Museum National d'Histoire Naturelle) in Paris achieves this by the scale and overall visual design of the displays. Here, for example, on the topic of the African savanna is a memorable display of mammals parading as if in some type of Serengeti migration. (For such displays, the skill of the taxidermist is still paramount.) The displays, covering approximately 6000 m<sup>2</sup>, emphasize the diversity of life in different marine and terrestrial environments as well as cover the principles of evolution and the role of mankind.

In contrast, the Life Galleries at The Natural History Museum in London have several themes with a focus on particular processes or concepts, whereas others are

devoted to particular animals and their habitats. A gallery on the subject of ecology impresses because of its architectural novelty, its use of modern methods of communication, and its integration of different topics, including the role of mankind. Another gallery deals with the principles of evolution. Others cover particular animal groups, such as the arthropods (in a gallery titled "Creepy Crawlies"), marine invertebrates, and mammals. In these galleries, attention is paid not only to highlighting descriptions of some of the representative organisms but also to habitats and behavior.

A related approach is adopted in the Zoological Museum of Copenhagen, Denmark (founded more than 350 years ago), at which a major display titled "From Pole to Pole" deals with animals in specific habitats, including the tundra, the Antarctic, the "bird cliff," the temperate forests, the desert, and the rain forest. Another display on animal life in the oceans covers subjects such as whales, the photic zone, deep-sea animals, and animal behavior. Much of the rest of the permanent exhibitions is devoted to animal life in Denmark. The overall aim is to show animals in their natural habitats and to educate about biodiversity in an integrating and aesthetically pleasing way, especially through the use of large modern dioramas. School parties comprise approximately 30–40% of the visits.

The American Museum of Natural History in New York was probably one of the first, if not the first, to mount a major exhibition specifically on biodiversity (titled "Hall of Biodiversity"). This permanent exhibition of approximately 1200 m<sup>2</sup> was opened in 1998 with the aim of alerting the public to the biodiversity crisis and its implications and also to show the beauty and abundance of life on Earth. There is significant use of new technology to aid communication, such as multiscreen video installations and use of high-resolution imagery. An electronic "BioBulletin" presents a regularly updated account of events that affect biodiversity and the threats it is currently experiencing. A large diorama (occupying approximately 23% of the total space) of a portion of rain forest of the Central African Republic emphasizes the interaction of animals and humans with the environment. By showing the rain forest in three different states—pristine, altered by natural forces, and degraded by humans—it is possible for the visitor to grasp readily the broad impact of current forces on forest ecology. The diversity of life is covered by a display of approximately 1500 specimens and models mounted along one side of the Hall. The visitor is able to learn more about the organisms, their distribu-

tion, their habitats, and ecological processes by using interactive computer stations. Finally, a so-called "Solutions Wall" enables visitors to learn how mankind can help to manage and protect biodiversity.

These four examples (necessarily highly selective) of the approaches adopted by some modern museums exemplify the importance attached to biodiversity and the issues surrounding it. The visiting public is treated to a visually interesting and informative experience of direct relevance to the modern world. The experience is essentially unique to museums and closely related institutions. As such, they are providing an important role in educating and influencing a wide range of people.

"Traveling exhibitions" form part of the program of some larger museums. The aims are to enable a greater audience to experience a particular exhibition and to recoup some of the production costs. They may be shown in several countries. The topics can also be diverse, although biodiversity is often featured in one form or another. Two recent examples are the "Ocean Planet" organized by the Smithsonian Institute (Washington, DC), which includes coverage of the plant and animal life of the oceans, and "Voyages of Discovery" produced by The Natural History Museum (London), which tells of several major voyages, from Hans Sloane's to Jamaica (1687–1689) to that of the oceangoing *Challenger* (1872–1876). These voyages were responsible for many major discoveries about the earth's biodiversity.

Rental, transport, and installation costs, however, can be a significant deterrent to an extensive use of traveling exhibitions. The costs may have to be met by an admission charge or through sponsorship. Exhibitions about biodiversity and related exhibitions therefore compete with a mass of other attractions for peoples' attention.

Finally, electronic media allow for the development and transmission of the "virtual museum" in which images of specimens with associated information can be remotely accessed, sometimes in an interactive way. Museums are experiencing a revolution in the ways that they display and disperse their information, representations of their holdings, and their skills.

#### IV. EDUCATION AND TRAINING

The educational role described previously extends in different ways once the greater functions of museums are considered. Many museums provide a special service for visiting school groups, such as activity programs,

schemes to help teachers guide children around the exhibitions so as to learn more from them, and special talks and demonstrations, sometimes involving live organisms. Local schools often find these services invaluable for teaching more about biodiversity and helping the students gain a genuine and long-term interest.

A community approach can be particularly beneficial by bringing together the aims of educating people about biodiversity while simultaneously achieving a useful product or new research on it. The Chicago Wilderness Project involves individuals from the community together with the Field Museum of Natural History and other U.S. organizations and NGOs to document and restore the wild areas in and around the city of Chicago. All are learning about, and contributing to, the environment through an interactive process.

Since biodiversity is a complex concept, it is recognized that careful thought should be given to the educational role. What should be the aims? In what way is the museum trying to give the visitor a greater awareness or new skills relating to biodiversity and other topics? One recent example of analysis of this issue is provided by the work of the Council for Environmental Education (1997), which is part of the University of Reading, England. The council researched, with government backing, site-based biodiversity education provision. Museums were included in its site reviews. The council's findings suggested that biodiversity education should enable people to

- Understand what biodiversity means
- Understand that biodiversity is a dynamic concept; that species, habitats, and ecosystems are part of a balanced system that changes naturally over time
  - Become more aware of biodiversity as part of their cultural and spiritual as well as economic heritage
  - Be more aware of, informed about, and understand the significance of biodiversity around them and define their own level of interaction with it
  - Recognize the relationship between biodiversity and the maintenance and quality of life
  - Know that factors influence biodiversity and understand that human activity can both damage and enhance biodiversity
    - Be aware of the impact of their own and others' actions on biodiversity, including lifestyle and consumer choices
    - Improve their own skills in relation to biodiversity, including those skills that enhance understanding and promote appropriate action

- Be aware of what actions they can take to preserve and enhance biodiversity and act on that awareness

Museums do not, as a matter of course, address all these issues in their educational programs, but in practice several programs are being implemented in some of the larger institutions. The research helped to formulate and bring together much of what had been recognized in other ways.

The teaching of systematics and the training of systematists have been in decline during the past 20–30 years in many parts of the world. The decline has occurred in both the schools and the universities at a time when the need for advances in the field is increasing so as to meet the challenges of biodiversity conservation, of environmental protection, and of sustainable development. The issue is compounded by the fact that there is an estimated decrease in the number of systematists throughout the world. Furthermore, the age distribution of those remaining is skewed toward older age groups.

Museums are attempting to counter these trends by providing courses and training in systematics and in related fields of direct relevance to biodiversity. They work with universities in postgraduate research training and also in 1-year "masters" courses to teach taxonomy. The scale of these operations is still limited and is insufficient to meet the current and predicted future needs.

Many larger museums and gardens provide training and workshops to meet specific needs, often including programs related to biodiversity. These inevitably change with time as demand changes, but these institutions are providing an important service, whether it be for training in biodiversity surveys and inventories or training in handling biological informatics.

## V. PUBLICATIONS AND OUTREACH

Museums and related institutions are increasingly the providers of biodiversity information and interpreters of that information. The forms of the information are numerous and range from monographs on particular genera or families of organisms to popular books for a wide readership. Interactive keys and guides through the medium of the CD-ROM have also gained prominence in recent years. They can be a fast and user-friendly way of identifying a specimen within a particular group. The CD-ROM has also been used for educational packages on insects, birds, and other wildlife. Development costs, however, can be high, and with the growth of the Internet as an effective competitor, the

CD-ROM is seen as having only a limited applicability in the marketplace.

The Internet provides a major challenge for collections-based institutions to distribute their relevant information to a wide range of potential users. As discussed in Section I, the information contained in the collections and in other related databases is of important use in many areas. The challenge in making that information available is primarily through the substantial resource demands (in staff, software, and hardware) not only to establish the large-scale databases but also to maintain them. It is also still very early to be able to assess realistically what the user requirements are likely to be and how these will change with time.

Because of the scale of the operation and the need for many institutions to play a role in these kinds of information dissemination so as to gain much greater benefits, some planned approaches are international in scope. Data sets held by different institutions and maintained by them will form part of a global network. Developing and other countries will be able to access information, held in developed countries, about their own biodiversity. In this sense, information is being "repatriated" to those who need it most.

Currently, the World Wide Web gives any user a considerable degree of access to information that was difficult to access only a few years ago. Much remains to be done, but many museums and institutions have their own web pages with associated databases and other packages. Inevitably, the quality of these pages varies markedly from one institution to another, primarily because we are at a relatively early stage of development and not all institutions have the necessary resources to take on these additional tasks. In 1999, a link list of "natural history museums" gave the web addresses of several institutions (including universities) in the Americas, Africa, Asia, Europe, Australasia, and Oceania: <http://www.lib.washington.edu/sla/natmus.html>.

## VI. ADVISORY

Museums have been contributing to broader issues in addition to the pursuit of systematics, the curation of collections, and the provision of exhibitions. They contribute through the provision of advice and consulting services on biodiversity to a wide range of users. This is achieved primarily in three broad ways: consulting work, aiding governments in policy formulation, and helping to implement regional and global initiatives or policies.

### A. Consulting

In an age when there is increasing awareness of and activity regarding conservation, environmental quality, and sustainable development, the role of natural history museums in servicing the associated demands is becoming more important. Museums, with sufficient resources, can provide the most fundamental of data in almost all environmental studies or developments, including information on the species present and their behavior as well as on the changes that an activity or process can bring to the local biodiversity system.

Assessments of the impact of any activity, such as mining, deforestation, or urban development, can be undertaken. The information will also be used to provide predictions for work in other areas. Planning authorities will be one of the prime clients of such a service, as will other government agencies and industrial companies. In some countries (e.g., Australia), it is a requirement that proper assessment of the biodiversity of an area is made, together with assessment of the likely impact, before a development can be given approval to proceed. Museums are often the only institutions that can provide much of the necessary information and therefore are increasingly being called on to undertake an advisory role with related research.

Other consulting work done by museums includes waste management, chemical analysis, resource exploration, health and veterinary studies, and collections management. Some are also helping others to develop their public exhibitions, including those on biodiversity.

### B. Aiding Governments on Policy

Governments call on museums because of their special skills, to advise them on many policy issues and their implementation. Implementation of the CITES is a good example because museums are key to sound identifications that can carry legal weight.

In the realm of biodiversity, museums have been used to help governments create their country's action plans under the Convention on Biological Diversity (CBD). Several museums have staff that serve on their official national delegation to the Conferences of the Parties of the CBD. Similarly, they may serve on the delegation to the CBD's Subsidiary Body on Scientific, Technical and Technological Advice. This body advises on the implementation of the convention. Governments recognize that museums have a special role to play which complements those of the other institutional and government members. This is particularly important



since taxonomy and systematics are currently topics on the agendas of the CBD.

The expertise of museum staff is being used in many ways by countries to implement their obligations under the CBD. Several are undertaking biodiversity inventories or assessments. In aiding this process, guides on procedures and organisms have been developed. One particularly relevant example is a guide to good practice in biodiversity assessment (Jermy *et al.*, 1995), the purpose of which "is to help those who have the responsibility to survey and assess the biodiversity of their own country."

### C. Regional and International Initiatives

The CBD operates at both national and international levels. Biodiversity information, through its sheer scale, also operates in this way. Natural history institutions are partners or contributors to several regional and international initiatives, especially those related to biodiversity. The European Union is currently supporting some European museums and botanic gardens in planning and subsequently developing an integrated collections databasing facility at the specimen level. Governments are also discussing how a major international

program bringing together many aspects of biodiversity information might be established.

### See Also the Following Articles

BIOPROSPECTING • CONSERVATION BIOLOGY, DISCIPLINE OF • CONSERVATION MOVEMENT, HISTORICAL • EDUCATION AND BIODIVERSITY • GOVERNMENT LEGISLATION AND REGULATION • TAXONOMY, METHODS OF • ZOOS AND ZOOLOGICAL PARKS

### Bibliography

- Carter, D., and Walker, A. K. (1999). *Care and Conservation of Natural History Collections*. Butterworth-Heinemann, London.
- Forey, P. L., Humphries, C. J., and Vane-Wright, R. I. (Eds.) (1994). *Systematics and Conservation Evaluation*. The Systematics Association Special Vol. No. 50. Oxford Univ. Press, Oxford.
- Hawksworth, D. M., Kirk, P. M., and Clarke, S. D. (Eds.) (1997). *Biodiversity Information: Needs and Options. Proceedings of the 1996 International Workshop on Biodiversity Information*. CAB International, Wallingford, UK.
- Jermy, A. C., Long, D., Sands, M. J. S., Stork, N. E., and Winsor, S. (Eds.) (1995). *Biodiversity Assessment: A Guide to Good Practice*. HMSO, London.
- Nudds, J. R., and Pettitt, C. W. (Eds.) (1997). *The Value and Valuation of Natural Science Collections*. Geological Society, London.
- Raven, P., *et al.* (1998). *Teaming with life: Investing in science to understand and use America's living capital*. Report of the President's Committee of Advisers on Science and Technology.



# MUTUALISM, EVOLUTION OF

Egbert Giles Leigh, Jr.  
*Smithsonian Tropical Research Institute*

---

- I. Why Is Mutualism So Important and So Puzzling?
  - II. The Evolution and Maintenance of Mutualism
  - III. Why Is Mutualism So Crucial to Progressive Evolution?
- 

## GLOSSARY

**ethology** The study of the behavior and social relations of animals.

**genetic terms** A locus is a particular place on a chromosome (strictly speaking, an “address” on a chromosomal map) where a gene resides. This gene may be any one of several alternative types or alleles. A diploid organism contains two genes at each autosomal locus—one from its mother and one from its father. A haploid organism contains one gene at each locus. A genome is the set of all an organism’s genes.

**mutualism** A mutually beneficial association or interaction, temporary or permanent, among organisms of the same or different species.

**parasitism** An association among members of different species in which members of one species (the parasite) live inside or on the resources of a body of the other.

**symbiosis** A mutualism among members of different species in which members of one species (the symbiont) live inside or on a body of the other (the host).

---

**MUTUALISMS**, mutually beneficial activities among organisms, are essential for the function of organisms and ecosystems. However, mutualists are often each other’s closest competitors. Natural selection is a competitive process. What keeps competition from annihilating the possibility of mutualistic cooperation? In this article, I outline conditions that allow mutualisms to evolve, describe various devices and circumstances that prevent one partner from becoming parasitic or overly exploitative of the others, and discuss why mutualism is so important in evolution.

## I. WHY IS MUTUALISM SO IMPORTANT AND SO PUZZLING?

Natural selection is a competitive process. The winners are those genes—and the individuals carrying them—which reproduce most successfully (Fisher, 1999). However, many animals live in groups because they depend in some way on each other’s activities. The partners in an animal’s group are likely to be its closest competitors—for food, mates, or living space. Life in groups thus poses two questions: What are the advantages of living in a group? and What factors prevent competition among group members from undermining their cooperation, thereby destroying the benefits of living together? Indeed, how the mutualism of social behavior could possibly evolve is a central problem of ethology. Solving this problem means understanding (i) the potential benefits of group life (Maynard Smith

and Szathmáry, 1995), (ii) the social mechanisms and ecological circumstances which minimize or prevent destructive competition among the partners in a group (Moynihan, 1998), and (iii) how such social mechanisms could evolve (Leigh and Rowell, 1995).

The central problem of ethology has a political analog. Human beings benefit from, and indeed depend on, social life. Nevertheless, some individuals can siphon off for their private advantage part of the common wealth created by social activity. A standard problem of political philosophy from the time of Aristotle onward is how to organize a society so that individual advantage coincides most nearly with the common good.

The "ethological" problem recurs at other levels of biology. Especially in the tropics, natural ecosystems are more luxuriant, productive, and diverse than the sterile grasslands which so often replace them when the land is abused (Leigh, 1999). Indeed, the human influence on natural ecosystems is called "disturbance" and considered damaging, as if ecosystems were organized for functions which include the maintenance of productivity and diversity—organization which a random alteration of the system would disrupt. This (rarely stated) assumption parallels Aristotle's argument that organisms are adapted because visibly mutant individuals are usually less functional than their normal counterparts. Indeed, despite the abundance of natural enemies (predators, parasites, and pathogens), organisms depend on their ecosystems for the necessities of life, even the air they breathe. Therefore, organisms share a common interest in their ecosystem's integrity, as if ecosystems are true "commonwealths." Does the common interest of organisms in the integrity of their ecosystem influence the evolution of individual species? If so, how? These are among the greatest mysteries of biology (Leigh, 1999).

Genes are the "ultimate" units whose self-interest drives evolution in the sense that no characteristic evolves nonrandomly unless it serves the self-interest of some gene (Dawkins, 1982). However, an individual gene, divorced from the rest of its genome and an organism appropriate for the genome's expression, is as useless as a piece of computer program without the rest of the program, the right computer for running the program, and an operator who can run the program on the computer. Thus, these ultimate units of self-interest are utterly dependent on each other. Chromosomes express the mutual advantage that different kinds of genes derive from each other's presence—their mutual dependence on each other's functions. In particular, chromosomes maintain their genes' continued association by ensuring their simultaneous replication (May-

nard Smith and Szathmáry, 1995). However, some types of genes, which do nothing to benefit the organisms that carry them, multiply independently of the chromosomes, threatening to fill their genome with selfish DNA. How is this sort of destructive competition prevented from draining away the common wealth of the genome?

Conflicts can also occur between chromosomal genes and their genome. Meiosis, the process whereby a gamete is assigned one gene at each locus, is usually as fair a lottery as nature can devise, in the sense that at each locus a gamete has equal chance of inheriting its grandmother's or grandfather's gene. When meiosis is fair, natural selection favors an allele only if it enhances the survival or reproduction of the organisms carrying it. On the other hand, a few alleles which injure their bearers spread by biasing meiosis of heterozygotes in their own favor. Such alleles are called segregation distorters. Some experimental populations have been wiped out by the spread of such alleles. Why is meiosis normally so fair when segregation distortion can spread alleles so effectively?

Eukaryotic cells are cells whose chromosomes are separated from the cytoplasm by a nuclear membrane. Nearly all eukaryotic cells contain organelles, such as mitochondria and chloroplasts. Organelles are essential to their host cells. Cells require mitochondria to obtain energy by oxidizing carbohydrates to carbon dioxide and water. Leaves cannot photosynthesize unless their cells contain chloroplasts. Organelles contain DNA and can reproduce with some help from nuclear genes. Their DNA reveals that organelles descend from free-living bacteria which invaded, or were ingested by, ancestors of their host cells more than a billion years ago (Margulis, 1993). Although organelles appear to be integral structures of their host cells, conflicts can arise between organelles and their hosts. Organelles are usually passed on by the mother. Natural selection thus favors organelles which cause female-biased, or all-female, sex ratios among the young of their host organisms, although selection usually favors hosts with equal numbers of young of each sex. Moreover, zygotes (fertilized eggs) which receive organelles from both parents may be impaired by the struggle for dominance between organelles from different parents. How did the symbiosis between cells and their organelles evolve? Why does harmony usually prevail between cells and their organelles?

Metazoans are complex multicellular animals. A metazoan's cells usually serve their organism well. However, cancer suggests that conflict between an individual and one of its cell lines is a real, often devastating,

possibility (Buss, 1987). How did the harmony among a metazoan's cells evolve? How are conflicts between metazoans and their cells reduced or suppressed?

## II. THE EVOLUTION AND MAINTENANCE OF MUTUALISM

A mutualism evolves only if the potential partners all benefit by cooperating. In other words, there must be advantages which are most easily or effectively obtained by cooperating, and all partners must benefit from cooperating, even if unequally. Cooperating must truly serve a common good.

### A. What Can Be Gained by Cooperating?

Among members of the same species, the most basic form of cooperation is sexual reproduction, by which two individuals jointly produce offspring more varied than those that either could produce alone (Maynard Smith and Szathmáry, 1995). This cooperation is extended in many species, including insectivorous birds and some fish, in which the parents share the labor or divide the tasks involved in feeding, sheltering, and defending the young.

Members of the same species initially join in larger groups, as a rule, if grouping enhances safety or effectiveness (Moynihan, 1998). They may combine for greater safety against predators because, with more eyes, predators are detected sooner or because members of a group can coordinate activities to confuse or repel predators. A group may also be able to overcome a competitor, or a large prey animal, which could defeat any one of its members. Among social insects, however, advantages of a suitable division of labor soon influenced the evolution of social life.

Mutualisms among different species usually involve complementation of different functions. Thus, a coral provides its symbiotic algae with nutrients from small animals that the coral polyps catch. In return, the coral receives carbohydrates from the photosynthesis of these algae. A plant provides its pollinators with nectar or pollen in hopes that these mobile animals will convey some of this pollen to another plant of the same species. Some plants provide food and shelter for specific kinds of ants, which repel most of their host plant's herbivores and destroy the growing tips of vines that would otherwise overgrow their host plant.

### B. What Keeps Mutualisms from Becoming Parasitisms?

Cooperation involves the pooling of labor to create a common good. What keeps a group member from exploiting the good without helping to create it? Why don't some group members cheat their partners in mutualism?

#### 1. Kin Selection

If members of a group are related to each other, and much less closely related to members of competing groups, an individual propagates copies of its own genes by helping fellow members of its group reproduce (Hölldobler and Wilson, 1990). A child inherits half its mother's genes: Thus, either mother or child can propagate copies of its own genes by helping the other reproduce.

Kin selection plays a crucial role in the evolution and maintenance of insect societies. A wasp lays an egg where the hatchling larva has immediate access to food. Often, this is a provisioned nest cell. Sometimes, solitary nesting is futile because unguarded nests are robbed of their eggs or provisions. Communal nesting makes for better defended nests but intensifies competition for nest cells and provisions, creating winners and losers. If winners and losers are related, it may be more profitable (as measured by the numbers of copies of the loser's genes propagated to offspring) for the loser to help the winner reproduce, perhaps by provisioning the winner's nest while the winner guards it. Relatedness to the winner enables the loser to benefit from this mutualism, but their common interest in a defended nest is equally essential to this mutualism.

In more complex insect societies, a single queen produces all the young (Fisher, 1999), and most of her daughters are workers who help the queen reproduce. Workers share a common interest in helping their queen if it is impossible, or at least much less effective, for them to reproduce on their own. Among ants, workers are either sterile or lay eggs so slowly that they do better to help their queen (Hölldobler and Wilson, 1990). A queen honeybee attains this end differently. She mates with many males and mixes their sperm thoroughly. Thus, most of a worker's colleagues are half-sisters. Because a worker is more closely related to her mother's egg than to a half-sister's egg, workers eat eggs laid by half-sisters, rendering worker reproduction futile (Seeley, 1995). The queen thereby creates a circumstance in which mutual policing among workers enforces their common interest in helping their queen. This community of interest among the workers makes

beehives models of self-organization, where workers perform their tasks, change from task to task, and adjust appropriately to changed circumstances automatically without the queen exerting direct control (Seeley, 1995). Kin selection is crucial to the evolution of these insect societies, but other arrangements must create community of interest among the workers for truly complex insect societies to evolve.

## 2. Selection among Groups

Selection among groups is one type of kin selection. Suppose that each individual mates only with fellow members of its group, no migrants are exchanged among groups, and each group is founded by emigrants from a single parent group. Then the groups are endogamous. This endogamy, or inbreeding, makes a group's members so much more closely related to each other than to outsiders that an individual's reproductive advantage coincides almost precisely with what enhances its group's reproduction (Leigh, 1999). In other words, the raw material of natural selection is genetic variation. If genetic variation among a group's members exceeds variation among groups (variance among group means), selection among individuals within a group usually overrides selection among groups. Only if endogamy is so strong that variation among groups exceeds variation within groups can selection among groups prevail. In the 1950s selection among groups was often invoked to explain cooperation. In 1966, however, George Williams showed that very stringent conditions were required for the good of the group to override individual advantage, and interest in selection among groups diminished greatly. Nevertheless, selection among groups sometimes enforces mutualism.

To understand how endogamy increases relatedness among a group's members, try the following experiment. Set 10 pennies on a table, 5 with heads up and 5 with tails up. Every minute, chose 2 pennies at random, 1 to "die" and the other to "reproduce." This is symbolized by turning over the first penny, if need be, so that it has the same side up as the second. In a short period of time, all pennies will have the same side up. If one uses more pennies, still starting with half heads and half tails, uniformity takes longer to achieve but eventually prevails. Next, represent migration from outside by choosing a penny at random every 5 or 10 min and tossing it to see which side should be up. Now, uniformity is achieved more rarely and briefly, if ever.

How endogamous must groups be to allow selection among groups to work? Let groups, as well as individuals, have finite lifetimes. The intensity of selection among groups on a characteristic is the proportionate

increase in the number of groups founded per preexisting group per group lifetime conferred by its members' possession of this characteristic, just as the intensity of selection among individuals on this characteristic is the proportionate increase that this characteristic confers on an individual's lifetime reproductive success, relative to fellow group members. Selection among individuals balances an equally intense selection among groups on the same characteristic if (i) each group descends from a single parent group, and one migrant is exchanged per two groups per group lifetime, or (ii) no migrants are exchanged but one of every  $N$  groups is founded by the joining of colonists from two parent groups, where  $N$  is the number of mature individuals per group (Leigh, 1999). Moreover, selection among groups is unlikely to override strong within-group selection unless there are many more groups than individuals per group. In sum, selection among groups is decisive only if groups are definite individuals in their own right.

Selection among groups plays a crucial role in maintaining mutualism between organelles and their host cells. Cells do not exchange organelles. Each cell receives its organelles from a single preexisting cell. This is true even of zygotes: Nearly all kinds of eukaryotic organisms have evolved means to ensure uniparental transmission of organelles. Organelles are therefore subject to a stringent selection among groups (each group being the organelles of given type in one cell) in their host's interest. In other words, organellar reproduction depends utterly on the reproductive success of their hosts.

How did organelles become susceptible to selection among their hosts? Consider mitochondria. The mitochondrial ancestors which first invaded host cells were probably parasites (Margulis, 1993). In some parasites, selection favors increased dispersal from one host to another, although this injures their current host's welfare. For others, selection favors "caring for" their current host even if this reduces the effectiveness of dispersal. It depends on how greatly increasing dispersal impairs the current host's usefulness, which differs from case to case. The complementation of functions between a host cell which procures food and an ancestral mitochondrion which metabolizes food far more effectively than the host can do so presumably favored symbionts which "cared" for their current hosts. As a house cat defends its new owner's house against conspecifics, so ancestral mitochondria defended their valuable hosts against conspecific invaders. Thus, migration of mitochondria between cells was prevented, thereby subjecting the mitochondria to selection among host cells.

Selection on hosts favored mechanisms ensuring uniparental transmission of organelles when their hosts reproduced sexually, thereby perfecting the mechanism maintaining mutualism between host cells and their organelles (Leigh, 1999).

Selection among groups also played an integral role in the evolution of metazoans. All metazoans descend from sexually reproducing ancestors, whose young began life as a fertilized egg or zygote. A zygote grows by dividing mitotically so that an animal's cells all have the same genotype. Each individual is therefore genetically unique, but (mutations excepted) an individual's cells are genetically identical. Moreover, selection favors individuals which can distinguish "self" from "nonself" and prevent invasion by unrelated cells, thereby preventing the exchange of migrant cells among individuals. These circumstances identify the reproductive interests of an animal even more closely with those of each of its cells (Maynard Smith and Szathmáry, 1995) than strict endogamy does for a group and its individuals.

Mutants do occur, however, and cancerous cell lineages can run riot at their organism's expense. Buss (1987) reviewed features of different species which enhance the harmony between individuals and their cells. In some phyla, the "germline" is sequestered, preventing a cancerous lineage from spreading to an individual's offspring. This circumstance is analogous to a honeybee or ant queen sterilizing her workers to create a common interest among them in helping her reproduce. In some species, maternal genes control the early stages of her offsprings' development, just as, by judicious distribution of food or hormones, the queen of a complex ant society apportions her workers among different "castes" or morphotypes, thereby programming a division of labor suitable for her colony (Hölldobler and Wilson, 1990).

In summary, selection among groups played a crucial role in the evolution of both eukaryotic cells and metazoans.

### 3. Mutual Enforcement

Mutual enforcement is vital for the maintenance of many mutualisms. I mentioned previously the mutual enforcement by honeybee workers of their common interest in helping their queen reproduce. By eating each other's eggs, they make it pointless for workers to lay eggs on their own rather than helping their queen.

An analogous mechanism enforces the fairness of meiosis. Fair meiosis ensures that selection favors only those alleles that benefit their carriers. Some "distorter alleles," however, bias meiosis in their own favor (Leigh and Rowell, 1995). Up to 95% of the sperm of an indi-

vidual heterozygous for a segregation distorter carry the distorter allele. If individuals homozygous for the distorter die childless, the distorter allele spreads until death of its homozygotes balances its spread from biased meiosis. Imagine a mutant at a locus on another chromosome whose only effect is to restore fair meiosis in all its bearers. Because alleles on different chromosomes assort independently at meiosis, no allele at this mutant's locus can benefit from the distorter's bias. This mutant spreads because it prevents the distorter's spread among its bearers and thus spares some of its bearers from death caused by being homozygous for the distorter. Because mutants which restore fair meiosis are favored at loci on every chromosome but the distorter's, fair meiosis expresses the common interest of the genome as a whole. This selection has been demonstrated empirically. Moreover, the rarity among birds and mammals of alleles affecting sex ratio suggests that this selection has closed off most of the "easy" ways to bias meiosis, ensuring that selection favors alleles only if, on balance, they benefit the individuals that carry them. Therefore, even though genes are the ultimate units of self-interest which drive evolution, we analyze the adaptation of whole animals.

The most familiar form of mutual enforcement is reciprocal altruism—help those who help you, and make it stick by retaliating against those who cheat you (de Waal, 1996). Reciprocal altruism is illustrated by the sexual behavior of hamlets, simultaneously hermaphroditic coral reef fish, studied by Eric Fischer (Leigh and Rowell, 1995). Even though a sperm is much smaller and "cheaper" than an egg, a successful sperm contributes as many genes to future generations as a successful egg. Thus, most hermaphrodites devote as much energy to male functions—making great quantities of sperm, fighting or otherwise competing with each other for mates, and so on—as to the female functions of bearing and raising young. The proportion of a population's reproductive effort devoted to male functions represents the "50% cost of sexual reproduction." Hamlets avoid this cost by pairing off at mating time and alternating sex roles in successive spawns (matings). The fish which releases eggs during one spawn fertilizes the other's eggs during the next. This "egg-trading" short-circuits competition for mates and permits much lower sperm production, greatly reducing the cost of male functions. Reciprocation is favored even though it is cheaper to play the male role. At successive rounds of reciprocal spawns, each partner releases more eggs for the other to fertilize, as if mutual trust were increasing, whereas a fish that tries to play the male role twice in a row is dismissed by its partner

to find another mate and begin the process of confidence building all over again.

Mutualism among members of a chimpanzee group is also based in large part on reciprocal altruism (de Waal, 1996). The community of interest among a group's chimpanzees is based on two fundamental features. First, a chimpanzee is in danger outside a group: It needs to belong to one. Second, a group's effectiveness depends on all its members. Thus, all must accommodate, even the dominant, because without his fellow members the dominant would be alone (W. J. Smith as cited in Leigh and Rowell, 1995). Chimpanzees help those who have helped them, and they retaliate against those who have attacked them or failed them in time of need. Chimpanzees who are most generous to others when they obtain food are most likely to be given food when they beg for it. Chimpanzees' strategies of deception show that they are sufficiently self-aware that they can predict another's behavior by "putting themselves in the other's shoes." A chimpanzee who has supported another in a fight expects support from that other if it is attacked in consequence, and it will show anger if that support is not forthcoming.

Chimpanzees, however, seem to have a sense of justice (defined as what promotes the group's common good) that transcends reciprocal altruism (de Waal, 1996). Females may gang up on an alpha male who is taking excessive revenge on a subordinate. The group supports alpha males who mediate fairly in disputes, protecting lower ranking animals. If an alpha male always favors his allies when he interferes in fights, females will keep him from interfering. Indeed, alpha males retain their dominance by consent of the dominated. Finally, chimpanzees try to reconcile with animals whom they have just attacked, while the group noisily celebrates such reconciliations. Because fights endanger the welfare of the whole group, reconciliations serve the "common good," including that of the reconciler, who needs its fellow group members.

#### 4. Mutualism and the Common Good

Mutualism evolves only if there is genuine community of interest among the potential partners. In his *Politics*, Aristotle expressed this point very clearly: States whose constitution and social organization serve the common good more nearly are less susceptible to overthrow by popular revolution or factional putsch. Often, a common interest permits mutualisms to develop that lack any means of enforcement. On the other hand, even coreplication of host and symbiont (whereby symbionts can propagate only to the offspring of the host), such as that maintaining the mutualism between eukaryotic

cells and their organelles, will not create a stable mutualism if host and symbiont do not share a stable common interest.

Mixed-species bird flocks are examples of mutualisms without enforcement (Leigh and Rowell, 1995). A typical Neotropical mixed flock contains one pair each of several nuclear species, some with attendant young. Adults of these nuclear species jointly defend a common territory. As the flock progresses regularly over its territory, certain birds with smaller territories join the flock as it passes through them, whereas other birds with larger territories move from flock to flock. Each species of bird eats different food, or feeds in different types of places, as if to minimize competition among flock members. The advantage of a flock is more eyes to watch for predators and more birds to "mob" a predator when need be (Moynihan as cited in Leigh and Rowell, 1995). Even an alarm signal benefits its giver—not only by helping to keep safe the group on which it depends but also by informing a potential predator that, since it has been seen in time to be avoided, it might as well go elsewhere. This mutualism of seemingly unenforced common interest has caused the evolution of certain colors and behaviors among members of certain mixed flocks (Moynihan, 1998).

One of the most striking mutualisms of the tropical forest is that between fig trees and their pollinating wasps (Herre, 1996). Each fig fruit is a flowerhead turned outside in to form a ball, lined with flowers on the inside, with a hole at one end. Each fruit is pollinated by one or more female wasps. These enter the fruit, pollinate its flowers, and lay eggs in up to half these flowers. Each wasp's larva grows within a single fig seed. When a fruit's adult wasps "hatch," they mate among themselves and the fertilized females fly out in search of new trees to pollinate.

These pollinator wasps are parasitized by nematodes. In some fig species, each fruit is pollinated by several wasps, each carrying nematodes into the fig. The young of these different nematodes compete to enter the fertilized females leaving the fig, reducing the wasps' ability to reach other figs and their reproductive success should they reach one. In other, "single-foundress" fig species, almost every fruit is pollinated by a single wasp. In these species, nematodes can only infect the young of their current host. Because of this coreplication, their reproductive success depends on that of their host. Selection thus favors nematodes which minimize the damage they inflict on their host. Such nematodes sometimes even enhance their host's reproduction.

Coreplication shaped a stable mutualism between eukaryotic cells and their organelles, which persists in

those few species whose members receive organelles from both parents. Nonetheless, coreplication of wasps and nematodes in single-foundress fig species has had no such effect. Some fig species with several pollinators per fruit are descended from single-foundress species. Nonetheless, their nematodes have reverted to the status of damaging parasites: They must lack a stable community of interest with their host wasps.

Another striking mutualism involves reef-building corals and their zooxanthellae, the symbiotic algae that supply them with carbohydrates in return for nutrients and a place in the sun. Many of these corals call their zooxanthellae from the surrounding water rather than inheriting them from their mothers by mutualism-enforcing coreplication. Wasps which inherit their nematodes from several hosts apiece suffer grievously from these parasites. Multiple origins of a coral's zooxanthellae do not disrupt the coral-algal mutualism, presumably because the complementation of functions between corals and their zooxanthellae establishes an effective community of interest between these organisms. Community of interest makes the mutualism, not the breeding system.

Similarly, a mutualism does not outlive the community of interest among its partners. When the queen dies in a colony of the social wasp *Metapolybia aztecoides*, no one member of the colony can lay enough eggs to maintain the colony's pool of workers. Several would-be queens thus join forces to undertake this task. Practice, however, increases their reproductive rate. Mary Jane West Eberhard found that when each reproductive no longer needs the others, they fight it out for the profits of their cooperation. In summary, a community of interest among the potential partners is the one thing needful for the evolution of mutualism. Coreplication, social system, reciprocal altruism, etc. can only shape a mutualism if there is a genuine community of interest among the potential partners.

### III. WHY IS MUTUALISM SO CRUCIAL TO PROGRESSIVE EVOLUTION?

Mutualism has allowed the successive formation of more complex and effective wholes from parts already tested by natural selection (Maynard Smith and Szathmáry, 1995). Moreover, wholes composed of modules already adapted to respond "constructively" to challenging circumstances are more capable of further adaptive evolution (Gerhart and Kirschner, 1997). Finally, the conflicts between selection at different levels, and the

mechanisms by which these conflicts are controlled or suppressed, represent a series of distinctive footprints of the decisive role played by natural selection in macroevolution. Indeed, study of the role of mutualism in evolution allows evolutionary history to testify decisively to the mechanisms of evolution.

#### A. Mutualism and the Major Transitions of Evolution

Mutualism played a crucial role in all the major transitions of evolution. Perhaps the first of these transitions, almost effaced by the mists of time, was the transformation of nucleic acid molecules which replicated themselves by parasitizing protein-based metabolism into genes programming the metabolism and development of discrete organisms (Dyson, 1985). Whether Dyson is correct or not, the origin of life as we know it did involve the transformation of mutually dependent genes into coherent genomes (Maynard Smith and Szathmáry, 1995). Later transitions include the evolution of bacterial cells containing a variety of smaller parasitic and commensal microbes into genuine eukaryotes (Margulis, 1993); the evolution of sexual reproduction, whereby two individuals cooperate to produce offspring more varied than either could alone (Maynard Smith and Szathmáry, 1995); the evolution of meiosis, which ensures an appropriate apportionment of the genes of a diploid cell to its haploid descendants (Maynard Smith and Szathmáry, 1995); the evolution of complex, multicellular organisms (Buss, 1987); and the evolution of animal societies (Leigh and Rowell, 1995).

All these transitions involve entities cooperating toward achievements which no one entity could accomplish unaided. Many of these transitions, such as the evolution of eukaryotes, metazoans, and complex insect societies, transformed groups of formerly independent entities into coherent, integral wholes. Insofar as evolutionary progress has occurred, the evolution of mutualisms has made it possible.

#### B. Mutualism, Modularity, and Evolvability

Herbert Simon compared two watchmakers, Tempus and Horus. Horus put together a watch by making subassemblies of 10 parts apiece and then combining them into 2d-level assemblies of 10 subassemblies apiece, and so forth, until his watch was finished. Tempus would try to put together a watch without benefit of subassemblies. Both were subject to interruptions. An interrup-



tion forced Tempus to begin his watch again from scratch, whereas it forced Horus to begin his current subassembly again. As a result of his subassemblies, only Horus could finish watches.

The relevance of this parable to evolution is that mutualism is a mechanism that can combine preexisting "subassemblies," each already tested by natural selection, into larger, more effective wholes (Dyson, 1985). Mutualism thereby allowed the evolution of complex organisms by manageable stages.

Moreover, when cooperating entities share a common interest strong enough for selection to transform groups of cooperators into better integrated, more effective wholes, the resulting combines benefit from two properties. First, the self-interested adaptability of the component entities makes for a more effective and adaptable combine. Second, the adaptability of its components enhances the combine's capacity for adaptive evolution.

The first of these propositions is illustrated by the behavior of the oldest of intracellular organelles. The microtubules and centrioles (centrosomes) of eukaryotic cells, responsible for the "asters" and "spindle" that form during mitosis and meiosis, are not programmed by nuclear DNA. Instead, these organelles appear to be remnants of spirochaetes which invaded the ancestors of their hosts countless ages ago (Margulis, 1993). When a cell divides in mitosis, the centriole, just outside the nucleus, divides first, and its two descendants move to opposite sides of the nucleus. Microtubules grow out from the centrioles, forming asters. A microtubule quickly dissolves unless it encounters a part of a chromosome called a "kinetochore," which stabilizes a microtubule touching it. This "random exploration" by microtubules allows chromosomes to be attached to their respective asters in a few minutes, accomplishing the essential first step in apportioning a complete set of chromosomes to each daughter cell (Gerhart and Kirschner, 1997). At the level of cells within organisms, the generation and multiplication of a great variety of "T cells" allows selection for and mass production of those antibodies required to protect their organism from disease (Gerhart and Kirschner, 1997). At the level of organisms within societies, the search by foraging honeybees for new sources of food and the communication of successful searches to the hive, thereby recruiting other foragers to the newly found food, allow honeybee colonies to allocate foraging effort effectively (Seeley, 1995). At all three levels, forms of random exploratory search, conjoined with mechanisms rewarding success, play a crucial role in the development of well-adapted morphology and

behavior. Search by quasi-autonomous organisms often serves these purposes more effectively than more centrally controlled processes (Gerhart and Kirschner, 1997).

Adaptability of cells and tissues greatly increases the probability that a major change, genetic or environmental, will not prove lethal. E. J. Slijper described a goat born without forelegs. This goat learned to hop about like a kangaroo. To suit this goat's peculiar two-legged hop, its hind legs became enlarged, its spine became curved and the sizes and shapes of its vertebrae changed, and the sizes and points of attachment of many muscles and ligaments also changed. These adjustments testify to the adaptability of this goat's cells and tissues. Similarly, N. Smythe (as cited in Leigh, 1999) was able to create a social strain of pacas, otherwise fiercely territorial animals, by rearing one generation of young under special conditions. These young passed on their new social "traditions" to their descendants. The adaptability of animals, developmental and social, allows them to adapt to some very unusual circumstances.

### C. Mutualism and Unmistakable Footprints of Natural Selection

I have shown how several evolutionary transitions involved smaller parts combining into larger, more integral wholes. Thus, certain parasites and commensals of bacteria were transformed into organelles of eukaryotic cells, certain multicellular aggregates were transformed into coherent multicellular individuals, and certain insects which lived together in groups were transformed into units of cohesive insect societies with a single reproductive and complex division of labor.

A trace of each of these transitions is left in the potential conflicts between the erstwhile parts and the whole they formed. Competition between organelles inherited from different parents can injure their hosts. The spread of cancerous cells can kill the individual to which they belong. Worker bees can lay eggs of their own rather than help their queen.

How a transition occurred is often revealed by the mechanisms which suppress conflicts between the resulting whole and its parts. Most organisms are designed to ensure uniparental transmission of organelles, thereby preventing competition between organelles from different parents and ensuring that the reproduction of organelles depends utterly on the reproduction of their hosts. In most metazoan species, sexual reproduction ensures that each individual is genetically unique, but that individual's cells are genetically identi-

cal save for the accidents of somatic mutation, a circumstance that ensures that selection discriminates among individuals and not the cells within a single individual. The other means by which metazoans reduce the threat of “rebel” cell lineages—sequestration of the germline (so that cancers of the body cannot be passed on to offspring), maternal control of the early stages of development of her young, and so on—are paralleled by the means whereby queens of complex insect societies maintain harmony within their colonies.

Since these conflicts are suppressed by the very mechanisms that joined the parts into larger wholes, the mechanisms involved represent unmistakable footprints of the critical role natural selection played in these transitions—the most important events of macroevolution. The conflicts involved in each such transition, and the means by which they are suppressed, reveal the historical path of evolution and the selective mechanisms that directed it. Thus, study of the role of mutualism in evolution enables evolutionary history to testify to the mechanisms of this evolution.

### See Also the Following Articles

BIODIVERSITY, EVOLUTION AND • COEVOLUTION •  
COMPETITION, INTERSPECIFIC • DIVERSITY, COMMUNITY/  
REGIONAL LEVEL • GENETIC DIVERSITY • PARASITISM •  
SOCIAL BEHAVIOR • SPECIES COEXISTENCE

### Bibliography

- Buss, L. W. (1987). *The Evolution of Individuality*. Princeton Univ. Press, Princeton, NJ.
- Dawkins, R. (1982). *The Extended Phenotype*. Oxford Univ. Press, Oxford.
- de Waal, F. (1996). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Harvard Univ. Press, Cambridge, MA.
- Dyson, F. J. (1985). *Origins of Life*. Cambridge Univ. Press, Cambridge, UK.
- Fisher, R. A. (1999). *The Genetical Theory of Natural Selection: A Complete Variorum Edition* (J. H. Bennett, Ed.). Oxford Univ. Press, Oxford.
- Gerhart, J., and Kirschner, M. (1997). *Cells, Embryos and Evolution*. Blackwell Scientific, Oxford.
- Herre, E. A. (1996). An overview of studies on a community of Panamanian figs. *J. Biogeogr.* 23, 593–607.
- Hölldobler, B., and Wilson, E. O. (1990). *The Ants*. Harvard Univ. Press, Cambridge, MA.
- Leigh, E. G., Jr. (1999). *Tropical Forest Ecology: A View from Barro Colorado Island*. Oxford Univ. Press, New York.
- Leigh, E. G., Jr., and Rowell, T. E. (1995). The evolution of mutualism and other forms of harmony at various levels of biological organization. *Écologie* 26, 131–158.
- Margulis, L. (1993). *Symbiosis in Cell Evolution*, 2nd ed. Freeman, New York.
- Maynard Smith, J., and Szathmáry, E. (1995). *The Major Transitions of Evolution*. Freeman/Spektrum, Oxford.
- Moynihan, M. H. (1998). *The Social Regulation of Competition and Aggression in Animals*. Smithsonian Institution Press, Washington, DC.
- Seeley, T. D. (1995). *The Wisdom of the Hive: The Social Physiology of Honeybee Colonies*. Harvard Univ. Press, Cambridge, MA.
- Smith, D. C., and Douglas, A. E. (1987). *The Biology of Symbiosis*. Arnold, London.





# MYRIAPODS

Alessandro Minelli\* and Sergei I. Golovatch†

\*University of Padova and †Russian Academy of Sciences

- I. Introduction
  - II. Number of Species Known and Expected
  - III. Geographical Patterns
  - IV. Habitats and Adaptations
- 

## GLOSSARY

- anthropochoric** Geographic distribution due to human agency.
- cerotegument** Hydrophobic secretion layer covering the body of several arthropods adapted to temporary submersion.
- disharmonic, fauna** A fauna in which a strongly reduced number of major lineages is represented, as in oceanic islands, due their improbable colonization from distant sources.
- laurisilva** Warm-temperate woodlands characterized by evergreen broadleaves such as laurel (*Laurus*) trees.
- pupoid** Earliest postembryonic stage with incompletely developed appendages, as in insect pupae, hence motionless.
- stenotopic** A species (or higher taxon) with very restricted geographic range.
- 

**MYRIAPODS** have long been treated as a natural class (Myriapoda) of the phylum Arthropoda, but many

doubts have been raised regarding the close affinity between the four main groups of terrestrial, tracheate, and multilegged arthropods traditionally classified as myriapods. The term, however, is still universally used as vernacular to contrast these arthropods to the insects, also terrestrial and tracheate, with these latter being easily identified by their smaller number of legs (three pairs), the strong differentiation of the trunk into thorax and abdomen, and the generalized presence of wings.

## I. INTRODUCTION

Myriapods are wingless terrestrial arthropods with at least nine pairs of walking legs in the adult and a trunk not distinctly subdivided into thorax and abdomen. Unlike many hexapods (insects), myriapods never undergo complete metamorphosis in their life cycle. Some of these and other features of myriapods have long been regarded as primitive within a large arthropod lineage called the Atelocerata (or Tracheata, Uniramia, or Labrata), traditionally recognized as including myriapods and hexapods, as opposed to the other main arthropodan lineages Chelicerata and Crustacea. However, recent molecular investigations (Friedrich and Tautz, 1995; Boore *et al.*, 1998) suggest that insects may be closer indeed to crustaceans than to myriapods, a scenario that invites fresh reinvestigation of the real affinities of the myriapod groups among themselves and to the remaining arthropods.

Four recent and several fossil classes of myriapods can be recognized. Centipedes (class Chilopoda) and millipedes (class Diplopoda) are the main myriapod groups, whereas the two remaining recent classes, the Symphyla and the Pauropoda, lack vernacular names. Of the extinct myriapodous arthropods, Arthropleurida and Kampecarida are the most remarkable, the former because of their very large size (up to 2 m long and 45 cm in breadth) and the latter because they possibly represented the earliest myriapods ever (Upper Silurian), although it is still debatable whether they were terrestrial or not. The extinct Archipolypoda were possibly true Diplopoda; they were conspicuously spiny and reached 30 cm in length. Remains of several other extinct orders are also known from the Paleozoic. The earliest centipedes recorded are the extinct Devonobiomorpha from the Devonian.

## A. Basic Morphology

### 1. Chilopoda

Centipedes range in length between a few millimeters and the ca. 30 cm of *Scolopendra gigantea* (Fig. 1). The centipede body is divided into head and trunk. Compound eyes are only present in Scutigermorpha, whereas groups of simple eyes (ocelli) are present in most Lithobiomorpha and many Scolopendromorpha; all Geophilomorpha and many Scolopendromorpha are blind, as are several cavernicolous representatives of the Lithobiomorpha. The antennae may be longer than the body, as in the Scutigermorpha and in some Lithobiomorpha. In these two groups, the number of antennal articles is mostly high and variable, but it is fixed and generally lower in the other groups (14 in all Geophilomorpha and 17 in many Scolopendromorpha). The first trunk segment bears a pair of specialized appendages—the poison claws (or forcipules), each of which contains a voluminous poison gland—which are used in the capture of prey and occasionally in defense. Each of the following segments bears a pair of legs, mostly of cursorial type, often with sexual or other specializations in the last pair(s). Curiously, the number of leg-bearing segments is always odd-numbered in the adults, with 15 pairs of legs in all Scutigermorpha, Lithobiomorpha, and Craterostigmomorpha, 21 or 23 in the Scolopendromorpha, and 27 (*Schendylops oligopus*) to 191 (*Gonibregmatus plurimipes*) in the Geophilomorpha. In most geophilomorph centipedes the number of trunk segments is variable within the species and higher in females (Minelli and Bortoletto, 1988).

The Scutigermorpha has a unique kind of respiratory system, with seven dorsal openings providing oxy-

gen via thick bundles of tiny tracheae to the dorsal vessel; the oxygen is further distributed to the whole body with the help of the circulatory apparatus: The circulating liquid (hemolymph) contains an oxygen-binding pigment (hemocyanin) related to the respiratory pigments found in several chelicerates (scorpions and spiders) and crustaceans. In all remaining centipedes there are tracheae of the same kind as those of insects, opening on lateral spiracles of most trunk segments (Geophilomorpha and the scolopendromorph genus *Plutonium*) or on alternate segments (the remaining groups). The genital opening is found on a subterminal segment at the posterior end of the body, near the anus.

### 2. Diplopoda

The size of millipedes spans between a few millimeters and 35 cm, one of the largest species being the spirostreptid *Sechelleptus sechellarum* (Fig. 2). The body includes a head and a trunk. Groups of simple eyes are present in most orders but are lacking in all Polydesmida and in some other groups as well as in all specialized cavernicolous species. The antennae are generally short and always comprise only eight articles. The mouthparts are generally adapted for cutting and chewing hard matter, such as wood or dead leaves, but some millipedes have evolved adaptations to sucking. The trunk is normally elongate, more or less flattened dorsoventrally, but subcylindrical in several orders. The body wall is rarely soft and flexible (subclass Pselaphognatha with the only order Polyxenida) and the exoskeleton is usually rigid (subclass Chilognatha, with all remaining orders) due to the presence of calcium salts in the internal layers (endocuticle). When the millipede is about to undergo a molt, these salts are dissolved. Accordingly, these arthropods need to obtain from the food sizable amounts of calcium. There are no waxes in their epicuticle, but some protection from evaporation is obtained through the lipid compounds in their exocuticle.

The first trunk segment (the collum) is legless; it is followed by three (four in the Spirobolida) “thoracic” segments with one pair of legs each and a further number, sometimes very high, of “abdominal” segments with two pairs of legs each. There are 11–17 pairs of legs in the Polyxenida and at least 17 (but usually many more) in the remaining groups, with the highest number being 375 pairs of legs recorded in *Illacme plenipes* (Engelhoff, 1990).

Most millipedes are provided with chemical defenses in the form of glands (ozadenes) producing noxious substances which pour out from lateral series of repug-

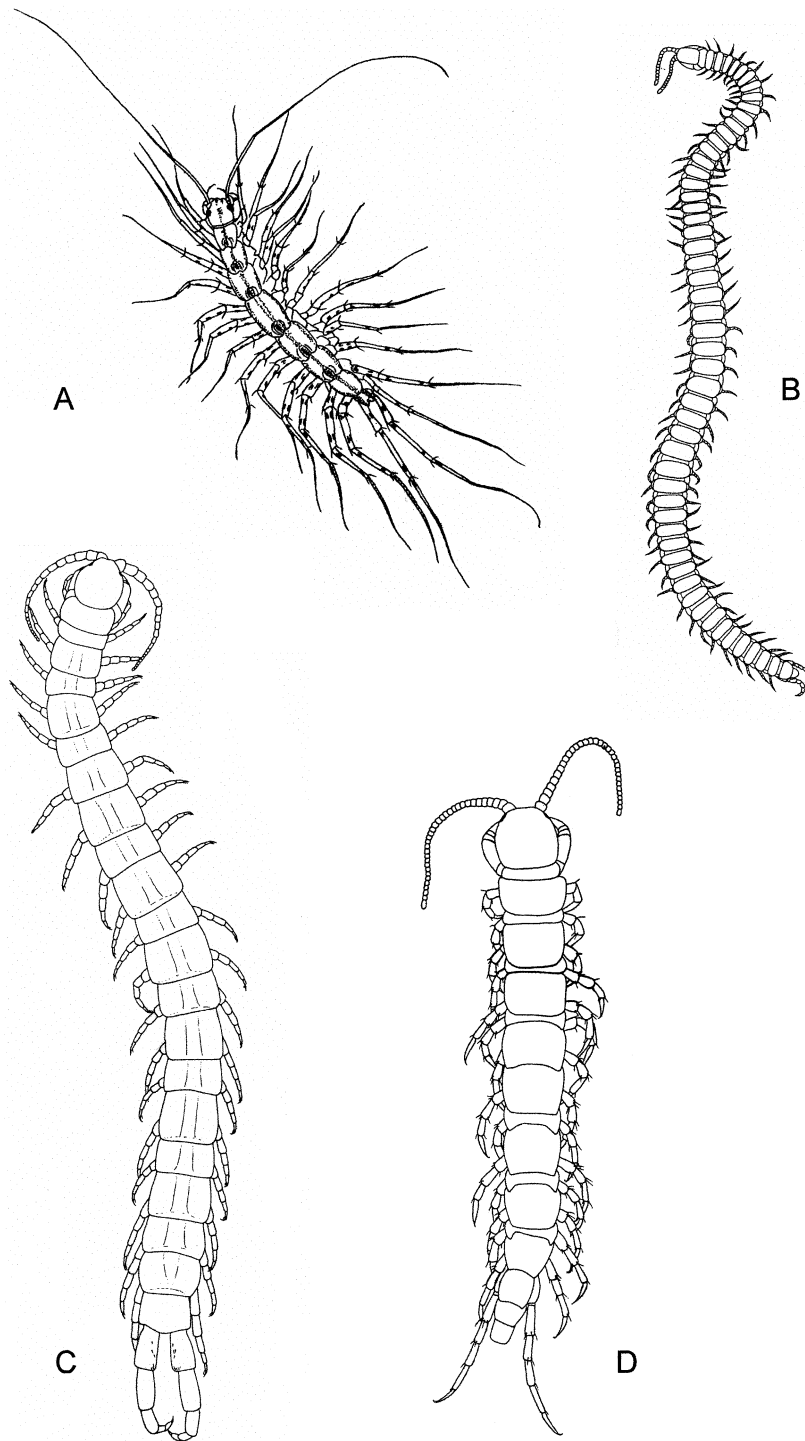


FIGURE 1 Habitus of representatives of the four main centipede orders: A, *Scutigera coleoptrata* (Scutigeroidea); B, *Geophilus carpophagus* (Geophilomorpha); C, *Scolopendra cingulata* (Scolopendromorpha) (adapted with permission of the Centre International de Myriapodologie); D, *Lithobius validus* (Lithobiomorpha).

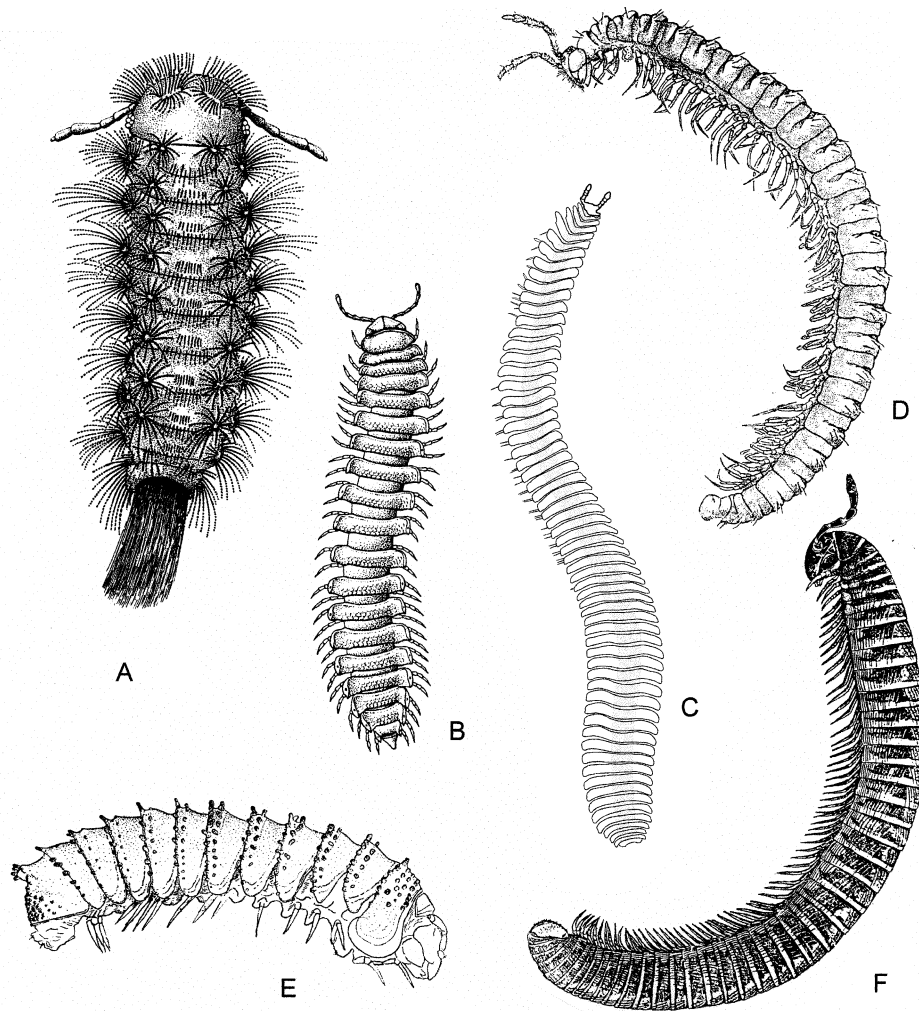


FIGURE 2 Habitus of representatives of the main millipede orders: A, *Macroxenus enghoffi* (Polyxenida) (after Nguyen Duy-Jacquemin, 1996); B, *Polydesmus* sp. (Polydesmida); C, *Brachycybe* sp. (Platydesmida); D, *Opisthocheiron canayerensis* (Chordeumatida) (adapted with permission of the Centre International de Myriapodologie); E, *Trachysphaera lobata* (Glomerida) (adapted with permission of the British Myriapod Group); F, *Stemmiulus adisi* (Stemmiulida) (adapted with permission of the Centre International de Myriapodologie).

natorial openings (ozopores). A diversity of defense substances have been found in these animals, with distinct classes of chemicals being characteristic of the different millipede groups: The juloids rely on benzoquinones, hydroquinones, phenol, and the acetates of some long-chain carboxylic acids; the polydesmoids rely on several carboxylic acids but especially on benzoic acid, benzaldehyde, hydrogen cyanide, and mandelonitrile; and the pill millipedes (Glomeridae) rely on very peculiar and complex heterocyclic compounds (glomerin and homoglomerin) and the Polyzoniidae on compounds known as polyzonimine and nitropoly-

zonimine (Eisner *et al.*, 1978). Several millipedes are adapted to coiling onto themselves (volvation), thus becoming a smooth ball difficult to seize; many others can roll themselves into a spiral flat coil. However, despite such precautions, many millipedes fall victim to insectivorous birds and predatory arthropods.

Special modifications of one or two leg pairs in the male may involve their transformation into clasping and/or genital organs. Posterior claspers are characteristic of the Pentazonia (the pill millipedes and their closer relatives). The male uses these claspers to fix the female

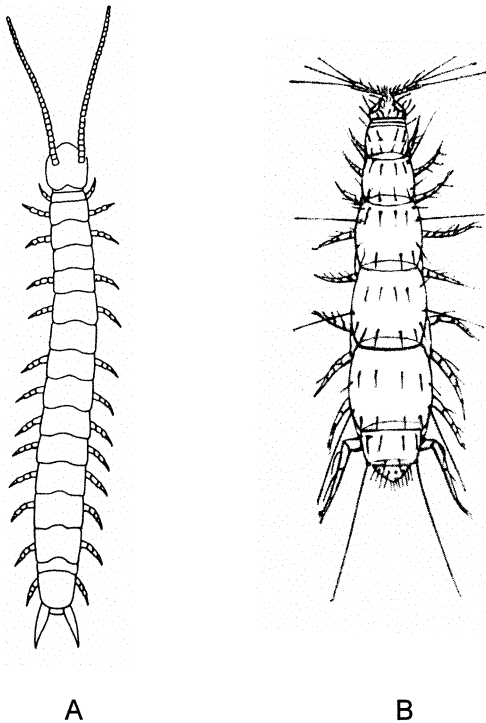


FIGURE 3 Habitus of (A) a symphylan and (B) a pauropod.

during mating. In the males of most other millipedes, one or, more commonly, two leg pairs of body segment 7, or of segments 7 and 8, are modified into gonopods. These are mostly very complex structures directly involved in sperm transfer. The use of the fifth pair of legs of the male *Chelojulus sculpturatus*, which are transformed into functional forceps, is unknown.

The tracheal respiratory system is provided with spiracles normally opening on the sternum near the coxa. The paired gonopores open on the coxae of the second pair of legs or in their proximity. Those of the male are found on a special structure incorrectly termed penis (or penes to emphasize the origin from the fusion of paired outgrowths), which has nothing to do with direct sperm transfer.

### 3. Symphyla

Symphylans are tiny myriapods, mostly 2–9 mm long (but *Hanseniella magna* is 25–30 mm long alive), generally with a whitish body (Fig. 3A). All species are blind and bear one pair of elongate multisegmented antennae, 12 (rarely 11) pairs of legs, and a pair of posterior unarticulated appendages called spinnerets. The mouthparts are of masticatory type. The genital opening is found on the fourth trunk segment. Symphylans are

the only arthropods with a pair of respiratory openings on the head, but the whole of their very soft and permeable cuticle is important for gas exchange. In these tiny myriapods there is no cuticular protection against evaporation.

### 4. Pauropoda

Pauropods are the smallest among myriapods; most species are only 0.5–0.7 mm long, and the largest is just 1.9 mm long (Fig. 3B). There are no eyes. The antennae are of unique structure, with four (order Tetramerocerata) or six (order Hexamerocerata) basal articles and two apical or subapical branches, each bearing in turn one or two flagella. The mouthparts are comparable to those of millipedes but are generally adapted to the suction of fluid aliment. The trunk bears 9–11 pairs of legs. Most species lack a respiratory system; one pair of tracheae, with spiracles on the coxae of the first pair of legs, is present only in the Hexamerocerata.

## B. Basic Biology

Nearly all myriapods are strictly terrestrial and most species are found in forest leaf litter or in rotten wood, under the bark of dead trees, in the soil, or in caves. Several millipedes also occur in dung, compost, and almost any other kind of plant debris; in nests of ants, termites, and birds; in worm and mammal burrows; near water; on the seashore; on open terrain; under stones; in buildings; etc. That is, they are found in virtually any terrestrial environment.

A few species of diplopods (e.g., *Blaniulus guttulatus*) and symphylans (*Scutigereilla immaculata*) are agricultural pests. Due to their swarming habits, a few millipedes can be of danger for transport when smashed in myriads on roads or railways or for housing, especially due to their smell, but such extremes are exceptional. In temperate regions, migrating swarms of many thousands of millipedes have been recorded, especially swarms of polydesmids (e.g., *Fontaria virginiensis*, *Parafontaria laminata*, and *Polydesmus collaris*). Causes for and mechanisms of this behavior are little known, but it has been suggested that heavy rainfalls may cause the swarming behavior of a large spirostreptid (*Plusioporus setiger*) in southern Brazil. Larger millipedes have quite long life spans lasting between 2 and 11 years, whereas their life histories usually take 2 or 3 years even in tropical environments. In high-montane habitats, the development is normally retarded by a year or two. However, some tiny tropical species live for less than



1 year and their life cycle takes only about 3 months (Hopkin and Read, 1992).

A few large scolopenders are generally considered of medical importance because they can have a painful bite, although virtually no authenticated lethal accident is on record. The repugnatorial secretions of several large millipedes appear so noxious that some Indian tribes in Central America use them for poisoning arrows prior to hunting. However, some of these tribes cook large millipedes and find them tasty. Finally, some scolopenders have found a place in traditional Chinese medicine alongside scorpions and other arthropods.

Among myriapods, only some centipedes have successfully colonized eremic environments, with many species of this class (mostly belonging to the Scolopendromorpha) being true deserticoles. This is due to their structural adaptations (waterproof cuticle and/or complex spiracles) allowing them to reduce water loss. The remaining myriapods are distinctly hygrophilous to mesophilous, often (the bulk of millipedes) calciphilous as well, which explains their difficulty in coping with conditions differing from those prevailing in forest litter.

Tree dwellers are not common among myriapods, being mainly represented by small or slender, if not flat-bodied, millipedes and by some representatives of lithobiomorph and geophilomorph centipedes. Species confined to tree crowns or living in suspended soil are even fewer. Some scolopendromorphs, however, are fairly common in suspended soils in the tropical forests of America.

Semiaquatic, littoral, or nidicolous myriapods are exceptional. Myriapod species capable of tolerating seawater are particularly few, although these are often quite common and widespread. Interestingly, no conspicuous morphological types can be discerned among myriapods as obvious adaptations to any of such definitely marginal environments. Again, no common morphological adaptation is shared by the relatively few anthropochoric myriapods that have attained more or less vast, sometimes worldwide, distributions due to human agency. However, several minor modifications of leg structure can be assumed as advantageous for better climbing performance, whereas other adaptations involving the mouthparts and spiracles indicate the species' semiaquatic habits (Adis *et al.*, 1998).

In their feeding habits, myriapods are quite diverse. Centipedes, as revealed by their poisonous forcipular fangs, are basically predatory. They are mainly solitary hunters; gregarious habits have rarely been observed, including feeding on barnacles by swarms of specimens of the geophilomorph *Strigamia maritima* in the tidal

region along the seashore. Some centipedes, however, are known to occasionally consume plant material. In contrast, most millipedes feed on vegetable matters, particularly dead leaves and wood, but also algae, fungi, lichens, moss, and pollen.

Millipedes, whose density in forest soil is sometimes higher than 1000 individuals per square meter, are estimated to consume 10–15% of the annual leaf fall in temperate forests. A few species may be qualified as omnivores; very few are carnivores—one of them being *Apfelbeckia lendensfeldii* (order Callipodida), which feeds on earthworms. Coprophagy is also quite widespread in this animal group.

Some suctorial myriapods (all tetramerocerate pauropods and some millipedes), all with mouthparts more or less strongly reduced, are known to live on fungi and/or semiliquid end products of decaying plant material. Most pauropods feed on fungal hyphae and spores, but *Millitauropus* (Hexamerocerata) feeds on small arthropods (springtails and their eggs). Symphylans are largely vegetarian, rarely saprophagous. Hence, the roles played by various myriapod groups in the terrestrial food chains differ, with millipedes being of particular importance as primary destructors and soil-formation factors and centipedes as active and restless predators. Representatives of only a few millipede orders are capable of manufacturing silk for cocoon production during molting and/or egg-laying.

### C. Reproduction

A few dozen cases of parthenogenesis are known (Enghoff, 1994). One is the lithobiomorph centipede *Lamyctes coeculus*, a synanthropic species known from several continents. Two small European millipedes (*Nemasoma varicorne* and *Polyxenus lagurus*), both mainly living under tree bark, exhibit geographical parthenogenesis, with bisexual populations in less disturbed areas and unisexual ones in areas where the species appear to be recent colonizers. Similarly, *Muyudasmus obliteratedus* is bisexual in its native South American habitats, but a stable population recently established in a hothouse in Germany is parthenogenetic. There is also evidence of parthenogenesis in pauropods.

All myriapods are oviparous. Two groups of centipedes (Scolopendromorpha and Geophilomorpha) have evolved brood care. In these animals, the female remains coiled (for as long as 3 months in some species) around her brood until hatching. Parental cares are virtually unknown in the other myriapod classes, with two of the rare exceptions being the millipedes *Polyzonium germanicum* (order Colobognatha) and *Bericostenus hu-*

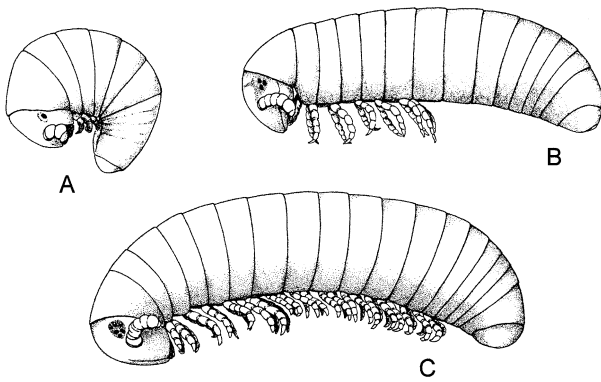


FIGURE 4 The first postembryonic stages of a millipede (*Anadenobolus leucostigma martinicensis*; Spirobolida): A, larva I with 3 pairs of legs; B, larva II with 7 pairs of legs; C, larva III with 21 pairs of legs (adapted with permission from Nguyen Duy-Jacquemin, 1992).

*micola* (order Platydesmida), whose habits are comparable to those of the brood-caring centipedes. Egg brooding by the male has been recorded in a few species of platydesmid millipedes.

#### D. A Diversity of Developmental Modi

Newborn millipedes have an incomplete number of segments and legs (Fig. 4). In most cases, only three pairs of legs are present in the first active or inactive postembryonic stadium. New segments and leg pairs are added at each molt, sometimes inside of constructed molting chambers, according to a precise schedule which is different in different orders, families, or genera. Three main developmental modi can be distinguished (Engelhoff *et al.*, 1994).

In hemianamorphosis (as in the pill millipede *Glomeris*) the final number of segments and leg pairs is obtained after some molts, but the millipede must undergo further molts without segmental increase before achieving adult size and sexual maturity. In teloanamorphosis (as in polydesmids) the achievement of the final number of segments and leg pairs coincides with obtaining the adult condition and no further molt occurs. In euanamorphosis (as in the cylindrical julids), new segments and leg pairs are added at every molt and the animals continue molting even after reaching sexual maturity.

The males of some western European representatives of the order Julida exhibit the peculiar developmental behavior known as periodomorphosis. Under specific adverse environmental conditions, sexually mature males of these species molt to intercalary males with

regressed gonopods and suspended reproductive activity. After one or more additional molts, these animals reach a further mature stage, with newly differentiated gonopods.

Hemianamorphosis is also known in centipedes: Scutigermorphs, lithobiomorphs, and craterostigmomorphs hatch with 4, 7, and 12 pairs of legs, respectively, and obtain the final number of 15 leg pairs following one (craterostigmomorphs) or more molts but continue molting until they reach maturity. The remaining centipedes (scolopendromorphs and geophilomorphs), however, develop by epimorphosis. That is, they reach the final number of body segments during the embryonic development; no segment or leg pair is thus added during the postembryonic life.

In symphylans, 6 or 7 of the final 11 or 12 pairs of legs are present in the first postembryonic stadium. Quite peculiar is the following developmental schedule in that first-order and second-order molts regularly alternate until maturity is reached. After each first-order molt a new two-segment unit is added, but only 1 more pair of legs is differentiated, whereas the second-order molts are accompanied by the differentiation of another pair of legs without addition of new segments. Mature symphylans continue to molt, up to 40 times in *Scutigrella immaculata*.

In pauropods, the first postembryonic stadium is a motionless pupoid with inarticulated traces of the antennae and of the first two pairs of legs. After a molt it changes into an active stadium with three pairs of legs. The development then proceeds by anamorphosis.

## II. NUMBER OF SPECIES KNOWN AND EXPECTED

Our knowledge of myriapods is highly fragmentary and incomplete (Table I). Not only does the bulk of existing species diversity remain to be described, but also the number of higher taxa appears to vary considerably as estimated by different taxonomists. For example, according to E. H. Eason (personal communication), of the approximately 150 genera and 1730 species/subspecies of Lithobiomorpha named to date, only approximately 17 and 950, respectively, are probably valid. In the Diplopoda, by far the largest of the myriapod classes, the situation concerning the number of orders and families, let alone genera and species, is badly equivocal to the point that Hoffman (1980) could equate the state of the art in millipede systematics to that in entomology in the middle of the nineteenth

TABLE I  
Myriapod Diversity (Approximate Data), Known and Estimated

Class	Recent orders	Families	Genera	Species	
				Described	Estimated
Chilopoda	5	21	325	3,300	6,000–10,000
Diplopoda	15–16	130	1800	11,000	50,000–80,000
Pauropoda	1	5	29	700	2,000–5,000
Symphyla	1	2	13	200	500

century. Thousands of taxa awaiting description are available in the collections. For example, according to P. Johns (personal communication), the collections of Australian millipedes already kept in Australia's natural history museums contain at least 2000 undescribed species. Indeed, it is mostly tropical myriapod faunas that appear to be particularly poorly explored.

Judging from the species described thus far, one millipede order outnumbers all others in terms of species counts—Polydesmida (up to one-half of the genus and species richness of the Diplopoda), followed by Spirostreptida, Chordeumatida, and Julida. The remaining orders are minor to marginal. Thus, the pentazonian order Glomeridesmida contains only a single genus with a handful of species in the Indo-Australian and Neotropical realms. The order Siphoniulida is extreme and obviously relict, being known by a single species each in Malaya and Central America. Because no male siphoniulidan has ever been described, and the structure of the male gonopods (if any) is absolutely unknown, the affinities of this order remain dubious.

Probably less than 20% of the world fauna has been described to date; this is particularly true of the largest class Diplopoda (Table I). Indeed, updated counts/checklists are only available for the millipede faunas of some European countries (Kime and Golovatch, 2000) and those of North and Central America and the Caribbean, whereas the fauna of Australia is assessed very crudely at the familial and ordinal levels only and that of Eurasia remains unpublished. The fact that the numbers of genera and species of Diplopoda currently known in Europe and the Mediterranean are similar to those known from the whole of Asia shows how poorly investigated these are is the Asian, largely tropical fauna. A world list of Pauropoda is available, but the distributional data concerning Chilopoda and Symphyla are still scattered in numerous local publications and no recent overview is available at the world or at least the continental level.

### III. GEOGRAPHICAL PATTERNS

The bulk of species richness is definitely confined to tropical countries, mainly woodlands (Black, 1997), with certain notable exceptions. Thus, some orders or families appear to be largely temperate and more or less restricted to the Northern Hemisphere, e.g., Lithobiomorpha among Chilopoda or Glomerida, Julida, Callipodida, and Polydesmidae among Diplopoda. For these groups, the hot spots of diversity are mountainous lands at the periphery of the Holarctic Region within the temperate to subtropical areas such as the Atlas, Pyrenees, Alps, Balkans, Caucasus, Tien-Shang, Himalaya, and Appalachians. Figure 5 shows the patterns of millipede generic diversity in Eurasia.

Harsh environments such as tundra or deserts are populated by myriapods patchily to marginally if at all, especially by Diplopoda and Symphyla. However, Chilopoda, Pauropoda, and even some Symphyla occur in permafrost areas, and a few scolopendromorph centipedes are quite characteristic desert dwellers. In contrast, high-montane habitats, however adverse, are regularly populated by myriapods, inclusive of millipedes, and sometimes even by endemic species. The highest altitudes recorded are approximately 4500 m above sea level for both millipedes and centipedes, the latter being encountered well within the nival belts of the Himalaya and the Andes.

Given the hundreds to thousands of species in each of the major myriapod groups, one may get the wrong impression that these basically cryptic animals should be highly diverse in most of the local environments, in particular in productive temperate or tropical forests. However, virtually no local fauna is known to include more than two or three dozen species, depending on habitat. Given that myriapods are basically composed of very ancient, widespread, but poorly vagile and largely stenotopic forms, most of the faunules, even among the

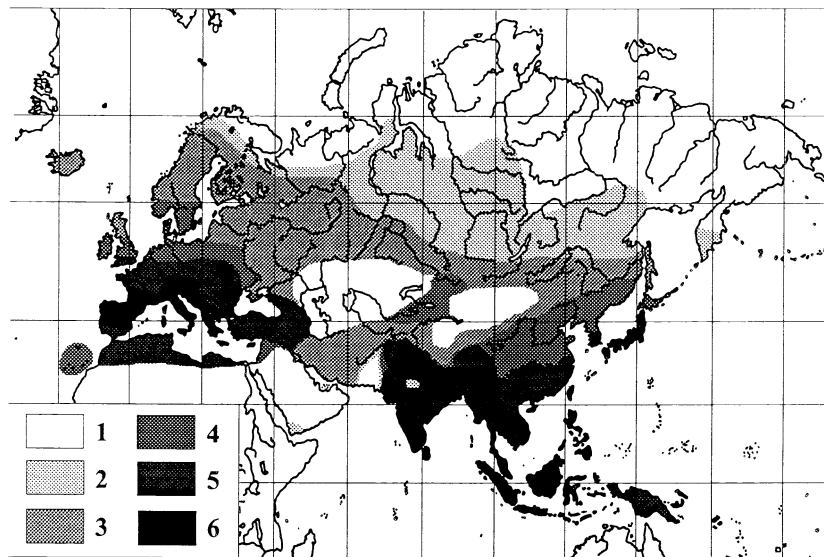


FIGURE 5 Patterns of millipede generic diversity in Eurasia (reproduced with permission from Golovatch, 1997). 1, millipede-free territories; 2, 1–4 genera; 3, 5–10 genera; 4, 11–25 genera; 5, 26–50 genera; 6, over 50 genera. The data exclude obvious anthropochores.

richest and monotonous-like tropical rain forest, appear different at least to some degree. In other terms, ancient and widespread endemism is generally characteristic of Myriapoda.

Cold areas host a very reduced fauna of millipedes. In Europe, only two species (*Polyxenus lagurus* and *Proteroiulus fuscus*) live north of the polar circle; no millipede is known from such areas of Siberia or from Greenland.

Only relatively few nonanthropochoric species display vast distributions; these tend to be particularly vast among predators (= Chilopoda), especially soil dwellers. Several geophilomorphs are examples of this pattern of distribution, e.g., *Arctogeophilus glacialis* and *Pachymerium ferrugineum*. Forms populating special but widespread habitats likewise appear widely distributed, e.g., the littoral geophilomorph *Strigamia maritima* or the huge, largely arboricolous, Neotropical scolopendromorph *Scolopendra gigantea*. Active colonizers of relatively young areas/habitats, such as the trans-Siberian *Angarozonium amurense*, the pan-European *Ommatoiulus sabulosus*, or the subcorticolous pan-European *Proteroiulus fuscus* among millipedes or the Euro-Siberian *Lithobius curtipes* and some other congeners among lithobiomorph centipedes, reach up to the taiga and, in part, tundra belts in the north, the youngest boreal biomes. The Mediterranean chilopod *Scolopendra cingulata* is a notable colonizer of semiarid badlands, whereas the European symphylan *Symphylella vulgaris* is a colonizer of fresh mines.

Worldwide distribution due to human agency is known for several myriapod species, mainly of European origin such as the millipedes *Cylindroiulus latestriatus* and *Ophiulus pilosus* and the centipede *Lithobius forficatus*. The origin of the synanthropic centipede *Lamycetes emarginatus*, also widely distributed, is unknown but it is possibly from the Southern Hemisphere.

### A. Endemism and Speciation

High numbers of endemic species and high percentages of endemics within the total millipede faunas are peculiar to some environments such as caves and some areas, such as islands and some mountain chains in the temperate regions.

The best investigated cases of intense speciation in millipedes are by far those of the genera *Cylindroiulus*, *Dolichoiusulus*, and *Acipes* in the Macaronesian area (Enghoff, 1992; Enghoff and Baez, 1993). On Madeira, of 60 millipede species known from the island, at least 29 belong to the endemic *Cylindroiulus madeirae* group (Julidae) and another 6 to an endemic lineage of the genus *Acipes* (Blaniulidae). Still more conspicuous is the *Dolichoiusulus* (Julidae) species swarm in the Canary Islands, which has 46 species of a total of 79 millipede species thus far recorded from the archipelago. These *Dolichoiusulus* species have colonized the most diverse habitats, from the laurisilva to the xeric habitats and the caves, whereas the Madeiran *Cylindroiulus* are all confined to laurisilva forest habitats.

TABLE II  
Distribution of Centipede and Millipede Diversity among the Major  
Continents/Regions<sup>a</sup>

Class	Continent/region				
	Eurasia	North and Central America	Afrotropical region	Australia and South Pacific	South America
Chilopoda	1.3	1.5	0.5	0.5	1.1
Diplopoda	7.5	3.5	3.6	0.4	2.8

<sup>a</sup> Approximate number of genera in hundreds.

In the Canarian *Dolichoilulus* species swarm, conventional allopatric (interisland) speciation is only marginally responsible for the genesis of the huge diversity of the group. According to Enghoff, speciation occurred mainly within individual islands, with distribution patterns governed mainly by habitat differences between closely related and nearly sympatric species. In several cases, those species which coexist in the same habitat differ conspicuously in size. The same is often seen in Madeiran *Cylindroiulus* and *Acipes* species; this rule, however, is far from universal.

## B. Distribution by Continents

In general, the overall distribution of myriapods is quite consistent with the world's biogeographic regionalization known since the times of J. Hooker and A. Wallace. In other words, the traditional division of the globe's territory into the Holarctic (= Palearctic + Nearctic), Afrotropical (including Madagascar and the Cape), Neotropical, Oriental (Indian + Malesian), and Australian (including New Guinea and the islands of the southern Pacific) realms is supported by the patterns demonstrated by myriapod higher taxa. Antarctica is the only continent harboring no Myriapoda, whereas, due to the particularly poorly explored tropical faunas, the distribution of genera and species between the remaining continents or biogeographic regions is highly provisional and appears extremely uneven and patchy (Table II).

## C. Hot Spots of Biodiversity and Endemism

The bulk of species richness is confined to tropical countries, mainly woodlands (Black, 1997), although

there are some notable exceptions. The opposite extreme is represented by numerous species endemic, e.g., to a single cave or mountain slope; this is especially characteristic of Diplopoda. Local cavernicolous endemics, notably troglobites, are particularly numerous in southern Europe, especially in the Balkan area, which is rich in karst. Similarly, nearly every patch of Amazonian primary rain forest near Manaus, Brazil, contains (much of) its own millipede faunule, with differences being particularly drastic between various kinds of inundation vs nonflooded (= terra firme) woodlands. Approximately the same pattern has been observed among Pauropoda.

## D. The Fauna of Islands

Analysis of insular biotas clearly shows the differences in colonizing capability and radiation patterns of centipedes vs millipedes. In most faunas, the ratio of millipede to centipede species is approximately 4:1, but this ratio is much lower in Hawaii, Tahiti, Samoa, and the Fiji Islands. This reveals that centipedes are apparently better dispersers than millipedes. The distribution patterns of millipedes on continental vs oceanic islands are those of poor dispersers. For instance, there are 18 millipede families represented by indigenous species in Sumatra, 17 in Japan, and 14 in Java, but there are only 10 in Australia, 9 in Madagascar, 5 in New Zealand, 3 in New Caledonia, and 1 in the Hawaii Islands. The number of endemic genera is very high in Indonesia and the Antilles, but these island groups do not have endemic families and suprageneric indigenous taxa are only known from Borneo.

The myriapod fauna of most oceanic islands is not simply disharmonic but also poor in species. Native myriapods are virtually absent from Kerguelen, Tristan da Cunha, and St. Helena.

## IV. HABITATS AND ADAPTATIONS

At least five morphological body plans can be distinguished among millipedes. The pin-cushion millipedes (i.e., the representatives of the worldwide distributed order Polyxenida) are soft bodied, very small (usually <5 mm long), extremely bristly, swift, and capable of withstanding much drier conditions than most other Diplopoda. Many polyxenidans live under loose tree bark or are characteristic inhabitants of microcaverns and small crevices under stones, in the uppermost soil, in litter, and in similar substrates.

The pill millipedes (the largely Holarctic order Glomerida and the Afrotropical + Indo-Australian order Sphaerotheriida) are collectively termed “rollers” because of their ability to roll themselves into mostly glossy balls. Pill millipedes are generally litter dwellers, but some smaller Glomerida are either troglobionts or geobionts.

The colobognathans (i.e., the “sucking millipedes” with reduced mouthparts) and some Chordeumatida, which mostly possess flexible, worm-like, strongly tapered bodies and shorter legs, are termed “borers.”

Many Polydesmida and numerous Chordeumatida, often conspicuously ornamented and relatively shorter, long-legged, and displaying more or less strong paranota (wing-like dorsolateral projections of the diplosegments), are referred to as “wedge types” and are characteristic of forest litter.

The lifestyle, or ecomorphological type, most common and widespread among Diplopoda is that of “bulldozers” or “rammers.” Their long, cylindrical, hard body with numerous diplosegments (hence, numerous pushing legs) penetrates the substrate like a bulldozer, using the broad head as a ram. Most juliform millipedes (Julida, Spirobolida, Spirostreptida, Callipodida, and some others) belong to this ecological type. This burrowing habit must have been critical in ensuring diplopods much or even most of their present-day high ecological and geographical performance. In particular, unlike the remaining ecomorphotypes, only juliforms, among Diplopoda, appear to have colonized virtually any suitable habitat. In fact, in Juliformia belong the few littoral dwellers and deserticoles, but this body plan also provides most of the millipede diversity in troglobionts and geobionts, in addition to anthropochores. Indeed, the northernmost record of a millipede belongs to the subcorticolous European species *Proteroiulus fuscus* (Julida) in the forest–tundra belt of Yamal Peninsula, Russia’s north, whereas perhaps the most characteristic deserticole among millipedes is *Orthoporus ornatus*

(Spirostreptida) in the southern United States and adjacent parts of Mexico. Both these species are rare except in their native environments.

There are sound reasons to believe that, due to their capability to escape adverse conditions by burrowing in the soil, rotten logs, and similar shelters, juliforms (namely, species of the order Julida) dominate in Europe the youngest, fully migratory nucleus of the millipede fauna, showing clear-cut inclinations to dwelling on open terrain. In contrast, all remaining millipede orders display very evident geographical trends in diversification. In Europe, even the relatively uniformly distributed Polydesmida have a remarkable center of secondary diversification in Slovenia, whereas the Chordeumatida are particularly species rich within the Atlantic climatic zone of western Europe (Kime and Golovatch, 2000). Furthermore, only a few species of Julida, apparently in response to the strongly adverse conditions that existed in Europe during the Ice Age, appear to have developed a population strategy unique among other terrestrial animals—the periodomorphosis strategy mentioned previously, i.e., the extension of male life by means of intercalary stadia (Hopkin and Read, 1992).

In summary, most millipede lineages are currently in a phase of rapid evolution and speciation; however, there are a few apparently relict groups (Hoffman, 1980). Polydesmida seem remarkably diverse largely due to their wedge type of burrowing allowing ecological niche partitioning mainly in the litter and at the soil–litter interface. This is particularly obvious in tropical and subtropical faunas. However, examples of troglobionts, geobionts, and myrmecophilous as well as arboricolous species are about as numerous among Polydesmida as among juliforms, whose type of burrowing and lifestyle seem the most characteristic, widespread, and ecologically progressive among all recent Diplopoda.

### A. Burrowing Myriapods

Several major myriapod groups tend to be soil dwellers. Usually, adaptations to geophily involve body elongation due to an increased number of segments. This is quite evident, for instance, in several lineages of geophilomorph centipedes, especially the Himantariidae and the Oryidae, which have a flexible worm-like body and peculiarly short appendages. This is also the case for certain scolopendromorph centipedes and some juliform and colobognathan millipedes, which are active burrowers. Body miniaturization also appears advantageous to dwelling in the crevices and/or burrows in the

soil. This adaptation is observed in most Symphyla, numerous Pauropoda, and some Diplopoda (e.g., some representatives of Glomerida, Julida, and Polydesmida). However, only more or less loose soils appear suitable for such myriapods because harder grounds are apparently too difficult to penetrate, especially by the actively burrowing species.

## B. Open-Country Myriapods

Short-living Polydesmida are dominant in the savanna: The adults are active during the rain season, whereas the juveniles spend the dry season in resting conditions. The resistance to desiccation varies much between species, with the females being mostly less sensitive than the males, for unknown reasons. Cuticular waxes seem to be present in very few millipede species. One of these species is the *Orthoporus ornatus* of the American semideserts; when the temperature rises over 40°C, it rests in a coiled condition to reduce water loss. Lowest lethal temperatures for millipedes are in the range of -5 to -7°C, and the highest temperatures are between 36 and 41°C. Preferred temperatures are clearly lower for hygrophilous forest species (4–18°C) than for xerophilous ones (e.g., 26–32°C for *Ommatoiulus sabulosus*, a species common in open environments in southern and eastern Europe).

## C. Special Habitat Myriapods

A few species have semiaquatic habits. One of them is the littoral julid *Thalassiosobates litoralis*, which can be found among stranded decaying matters on the seashore. However, this species does not seem to spend much time under water, as do some cavernicolous species from southern Europe (several julids and the polydesmid *Serradium semiaquaticum*; Enghoff *et al.*, 1997) and some Amazonian species, including *Myrmecodesmus adisi*, whose juveniles spend up to 11 months under water feeding on algae.

Both morphological and behavioral adaptations have been observed in myriapods that spend prolonged periods of time under water. Conspicuous are the microtrichia in their spiracles and the cerotegument covering their body which enable flood resistance for weeks and even months compared to the simple flood tolerance (up to a few days) exhibited by nonspecialized forms. These morphological adaptations are present in a few millipedes of the seasonally inundated forests of Amazonia; they are also present in the polydesmid *Serradium semiaquaticum*, which inhabits caves near Verona (northern Italy), and its sister species *S. hirsutipes*,

which does not have spiracular microtrichia and, as expected, does not have the flood resistance of its congener.

Regarding centipedes, resistance of seashore geophilomorphs to submersion in seawater is not very conspicuous, being 12–24 hr in *Strigamia maritima* and up to 36 hr in *Hydroschendyla submarina*. Longer resistance to submersion in fresh water has been shown for the lithobiomorph *Lamyctes emarginatus* from European floodplains and from populations living along the Amazon River (J. Adis, personal communication).

Behavioral adaptations to the seasonally flooded forest of Amazonia have been also recorded in the other myriapod groups. Regarding Symphyla, members of the Scolopendrellidae have a dormant stage of 5–7 months duration, which in the case of *Ribautiella amazonica* is spent among roots. Due to the anoxic conditions of the flooded soil, these arthropods are likely to recur to anaerobic metabolism. A similar situation is presumed for Pauropoda living in the same environment because *Scleropauropus tarumamirimi* has been found to spend 5–7 months under water (J. Adis and U. Scheller, personal communication). Myriapods lacking similar morphological and/or physiological adaptations must rely on vertical migrations along the tree trunks to survive the flood period in the upper trunk or canopy region [e.g., *Cutervodesmus adisi* (Diplopoda), *Hanseniella arborea* (Symphyla), and some Geophilomorpha]. The migratory behavior of these species, however, is the most conspicuous aspect of their complex ethological and physiological adaptations. No reproduction occurs during the time they spend on the trunks, and trunk ascent is associated with changes in local climatic conditions (“dry” versus rainy season) and with macroclimatic influences such as the El Niño Southern Oscillation.

A large fraction of myriapods are represented by cave dwellers. Adaptations to life in caves are often opposite those to geophily, often including larger size compared to their epigeic counterparts (Causey, 1960), frequently with strong paranota and/or hairs, elongated antennae and appendages, and reduced or missing ocelli. Among myriapods, troglobites are known mostly among millipedes and, to a lesser degree, centipedes. Many myriapods have developed adaptations to cave life and can be defined as troglobionts.

Regarding centipedes, approximately 50–60 troglotic species of lithobiomorphs are known from caves of southern Europe: the Pyrenees, Sardinia, Italy, the Balkans, Greece, and Turkey, Algeria, and Morocco. A troglotic geophilomorph (*Geophilus persephones*) has recently been described from the cave Gouffre de la Pierre Saint-Martin in the French Pyrenees; it is blind,

as expected, but in this case the blindness is not a specific adaptation to cave life because all geophilomorphs lack eyes (Foddai and Minelli, 1999).

Modified mouthparts have evolved convergently in several unrelated genera of hygrophilous cave-dwelling millipedes from southern Europe and the Caucasus (Julidae: *Leucogeorgia*, *Typhloiulus*, and *Trogloiulus*; Blaniulidae: *Vascoblaniulus*; Polydesmidae: *Serradium*). In these forms, the biting or masticatory part of the mandible is reduced, whereas its pectinate lamellae are hypertrophied, suggesting a function like a filter, used in collecting organic material suspended in water. Recent investigations on the gut content of *Serradium semiaquaticum* from Italian caves, however, have failed to substantiate this expectation (Enghoff *et al.*, 1997).

### Acknowledgments

We gratefully acknowledge the help provided by Joachim Adis with his precious comments on a draft of the manuscript and by Edward Holt Eason (deceased), Peter Johns, and Ulf Scheller, who shared with us unpublished information.

### See Also the Following Articles

ARTHROPODS, AMAZONIAN • INVERTEBRATES, TERRESTRIAL, OVERVIEW

### Bibliography

- Adis, J., and Messner, B. (1997). Adaptation to life under water: Tiger beetles and millipedes. In *The Central American Floodplain. Ecology of a Pulsating System* (W. J. Junk, Ed.), Ecological Studies No. 126, pp. 318–330. Springer, Berlin.
- Adis, J., Golovatch, S. I., Hoffman, R. L., Hales, D. G., and Burrows, F. J. (1998). Morphological adaptations of the semiaquatic millipede *Aporodesminus wallacei* Silvestri 1904 with notes on the taxonomy, distribution, habitats and ecology of this and a related species (Pyrgodesmidae Polydesmida Diplopoda). *Trop. Zool.* 11 (2), 371–387.
- Black, D. (1997). Diversity and biogeography of Australian millipedes (Diplopoda). *Mem. Mus. Victoria* 56, 557–561.
- Boore, J. L., Lavrov, D. V., and Brown, W. M. (1998). Gene translocation links insects and crustaceans. *Nature* 392, 667–668.
- Causey, N. B. (1960). Speciation in North American cave millipedes. *Am. Midl. Nat.* 64, 116–122.
- Eisner, T., Alsop, D., Hicks, K., and Meinwald, J. (1978). Defensive secretions of millipedes. In *Arthropod Venoms* (S. Bettini, Ed.), Handbook of Experimental Pharmacology No. 48, pp. 41–72. Springer, Berlin.
- Enghoff, H. (1985). Modified mouthparts in hydrophilous cave millipedes (Diplopoda). *Bijdr. Dierkde.* 55, 67–77.
- Enghoff, H. (1990). The ground-plan of the chilognathan millipedes. In *Proceedings of the 7th International Congress of Myriapodology* (A. Minelli, Ed.), pp. 1–21. Brill, Leiden.
- Enghoff, H. (1992). Macaronesian millipedes (Diplopoda) with emphasis on endemic species swarms on Madeira and the Canary Islands. *Biol. J. Linnean Soc.* 46, 153–161.
- Enghoff, H. (1994). Geographical parthenogenesis in millipedes (Diplopoda). *Biogeographica* 70, 25–31.
- Enghoff, H., and Baez, M. (1993). Evolution of distribution and habitat patterns in endemic millipedes of the genus *Dolichoulus* (Diplopoda: Julidae) on the Canary Islands with notes on distribution patterns of other Canarian species swarms. *Biol. J. Linnean Soc.* 49, 277–301.
- Enghoff, H., Dohle, W., and Blower, J. G. (1994). Anamorphosis in millipedes (Diplopoda)—The present state of knowledge with some developmental and phylogenetic considerations. *Zool. J. Linnean Soc.* 109, 103–234.
- Enghoff, H., Caoduro, G., Adis, J., and Messner, B. (1997). A new cavernicolous, semiaquatic species of *Serradium* (Diplopoda, Polydesmidae) and its terrestrial, sympatric congener. With notes on the genus *Serradium*. *Zool. Scripta* 26, 279–290.
- Foddai, D., and Minelli, A. (1999). A troglomorphic geophilomorph centipede from southern France (Chilopoda: Geophilomorpha: Geophilidae). *J. Nat. Hist.* 33, 267–287.
- Friedrich, M., and Tautz, D. (1995). Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature* 376, 165–167.
- Golovatch, S. I. (1997). On the main traits of millipede diversity in Eurasia (Diplopoda). *Senckenbergiana Biol.* 76, 101–106.
- Hoffman, R. L. (1980). *Classification of the Diplopoda*. Mus. Hist. Nat., Genève.
- Hopkin, S. P., and Read, H. (1992). *The Biology of Millipedes*. Oxford Univ. Press, Oxford.
- Kime, R. D., and Golovatch, S. I. (2000). Trends in the ecological strategies and evolution of millipedes (Diplopoda). *Biol. J. Linnean Soc.* 69, 333–349.
- Minelli, A., and Bortoletto, S. (1988). Myriapod metamerism and arthropod segmentation. *Biol. J. Linnean Soc.* 33, 323–343.
- Nguyen Duy-Jacquemin, M. (1992). Contribution à l'étude du développement postembryonnaire d'*Anadenobolus leucostigma martinicensis* (Chamberlin, 1918). Note préliminaire. *Ber. Nat.-med. Verein Innsbruck, Suppl.* 10, 133–140.
- Nguyen Duy-Jacquemin, M. (1996). Systématique et biogéographie des diplopedes pénicillates des Iles Canaries et du Cap Vert. *Mém. Mus. Natl. Hist. Nat.* 169, 113–126.







# NATURAL EXTINCTIONS (NOT HUMAN INFLUENCED)

Christopher N. Johnson  
*James Cook University*

---

- I. Introduction
  - II. Natural Causes of Extinction
  - III. Lifetimes of Species
  - IV. The Selectivity of Extinction
  - V. Geographic Patterns in Extinction
  - VI. Natural Extinction and Trends in Biodiversity
- 

## GLOSSARY

**conditions** Physical features of the environment, such as substrate type, ambient temperature, or salinity, that affect the ability of organisms to survive, grow, and reproduce.

**dispersal** The movement of an individual organism away from its place of origin to the place where it breeds (also, movement by an adult from one breeding location to another).

**pseudoextinction** The disappearance of a species from the fossil record due not to the death of all its members but to an evolutionary change that results in it being classified as a new species.

**quaternary** The past 2 million years (approximately) of Earth history, including the Pleistocene and Holocene (or Recent) epochs, and characterized by the extreme fluctuations in global temperature that produce the ice ages.

**resources** Physical and biotic features of the environment, such as shelter sites or foods, that are required by organisms and are consumed such that use by one individual reduces their availability to others.

**secondary extinction** Extinction of a species resulting from the extinction of another species on which it relies.

---

**EXTINCTION** is the ultimate fate of all species, but the vulnerability of different species to extinction is strongly affected by their biological characteristics. As a result, the durations of species vary widely. This variation helps to shape patterns of diversity among different types of organisms.

## I. INTRODUCTION

Most species have durations that are very short relative to the history of life, and it is likely that 99% of all species that have ever lived are now extinct. Extinction, therefore, is very much a natural process that produces continuous turnover in the membership of biological assemblages as species are steadily lost by extinction and replaced by speciation. Natural extinctions may be divided into two kinds: those that happened during geologically brief periods of crisis when many taxa disappeared (the mass extinction events) and those during the long intervals of “background” time between mass extinctions. Mass extinctions have attracted a great deal of interest because they bit so deeply into standing biodiversity and because the biotas that reformed after mass extinctions were often quite different from those

that existed before; they dramatically reshaped world patterns of biodiversity. However, 95% or more of the total number of extinctions have taken place outside mass extinctions (May *et al.*, 1995). These background extinctions have been most important in producing turnover in species assemblages.

Differences in the probability of extinction for different types of organisms interact with rates of origination of new lineages to shape patterns of biodiversity. To understand patterns of biodiversity it is therefore essential that the process of extinction be understood. In particular, we need to know whether extinction rates vary among different types of organisms and to identify the characteristics of taxa that make them more or less vulnerable to extinction. Extinction can be studied at the level of higher taxa—genera, families, and so on—but this article is concerned primarily with extinctions of species.

## II. NATURAL CAUSES OF EXTINCTION

The causes of extinction can be divided into two types: changes in the environment and interactions with other species.

### A. Changes in the Environment

Populations of all species fluctuate in abundance. A large part of this variation is due to chance variation in environmental conditions, such as variations in weather or events such as cyclones, which can be regarded as environmental “accidents.” Species vary in their ability to resist accidents, but for any species there will eventually come an environmental event of sufficient magnitude to wipe it out, or an unlucky succession of smaller blows that drive its abundance down to zero, even when there is no trend in average conditions.

Additionally, species may be driven extinct by directional changes in the environment, such as changes in temperature to levels outside their tolerance or the disappearance of key habitats. Species can respond to such changes in the environment by (i) acquiring adaptations to the changed conditions (“evolving out of trouble”), (ii) shifting their geographic ranges to remain within a set of suitable conditions (“moving out of trouble”), or (iii) going extinct. If the pace of environmental change is fast relative to the rate of evolution, the first of these responses is much less likely than the other two. The fossil record provides many examples of extinctions apparently driven by environmental changes of many different kinds, especially during mass

extinctions events. Environmental changes have been much more extreme and rapid during some periods of Earth history than during others, but environmental conditions have never been truly static. Global temperatures have trended downwards for much of the past 50 million years, and at a finer temporal scale variations in the earth’s orbit produce fluctuations in climate with periodicities in the range of 10,000 to 400,000 years. These are called Milankovitch cycles, and during the past 2 million years they have produced the succession of ice ages, but they must have forced rapid swings in climate throughout Earth history (Bennett, 1997).

### B. Interactions with Other Species

The population growth rates of species are constrained by the species predators, parasites, and competitors. The abundance of a species will be reduced if its natural enemies become more abundant or evolve greater efficiency, or if its geographic range is invaded by a new enemy to which it has not evolved defenses.

It is generally difficult to reconstruct such interactions in the fossil record, but some patterns may reflect the impact of predators on prey species. For example, the number of species of endobysate bivalves gradually declined through the Mesozoic. These species were abundant, immobile, and lay partly exposed on the open seafloor. Their decline coincided with radiations of marine crabs, teleost fish, and carnivorous snails, groups that account for most predation on modern bivalves and to which the endobysates must have been vulnerable (Stanley, 1979). The few present-day survivors of the group live in conditions with low predator pressure. Recent experience shows that prey species can be rapidly driven extinct by unfamiliar predators that invade their habitat. The carnivorous snail *Euglandina rosea*, for example, has caused the extinction of hundreds of species of endemic snails, including 600 of more than 1000 original species from the Hawaiian Islands, since its introduction to many Pacific islands in the 1970s. Such a rapid course of events would be unresolvable in the fossil record.

Probably, the impacts of environmental changes and of other species interact in subtle ways to cause extinction. For example, a species might experience a slight environmental change that reduces its population without driving it extinct and to which it could readily adapt given sufficient time. However, the same change might favor a competitor or trigger a range expansion by an unfamiliar predator suited to the new conditions. Interactions of this kind could mean that quite small envi-

ronmental changes could have dramatic consequences leading ultimately to extinction.

### III. LIFETIMES OF SPECIES

#### A. Variation among Taxonomic Groups

The oldest extant species known is the tadpole shrimp, *Triops cancriformis*, a small freshwater crustacean found in temporary pools in arid regions of Eurasia and north Africa that is indistinguishable from 180-million-year-old fossils bearing the same name. Most species do not live to this age, as shown in Table I.

The lifetime of a species begins with its origin in a speciation event and ends either when it dies out, leaving no descendants, or when its characteristics have been sufficiently changed by evolution that it is classified as a new species. The first type of disappearance of species is "real" extinction, and the second is referred to as pseudoextinction. This distinction is important but not easy to draw in practice. Our understanding of species life spans derives from study of the fossil record, and when a recognizable species disappears from the record it can be difficult to determine whether it has died out or evolved into something else. Information in Table I is based on real extinctions when the distinc-

TABLE I  
Durations of Species and Extinction Rates for Major Groups of Organisms in the Fossil Record<sup>a</sup>

Taxon	Estimated mean species duration (millions of years)	Extinction rate (extinctions per million species years)
Single-celled organisms		
Diatoms	8	0.12
Dinoflagellates	13	0.08
Planktonic foraminifera	7–20	0.14–0.05
Benthic foraminifera	25	0.04
Plants		
Early vascular plants	10–14	0.1–0.07
Pteridophytes	7–15	0.14–0.07
Gymnosperms	2–15	0.5–0.07
Monocots	4	0.25
Dicots	3	0.33
All invertebrates	11	0.09
Reef corals	20	0.05
Mollusks:		
Marine gastropods	10	0.1
Marine bivalves	15	0.07
Mesozoic ammonites	1	1
Upper Cambrian trilobites	1.3	0.77
Marine ostracods	8	0.12
Silurian graptolites	2	0.5
Echinoderms		
Echinoids	6	0.17
Crinoids	6.7	0.15
Bryozoans	12	0.08
Freshwater fish	3	0.33
Birds	2.5	0.42
Mammals		
Cenozoic mammals	1–2	1–0.5
Horses	4	0.25
Primates	1	1

<sup>a</sup> Compiled from Stanley (1979), Niklas *et al.* (1993), May *et al.* (1995), and McKinney (1997).

tion can be made, but most estimates include an unknown combination of real and pseudoextinctions.

There is considerable variation in species life spans among different groups of organisms. In general, it seems that species of unicellular organisms last longer than species of multicellular organisms, and invertebrates appear to last longer than vertebrates. The most long-lived animal groups tend to be marine rather than terrestrial, although too little is known about species durations of marine vertebrates or terrestrial invertebrates to judge differences between marine and terrestrial species independent of the difference between vertebrates and invertebrates. Some marine invertebrate groups that are now entirely extinct (the trilobites, ammonites, and graptolites) had short species durations, even though they were successful, abundant, and species rich for long periods of geological time. Among plants, species durations were longer in groups that appeared early in the evolutionary history of plants but have tended to shorten in recently evolved groups.

The inverse of the typical duration of species in a given taxonomic group can be used as a measure of the probability of extinction per unit time for species belonging to that group: Long durations equate to low probabilities of extinction. In Table I, probabilities of extinction have been expressed as the rates of extinction per million species years. Thus, marine gastropods typically last 10 million years, so a sample of 1 million marine gastropod species would be expected to experience 0.1 extinctions per year.

## B. Variation within Taxonomic Groups

The distributions of species lifetimes within higher taxa are typically right-skewed: Most species have short durations, but a small minority are long-lasting (Fig. 1). There could be two causes for this pattern. First, only a few species might have characteristics that make them intrinsically resistant to extinction, whereas most are sensitive to extinction risk. Second, species might not vary in their susceptibility to extinction, but if many independent factors can cause extinction only a small minority of species will be lucky enough to avoid extinction for long. Probably both factors play a part in shaping distributions of species durations.

## C. Risk of Extinction in Relation to Species Age

Does the likelihood that a species will go extinct depend on its age? One might think that it should because the longer a species stays in existence the more adaptations to its environment it can accumulate; older species should therefore be better at avoiding extinction. This idea can be tested by examining species survivorship curves. A species survivorship curve shows the relationship between age and the proportion of species in a taxon surviving to each age. Such curves are typically log linear; that is, when the proportion surviving is plotted logarithmically its decline with age approximates a straight line. This shows that the proportion

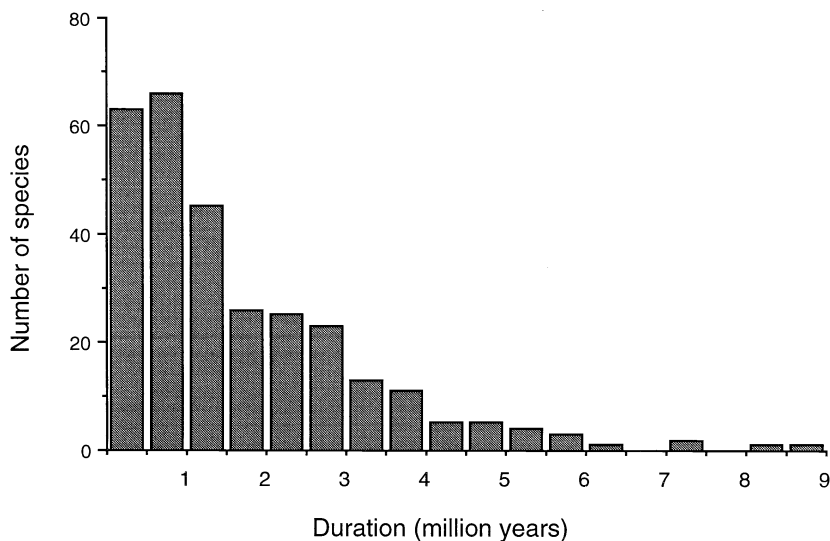


FIGURE 1 Species durations of Silurian graptolites of the British Isles (From *Macroevolution: Pattern and Process*, Stanley, S. M. © 1979 by W. H. Freeman and Company. Used with permission.)

of species going extinct is approximately constant with respect to age.

The constancy of extinction probability with species age can be explained by assuming that the environment of any species is continually changing so that even if the species continually adapts, its fitness lags behind the current condition of the environment. Thus, no matter how long a species lasts it can never reach a condition of optimal adaptation. There are two forms of this explanation. The first assumes that the probability of extinction for a species is determined primarily by its interactions with other species and that each species is under constant selection pressure to increase its fitness relative to these other species. However, adaptations in one species that increase abundance at the expense of other species will be met by counteradaptations from them so that, on average, no species improves its fitness with time. This is the Red Queen hypothesis, named after the character in Lewis Carroll's *Alice Through the Looking Glass* in whose world "it takes all the running you can do, just to stay in one place." The Red Queen hypothesis predicts that extinction rates will be approximately constant with respect to absolute time. The second form of the explanation assumes that it is the abiotic environment that is continually changing. Changes in the abiotic environment will tend to be episodic and will therefore produce fluctuations in extinction rates with time, but their effects will still be independent of species age. It is not clear which form of the hypothesis is closer to the truth.

#### IV. THE SELECTIVITY OF EXTINCTION

Extinction may be largely a matter of the bad luck of environmental change, exacerbated perhaps by pressure from other species; however, the fact that species durations vary so widely among higher taxa suggests that biological characteristics of species influence their susceptibility to extinction. Three types of studies have helped to identify these characteristics.

First, the attributes of species that have short versus long durations in the fossil record can be compared. A common variant of this technique compares species that persisted through mass extinction events with those that did not. Second, the living faunas of "land-bridge islands" can be compared with their presumed original faunas. A land-bridge island is an island formed as a result of isolation from the mainland by rising sea levels during the Pleistocene. Such islands typically have fewer species than do similar areas of mainland habitat to which they were once connected, reflecting

extinctions caused by the pressure of reduced habitat area. Third, extinction of local populations can be observed directly in contemporary ecological studies. The second and third approaches focus on population rather than species extinctions, but the extinction of a species is the end point of a series of population extinctions, so such studies may still reveal traits that correlate with risk of extinction at the species level.

The following traits consistently emerge from many studies of the selectivity of extinction: rarity, dispersal ability, body size, and specialization.

##### A. Rarity

The commonness or rarity of a species is a function of two factors: its geographic range and its population density where it occurs. Both components of rarity influence extinction risk.

###### 1. Geographic Range

There is strong evidence from the fossil record (mainly for marine invertebrates) that species with large ranges have lower extinction rates (Jablonski, 1995; McKinney, 1997). There are probably two causes for this effect. First, at a given scale, a disturbance will affect a smaller proportion of the range of a widespread rather than of a geographically restricted species. In the extreme, a single catastrophic event may wipe out a localized species but have little impact on a widespread species. Second, even if an environmental change affects a very large area, it is likely that some populations of widespread species will persist in isolated refuges that provide some local protection from the change. Species that hang on in such refuges may reinvade their original ranges once conditions improve. This is probably part of the explanation for the existence in the fossil record of "Lazarus species" that vanish during periods of environmental crisis and then reappear much later. The effect of range size on extinction risk tends to be strongest for background extinctions, and it may weaken or disappear in mass extinctions. This seems to have been the case for marine mollusks, for example (Jablonski, 1991), and is probably because mass extinctions were caused by events that affected such large areas and had such profound impacts that even widespread species were susceptible to them.

###### 2. Local Abundance

Population density is generally not well represented in the fossil record, but patterns of extinction of populations on land-bridge islands and in the present day show that local extinction is more likely for species

with low population densities. For example, Foufopoulos and Ives (1999) found that reptile species with low population densities were more likely to go extinct from land-bridge islands in the Mediterranean Sea. The causes of extinction of small populations have been widely discussed in the literature on conservation. Briefly, small populations are more vulnerable than large populations because (i) they may go extinct more quickly when chance environmental variation causes fluctuations in abundance; (ii) they are affected by chance demographic fluctuations, such as occasional production of biased sex ratios of offspring, that would be averaged out in large populations; (iii) they may lose genetic variation and experience high levels of inbreeding and inbreeding depression; and (iv) they may be subject to Allee effects—that is, social or reproductive dysfunction as a direct result of low numbers.

### 3. The Relationship between Range and Abundance

The components of rarity—geographic range size and population density—are generally positively related among species: Species with small geographic ranges also tend to have low population densities, and wide-ranging species have high population densities. This pattern has been found in many terrestrial plant and animal taxa, although it remains almost unstudied in marine organisms. Why there should be a positive relationship between range and abundance is not clear. Several ecological mechanisms have been proposed as its cause, but two ideas have been especially influential. First, Brown (1995) argued that niche breadth is positively correlated with both range size and population density because species with broad niches can exist under a wider range of conditions and use more types of resources than can species with narrow niches; variation among species in niche breadth therefore produces a positive relationship between range and density. Second, species that reach high local densities should be both more resistant to extinction on habitat patches because their populations are larger and produce more migrants that are able to recolonize habitat patches after local extinctions (Hanski, 1999). High local abundance therefore results in wide geographic distribution.

To the extent that geographic range and local abundance affect extinction risk independently of one another, the correlation between them should exaggerate differences among species in extinction risk. Rare species face double jeopardy: The vulnerability that comes from having a small range is compounded with the vulnerability due to low abundance. Common species, on the other hand, are likely to be highly resistant to

extinction because they combine high abundance with large ranges. For example, Johnson (1998) showed that among Australian marsupials, ancient species (those that diverged from their closest relative more than 4 million years ago) are very unlikely to have both small ranges and low population densities, although this combination occurs frequently among young species (Fig. 2). This suggests that species with low abundance and small ranges tend to be short lived. There are ancient marsupials with small ranges but they have unusually high densities, and conversely there are species with low densities but they have unusually large ranges,

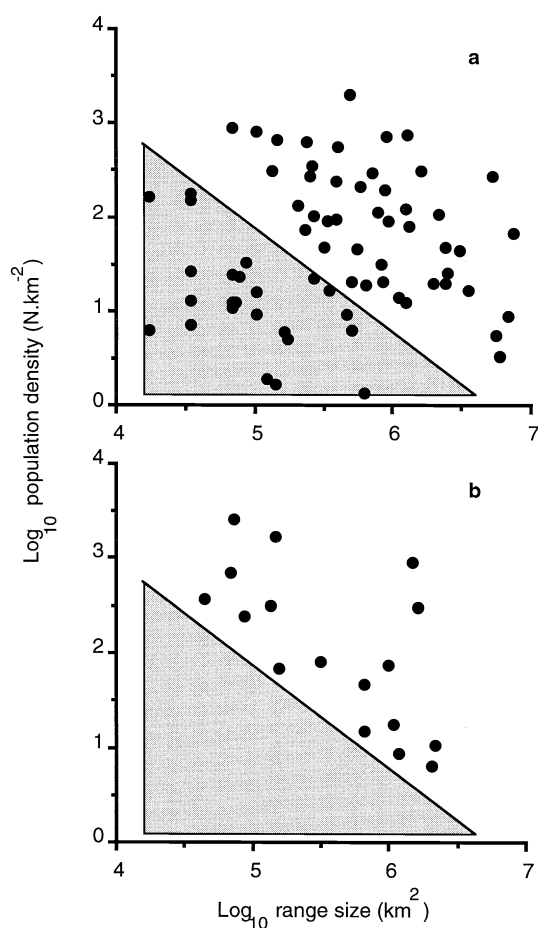


FIGURE 2 Relationships between size of geographic range and population density for species of Australian marsupials: (a) all species other than those defined as ancient and (b) ancient species only. Ancient species are those that diverged from their closest living relative more than 4 million years ago and have therefore demonstrated resistance to extinction. The shaded triangle defines the region of distribution-abundance space from which species would need to have gone extinct to produce the pattern observed among ancient species (reproduced with permission from *Nature*, Johnson, copyright 1998 Macmillan Magazines, Ltd.).

implying that a high density can compensate for the vulnerability that comes from having a small range and vice versa.

#### 4. Spatial Structure of Populations

Another aspect of rarity is ubiquity or patchiness of distribution within the geographic range. The distributions of practically all species are discontinuous at some scale, but the degree of patchiness varies, reflecting specialization for habitats or resources that are themselves patchily distributed. In a species that has a patchy distribution, metapopulation dynamics are likely (Hanski, 1999). That is, the risk of extinction of populations on discrete patches of habitat may be high, but local populations are eventually reestablished by migrants from other populations. Recolonization will involve a time lag that depends on the distance between patches and the dispersal ability of the species.

A general deterioration in the quality of habitat for a species is likely to cause contraction of larger habitat patches and the disappearance of smaller ones, causing local extinctions and simultaneously increasing the distance between surviving patches and thus reducing the probability of migration. Patchily distributed species should therefore be more likely to go extinct as a result of environmental change than should continuously distributed species. The sensitivity of patchy species to environmental change has been demonstrated for insects in Great Britain subject to human-caused changes (Webb and Thomas, 1994), but it is likely to be general to any form of shift in environmental conditions. Patchiness of distribution is loosely correlated with the other components of rarity so that widespread and locally abundant species tend to be continuously distributed within their ranges, whereas rare species are likely to be patchy.

### B. Dispersal Ability

Dispersal ability is positively associated with species longevity for two reasons. First, species that disperse widely tend to have large geographic ranges, and their resistance to extinction can be a direct result of large range. Marine gastropod snails, for example, have two different forms of development. In some, the egg is released into surface waters and develops into a larval form that feeds in the plankton and drifts widely before settling and developing into an adult snail. This form of development promotes wide dispersal. Others have direct development in which eggs and young grow up close to the parent. Species with planktonic development have larger geographic ranges and longer dura-

tions in the fossil record than do species with direct development (Jablonski, 1995).

Second, species that disperse widely are better able to recolonize after going extinct from part of the range, and they can more rapidly shift their ranges to track changes in the distribution of their preferred climates and habitats. During the Quaternary in the Northern Hemisphere, the distribution of habitats changed very rapidly as ice sheets repeatedly advanced and receded. These changes produced remarkably few extinctions among Northern Hemisphere beetles. Instead, species shifted their geographic ranges to keep pace with the changing distribution of habitats, an effect that was more dramatic in flighted than in flightless species (Coope, 1995). Probably, dispersal ability and local abundance interact to confer high resistance to extinction because high local abundance means that large numbers of dispersers are produced, resulting in the potential for rapid shifting of range boundaries as conditions change. This combination of characteristics might protect species that have specialized habitat requirements and small geographic ranges and would otherwise be vulnerable to extinction.

### C. Body Size

It often seems to be the case that large-bodied species are at higher risk of extinction than small-bodied species. Although there are exceptions, this pattern is reasonably consistent in the fossil record both for background extinctions and for mass extinctions (McKinney, 1997), and it also emerges in extinctions on land-bridge islands (Brown, 1995). There are several reasons why large-bodied species might be particularly vulnerable to extinction:

1. Potential rate of population increase declines with body size because large-bodied species generally have longer generation times and lower fecundity than do small-bodied species. This means that large-bodied species will generally recover more slowly from population declines, and because they produce fewer offspring they may be slower to track habitats than will small-bodied species.

2. Individuals of large-bodied species generally need larger areas, and large-bodied species are therefore strongly affected by declines in habitat area.

3. Population density tends to decline with body size so that large species are often naturally rare. This is partly compensated by a tendency for size of geographic range to increase with body size, but this relationship is often weak or absent so that total population size is



usually much less for large than small species. The decline in population density with body size is clearest when species of very different size (e.g., mice and elephants) are compared, but in guilds of ecologically similar species it is often the case that density increases with body size, possibly because larger species are better competitors for resources (Cotgreave, 1993). At this finer scale of comparison, therefore, large-bodied species might be more resistant to extinction than small-bodied species.

## D. Specialization

Species can be classed as “specialists” or “generalists” on three criteria: the range of conditions that they are adapted to tolerate, the range of resources that they are able to use, and the degree of their evolved dependence on a small number of other species.

### 1. Conditions

Species that are narrowly adapted to environmental conditions are likely to be the first to go extinct when the environment changes. This is difficult to demonstrate in the fossil record, however, because it is not possible to measure directly the environmental tolerances of fossil species. Instead, environmental tolerances are inferred from geographic distributions. Species with small geographic ranges will necessarily occupy a narrow range of climate zones and habitats, but it does not follow from this that their environmental tolerances are narrow. Some species that could potentially be widely distributed have small ranges because of population history, geographic barriers to movement, or poor dispersal ability or because range expansion is prevented by interactions with other species (competitors, predators, and so on). Therefore, although there is abundant evidence from the fossil record that species with small geographic ranges are extinction-prone, there is much less evidence for the commonsense view that the breadth of tolerance of environmental conditions directly affects extinction risk.

### 2. Resources

Species that depend on a narrow range of types of resources are likely to be more sensitive to changes in resource abundance than generalists that can easily switch resources. This vulnerability may be partly compensated by the fact that specialization on a particular resource may be more likely to evolve if that resource is abundant and widespread. For example, feeding on grasses by mammals requires extensive specialization of the teeth and digestive system, and increases in body

size, to overcome the abrasiveness and poor nutritional quality of grasses. However, because of the abundance of grasses many mammals have evolved these adaptations, and grazing mammals typically have high local abundance and large geographic ranges. Nonetheless, grazing mammals in Africa have suffered higher extinction rates since the Miocene than mixed grazer/browsers, as the extent of grasslands has fluctuated during the Pliocene and Pleistocene (Vrba, 1987).

### 3. Interactions

Specialization is taken a step further in species that have evolved a close dependence on one or a small number of other species. For example, many herbivorous insects feed on only one species of plant, and many predators attack only one or a small number among many possible species of prey. Of particular interest are mutualistic interactions, in which species provide benefits to one another. Figs, for example, rely on fig wasps for pollination, and fig wasps in turn lay their eggs only in the flowers of figs. This interaction tends to be highly species specific, with each species of fig visited by only one species of fig wasp and each partner in the interaction completely dependent on the other for reproduction. Such tight species specificity results from coevolution, in which each species in the interaction evolves special characteristics in response to evolutionary changes in the other to increase its benefit from the interaction.

Specialization of this kind is classically regarded as an extinction trap because the specialist will inevitably go extinct if the species that it depends on goes extinct or becomes very rare. This view is probably an oversimplification. Careful study of some species-specific interactions has revealed more flexibility and greater potential for rapid evolutionary response to changes in the abundance of interacting species, including the ability to switch to new partners, than was previously assumed (Thompson, 1994). Also, mutualistic interactions have the general effect of increasing the geographic range and abundance, and stabilizing the population dynamics, of both partners in the interaction. These ecological benefits may at least partly compensate for the vulnerability caused by dependence on another species.

Because interactions between species are not revealed in detail in the fossil record, there is little direct information on rates of secondary extinction. It is sometimes possible, however, to evaluate the risks of secondary extinction from study of living communities. Many plants cannot set seed without cross-pollination and rely on animals to transfer pollen. Such plants should be vulnerable to reproductive failure, and possibly ex-

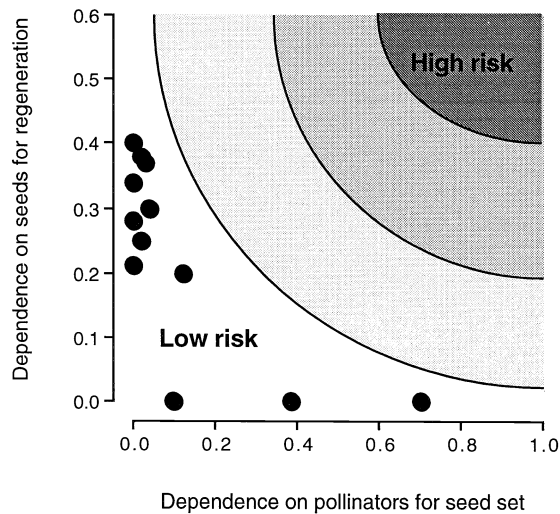


FIGURE 3 The relationship between an index of dependence on animal pollinators for seed set and dependence on seeds for regeneration for species of spring flowering herbs in a temperate deciduous forest. Any species that was both dependent on pollinators for seed set and dependent on seeds to regenerate would be considered at high risk of extinction due to pollinator failure (redrawn from Bond, 1995).

tion, if their pollinators decline, and species that have only one or a small number of pollinators are likely to be especially vulnerable. This vulnerability can be reduced by traits such as the ability to propagate vegetatively that reduce demographic dependence on seeds. Bond (1995) showed that in some plant communities there are no species that have both a high dependence on animal pollinators for seed set and a high demographic dependence on seeds, despite wide variation among species in both sets of characteristics (Fig. 3). One explanation for this pattern is that such species are unusual because their high dependence on other species means that they are very likely to go extinct.

### E. Reproductive and Life History Traits

A wide variety of reproductive and life history traits may influence the likelihood of extinction in some circumstances. Perhaps the most important is sexual reproduction. Sexual lineages appear able to evolve more quickly than parthenogenetic ones because genetic recombination generates more variability among individuals on which natural selection can act. Sexual species may therefore be better able to adapt to changed environmental conditions or to meet new challenges from other species. The taxonomic distribution of parthenogenetic lineages suggests that they are short lived: With few exceptions, no higher taxa (genera, families, etc.)

consists solely of parthenogens; almost all parthenogens have close sexual relatives and are likely to be recently derived from sexual species (Maynard Smith, 1978).

Species with resting or resistant stages in their life cycle may be better able to pass through environmental catastrophes than species in which all life stages are vulnerable to environmental changes. In the islands of the Lesser Antilles, for example, butterflies are better able to persist on small islands than are birds or mammals, perhaps because they are able to pass through unfavorable seasons as diapausing eggs and pupae. Similarly, reptiles and amphibians are resistant to extinction, possibly by virtue of their ability to survive for long periods in protected microenvironments without having to feed (Ricklefs and Lovette, 1999).

## V. GEOGRAPHIC PATTERNS IN EXTINCTION

Natural extinction rates vary along two major geographic axes: Extinction rates are lower for marine than terrestrial organisms, and extinction rates may be higher near the equator than at high latitudes. The low extinction rates of marine organisms are presumably due to their large geographic ranges—the oceans are bigger and less subdivided than the continents—and the high dispersal ability of many species that have larval life stages that drift in the plankton. The effects of latitude on extinction rates are less distinct, but where extinction rates have been shown to vary with latitude they tend to be higher in the tropics. Tropical species often have smaller geographic ranges and lower population densities than species at higher latitudes and may thus be more vulnerable to the effects of environmental change (Lawton, 1995). Also, some particularly species-rich tropical environments, such as the low-sediment, low-nutrient, shallow-water platforms that support reef and related communities, are easily disrupted by changes in climate and sea level (Jablonski, 1995).

## VI. NATURAL EXTINCTION AND TRENDS IN BIODIVERSITY

The diversity of any lineage of organisms through time reflects the balance of the rates of loss of species by extinction and gains by speciation. Because extinction rates vary among lineages, extinction is an important factor shaping patterns of biodiversity. Much of the explanation for the very high diversity of insects is that,

owing presumably to traits such as small size, high abundance, and high mobility, they have low extinction rates. Essentially no insect families have gone extinct during the past 50 million years, and some living genera and species with good fossil records are of Miocene age or older (Labandeira and Sepkoski, 1993). Species diversity of insects has therefore accumulated over long periods of time. Mammals, on the other hand, appear to evolve and speciate rapidly, but they are a minority taxon because their extinction rates are high.

Insects and mammals provide an extreme example of contrasts in the balance of extinction and speciation rates, but in general extinction and speciation rates are positively correlated among different lineages (Fig. 4). Such a relationship could arise if total species number in a lineage is held at equilibrium by competition between species so that a new species can only establish by occupying a niche left vacant by a species that has gone extinct. However, the relationship was first shown for groups of animals undergoing rapid increase in diversity (Stanley, 1979). There are three possible causes of the correlation between rates of speciation and extinction:

1. The effects of dispersal and gene flow: Dispersal rates should be correlated with resistance to extinction, but the gene flow resulting from high dispersal should also prevent genetic divergence of populations, thus impeding allopatric speciation. This explanation may apply to marine gastropods during the Late Cretaceous, when lineages with planktonic larvae and large ranges not only were less likely to go extinct but also were less likely to produce new species than those with direct development and restricted dispersal (Hansen, 1983).

2. The effects of niche breadth: Ecologically specialized species are likely to be vulnerable to environmental change, but they also tend to have patchy distributions, and this spatial subdivision should promote divergent evolution of local populations and the generation of many species. Generalists may be resistant to extinction, but their lack of spatial population structure also impedes divergent evolution among local populations.

3. The effects of rarity: Species with small populations are prone to extinction. Somewhat controversially, it is believed that speciation may be enhanced by small or fluctuating population sizes (Futuyma, 1998).

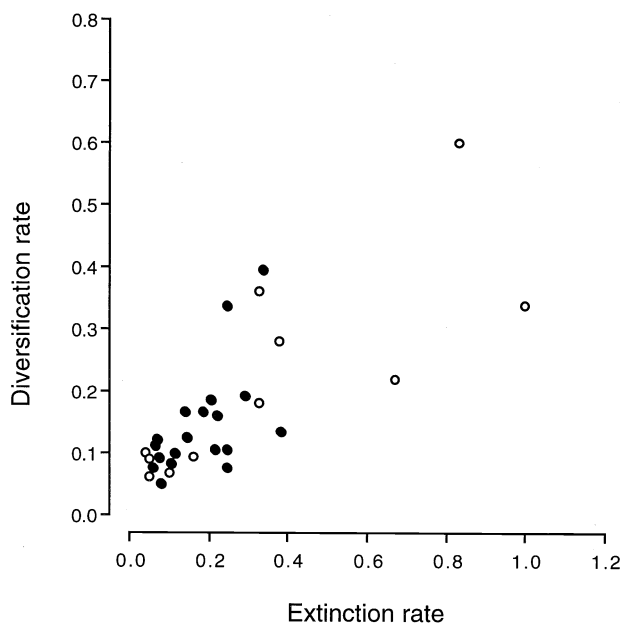


FIGURE 4 The relationship between extinction rate and diversification rate for lineages of animals (○) and plants (●). Because diversification rate equals speciation rate minus extinction rate, the positive relationship shows that rates of extinction and speciation must be positively correlated (if not, the relationship on the graph would be negative). Extinction rates are measured per million species years and calculated as the reciprocal of species durations [from data in Stanley (1979) and Niklas *et al.* (1983)].

Because extinction is selective for particular characteristics of species, it can act as a filter through which certain types of species are more likely to pass. Selective extinction therefore shapes the composition of ecological assemblages of species. This filtering effect can be seen most clearly in mass extinctions, which typically leave in their wake "survival faunas" dominated by small numbers of wide-ranging and abundant generalist species of a few simple ecological types. The examples shown in Figs. 2 and 3 illustrate some more subtle effects of the continuous operation of extinction filters, shaping respectively the patterns of distribution and abundance in living marsupials and the combinations of demographic characteristics of living plants.

In some cases, directional trends in evolution and the development of communities can be opposed or curtailed by selective extinction. This effect can produce taxon cycles, which occur if derived species tend to evolve characteristics that make them vulnerable to extinction. For example, species that colonize islands are often generalists that are abundant and competitively dominant and quickly become widespread. Their descendants, however, tend to become more specialized and less mobile, occupy fewer habitats, lose their competitive ability, and sustain smaller and more subdivided populations. These trends eventually lead to extinction and the repetition of the cycle as sets of derived species are replaced by newly invading generalists (Ricklefs and Miller, 1999). Taxon cycles are most visi-

ble on island chains, where ecological communities are relatively simple and the patterns of distribution of related species on different islands provide clues to the direction of evolution. However, similar processes probably operate over larger areas on continents and in the oceans.

What can the study of natural extinctions teach us about the coming wave of human-caused extinctions? There are at least two messages. First, species durations among different taxa in the fossil record are positively related to rates of endangerment of living species in the same taxa (McKinney, 1997). This suggests that the characteristics that have made species vulnerable to extinction under natural conditions also make them sensitive to human impacts on the environment, and in the absence of better information these characteristics could be used to identify species likely to be at risk in the near future. Second, the selectivity of extinction means that extinction rates will be higher in certain ecological types and taxonomic groups than in others, and if environmental pressures continue for a sufficient amount of time some groups may be removed entirely. The result will be a much greater impoverishment of biodiversity than would be the case if extinctions were randomly distributed among species.

### See Also the Following Articles

EXTINCTION, CAUSES OF • EXTINCTION, RATES OF • MAMMALS, LATE QUATERNARY, EXTINCTIONS OF • MAMMALS, PRE-QUATERNARY, EXTINCTIONS OF • MASS EXTINCTIONS, CONCEPT OF • POPULATION DENSITY • SPECIES INTERACTIONS

### Bibliography

- Bennett, K. D. (1997). *Evolution and Ecology: The Pace of Life*. Cambridge Univ. Press, Cambridge, UK.
- Bond, W. J. (1995). Assessing the risk of extinction due to pollinator and disperser failure. In *Extinction Rates* (J. H. Lawton and R. M. May, Eds.), pp. 131–147. Oxford Univ. Press, Oxford.
- Brown, J. H. (1995). *Macroecology*. Univ. of Chicago Press, Chicago.
- Coope, R. G. (1995). Insect faunas in ice age environments: Why so little extinction? In *Extinction Rates* (J. H. Lawton and R. M. May, Eds.), pp. 55–74. Oxford Univ. Press, Oxford.
- Cotgreave, P. (1993). The relationship between body size and population abundance in animals. *Trends Ecol. Evol.* 8, 244–248.
- Erwin, D. H. (1998). The end and the beginning: Recoveries from mass extinctions. *Trends Ecol. Evol.* 13, 344–349.
- Foufopoulos, J., and Ives, A. R. (1999). Reptile extinctions on land-bridge islands: Life-history attributes and vulnerability to extinction. *Am. Nat.* 153, 1–25.
- Futuyma, D. J. (1998). *Evolutionary Biology*. Sinauer, Sunderland, MA.
- Hansen, T. A. (1983). Modes of larval development and rates of speciation in early Tertiary neogastropods. *Science* 220, 501–502.
- Hanski, I. (1999). *Metapopulation Ecology*. Oxford Univ. Press, Oxford.
- Jablonski, D. (1991). Extinction: A paleontological perspective. *Science* 253, 754–757.
- Jablonski, D. (1995). Extinction in the fossil record. In *Extinction Rates* (J. H. Lawton and R. M. May, Eds.), pp. 25–44. Oxford Univ. Press, Oxford.
- Johnson, C. N. (1998). Species extinction and the relationship between distribution and abundance. *Nature* 394, 272–274.
- Labandeira, C. C., and Sepkoski, J. J. (1993). Insect diversity in the fossil record. *Science* 261, 310–315.
- Lawton, J. H. (1995). Population dynamic principles. In *Extinction Rates* (J. H. Lawton and R. M. May Eds.), pp. 147–163. Oxford Univ. Press, Oxford.
- May, R. M., Lawton, J. H., and Stork, N. E. (1995). Assessing extinction rates. In *Extinction Rates* (J. H. Lawton and R. M. May, Eds.), pp. 1–24. Oxford Univ. Press, Oxford.
- Maynard Smith, J. (1978). *The Evolution of Sex*. Cambridge Univ. Press, Cambridge, UK.
- McKinney, M. L. (1997). Extinction vulnerability and selectivity: Combining ecological and paleontological views. *Annu. Rev. Ecol. Syst.* 28, 495–516.
- Niklas, K. J., Tiffney, B. H., and Knoll, A. H. (1983). Patterns in vascular land plant diversification. *Nature* 303, 614–616.
- Ricklefs, R. E., and Lovette, I. J. (1999). The roles of island area *per se* and habitat diversity in the species–area relationships of four Lesser Antillean faunal groups. *J. Anim. Ecol.* 68, 1142–1160.
- Ricklefs, R. E., and Miller, G. L. (1999). *Ecology*, 4th ed. Freeman, New York.
- Stanley, S. M. (1979). *Macroevolution: Patterns and Process*. Freeman, San Francisco.
- Thompson, J. N. (1994). *The Coevolutionary Process*. Univ. of Chicago Press, Chicago.
- Vrba, E. S. (1987). Ecology in relation to speciation rates: Some case histories of Miocene–Recent mammal clades. *Evol. Ecol.* 1, 283–300.
- Webb, N. R., and Thomas, J. A. (1994). Conserving insect habitats in heathland biotopes: A question of scale. In *Large-scale Ecology and Conservation Biology* (P. J. Edwards, R. M. May, and N. R. Webb, Eds.), pp. 129–152. Blackwell, Oxford.





# NATURAL RESERVES AND PRESERVES

Alexander N. Glazer  
*University of California at Berkeley*

---

- I. Origins of Protected Areas
  - II. Classification
  - III. Nature Alongside Humanity
  - IV. Research in the National Parks
  - V. Biological Field Stations
  - VI. The Long-Term Ecological Research Network
  - VII. The International Long-Term Ecological Research Network
  - VIII. Sustained Ecological Research
  - IX. GAP Analysis
- 

change and of undesirable outcomes of human activities such as acid rain. There are many different types of protected areas: national parks, refuges, and sanctuaries, both terrestrial and marine. There are also numerous biological field stations whose primary mission is to support university-level research and instruction. This article describes protected areas as well as such research sites. Examples are also given of the types of scientific research carried out in the various protected areas and in field stations and of the contribution of research to the understanding and management of the biosphere.

## GLOSSARY

- endemic** Confined to a particular region or locality; native (not introduced).
- hectare** A metric measure of surface equal to 10,000 square meters or 2.471 acres (U.S.).
- 

**NATURAL BIOLOGICAL COMMUNITIES**, undisturbed by humans, are very valuable. Our understanding of ecological processes is still rudimentary and areas in which natural communities are protected are irreplaceable sites for the scientific study. In total, such areas also make a very important contribution to slowing the loss of the earth's biological diversity. Protected sites serve as ecological and baseline control areas where long-term studies provide early warnings of climatic

## I. ORIGINS OF PROTECTED AREAS

Preserves, such as sacred groves or royal hunting parks, were established in ancient times, but protection of natural habitat in those areas was incidental. Preservation of natural biological communities for their own value originated in the United States relatively recently. The world's first national park, Yellowstone, was established in 1872. It occupies more than 8000 km<sup>2</sup> mainly in the northwestern corner of Wyoming, and it extends into Idaho and Montana. The Yellowstone Act of 1872 reflected the contemporary faith in the inexhaustibility of natural resources. The act prohibited the wanton destruction of fish and game for profit, but it permitted hunting, trapping, and fishing for recreation or to provide food for park residents or visitors. Soon after, several other nations established national parks: Australia (Royal National Park, 1879), Canada (Banff National

Park, 1885), and New Zealand (Tongariro National Park, 1894). There are currently approximately 1400 national parks in the world, which are among the most popular tourist destinations. In 1998, more than 275 million people visited the national parks in the United States alone.

In Africa, and to a lesser extent in the Indian subcontinent, steps toward conservation during the colonial era emphasized large mammals, particularly those which had been intensively hunted. For example, Jim Corbett National Park, a haven for tigers in the Himalayan foothills of Uttar Pradesh, was established in 1936. It is named in honor of Jim Corbett, a well-known hunter-naturalist, who became an author and a photographer and helped define the park's boundaries. In contrast, in long-settled parts of Europe, protection was extended to humanized landscapes.

Protection has also been given to mangroves and coral reefs. Mangroves are found in the tropics and subtropics on river banks and along coastlines. Mangrove plants include trees, shrubs, ferns, and palms, which have adapted to the largely anaerobic conditions of muddy, aquatic environments by producing stilt roots which project above the water to absorb oxygen. Approximately 900 protected areas include mangroves. Coral reefs, many formed over periods of thousands of years, cover between 300,000 and 600,000 km<sup>2</sup> mainly between the Tropics of Cancer and Capricorn. The Great Barrier Reef Marine Park off the coast of Queensland, Australia, covers 345,000 km<sup>2</sup> of reefs and surrounding waters.

The diverse national parks, reserves, and preserves throughout the world, now referred to as "protected areas," comprise more than 5% of the earth's surface.

## II. CLASSIFICATION

The International Union for the Conservation of Nature (IUCN) has defined a protected area as "an area of land and/or sea especially dedicated to the protection and maintenance of biological diversity, and of natural and associated cultural resources, and managed through legal or other effective means." IUCN noted that the purposes for which particular protected areas are managed vary widely and noted specifically (i) scientific research, (ii) wilderness protection, (iii) preservation of species and genetic diversity, (iv) maintenance of environmental services, (v) protection of specific natural and cultural features, (vi) tourism and recreation, (vii) education, (viii) sustainable use of resources from natural ecosystems, and (ix) maintenance of cultural and tradi-

tional attributes. IUCN classified protected areas into the six categories listed in Table I reflecting the management purpose. Note that IUCN categories IV–VI encompass areas where limited or extensive use of biological resources is allowed. The following discussion of the characteristics of representative protected areas belonging to each of the categories in Table I clarifies the distinctions.

The subantarctic Snares Islands are located 209 km southwest of Bluff on South Island of New Zealand. The Snares are classified as Ia (Strict Nature Reserve) because the vegetation is nearly pristine and the islands are free of introduced mammals. The total area of the Snares Islands Nature Reserve is 328 ha, covering the islands and their foreshores. The islands are home to millions of breeding seabirds, including the endemic Snares crested penguin, and are strictly protected. Research and monitoring of natural plant and animal communities that does not cause long-term disturbance or damage is permitted. There is no access for tourists.

NorthEast Svalbard Nature Reserve, owned by the Norwegian government, exemplifies category 1b (Wilderness Area). The reserve includes the islands of Nordaustlandet, Kvitoya, and Kong Karls Land and the surrounding territorial waters, a total of 1.9 million ha, in the northeast of the Svalbard archipelago. The archipelago extends approximately from 74° to 81°N and 10° to 35°E, an Arctic area about one and a half times the size of Switzerland. This wilderness is covered by high arctic tundra vegetation, with many species near their northernmost distribution in Europe. The large mammals include polar bear, reindeer, and walrus. Access is limited to nonintrusive scientific research and recreation.

Canaima National Park (category II, National Park), owned by the Venezuelan government, covers 3 million ha centered on the Guayana Shield in Bolívar State, south of the Orinoco River. Canaima includes the uplands of the Grand Sabana and the eastern tepuis of the Roraima Range. The tepuis are flat-topped mountains with almost vertical slopes. The world's highest waterfall, Angel Falls, with a fall of 102 m, descends from one of these mountains. The summits of the tepuis have many endemic species. A small population of traditional occupants of the area, the Pémon, continues swidden (slash-and-burn) agriculture, hunting, and gathering. The park preserves representatives of the geological, biological, and cultural features specific to the Guayana Shield. Hunting and collecting of wildlife is forbidden. Recreational activities, research, and education are encouraged.

Ngerukewid Islands Wildlife Preserve (category III,

TABLE I  
International Union for the Conservation of Nature Management Categories of Protected Areas (1996)<sup>a</sup>

Category	Description	Approximate number	Approximate area (ha)
I	Ia. Strict Nature Reserve: managed mainly for science Ib. Wilderness Area: managed mainly for wilderness	1460	86,474,000
II	National Park: managed mainly for ecosystem protection and recreation	2040	376,784,000
III	Natural Monument: managed mainly for conservation of specific natural features	250	13,686,000
IV	Habitat/Species Management Area: managed mainly for conservation through management intervention	3810	308,314,000
V	Protected Landscape/Seascape: managed mainly for conservation and recreation	2275	141,092,000
VI	Managed Resource Protected Area: managed mainly for the sustainable use of natural ecosystems	Not available	

<sup>a</sup> See <http://www.wcmc.org.uk> for updated information and information by country.

National Monument) is an archipelago of high limestone islands sharing a common reef platform. The islands belong to Koror State, Palau. The reserve covers 1200 ha, of which 90 ha is terrestrial. These islands have never been inhabited and are home to many endemic and threatened species. Management is focused on protection of the virtually pristine habitat and on promoting research and public education. Tourist visitation is frequent, but there are no facilities.

Situated in southern Tanzania, the 5 million-ha Selous Game Reserve (category IV, Habitat/Species Management Area) is one of the largest protected areas in Africa, with altitudes ranging from 100 to 1200 m. The reserve protects the world's largest area of miombo woodland. This woodland requires a particular burning regime for its preservation. Selous is famous for its elephant, hippopotamus, and black rhinoceros populations, although only a few of the rhinoceros remain. Sizable populations of other large animals include buffalo, Nyasaland gnu, brindled gnu, hartebeest, Greater Kudu, sable antelope, eland, reedbuck, bushbuck, waterbuck, warthog, zebra, giraffe, and wildebeest. Predators include lion, leopard, spotted hyena, hunting dog, and a small population of cheetah. There are more than 350 species of birds and reptiles, including crocodiles. Ecological monitoring and research provide information for the quotas for regulated sport hunting, the main use of the reserve.

Dartmoor National Park (category V; Protected Landscape/Seascape) was designated one of the national parks of England and Wales in 1951. It is a moorland landscape with tors of exposed granite and wooded

valleys. Approximately 31,000 people live within the 91,300-ha area of the park. There are approximately 10 million visits for recreation each year. Dartmoor is a rich habitat for wildlife and includes lowland farming areas of meadow, pasture, and woodland, deep valleys, and upland moors. There is a wealth of archaeological remains. Management of the area aims to ensure preservation of the natural habitat, the cultural heritage of the area, and continuation of traditional agricultural uses.

Tamshiyacu-Tahuayo Communal Reserve (category VI, Managed Resource Protected Area), established in 1990, is an area of 332,500 ha of Amazonian rain forest in the state of Loreto in northeast Peru and encompasses a rich flora and fauna. The reserve is divided into three land zones: a totally protected core, a subsistence zone, and an area of permanent settlement. More than 6000 people exploit the resources of the subsistence zone and the bordering areas. These people, known as *riberenos*, depend on hunting, fishing, swidden agriculture, and gathering of nontimber plant products.

### III. NATURE ALONGSIDE HUMANITY

Do the protected areas represent adequate "Noah's Arks" for Earth's currently known biological diversity? Regrettably, this question must be answered in the negative. Natural communities vary in time and space and are vulnerable to introduced animals and plants.

Moreover, existing protected areas cover only a small fraction of biological diversity. Many of these areas are small and isolated and lose species over time. Many



places designated as "protected areas," even those in IUCN categories I–III, are either inadequately protected or not protected at all and are referred to as "paper parks." Such parks are subject to rapid exploitation for subsistence or commercial gain. Conflicts between conservation and exploitation of natural resources impact the establishment of new protected areas and affect existing ones. Finally, even adequately protected areas suffer impacts from visitation and from changes that take place in adjoining lands. Glimpses of these global problems are offered here.

There are approximately 7000 nationally protected areas in more than 125 nations encompassing ~5% of the earth's surface. Table I provides an estimate of the approximate number of sites and of the area worldwide protected within each of the six IUCN categories, whereas Table II indicates the percentage of land area protected for countries in which that area exceeds 10%. Many other nations have stated their intention to protect 10% or more of their land by the Year 2000. Even the realization of such goals is unlikely to stem the rapid loss of critical habitat. Protected areas are increasingly becoming islands encircled by intensively exploited landscapes.

The well-founded concern about lack of appropriate representation of biological diversity in reserves is illustrated by a few examples. Consider the Chilean National Park system. This system, first established in the early 1900s, was one of the earliest in Latin America. In the temperate forest region of south-central Chile, parks and reserves protect 13 million ha, 29% of the land in this region. However, more than 90% of the protected land is at high latitudes ( $>43^\circ$ ), largely outside the areas of the highest tree species richness, endemic woody genera, and maximum species richness of native mammals, amphibians, and freshwater fishes, which lie in the region of  $35.6^\circ$  to  $41.3^\circ\text{S}$ . Areas of high endemism and species richness in the latter region correspond with areas of high human population density and intense land use. Moreover, forest protected in the reserves in the  $35.6\text{--}41.3^\circ\text{S}$  region lies at high Andean locations (above 600 m), often on large active volcanoes. This example highlights a worldwide problem. Reserves larger than  $100,000\text{ km}^2$  protect species-poor high mountain, tundra, and the driest desert areas.

Utilization of natural resources, particularly in the poorer regions of the world, takes precedence over protection of habitat. In the Philippines, for example, the market demand for hardwoods and plywood led to a high rate of deforestation; less than one-fourth of the rain forest remains and  $<1\%$  of the original rain forest area is officially preserved. Even for countries with a

TABLE II  
Countries with  $>10\%$  Land Area in Protected Areas<sup>a</sup>

Country	Area protected (ha)	% land area protected <sup>b</sup>
Antigua and Barbuda	6,128	13.86
Australia	93,545,457	12.18
Austria	2,005,475	23.92
Belize	323,121	14.07
Bhutan	966,100	20.72
Botswana	10,663,280	18.54
Brunei Darussalam	115,133	19.97
Cameroon	2,997,750	16.56
Chile	13,725,125	18.26
Costa Rica	638,564	12.55
Czech Republic	1,066,808	13.53
Denmark	1,388,750	38.24
Dominican Republic	1,048,284	21.64
Ecuador	11,113,893	24.08
France	5,601,486	10.30
Germany	9,195,702	25.77
Greenland	98,250,000	44.95
Israel	307,835	14.82
Lao P.D.R.	2,440,000	10.30
Latvia	774,724	12.16
Liechtenstein	6,000	37.50
Luxembourg	36,000	13.93
Malawi	1,058,500	11.25
Namibia	10,217,777	12.40
New Zealand	6,147,794	23.19
Norway	5,536,512	17.09
Oman	3,736,250	13.74
Panama	1,326,332	16.89
Rwanda	327,000	12.42
St. Kitts and Nevis	2,610	10.00
St. Vincent and the Grenadines	8,284	21.30
Senegal	2,180,709	11.09
Seychelles	37,893	93.79
Slovakia	1,015,509	72.36
Sri Lanka	795,953	12.13
Switzerland	730,707	17.70
Taiwan	426,597	11.54
Tanzania	13,889,975	14.78
Thailand	7,020,276	13.66
Togo	646,906	11.39
United Kingdom	5,127,966	20.94
United States of America	104,238,016	11.12
Venezuela	26,322,306	28.86

<sup>a</sup> IUCN 1996 global protected areas summary statistics.

<sup>b</sup> Values may be inflated in some instances due to the inclusion of marine protected areas in the calculation (e.g., Ecuador, which includes the Galapagos marine reserve).

relatively high per capita income, detailed biota surveys and resulting conservation plans do not guarantee appropriate representation of biological diversity in protected areas. Such detailed planning was carried out during the 1990s for the diverse forest ecosystem of northwestern New South Wales in Australia. However, extensive logging is to take place in this area and land to be included in reserves is mostly in unloggable escarpment forest well represented elsewhere. Struggles between development and preservation are commonplace. In South Africa, Greater St. Lucia Wetlands Park may become the site of a titanium mine; construction of a uranium mine is under way in Australia's Kakadu National Park.

An increasing number of parks appears on the IUCN's list of threatened protected areas of the world. Many of the threats arise from internal use and external development. The density of automobile traffic in Yosemite valley approaches that on urban freeways. In the Everglades, decades of water diversion have greatly reduced the wetlands, with a decrease in the number of wading birds by approximately 90% since the 1900s.

Some protected areas have sustained long-term, or even irreparable, damage from massive accidental contamination, as illustrated by a recent example from Spain. The 200,000-ha Guadalquivir ecosystem, the largest wetlands in Europe, encompasses the World Heritage site of Doñana National and Natural Parks. This is one of the most important bird breeding and overwintering sites in Western Europe. A dam on a massive tailings pond used by a zinc mine on one of the tributaries of the Guadalquivir River collapsed in April 1998 releasing approximately 5 million cubic feet of acid sludge from the processing of pyrite ore. The waste entered ecologically sensitive areas of the park, resulting in massive contamination of the wetlands with high concentrations of lead, zinc, arsenic, and other heavy metals. The accident caused considerable fish and invertebrate kills and an impact of unknown extent on the protected bird species. High-level heavy metal contamination is likely to persist for many years.

#### IV. RESEARCH IN THE NATIONAL PARKS

Each year, 1000 scientific papers or more report results of research projects which examine questions concerning natural communities or particular organisms in national parks throughout the world. Surveys of flora and fauna, as well as of microbial communities, continue

TABLE III  
Estimated Alteration in the Fire Regime in Four Canadian National Parks<sup>a</sup>

National park (historic period)	Historic burned area (km <sup>2</sup> per decade)	Burned area, 1940–1955 (km <sup>2</sup> per decade)
Jasper (1510–1930)	590	5.8
Banff (1488–1928)	267	8.6
Kootenay (1491–1931)	91	4.5
Yoho (1700–1980)	92	10.6

<sup>a</sup> Cited in Woodley S. (1997, p. 17).

to reveal the existence of new species. Many studies address narrow questions specific to a particular park or species. For example, the Bwindi-Impenetrable Great Ape Project studies the ecological relationship between the mountain gorillas (*Gorilla gorilla beringei*) and chimpanzees (*Pan troglodytes schweinfurthii*) in Bwindi-Impenetrable National Park in southwestern Uganda, the only forest in the world in which these two apes occur together. Kanha National Park, a wildlife conservation area in Central India, is home to the highly endangered swamp deer (*Cervus duvauceli branderi*). Analysis of the evolutionarily conserved repeat sequence motifs in the genome of the swamp deer at Kanha National Park indicates a high level of genetic homogeneity that may underlie the gradual extinction of this species.

In the United States, the National Park Service has focused on six issues that threaten park resources: loss of ecosystem integrity and aesthetic degradation, polluted air, altered water quality or quantity, resource consumption, invasion by nonnative species, and visitor impacts. As demonstrated in the examples that follow, much of the research effort in the U.S. national parks addresses such concerns.

Until the latter half of the twentieth century, protection of forests from naturally occurring fires was considered essential to their preservation. A dramatic illustration of the impact of this policy is provided by the data for four Canadian national parks in which suppression of fires led to 10- to 100-fold decrease in burned area per decade relative to periods preceding fire suppression (Table III). Years of fire suppression have allowed unnatural, dangerously high accumulations of fuels as well as shifts in the proportion of fire-tolerant to fire-intolerant tree species. Research in Yosemite and in Sequoia/Kings Canyon National Parks strongly supported the need to reintroduce fire into these ecosystems. The

past 30 years have seen the introduction of a closely monitored prescribed burn regime. This is a clear example of a major change in management driven by the outcome of scientific research.

Removal of predators to protect “desirable” prey species was among the first resource management actions in the U.S. national parks. This practice was driven by an a priori belief-based consensus and was ultimately abandoned because of objections based on scientific research. The 30-year study in Isle Royale National Park of the interaction of wolves (*Lupus canis*), moose (*Alces alces*), and the vegetation on which moose feed was a particularly influential multifaceted study which demonstrated the long-term stability of this predator-prey system.

Mammoth Cave National Park lies in a classic karst terrane, an irregular limestone region, with sinks, underground streams, and caverns. Concerns about pollution of water at Mammoth Cave led to an extensive hydrological research program. An essential element of this program was the development of methods to measure and monitor conduit flow characteristics within the park and the surrounding area. This research documented the existence of an extensive underground conduit network through which water could flow several miles in a few hours. The source of pollution was shown to be drainage of untreated or inadequately treated sewage from communities well outside the park into the underground water system that enters the cave system. The hydrologic research influenced sewage treatment both within the park and in communities outside the park. This example provides another illustration of the critical importance to management of understanding the relationships between protected areas and adjacent lands.

## V. BIOLOGICAL FIELD STATIONS

The several hundred field stations throughout the world are located in a wide variety of natural settings and explore a huge array of topics in the field sciences. Most are supported by or affiliated with universities, whereas others are operated by governmental organizations or by private foundations and institutes. Field stations are outdoor laboratories, the bases of research and teaching in environmental sciences. Field stations also fulfill important long-term monitoring functions. In a particularly apt description, these sites have been termed “sense organs” in nature. Descriptions of many individual stations and of their activities are available on the World Wide Web (for links, see [\*iobfs\* and <http://www.ekoforsk.uu.se/ffslink.html>\). Many field stations are designated International Long Term Ecological Research Network sites \(see Sections VI and VII\).](http://www.capital.net/com/</a></p></div><div data-bbox=)

The research at a particular field station is strongly dependent on the biogeographic features of the area around the station, but it is also strongly influenced by the size of the station, the past history of the research at the site, and the interests of the researchers at the institution(s) with which it is affiliated. As the following examples show, research at some field stations covers a very broad spectrum of fields. Other stations support highly specialized research. The majority lie in-between.

Kristineberg Marine Research Station, the largest field station for marine research in Sweden and one of the oldest such stations in the world, is located on the west coast of Sweden, at Fiskebäckstil, 120 km north of Gothenburg. The station, founded in 1877 by the Royal Swedish Academy of Sciences, is operated by the academy and by Göteborg University. Located on the shore of a protected bay, a nature reserve within the Gullmarsfjord, the station has easy access to coastal and offshore marine habitats—sand and mud flats, steep rocky cliffs, and deep basin sediments. The brackish surface waters in the bay originate from the Baltic and the deep oceanic waters from the North Sea. These features endow the area with very high biodiversity. Research at Kristeneberg encompasses a broad spectrum of marine ecology and environmental research. Specific research areas include behavioral ecology, benthic ecology and monitoring, biological oceanography, ecophysiology, ecotoxicology, functional morphology, larval ecology, pelagic monitoring, physiology of macroalgae, plankton research, and trophic relationships in shallow coastal communities.

The Konrad Lorenz Research Station in the northern Alps of Austria, approximately 250 km west of Vienna, offers an example of a much narrower research program. The station was established by Konrad Lorenz in 1973 to continue his behavioral studies on the ecoethology of social life in greylag geese (*Anser anser*) and other animals. Such research continues. Projects examine status dependence of within-flock competition, risk-sensitive foraging, and noninvasive behavioral endocrinology. Other research deals with the behavioral ecology and aspects of cognition in ravens (*Corvus corax*) and the functions and ecological roles of chemosensory organs in fish.

An intermediate situation is exemplified by the Alpine Research Center—Finse owned and operated by the universities of Bergen and Oslo. The center is 1222

m above sea level, approximately 250 m above the current tree line, on the northwestern end of Hardangervidda in south-central Norway (60°36'N, 7°30'E). Annual mean temperature is  $-2^{\circ}\text{C}$  and the mean precipitation  $\sim 1030$  mm, much of which falls as snow. The focus of biological research at the center has been on the population dynamics of birds and small mammals and on alpine plant ecology. The center is also a base for geological and glaciological studies. Ongoing research includes a 30-year series of rodent trapping data, 15 years of glacier monitoring, as well as meteorological monitoring.

Field stations play an indispensable role in the training of undergraduate and graduate university students in a wide variety of disciplines. For example, the universities of Bergen and Oslo both conduct undergraduate and graduate-level courses at the Alpine Research Center–Finse. Two of the longest running courses are alpine ecology and snow and winter ecology.

In the United States, the University of California Natural Reserve System manages 33 reserves that encompass more than 50,000 ha across 12 ecological regions in one of the most physiographically diverse regions in the United States. It is the largest university-operated reserve system in the world. Approximately 3500 University of California students utilize reserve system sites each year in courses in disciplines such as botany, entomology, zoology, geology, geography, meteorology, archaeology, paleontology, ecology, environmental planning, and wildlife management. When students from other educational institutions are included, the annual number increases to approximately 10,000. Hundreds of doctoral thesis projects worldwide depend on the use of the resources of a field station or on the use of a field station as a base of operations in a particular region.

## VI. THE LONG TERM ECOLOGICAL RESEARCH NETWORK

The U.S. National Science Foundation initiated the establishment of the Long Term Ecological Research Network (LTER; <http://www.lternet.edu>) in 1980. The 21-site network (Table IV) carries out long-term studies of phenomena with broad spatial and temporal scales on the following core topics: (i) the pattern and control of primary production, (ii) the spatial and temporal distribution of populations selected to represent trophic structure, (iii) the pattern and control of organic matter accumulation in surface layers and sediments, (iv) the

patterns and movements of inorganic inputs through soils and ground- and surface waters, and (v) the patterns and frequency of disturbances. A few examples provide a glimpse of the diversity of the LTER sites and of the broad range of research they make possible. Of the LTER sites described here, H. J. Andrews Experimental Forest, Coweeta Hydrologic Laboratory, and Konza Prairie Research Natural Area were among the initial group of six sites established in 1980.

The H. J. Andrews Experimental Forest LTER site in the Cascade Mountains of Oregon comprises a temperate coniferous forest biome. The main communities are Douglas fir, western hemlock, western red cedar, true fir, and mountain hemlock, interlaced with streams. The research centers on succession changes in ecosystems, forest–stream interactions, population dynamics of forest stands, patterns and rates of decomposition, carbon sequestration, and disturbance impacts on hydrologic response.

The Arctic Tundra LTER site at Toolik Lake, Alaska, includes tussock and heath tundra, riverine willows, oligotrophic lakes, and headwater streams. The research includes studies of the movement of nutrients from stream to lake, changes due to anthropogenic influences, and control of ecological processes by nutrients and predation.

The Coweeta Hydrologic Laboratory LTER site in the southern Appalachian Mountains of North Carolina is an eastern deciduous forest biome. The research includes long-term hydrology, nutrient cycling, and productivity responses to management practices and natural disturbances (such as drought, flood, wind, and insects), impacts of atmospheric deposition on forest ecosystems, and physiological studies of carbon balance and competition.

The Konza Prairie Research Natural Area LTER site in Flint Hills, Kansas, is a tallgrass prairie biome. Research at Konza focuses on the effect of fire, grazing, and climatic variability. Data obtained by remote sensing and geographic information systems (GIS) are used to evaluate grassland structure and dynamics.

The Sevilleta National Wildlife Refuge LTER site in central New Mexico lies at the intersection of diverse communities: montane mixed-conifer forest and meadows, riparian Rio Grande cottonwood forest, interior chaparral, Great Plains grasslands, Colorado Plateau shrub–steppe, Chihuahuan Desert, juniper savanna, and pinyon–juniper woodlands. The primary emphasis in the research program is to examine long-term changes in ecosystem properties, such as population dynamics of plants and animals, nutrient cycling, hydrology, and productivity and species diversity, re-

TABLE IV  
Long Term Ecological Research (LTER) Network Sites in the United States

LTER site	Location	Characteristics
H. J. Andrews Experimental Forest	Oregon	Temperate coniferous forest
Arctic Tundra	Alaska	Tundra, lakes, streams
Baltimore Ecosystem Study	Maryland	Urban and agricultural watershed
Bonanza Creek Experimental Forest	Alaska	Taiga
Central Arizona–Phoenix Urban	Arizona	Sonoran desert scrub, urban environments, regulated river and floodplain
Cedar Creek Natural History Area	Minnesota	Eastern deciduous forest and tall grass prairie
Coweeta	North Carolina	Eastern deciduous forest
Jornada Experimental Range	New Mexico	Hot desert
Kellogg Biological Station	Michigan	Row-crop agriculture
Konza Prairie Natural Research Area	Kansas	Tallgrass prairie
Luquillo Experimental Forest	Puerto Rico	Tropical rain forest
McMurdo Dry Valleys	Antarctica	Polar desert oases
North Temperate Lakes	Wisconsin	Lakes, eastern deciduous forest
Niwot Ridge	Colorado	Alpine tundra
Palmer Station	Antarctic	Polar marine
Plum Island Sound	Massachusetts	Estuarine
Sevilleta	New Mexico	Subalpine mixed-conifer forest/meadow, riparian forest, dry mountain-land, grassland, cold desert, hot desert
Shortgrass Steppe	Colorado	Shortgrass steppe
Virginia Coast	Virginia	Coastal barrier islands

sulting from both natural and anthropogenic disturbances.

## VII. THE INTERNATIONAL LONG TERM ECOLOGICAL RESEARCH NETWORK

By 1998, 14 nations had initiated long-term ecological research programs (International Long Term Ecological Research Network; <http://www.lternet.edu/ilter>) with more than 200 research sites, modeled after the U.S. LTER: Brazil, Canada, China, China-Taipei, Costa Rica, Czech Republic, Hungary, Israel, Korea, Mexico, Poland, United Kingdom, Uruguay, and Venezuela. Other countries are working to establish similar programs.

## VIII. SUSTAINED ECOLOGICAL RESEARCH

Establishment of long-term ecological research networks worldwide connotes a commitment to continuous research and monitoring to study long-term changes in ecosystems and to understand ecological

processes that vary over long periods of time. One characteristic of such “sustained ecological research” is that it must last at least as long as the phenomenon under study or must be scaled to the frequency of the events being studied. The indispensability of sustained ecological research is well illustrated by considering some of the outcomes of 30 years of such research at the Hubbard Brook Experimental Forest (HBEF).

HBEF, a 3160-ha reserve in the White Mountains of New Hampshire, was established by the U.S. Department of Agriculture Forest Service in 1955 and designated as a LTER site in 1988. The reserve represents a northern hardwood forest ecosystem with streams and lakes. HBEF has one of the most extensive and longest continuous databases on the hydrology, geology, chemistry, and biology of natural ecosystems in the world.

Measurements at HBEF as early as 1964 were among the first to direct attention to acid rain in North America. More than 90% of the sulfur and nitrogen oxide emissions to the atmosphere in North America are anthropogenic and originate primarily from the combustion of fossil fuels by electrical power plants and from smelters. Sulfuric acid contributes approximately 65% and nitric acid approximately 35% to the acidity of the precipitation in the eastern United States. Studies

at HBEF revealed that dry deposition of acidic substances (gases and particles) is also a quantitatively important atmospheric input to aquatic and terrestrial ecosystems. Long-term monitoring at HBEF showed that both the hydrogen ion and sulfate concentrations in the precipitation gradually decreased from about 1970 on, after passage of clean air legislation by the U.S. Congress. These monitoring data provide invaluable measures of the outcomes of emission controls and reveal shortcomings. For example, the concentration of nitrate in the precipitation has remained virtually unchanged.

Dispersal of toxic metals, particularly lead, from anthropogenic activities is also an important environmental problem. In 1980, global emissions of lead from human activity were estimated at 2000 metric tons compared with approximately 6 metric tons derived from natural sources. In 1975, the concentration of lead in rain and snow at HBEF averaged  $\sim 25 \mu\text{g}$  per liter, a concentration deemed unsafe in drinking water. The use of lead additives in gasoline was sharply restricted in the United States in 1977. By 1989, the average concentration of lead in rain and snow at HBEF had decreased to  $\sim 2 \mu\text{g}$  per liter.

Monitoring at HBEF has also provided much valuable information on the functioning of natural systems that is unrelated to the assessment of anthropogenic impacts. For instance, long-term data on the hydrologic cycle at HBEF revealed that the annual amount of evapotranspiration from this forested ecosystem was relatively constant from year to year even though the annual precipitation varied by about twofold. These data reveal that the vegetation utilizes about the same amount of water whether it is a wet or a dry year. This was an unanticipated finding.

## IX. GAP ANALYSIS

The current rate of species extinction is estimated to be 100 times the natural background rate, and it is generally accepted that species loss correlates with habitat destruction. Historically, many protected areas were selected for reasons other than the optimal protection of biological diversity. Recently, much attention has been given to saving endangered species. However, a set of such disconnected, highly directed efforts does not address the primary ongoing causes of species extinction: habitat loss and fragmentation and also degradation of natural landscapes. Application of computer-based geographic information system (GIS) to the plan-

ning for protection of biological diversity offers the best hope for the future.

The method, called "gap analysis," is a powerful systematic approach for assessment of the protection for biodiversity in a given area. Gap analysis consists of three primary data layers: (i) delineation of the distribution of vegetation types from satellite imagery data, (ii) land ownership (which provides information on management practices), and (iii) distributions of terrestrial vertebrates as predicted from the distribution of vegetation. Refinements to the individual data sets are provided by aerial photographs, animal distribution maps for particular species, data on the natural history of plants and animals, and so on. For a given region, the map that results from the layered data sets provides information on the geographic distribution of "species richness" versus the location of existing protected areas. Figure 1 shows a subset of layers utilized to produce such a map. The analysis provides information on the gaps where particular species are not included in protected areas and guides assessment for siting of additional protected areas and of corridors to allow migration between protected areas.

A study carried out in the mid-1980s on endangered birds in Hawaii provides a striking early demonstration of the power of gap analysis. Extensive field inventories were used to plot the distribution of each endangered forest bird species. Individual maps were then combined to obtain a map of species richness for this important group. Comparison of this map with a map of the existing reserves revealed that  $<10\%$  of the ranges of endangered forest birds were within the reserves. This analysis led to the establishment of the 6693-ha Hakalau Forest National Wildlife Refuge in one of the areas of highest species richness.

Gap analysis leads to natural resource management planning on a regional, national, or even global scale because it allows an integrated examination of species, habitats, and human ownership patterns and management practices over any desired area. The value of gap analysis depends on the quality of the data. Experience has shown that the available biological data for both public and private lands is generally fragmentary. Data on species distribution and habitat is frequently lacking. Finally, ready incorporation of biodiversity data into GIS requires that it be spatially referenced, a criterion not always met.

Such shortcomings can be addressed, and gap analysis has received widespread acceptance as the approach of choice to guide conservation planning. The desired outcome is that large-scale planning will lead to a world in which humans and the other

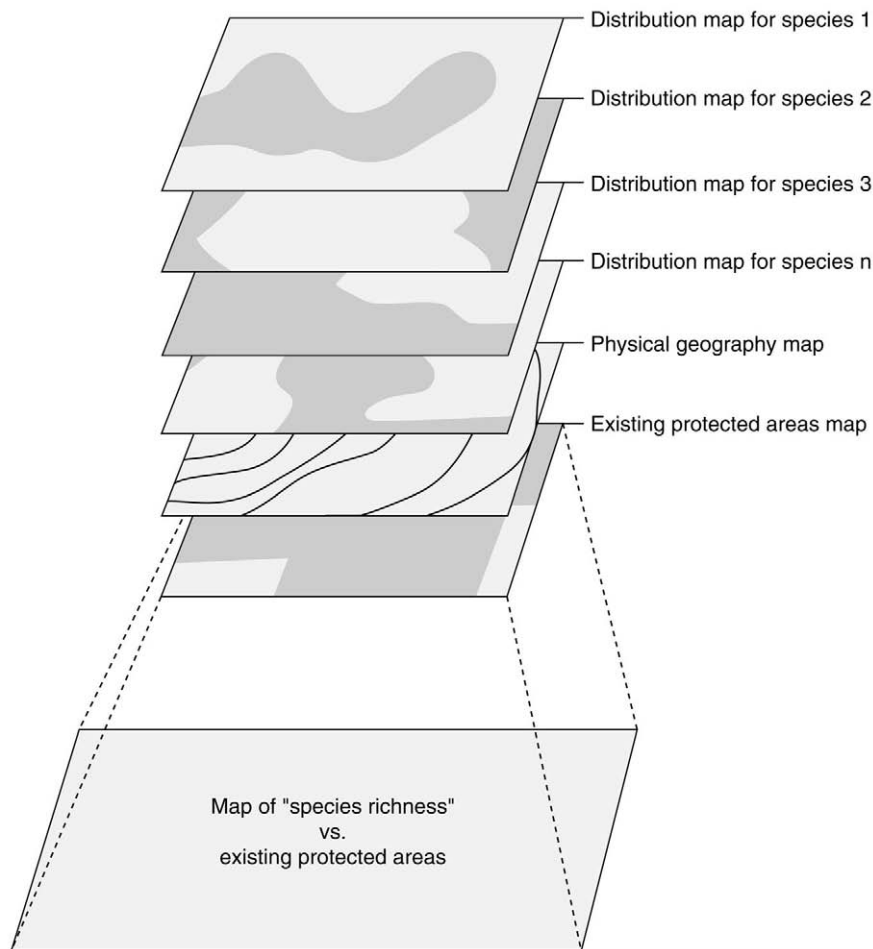


FIGURE 1 A step in gap analysis in which distribution maps for individual species are overlaid in the GIS to produce maps of species richness. An additional GIS layer for existing protected areas allows identification of gaps in species representation in these areas.

inhabitants of the biosphere live side by side in relative harmony.

### See Also the Following Articles

CONSERVATION EFFORTS, CONTEMPORARY • EDUCATION AND BIODIVERSITY • FIRES, ECOLOGICAL EFFECTS OF • PREDATORS, ECOLOGICAL ROLE OF • ZOOS AND ZOOLOGICAL PARKS

### Bibliography

Bormann, F. H. (1996). Ecology: A personal history. *Annu. Rev. Energy Environ.* 21, 1–29.

Brandon, K., Redford, K. H., and Sanderson, S. E. (Eds.) (1998). *Parks in Peril. People, Politics, and Protected Areas*. Island Press, Washington, D.C.

Grumbine, R. E. (Ed.) (1994). *Environmental Policy and Biodiversity*. Island Press, Washington, D.C.

Halvorson, W. L., and Davis, G. E. (Eds.) (1996). *Science and Ecosystem Management in the National Parks*. Univ. of Arizona Press, Tucson.

International Union for the Conservation of Nature (IUCN). (1994). *Guidelines for Protected Area Management Categories*. CNPPA with the assistance of WCMC, IUCN, Gland, Switzerland/Cambridge, UK.

Kramer, R., van Schaik, C., and Johnson, J. (1997). *Last Stand. Protected Areas and the Defense of Tropical Biodiversity*. Oxford Univ. Press, New York.

Likens, G. E. (1992). *The Ecosystem Approach: Its Use and Abuse*. Ecology Institute, Oldendorf/Luhe, Germany.

- Nelson, J. G., and Serafin, R. (Eds.) (1997). *National Parks and Protected Areas. keystones to Conservation and Sustainable Development*. Springer-Verlag, Berlin.
- Pace, M. L., and Groffman, P. M. (Eds.) (1998). *Successes, Limitations, and Frontiers in Ecosystem Science*. Springer-Verlag, New York.
- Saunders, D. A., Craig, J. L., and Mattiske, E. M. (Eds.) (1995). *The Role of Networks*. Surrey Beatty, Chipping Norton, NSW, Australia.
- Scott, J. M., and Jennings, M. D. (1998). Large-area mapping of biodiversity. *Ann. Missouri Botanical Garden* 85, 34–47.
- Scott, J. M., Davis, F., Csuti, B., Noss, R., Butterfield, B., Groves, C., Anderson, H., Caicco, S., D'Erchia, F., Edwards, T. C., Jr., Ulliman, J., and Wright, R. G. (1993). GAP analysis: A geographic approach to biological diversity. *Wildlife Monogr.* 123, 1–41.
- Sellers, R. W. (1997). *Preserving Nature in the National Parks. A History*. Yale Univ. Press, New Haven, CT.
- Wildland Resources Center (1996). Summary of the Sierra Nevada Ecosystem Project Report, Report No. 39. Davis: University of California, Centers for Water and Wildlands Resources.
- Woodley, S. (1997). Science and protected area management: an ecosystem-based perspective. In *National Parks and Protected Areas* (J. G. Nelson and R. Serafin, Eds.). pp. 11–21. Springer-Verlag, Berlin.







# NEAR EAST ECOSYSTEMS, ANIMAL DIVERSITY

Joseph Heller

*The Hebrew University of Jerusalem*

---

- I. The Levant
  - II. The Extent of Biodiversity
  - III. Historic Zoogeography
  - IV. Endemic versus Widespread Animals
  - V. Factors Determining Diversity
  - VI. Freshwater Diversity: Historic Factors
  - VII. Biodiversity: The Human Impact
- 

## GLOSSARY

**allopatric** Describing two taxonomic entities whose geographic ranges do not intersect with each other.

**endemism** The fact of being found in a specific location of limited size, rather than being widely distributed.

**eremic** Relating to or occurring in desert regions.

**parapatric** Describing two taxonomic entities whose geographic ranges are distinct but adjacent to each other.

**xeromorphic** Describing a form that has developed in response to highly arid conditions.

---

**THE ANIMAL FAUNA OF THE LEVANT** (the eastern shoreland of the Mediterranean Sea) is exceptionally rich and diverse. To what extent is this high diversity due to geological history, climate, and substratum? What is the human impact on the fauna of the Levant? It is these and other questions that I address in this article.

## I. THE LEVANT

The Levant is the eastern shoreland of the Mediterranean (Por, 1975), a stretch of land approximately 800 km long and approximately 150 km wide (Fig. 1). It is wedged in between the Mediterranean Sea in the west and the Arabo-Syrian Desert in the east, stretching from the mouth of the River Orontes in the north to the Isthmus of Suez in the south. It consists of four basic north–south-oriented features: the coastal plain, the western (cis-rift) mountains, the rift valley, and the eastern (trans-rift) mountains. In the east it merges gradually into the Arabo-Syrian Desert.

The 800-km-long Levant has several climatic and floral belts. In its northern (and high-altitude) areas the overall habitat is Mediterranean: Summers are dry and warm, and winters are rainy and mild with occasional snow in the higher mountains. Natural plant associations consist of xerophytic shrubs and trees. In its southern and eastern areas (the Negev, Sinai, and fringes of the Syrian desert) the overall habitat is eremic: Temperatures are high, precipitation is low and very irregular, and the vegetation cover is poor. Between these two major habitats there is an intermediate habitat of steppe character. The combination of the four major north–south relief patterns with the three climatic belts imparts an unusual heterogeneity to this relatively small zoogeographic province and has resulted in a highly dynamic faunal history.

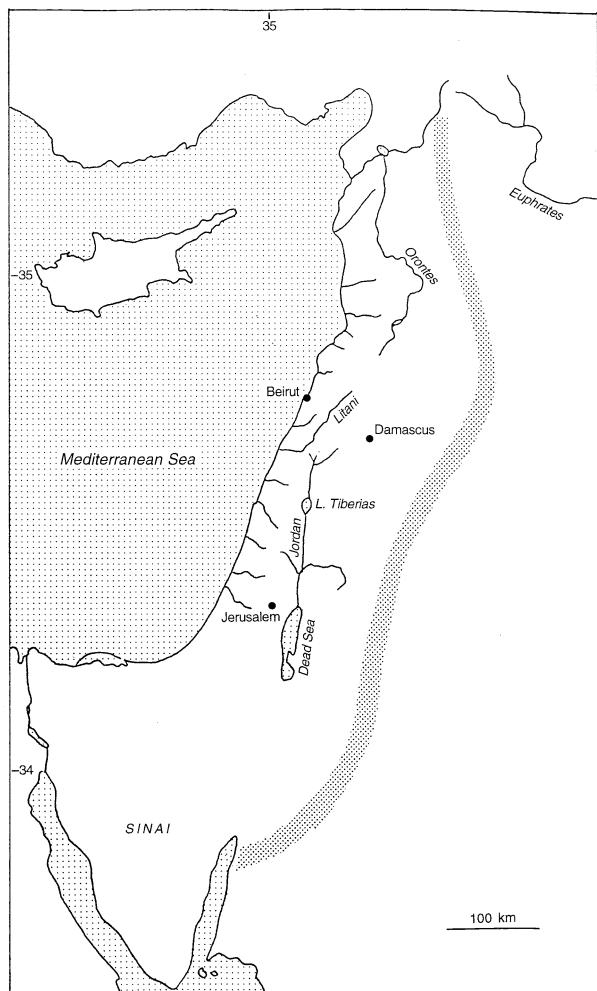


FIGURE 1 The Levant; its eastern boundary (stippled line) is hazily defined (adapted with permission from Por, 1975).

## II. THE EXTENT OF BIODIVERSITY

It is difficult to estimate animal diversity throughout the whole Levant because there is better knowledge about the fauna of its southern than of its northern areas. Unfortunately, the term "southern Levant" is loosely applied by various authors to include the area of Israel, and sometimes also the Palestinian Autonomy, and/or Sinai, Jordan, and the Golan Heights. These limitations should be taken in account when considering the data (of representative groups) in Table I. They suggest that animal diversity in the Levant is high and perhaps that this area is one of the richest and most diverse natural regions among temperate regions of the world (relative to its size). For example, the area between the Mediterranean Sea and the Jordan River, to-

gether with the Golan Heights, is only approximately 28,000 km<sup>2</sup>, but it contains approximately 170 bird and 100 mammal species (breeding). These numbers are not markedly fewer than those in California (210 birds and 110 mammals in an area of 404,000 km<sup>2</sup>), whose southern part lies at the same latitude as Israel.

Several factors contribute to this high diversity. One is the lack of a major disastrous disturbance; the general pattern of the landscape in the Levant has persisted since the early Pliocene. Only minor areas of the Levant (the tops of Mt. Lebanon and Mt. Hermon) were ever glaciated to any considerable extent, and only small regions (e.g., the Golan) were subject to volcanic eruptions that covered large areas with basalt. Consequently, the Levant never suffered the catastrophic wipeouts that hit northern and central Eurasia many times throughout the Pleistocene.

Local, region-scale changes did occur, however, repeatedly separating and then uniting different patches of landscape, thus permitting possible speciation under allopatry. These local geological events probably contributed, through speciation and through invasion, to enrich rather than to impoverish the Levantine fauna. This long, generally eventless but locally eventful history of the Levant is one of the reasons for its rich fauna.

A further reason is the heterogeneous climate. Mean annual rainfall may decrease from 1000 mm to less than 100 mm over a short distance of 150 km. The landscape therefore changes from Mediterranean oak maquis to batha and then to barren deserts, and the fauna gradually changes as one species replaces another. The net result is a fauna richer than would be expected were the landscape more uniform.

In contrast to factors contributing to biodiversity of the Levant, certain environmental factors limit it. The major limiting factor is that southern and eastern areas of the Levant are an extreme desert that, for many animals, is too hostile a habitat for survival.

## III. HISTORIC ZOOGEOGRAPHY

Levantine biota constitute one of the most complex ecosystems in the world: Ethiopian and Oriental taxa coexist with Euro-Siberian, central Asiatic, and Mediterranean species, mixed with Saharan and Arabian desert faunal elements. This kaleidoscopic admixture of Palearctic, Paleotropical, and Saharo-Arabian elements has been constantly changing during the Neogene and the Quaternary periods, disposing a new biogeographic configuration during each geological period. Consequently, the Recent fauna is a living monument to a

TABLE I  
Animal Diversity in the Levant or in Regions of It

Animal group	No. of species	Region	Reference
Crustaceans			
Ostracods (Ostracoda)	53	Israel	Martens and Ortal (1999)
Arachnaeids			
Scorpions (Scorpiones)	16	Southern Levant	Levy and Amitai (1980)
Camel spiders (Solifugae)	54	Israel	Levy and Shulov (1964)
Crab spiders (Thomisidae)	40	Southern Levant	Levy (1985)
Cobweb spiders (Theridiidae)	62	Southern Levant	Levy (1998)
Insects			
Dragonflies (Odonata)	82	All Levant	Dumont (1991)
Termites (Isoptera)	11	Southern Levant	Kugler (1988)
Bushcrickets (Tettigoniidae)	42	Israel	Ayal <i>et al.</i> (1999)
Grasshoppers (Acridoidea)	110	Southern Levant	Fishelson (1985)
Flies (Diptera)	3500	Southern Levant	Freidberg (1988)
Asilidae	158	Southern Levant	Theodor (1980)
Tephritidae	85	Southern Levant	Freidberg and Kugler (1989)
Butterflies (Rhopalocera)	106	Israel	Benyamini (1988)
Caddisflies (Trichoptera)	73	All Levant	Botosaneanu (1992)
Wasps (Vespidae)	7	Southern Levant	Kugler (1988)
Ants (Formicidae)	140	Southern Levant	Kugler (1988)
Mollusks			
Land snails (Pulmonata)	100	Southern Levant	Heller (1988)
Vertebrates			
Fish, freshwater (Pisces)	65	All Levant	Krupp (1987)
Amphibia + Reptilia	96	Southern Levant	Werner (1988)
Birds (Aves), breeding	170	Southern Levant	Yom Tov (1988)
Mammals (Mammalia), breeding	100	Southern Levant	Yom Tov (1988)

fascinating animal history and ecology. Current zoogeographic patterns in the Levant are the result of four major formative events (Por, 1975; Tchernov, 1988):

1. The breakup of the Tethys Sea during the Miocene and the consequent establishment of the Eurasian–African land bridge: Approximately 18–24 million years ago (late Oligocene–early Miocene) northern regions of today's Levant were still under the large Tethys Sea, whereas southern regions were terrestrial and part of the vast Afro-Arabian continent, with its unique faunal realm (Fig. 2). Approximately 17 million years ago (Ma) the Tethys contracted, the emerging lands combined the southeastern areas of Asia Minor with the northwestern areas of Arabia, and the Levant was formed (Fig. 3).

Bridging the two continents, the newly formed Levant triggered an interchange of terrestrial biota between Africa and Eurasia. A wide spectrum of habitats

were present in the Levant during much of the Miocene, including rivers, lakes, marshy habitats, dense woods, and Paleotropical rain forest, together with open country.

Early Miocene deposits in the southern Levant contain a diverse community of animal taxa, most of which were of African origin (proboscoideans, crocodiles, soft-shelled turtles, and catfish) but with others representing Eurasian elements (viverrines and cricetines). It seems that during this period African forms found it relatively easy to emigrate into the Levant, whereas contemporaneous Eurasian forms found it more difficult.

2. The late Miocene and Pliocene drying out of the Old World subtropics and the formation of the vast belt of the Saharo-Arabo-Syrian deserts, which restrained terrestrial migration between Africa and the Levant: The desertification process was very gradual. At first, approximately 12–15 Ma, open-land biota of savanna and semisteppe began to expand, and in doing so they

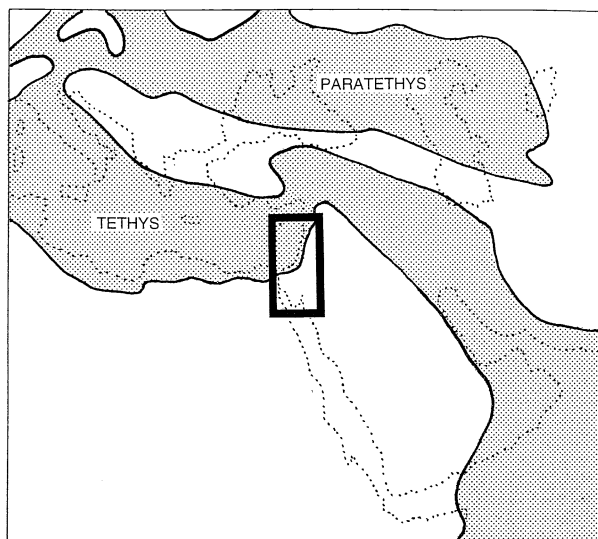


FIGURE 2 Paleographic map, 20 Ma. The Tethys is connected with the Indian Ocean, thereby separating Afro-Arabia from Eurasia (the Levant is represented by the rectangle). The southern Levant is part of the Afro-Arabian domain, and other parts are submerged (adapted with permission from Tchernov, 1988).

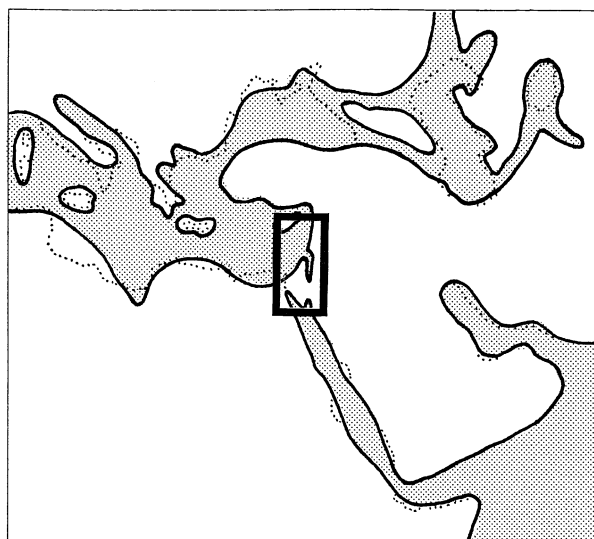


FIGURE 4 Paleographical map, 5 Ma. The marine connection established between the Red Sea and Indian Ocean results in relative isolation of the Levant from Africa (adapted with permission from Tchernov, 1988).

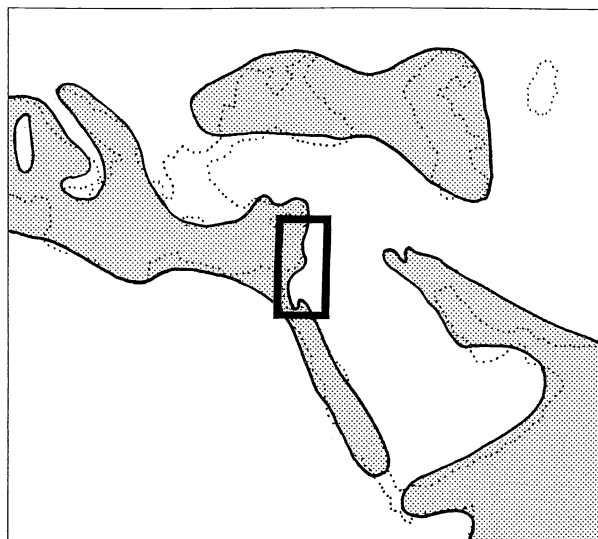


FIGURE 3 Paleographic map, 10 Ma. The Tethys is no longer connected with the Indian Ocean. In the north, extensive land connections with Eurasia enable the invasion of Palearctic elements into the Levant. In the south, land connections with Africa gradually break up due to the development of the Red Sea (adapted with permission from Tchernov, 1988).

created an extremely diversified mosaic of biotopes that ranged from the East African highlands through (still connected) Arabia to the Levant and enabled large-scale biotic interchange. Gradually, however, aridity along the Saharo-Arabian latitudes increased (associated with intensification of seasonality) so that by approximately 5.5–6 Ma (late Miocene) a vast arid belt of almost continental size formed which became a severe biogeographic barrier. Eventually, this desert barrier was settled by strongly eremic-adaptive species recruited from its immediate surroundings. The desert fauna of the Levant thus originates from both Eurasia and Africa and is established along adaptive rather than historical lines. It is more between than within faunal realms.

3. Quasi-isolation and biogeographical provincialism of the Levant during the Pliocene: Beginning approximately 5 Ma three geomorphological events had profound effects on the biota of the Levant. To the north, orogeny of the Taurus–Zagros mountain chains restrained many northern Palearctic elements from migrating southwards. To the south, the Red Sea opened up (Fig. 4) and connected with the Indian Ocean, thereby establishing, for the terrestrial fauna, a marine barrier between Africa and the Levant (in addition to the already existing desert). The third event was the brief flooding of the Mediterranean by oceanic waters through the Gibraltar straight. All along the Levant coastal plains were submerged, and deep marine embayments penetrated into such lowlands of the Levant as

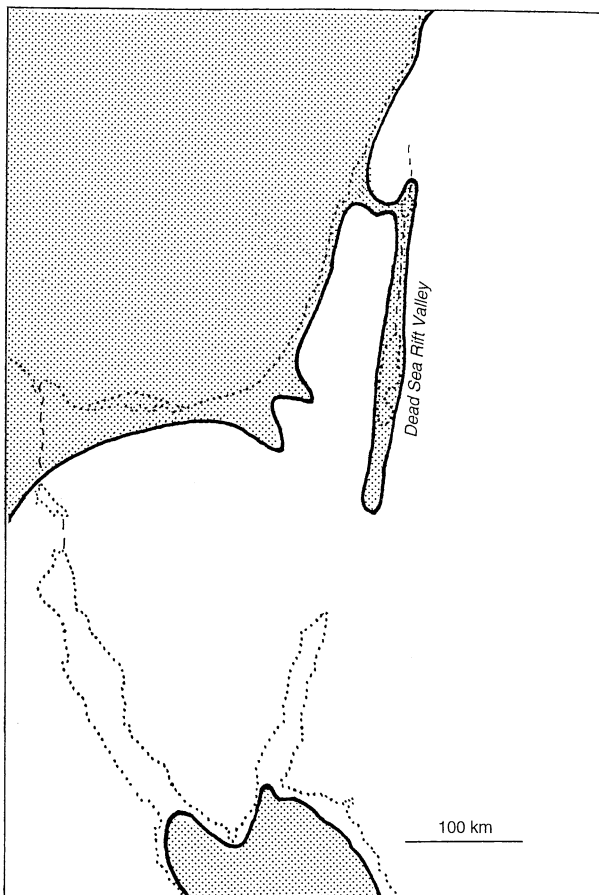


FIGURE 5 Paleogeographical map, 4–5 Ma. Marine ingressions of the Mediterranean into the newly formed Jordan–Dead Sea Rift Valley brings marine elements into the Jordan Valley [adapted with permission from Tchernov (1988), based on Bernor (1987), Gvirtzman and Buchbinder (1977, 1978), and Rögl and Steiner (1983)].

(today's) Dead Sea Rift Valley and Plain of Beer-Sheva (Fig. 5). These marine embayments of the Mediterranean (which gradually retreated later in the Pliocene) formed biogeographic barriers that enabled local speciation.

Pliocene mammals of the Levant (from a site near Bethlehem) are of open-country species, suggesting a lowland African savanna-like landscape, with some freshwater bodies nearby. The absence of deer, beavers, dormice, and bears (common in the Pliocene of Turkey) is noteworthy.

4. The successive creation of the Jordan–Orontes rift valley in the late Pliocene and early Pleistocene: During the Plio-Pleistocene transition (approximately 1.8 Ma) large-scale tectonism occurred in the southern Levant, which established a longitudinal geomorpho-

logical feature that opened initially in the south and gradually northward. The Jordan–Orontes valleys, with their chains of lakes and swamps, were then created, the cis- and trans-rift mountain ranges were then lifted, and basalts were extruded in the eastern regions of the Levant. It was during the Pliocene–Pleistocene that the contemporary relief structure of the Levant was molded (Por, 1975; Tchernov, 1988).

Levantine faunas during the lower Pleistocene were characterized by a bloom of Palearctic elements and a deterioration of tropical ones, suggesting that the Zagros Mountains filter barrier had by then been overcome and that the Saharo-Arabian barrier, between the Levant and Africa, had already become highly effective (Tchernov, 1988). By 1.4 Ma, two-thirds of the mammal genera (from a site in the Jordan Valley) were of Eurasian origin. The remaining one-third were mainly of tropical origin, and many African mammals entrapped in the Levant underwent local speciation (*Hippopotamus behemoth*, *Praomys minor*, and *Arvicanthis ectos*). By the mid-Pleistocene (800,000–150,000 years ago) additional Palearctic immigrants had invaded the Levant (squirrel, hare, Syrian bear, and wild cat), various tropical elements had become extinct (giraffe, monkeys, the fossil artiodactyls *Pelorovis* and *Kolpochoerus*), and there were no further massive invasions of tropical elements; the fauna was becoming similar to that of today. From 125,000 years ago onwards (upper Pleistocene) there were no drastic changes in faunal turnover rates. In many species of snails, voles, mole-rats, wolves, bear, and gazelle, however, there were considerable changes in body size to adjust to environmental fluctuations. These size alterations suggest that Pleistocene climatic changes in the Levant affected population levels more than species extinctions (Heller, 1988; Tchernov, 1988).

Today, the majority of animal genera in the Levant (more than three-fourths) occur outside the Levant exclusively to the north or northeast. Faunal connections with Eurasia are thus obvious. The extent of the Eurasian element varies from group to group. Almost all species of land snails are of Palearctic origin, as are three-fourths of the fruit flies (Tephritidae), two-thirds of the butterflies, and one-fifth of all chironomids. Furthermore, the extent of the Eurasian element within each group may vary from one region of the Levant to another. From Mt. Hermon to Sinai frequencies of Palearctic butterflies decrease from 88 to 43%; within the Palearctic ant genus *Formica*, the number of species decreases from Turkey (12) through Lebanon (1) to

TABLE II  
Oriental Representatives in the Levant

Group	Species	Reference
Spiders (Philodromidae)	<i>Thanatus fornicatus</i>	Levy (1991)
Ants (Formicidae)	<i>Polyrachis simplex</i>	Kugler (1988)
	<i>Monomorium mayri</i>	
	<i>Lophomyrmex quadrispinosus</i>	
Wasps (Vespidae)	<i>Vespa orientalis</i>	
Flies (Diptera)		
Chloropidae	<i>Anatrichus pygmaeus</i>	Friedberg (1988)
Culicidae	<i>Culex mimeticus</i>	
Butterflies (Rhopalocera)	<i>Madais fausta</i>	Benyamini (1988)
	<i>Zizeeria karsandra</i>	
	<i>Limenitis reducta</i>	
Caddisflies (Trichoptera)	<i>Hydroptila adana</i>	Botosaneanu (1992)
	<i>H. fonsorontina</i>	
	<i>H. hirra</i>	
	<i>H. libanica</i>	
	<i>H. palaestinae</i>	
	<i>Chimarra lejea</i>	
	<i>Setodes alala</i>	
	<i>Stactobia pacatoria</i>	
	<i>Pseudoneureclipsis palmonii</i>	
	Fish (Pisces)	
<i>Garra ghorensis</i>		
<i>Barbus canis</i>		
<i>Nemachilus insignis</i>		
<i>N. leontine</i>		
Birds (Aves), breeding	<i>Ketupa zeylonensis</i>	Yom Tov (1988)
	<i>Francolinus francolinus</i>	
	<i>Halcyon smyrnensis</i>	
	<i>Merops orientalis</i>	
Mammals (Mammalia)	<i>Hystrix indica</i>	Yom Tov (1988)
	<i>Nesokia indica</i>	

Israel (absent; Kugler, 1988; Benyamini, 1988; Heller, 1988).

Tropical elements comprise approximately 5–15% of the fauna, but again this varies from group to group. Approximately 12% of the ant species are tropical or have strong tropical affinities; Diptera (in general) comprise approximately 5–10%, chironomids 40%, land snails 1%, and there are no aquatic oligochaetes. Within the Levant, tropical components in butterflies increase from approximately 8% on Mt. Hermon to 32% in Sinai.

Although most tropical elements are either Ethiopian or Palearctic, a small Oriental element is also present (Table II). This is noteworthy because the Orient is distant from the Levant and separated from it by a vast

desert barrier. The influence of the Oriental species is sporadic and does not reach sizable percentages in the whole fauna. Freshwater fishes and aquatic insects are exceptions, however.

In conclusion, the Levantine land bridge serves as an important crossroads of biotic exchange. Along the tropical wet and savanna-like landscapes of the rift valley, African animals entered Palearctis. Along the Mediterranean woodland and dry steppe landscapes of the bordering mountains, Eurasian animals invaded Africa. Among the dunes of the shoreland, eremic, Saharan biota advanced northward. The Levant thus functions as a complex biogeographic corridor in which representative species of the Ethiopian, Palearctic, and even

TABLE III  
Endemism among Groups of the Levant or Regions of It

Animal group	Endemic species	Reference
Dragonflies (Odonata)	One-third	Dumont (1991)
Termites (Isoptera)	One-third	Kugler (1988)
Bushcrickets (Tettigonidae)	Half	Ayal <i>et al.</i> (1999)
Flies; asilids (Asilidae)	Half	Theodor (1980)
Butterflies	None, or negligible	Benyamini (1988)
Caddisflies (Trichoptera)	Two-fifths	Botosaneanu (1992)
Wasps (Vespidae)	None	Kugler (1988)
Ants (Formicidae)	One-fourth	Kugler (1988)
Land snails (Pulmonata)	Two-thirds	Heller (1988)
Fish (Pisces)	One-fourth	D. Golani (personal communication)
Reptiles and Anurans	One-tenth	Werner (1988)
Birds (Aves)	None, or negligible	Yom Tov (1988)
Mammals (Mammalia)	~2%	

Oriental biogeographic realms are encountered side by side. This mixture of faunas is outstanding and perhaps unique.

#### IV. ENDEMIC VERSUS WIDESPREAD ANIMALS

The Levant is characterized by the high frequency of marginal populations of species which range beyond the Levant. However, it is also rich in endemics (Table III). Endemism is a function of a group's low mobility, combined with its tendency to speciate. It is therefore not surprising that within the vertebrates, endemism is higher among (the more sedentary) fish and reptiles of the Levant than among (the more mobile) mammals and birds; within arthropods it is higher among (the more sedentary) ants than among the (highly mobile) butterflies and wasps.

Comparisons within and between groups should not be carried too far, but the very high extent of endemism among land snails is noteworthy. This may be because they are, as a whole, very ineffective in crossing even minor ecological barriers of substratum or aridity. Consequently, many land snail species of the Levant have a very small range, sometimes only a few square kilometers. The endemic snail *Pene galilaea* is confined to a very small area of approximately  $1.5 \times 2.5$  km in northwestern Upper Galilee, where it is surrounded by

another species, *P. sidoniensis* (Figs. 6 and 7); also, the endemic hygromiine *Trochoidea picardi* is confined to an area of  $3 \times 5$  km in the coastal plain, where it is completely surrounded by *T. davidiana*. Similarly, the high level of endemism among bushcrickets is probably due to limited dispersal ability due to their loss of ability to fly. Approximately 45% of the Israeli bushcricket species are brachy- or micropterous, and 84% of these flightless species are endemic compared to 13% of the fully winged flying species.

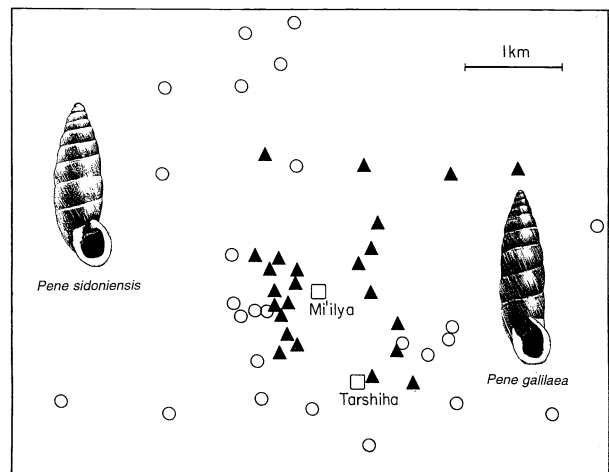


FIGURE 6 Narrow-ranged endemism in the Levant: total global distribution of the land snail *Pene galilaea* (▲) is only 8 km<sup>2</sup> in the western Galilee (○, sites of *P. sidoniensis*) (adapted with permission from Heller, 1993).



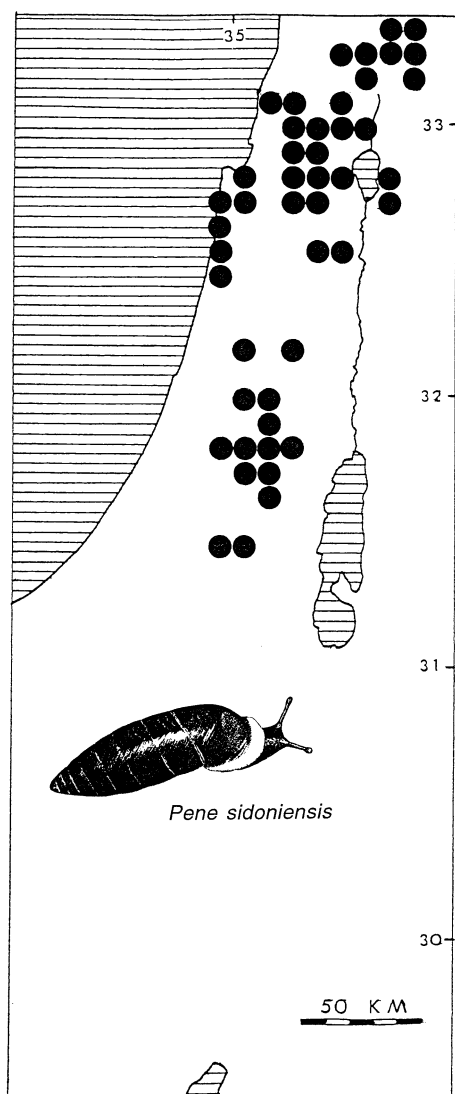


FIGURE 7 Wide-ranged distribution in the Levant: *P. sidoniensis* ranges over ca. 1200 km<sup>2</sup> in the southern Levant alone; it also occurs in the northern Levant and southern Turkey (adapted with permission from Heller, 1993).

It is tempting to seek "hot spot" regions, in which the extent of endemism is high and cuts across many animal groups. For the Levant, it is difficult to single out such regions. Endemics occur in many different regions.

In contrast to the endemics, several genera of the Levant enjoy an incredibly wide, intercontinental distribution. The land snail genera *Truncatellina*, *Lauria*, *Cecilioides*, and *Punctum*, minute snails of the litter or subterranean environment, are found outside the Levant not only throughout Eurasia but also in much of the African continent. *Oxyloma*, an amphibious genus

that lives on the vegetation above water, is also included in this list. All these taxa probably reached the Levant by aerial dispersal that, whether by wind or by birds, puts a premium on small size and light weight.

Aerial dispersal separates these widespread animals from many genera of Palearctic origin, which migrated into the Levant by the diffused, gradual movement of populations across hospitable terrain over many generations. It enables constant colonization. In the Levant, for land snails, birds are probably a more important aerial transporter than the wind. They migrate very regularly over considerable distances and are large enough for the accidental transport of minute animals without inconvenience. Birds migrating from Eurasia to Africa via the Levant could thus be responsible for at least some of the minute fauna of the Levant.

## V. FACTORS DETERMINING DIVERSITY

Diversity fluctuates considerably in the different geographic regions of the Levant. The following sections focus on rain and substratum, the major physical environmental forces that, by determining distribution, influence diversity; and on vegetation, an important biotic factor.

### A. Rain

The Levant is situated on the margin of the extreme desert and the mean annual rainfall decreases from 1000 mm to less than 100 mm over the short distance of 150 km. This sharp gradient is reflected in animal ranges, with borderlines coinciding with isohyets so closely that there can be little doubt that rainfall is a major factor and of paramount importance in regulating animal distributions (Fig. 8).

In rain-regulated distributions the species are classified into two categories: allopatric and parapatric. Allopatric species do not share borders with any other species of their genus. The dependence of such a species' range on rain, directly or indirectly, seems unequivocal. (For example, the Mediterranean-dwelling snails *Paramastus episomus* and *Pene sidoniensis* are the southernmost species of their genera. *Paramastus episomus* does not range below the 500-mm isohyet and *P. sidoniensis* not below the 400-mm isohyet.) Parapatric species pairs, on the other hand, share a common border of distribution. The location of the frontier within such pairs may well be influenced by interspecific competition, and rain could simply be the factor determining the point of equilibrium between the species (*Sphinc-*

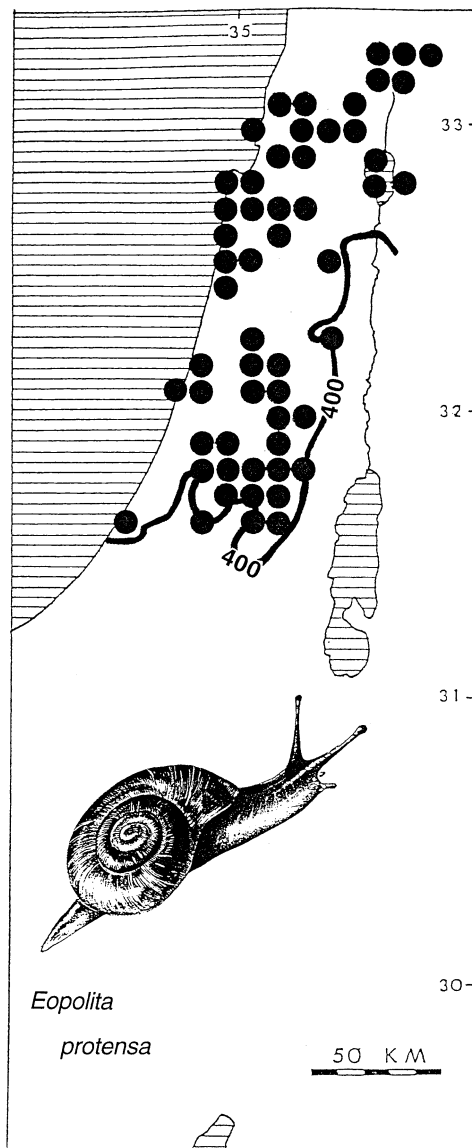


FIGURE 8 Distribution patterns that correlate with rain: The land snail *Eopolita protensa* is delimited by the 400-mm isohyete. Each dot represents all samples collected within a  $5 \times 5$  km area (adapted with permission from Heller, 1993).

*terochila cariosa* and *S. fimbriata*, on the 500-mm isohyete, exemplify this group).

In many groups, overall biodiversity is a function of the amount of rain. In this context, the 200-mm isohyete, representing the end of the Mediterranean and the beginning of the desert landscape, appears to represent a barrier that many animal species cannot cross. In the Negev and Judean Deserts (areas receiving less than 200 mm of rain annually) only 25 land snail species

are found compared to 73 species in the Mediterranean landscape. Of these 25 desert species, 7 occur in both the desert and the Mediterranean habitats. A generally similar pattern occurs at the generic level: Only 10 genera are recorded from the deserts of the Levant compared to 40 genera in the Mediterranean landscape; of these 10 desert genera, 8 occur in both desert and Mediterranean landscape and only 2 occur exclusively in desert areas. These comparative data indicate that adaptations to desert conditions are generally restricted to the species level rather than to higher taxonomic categories.

A similar trend of decrease in diversity also occurs among dipteran families (Table IV). In four latitudinal sections of Israel that broadly represent a decreasing gradient in rainfall, Freidberg (1988) noted that of 77 families, 20 were progressively restricted to the more northern sections, with 2 families occurring only in the extreme north. Only 57 families were detected from all four sections.

Among arachnids, the 350-mm isohyete (rather than the 200-mm one) demarcates the end of the Mediterranean and beginning of the desert fauna.

Studies of species distributions along climatic gradients may aid in predicting potential faunal responses to global climatic change. However, although a "univariate" approach in which a separate model is constructed for each species is more accurate than a multivariate one, it is not very efficient for analyzing climatic responses of large species assemblages or high taxonomic groups. However, scaling up from the level of individual species to higher taxonomic levels is crucial for evaluating faunal responses to global climatic change. Early suggestions (Heller, 1988) that rainfall is important in determining patterns of land snail diversity in Israel were based on nonquantified generalizations deduced

TABLE IV

Occurrence of Diptera Families in Latitudinal Sections of Israel<sup>a</sup>

Latitudinal section	No. of families common to various latitudinal sections				Total no. of families
	A-D	A-C	A-B	A	
A (N of 33°)	57	10	8	2	77
B (32-33°)	57	10	8	—	75
C (31-32°)	57	10	—	—	67
D (S of 31°)	57	—	—	—	57

<sup>a</sup>From Freidberg (1988).

from distribution maps of single species. Recent research (Kadmon and Heller, 1998) analyzes the response of the entire land snail fauna of Israel to gradients in mean annual rainfall by integrating geographical information system (GIS) tools and standard multivariate techniques. Faunal variation of land snails was found to be significantly correlated with underlying variation in rainfall (as expected), but the effect of rainfall on the fauna was much greater in drier regions than in more rainy areas. Above 450 mm, no relationships could be detected between the patterns of faunal variation and rainfall.

### 1. Adaptations to Dry Conditions

Many desert species of the Levant differ from closely related Mediterranean species, or subspecies or populations, in their physiological adaptations to stress (Shkolnik, 1988). In the harsh environment of the desert, shortage of water and scarcity of adequate food are the major factors that pose a threat to many animals. I briefly discuss the various strategies that land snails and mammals employ to overcome these threats.

Concerning water economy, important adaptive features in determining a snail's ability to inhabit dry deserts include not merely resistance to desiccation but also instant adjustment of the rate of water loss to oncoming desiccation. Within the snail genus *Sphincterochila* (Sphincterochilidae), desert-dwelling *S. zonata* is capable of immediate response to desiccating conditions compared to Mediterranean species that take 3 or 4 days to recruit their water-preserving mechanisms; the steppe-dwelling species *S. fimbriata* is intermediate in its speed of adjustment (Arad *et al.*, 1989). A similar picture emerges among hygromiid and helioid land snails of the Levant. Desert-dwelling *Trochoidea simulata* and steppe-dwelling *Xeropicta vestalis* respond rapidly to desiccation, whereas the Mediterranean *Theba pisana* and *Monacha haifaensis* require up to 4 days to adjust their rate of water loss (Arad *et al.*, 1992).

Barriers to water loss in land snails of Levantine deserts include the shell; consequently, land snails in which the shell is missing, internal, reduced to a small external remnant, or without considerable calcium enforcement do not enter the desert of the Levant (*Limax*, *Limacus*, *Milax*, *Deroceras*, *Daudebardia*, *Eopolita*, *Oxychilus*, *Vitrea*, and *Oxyloma*). The epiphragm (a mucous-calcareous sheet spread over the shell aperture during seasons of inactivity) is also an important barrier to water loss, and *S. zonata* has the thickest epiphragm and also lower water loss than the Mediterranean species of *Sphincterochila*. Another adaptation is the extrabody water reservoir that snails carry inside their shell: Spe-

cies that lose significant amounts of water during desiccation do so almost equally from both body and extrabody compartments. However, in desert species the water content of the body is more closely controlled at the expense of extrabody water, thereby avoiding severe dehydration of soft body tissue.

Superiority in resistance to desiccation, however, is not always unique to desert animals. Mediterranean-dwelling *X. vestalis* enjoy a water loss of only 0.13% per day and cope better with desiccation than do desert-dwelling *T. simulata* (Arad, 1990). However, *X. vestalis* is a semelparous, annual species. It cannot survive even one single rainless year in the desert since all the populations would be wiped out in such an event. *Trochoidea simulata*, with somewhat less efficient mechanisms, reaches a life span of at least 3 years and can survive a rainless year (Heller, 1988; Arad, 1990).

Furthermore, within the genus *Sphincterochila*, desert-dwelling *S. prophetarum* has water regulatory capacities similar to those of the Mediterranean species of this genus. *Sphincterochila prophetarum* dwells underneath stones, where humidity is high compared to other microhabitats of the Levantine desert.

Desert snails of the Levant, to conclude, are adapted to withstand desiccation stress in that they usually have lower and slower rates of water loss and are capable of closer regulation of stable water content of the body compared to Mediterranean snails.

Desert mammals of the Levant differ from Mediterranean mammals in that they have a lower energy metabolism. Among murid rodents of Mediterranean landscapes, *Apodemus sylvaticus* and *A. mystacinus* possess the basic metabolic rate as expected from their body mass (according to the "mouse to elephant curve"). However, in the common spiny mouse *Acomys cahirinus*, a species of both Mediterranean and desert habitats, this rate is only 75% of the expected value, and in the golden spiny mouse *A. russatus*, a species of the extreme desert and regularly active during the day, it is only 55%. Similar comparative patterns have been found among gerbils, hedgehogs, carnivores, and ruminants of the Levant. A low metabolic rate, implying a lower rate of heat generation, saves the water otherwise needed to dissipate heat in a hot environment (Shkolnik, 1988).

In addition to a shortage of water, food in the desert is also at a premium. The frugal food requirements of desert animals may be related not just to a low demand for metabolizable energy. An efficient digestion of the food consumed may also reduce the amount of food required for maintenance. Desert mammals require less food for their maintenance than do nondesert ones.

Bedouin goats herded in the southern Levant consume less food, retain it in the gut for a longer time, and digest it more efficiently than do European breeds of goat, thereby gaining more energy from a given mass of food. In addition, the Bedouin goat has a remarkably spacious rumen, which functions not only as a fermentation vat but also as a voluminous water reservoir, enabling the goat to graze in the water-depleted desert without depending on frequent drinking (Shkolnick, 1988).

Balancing their nitrogen metabolism is as much a challenge for desert mammals as the maintenance of a balanced energy metabolism. The camel, instead of wasting the nitrogenous end products of protein catabolism, first retains them in its kidney and later allows them to be recycled through the gut. Here, the urea and other nitrogenous wastes are used by the microbial symbiotic population for resynthesis of protein, from which the host animal will eventually benefit. The potential capacity for recycling urea is far greater in desert than in Mediterranean species. Recycling of urea, in addition to helping the animal survive on low-protein feed, attenuates the load on the water otherwise required for the elimination of that waste (Shkolnik, 1988).

In conclusion, rain is the major environmental factor determining the diversity of animals within the Levant. The (approximately) 200-mm isohyet marks the arid limit beyond which many species of the Levant cannot exist.

## B. Substratum

Within the boundaries of the relevant isohyets, animal distribution is further limited by lithic factors: The checkered geomorphologic map of the Levant presents a wide range of substrata, including granite, basalt, limestone, chalk, marl, heavy alluvium, calcareous sandstone, and light sand dunes.

Calcium-rich environments are very abundant in the Levant. Concerning one particular animal group, land snails, the abundance of limestone, chalk, or calcium-rich soils derived from them is a prerequisite for the flourishing of any testaceous fauna. Acid environments, which restrict the diversity in many regions of Europe and which in many evergreen forests throughout the world exist due to the gradual addition of acidic elements to the litter, are not known to exist in the Levant to any sizable extent. Granite and basalt, from which the extraction of calcium is difficult, are restricted to small areas. This lithic composition is an important reason for the high diversity of land snails in the Levant.

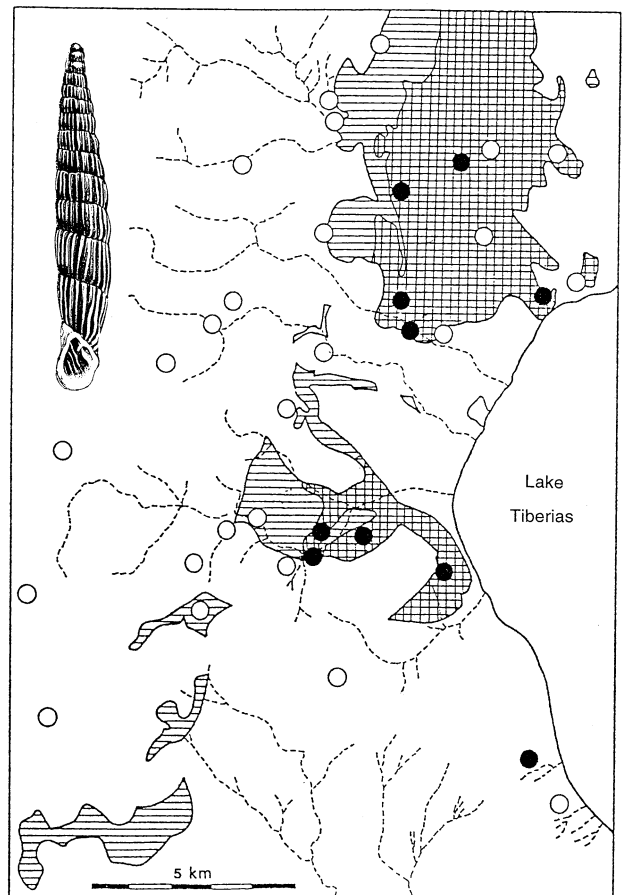


FIGURE 9 Distribution patterns that correlate with substratum: The land snail *Cristataria genezarethana* is restricted to the mid-Eocene formation (squares) and absent from the lower Eocene formation (lines). ●, Sites in which snails were found; ○, sites in which they were not found (adapted with permission from Heller, 1993).

In some cases, the distribution of land snails varies in close association with geomorphology. The endemic *Cristataria genezarethana* occurs almost exclusively on mid-Eocene formations and is absent from nearby lower Eocene ones. Both formations consist of calcium, but only the mid-Eocene rocks are rich in crevices, which *C. genezarethana* inhabits (Fig. 9). However, among many groups, only two major categories of substratum-dependent groups can be distinguished: the fauna of the limestone mountains and that of the loose sands. Limestone mountains in the Levant are the substratum that is richest and most diverse in species because they produce a rich mosaic of habitats and provide many shelter niches. Complex systems of rock crevices, moderately dense litter, soil pockets, generous vegetation cover, and a variety of exposed versus shaded sites are

all to be found on limestone but not always on other substrata. Accordingly, a  $100 \times 100$  m of limestone in the Judean Hills (in a Mediterranean landscape that receives approximately 570 mm of rain annually) yields 15 species of snails. A similar nearby area consisting of chalk, marl, or alluvium produces only 4 species, all of which also occur on limestone. Chalk, marl, or alluvium merely contain an impoverished, depauperated element of the rich limestone fauna but without any characteristic component of their own. Crevices provide shelter from the severe heat and drought prevailing throughout the summer. Whereas limestone has a well-developed system of crevices, chalkstone, marl, and alluvium do not and therefore, presumably, crevice dwellers cannot inhabit them. Also, the rich assemblage of litter habits, inhabited mainly by small invertebrates, is often well developed in limestone but nonexistent in nearby chalk or marl. Accordingly, there are no litter dwellers in such substrata. Extensive alluvial plains, such as Yizre'el, Zevulun, or the hinter parts of the Levant's coastal plain, harbor little diversity.

Sands of the Levant contain a fauna that differs considerably from that of the nearby calcareous mountains and alluvial soils (Kadmon and Heller, 1998). Furthermore, sands of the Mediterranean climatic region may differ from those of the Negev (Fig. 10), which is arid.

### 1. Adaptations to Sand

Sand differs from other substrata in its "near-fluid" texture, which makes locomotion and burrowing more difficult than in other habitats. However, it holds more water than many other substrates so that more water is eventually translated into a more productive biotic community.

Many taxa of the animals restricted to sands converge in their morphological adaptations, thereby signifying the importance of sand as an evolutionary agent. Among reptiles of the Levant, a suite of morphological, physiological, and behavioral adaptations enables certain taxa to live in the sands or endows them with a competitive advantage (Werner, 1985). Locomotion over sand is facilitated in lizards by expansion of the feet. In *Acanthodactylus*, the extent of fringing along the toes varies with the looseness of the substrate intraspecifically and interspecifically (Fig. 11A). Also, the toes of the gecko *Stenodactylus* ssp. are fringed, and those of the skink *Scincus scincus* are flatly expanded (Fig. 11B). Sand-dwelling viviperids locomote by sidewinding, as do *Cerastes* spp. and *Pseudocerastes fieldi*. *Lytorhynchus diadema* employs a unique variety of serpentine locomotion mechanisms wherein the loops push the sand down rather than sideways. Since burrows tend to collapse,

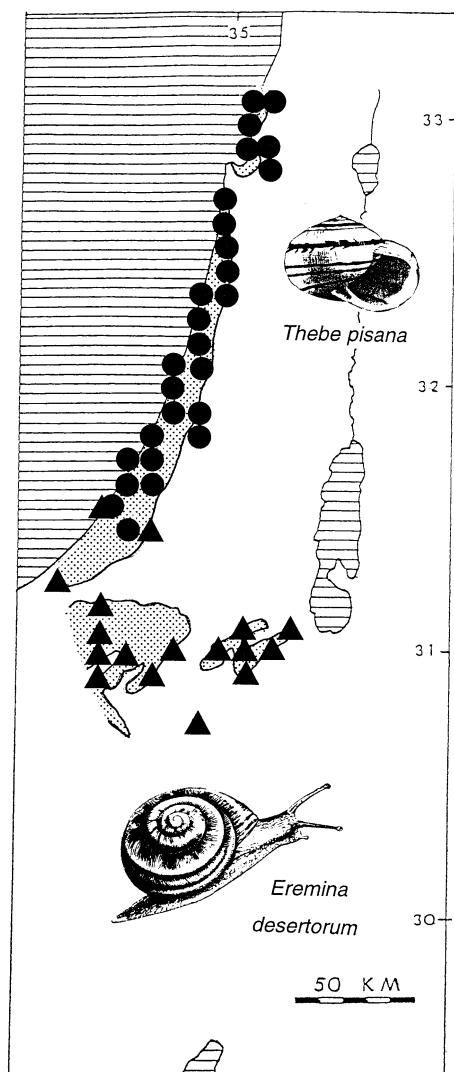


FIGURE 10 Distribution patterns that correlate with substratum: Land snails restricted to sand. *Theba pisana* (●) occurs in sands of more northern, Mediterranean climatic regions, whereas *Eremina desertorum* (▲) occurs in the more southern sands of the Negev. Each dot represents all samples collected within a  $5 \times 5$  km area (adapted with permission from Heller, 1993).

on the one hand, and loose sand enables instantaneous submergence and sand swimming, on the other hand, problems arise from submergence and locomotion within the sand medium. Wedge-shaped snouts, frequent in burrowing reptiles, are prominent in the sand-swimming skinks *S. scincus* (Fig. 11B) and *Sphenops sepsoides*. In these and in the snake *Lytorhynchus* the mouth opening is protected from sand during submerged progression by its ventral position. *Lytorhynchus* has an expanded rostral, common but not unique

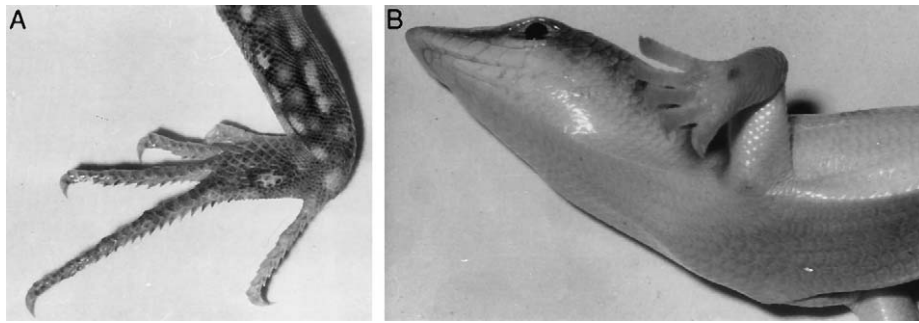


FIGURE 11 Some adaptations of Levantine psammophilous reptiles to sand. (A) Foot of *Acanthodactylus schreiberi syriacus* showing fringe of spiny scales along third and fourth toes. (B) *Scincus scincus* showing shovel-shaped snout, countersunk jaw, angularly set-off venter, and expanded toes (photographs by Y. L. Werner).

to sand-dwelling snakes. *Cerastes* spp. can submerge stationally by shuffling the flattened body sideways. The vertical flanks of *Scincus* (and *Sphenops*) may improve the lateral thrust during sand swimming. The relatively high species density of Gekkonidae on the sand may reflect the preadaptation of spectacled forms to sand. Breathing below sand is problematical not only at the nasal passage level but also at the thoracic: Upon exhalation the sand caves in and stifles inhalation. *Scincus* and *Sphenops* are structured for breathing with the protected ventral surface angularly set off from the lateral ones (Fig. 11B; Werner, 1985).

Sand-dwelling scorpions of the Levant can also be distinguished by their morphological adaptations. The most well-known are bristles to aid traction, locomotion, and burrowing in sand. These bristles occur on the tarsomeres and tibiae of all sand-living scorpions. Another leg modification in sand scorpions that acts to increase the effective surface area of the tarsi is long claws (Fet *et al.*, 1998). The bristles and claws create an expanded surface area on the tip of the legs and prevent sinking into the sand. The genus *Buthacus* (represented by *B. arenicola*, *B. yotvatensis*, *B. l. leptochelys*, and *B. l. nitzani*) is a typical sand dweller in the southern Levant, with up to 25 bristles on its walking legs (Levy, 1980).

Sand-dwelling grasshoppers can be distinguished from those of other biotopes by their behavioral adaptations. Sand-dwelling *Hyalorrhapis calcarata*, *Leptopternis* spp., and *Eremogryllus hammadae* dig into the sand with their hindlegs, throwing the sand behind them. They are thus able to cover themselves, leaving only the upper part of the head and the antennae exposed above the surface (species of other substrata usually settle "head to sun," thus casting the smallest shadow; Fishelson, 1985).

In conclusion, substratum is the second major environmental factor determining animal diversity in the Levant. Limestone rocks offer calcium, shelter, and an additional microhabitat—that of litter. Sands form a separate substratum category that usually has a unique fauna.

### C. Vegetation

Both animal and plant distribution are influenced by rainfall and substratum, and consequently their distribution patterns may overlap. Whether vegetation influences animal distribution directly, in that different animal species have food preferences for different higher plants, varies among groups. Among fruit flies (Tephritidae), the larvae of practically all species in the southern Levant are phytophagous. All representatives of the Myopitinae, Oedaspidinae, Tephritinae, and Schistopterinae (65–70 species, 25 genera) develop only in plants of the composites (Compositae), whereas none of the Aciurinae and Dacinae do so. The daciine *Dacus oleae* attacks olives (Oleaceae), whereas other daciines are associated with fruits of milkweed (Asclepiaceae). Many tephritid species are oligophagous; a few are polyphagous, using plant hosts of different families. The parts of plants affected by tephritid larvae also vary considerably among groups. Larvae of Dacinae and Trypetinae develop in fleshy or juicy fruits. However, *Euleia heracleii* is a leaf miner, and *Capparimyia savatani* (Fig. 12) produces larvae that develop in the flower buds of the caper *Capparis*. Most species of the five subfamilies that attack Asteraceae in Israel develop in fruit heads. The larvae of different species are usually restricted to feeding on certain parts of the flower head, such as flowers, achenes, or the receptaculum. Some species induce the formation of galls (e.g., *Myopites*

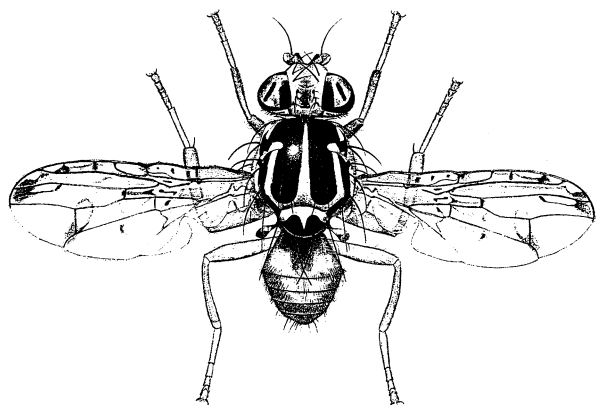


FIGURE 12 Distribution patterns that correlate with vegetation: *Caprimyia savastani* (Tephritidae), the larvae of which develop in flower buds of *Capparis* (reproduced from Freidberg and Kugler (1989) with permission from the Israel Academy of Sciences and Humanities, frontispiece).

spp. and *Urophora* spp.). Several species bore into the stems of their hosts, often forming galls (e.g., *Spathulina* spp. and *Oedaspis* spp.). *Orellia falcata*, unlike most members of the Terelliinae, mines along the stem of the goat's beard *Tragopogon* and pupates in the upper part of the root (Freidberg and Kugler, 1989).

Some grasshopper species use plants as a food source, resting source, and hideout. Morphological adaptations to this habitat include large arolia to serve for attachment, streamlined and smooth body structures, heads with an acute or subacute angle, long and delicate colorless wings, and usually green or greenish-gray or brown coloration. Behavioral adaptations include resting along branches or stems and, if flying up, moving from plant to plant. When disturbed, many of these species crawl around the stem or branch, remaining on the opposite side to the intruder; others drop from the plant and dig among the dense branches close to the ground. There is a strong correlation between special types of vegetation and their fauna of plant-dwelling acridoids within the Levantine fauna: The Tropidopola–Ochridia association of grasshopper species is found on reed *Phragmites*, reed mace *Typha*, and rush *Juncus*; the *Heteracris*–*Eyprepocnemis* association shows a great affinity for knotweed *Polygonum* sp. and wormwood *Artemisia monosperma*; and the *Dociostaurus*–*Morphacris* association occurs mainly in areas with low, usually dry, ephemeral grass (Fishelson, 1985). It is not known if these associations are obligatory.

In contrast to fruit flies, among land snails of the Levant a direct influence of vegetation on diversity is unlikely. Snails feed on a wide spectrum of decaying

vegetable matter and saprophytic fungi, occasionally on green tissues, and there is no evidence that food requirements for specific plants determine their wide-scale distribution patterns. Vegetation, however, may play an important role in microgeographic distribution, especially in preventing some snails from occupying certain habitats. Some species of the Mediterranean region (*Buliminus labrosus*, *Sphincterochila cariosa*, *Euchondrus septemdentatus*, *E. saulcyi*, *Paramastus episomus*, and *Xeropicta vestalis*) are found on the southern slopes of hills, where annual vegetation prevails, and are very uncommon on the northern slopes that are covered with dense oak maquis. In regions in which the vegetational distinction between northern and southern slopes is not obvious, the distinction of the snails between the slopes also becomes obscure, and they also occupy the northern slopes. A preference for the south slopes could perhaps be due to the annual vegetation prevailing on it. Annual plants do not develop xeromorphic characters and are therefore easily eaten and digested. The oak maquis of the north-facing slope may be cool and damp and offer more shelter, but it consists mainly of perennial vegetation. The development of xeromorphic characters (such as the thick cuticle of the oak leaves) necessary for plant survival in the dry season could perhaps be an obstacle for snail feeding and digestion. Too little information is available on the precise diet of many animal groups. Once this information is gained, I believe that a major breakthrough will occur in our understanding of biodiversity in the Levant.

## VI. FRESHWATER DIVERSITY: HISTORIC FACTORS

The inland water system of the Levant is dominated by the north–south-oriented topography of the Rift Valley. The three major rivers of the Rift (the Jordan, Litani, and Orontes) run along successive segments of the valley and create a “steep chase” waterway. Between the Rift basin and the Mesopotamian basin, watershed shifting and headwater capture may well have enabled a faunal transition of transcontinental dimensions.

During the Pliocene the Mediterranean flooded the southern Rift Valley and upon its retreat it left, in the freshwater of the Jordan Valley, several marine relicts. Crustacean species of the Jordan Valley suggested by Por (1975) as marine relicts of the Pliocene invasion include *Loxoconcha galilea*, *Pseudobradya barroisi*, *Nitocra balnearia*, *Monodella relictia*, *Typhlocirolana reichi*,

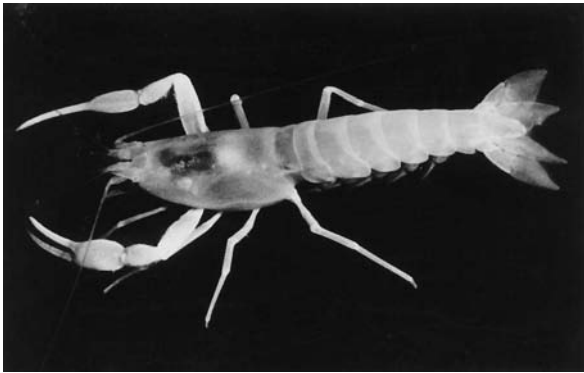


FIGURE 13 The blind prawn *Typhlocaris galilaea*, a relict of the Pliocene marine ingression into the Jordan Valley (photograph by Dr. M. Tzurnamal).

*Typhlocaris galilea*, and *Bogidiella hebraea*. The blind prawn *Typhlocaris galilea* (Fig. 13) is endemic to a subterranean warm, sulfide-rich spring near the shores of Lake Kinneret, where it feeds on oligochaetes (*Isochaeta israelis*) and snails (*Theodoxus jordani*).

From the Pleistocene onwards there was a continuous existence of freshwater or saline lakes in the Jordan Rift Valley. In this large endorheic system, the smallest climatic change and the changing balance of evaporation and precipitation (in addition to any slight tectonic movement) resulted in considerable fluctuations of water levels, with the creation, fusion, separation, and disappearance of lakes and rivers. This resulted in richly diverse faunas that partly replaced one another throughout the Pleistocene. Some old lacustrine freshwater endemics have survived in the Jordan system since the early Pleistocene (Table V).

The fish fauna of the Levant is a mixture of Palearctic, Oriental, and Ethiopian species. Thirty-four species of freshwater fishes occur in the Orontes, the largest river of the Levant. Of these, approximately two-thirds are Palearctic, one-third are Oriental, and only 3% are African elements. The Orontes shares 18 of its species with the Tigris–Euphrates but only 7 with the Jordan River. The number of species occurring in the Jordan, the second largest river, is only slightly lower, but the species composition is totally different: Of 28 species, the Palearctic, Oriental, and African elements each comprise about one-third. The Jordan River shares only 3 species with the Tigris–Euphrates, but 6 additional species have their closest relatives in Mesopotamia. Apparently, the Jordan and various courses of the Orontes were separately colonized via branches of the Mesopotamian river system, and the Jordan drainage basin be-

came an important center of speciation for several fish lineages with Palearctic affinities. Later, faunal elements of the Jordan reached the Orontes (through the Litani). Some freshwater fish of African origin reached the Jordan Valley by way of freshwater connections to the south. They could have done so up to Pliocene times, and in the Jordan Valley they differentiated specifically or even generically (*Astatotilapia flavijosephi* and *Tristramella* spp.). During later periods, some African fish may have reached the southern Levant from the Nile via the Mediterranean Sea (the euryhaline cichlid *Tilapia zilli* survives well in brackish and even marine waters) (Fig. 14). The fish fauna of the Rift's lakes and rivers was thus recruited from the north (Palearctic), east (Euphrates), and south (the Nile; Krupp, 1987).

The headwaters of Jordan constitute a typical temperate Palearctic freshwater fauna, with a "normal" level of diversity that includes pulmonate snails, cold-water copepods, and many stream-living insects such as stone flies (Plecoptera), Elmidae among the Coleoptera, and typically cold-water-like *Rhyacophila* (Trichoptera). There are no less than 12 species of caddis flies (Trichoptera) in the springs of the River Dan; there are 13 species of Ephemeroptera in the River Hatzbani.

TABLE V  
Endemic Species of the Jordan Rift Valley<sup>a</sup>

Group	Species
Porifera	<i>Cortispongilla barroisi</i>
Tricladida	<i>Dugesia salina</i> <i>D. biblica</i>
Oligochaeta	<i>Isochaeta israelis</i>
Halacarida	<i>Limnohalacarus capernaumi</i> <i>Lohmanella heptapegoni</i>
Crustacea	<i>Loxococoncha galilea</i> <i>Ilyocypris harmanni</i> <i>I. nitida</i> <i>Pseudobradia barroisi</i> <i>Nitocra incerta</i> <i>Nannopus palustris tiberiadis</i> <i>Schizopera taricheana</i> <i>Parabathynella calmani</i> <i>Monodella relict</i> <i>Typhlocirolana reichi</i> <i>T. steinitzi</i> <i>Typhlocaris galilea</i> <i>Bogidiella hebraea</i>

<sup>a</sup> From Por (1975), select groups.



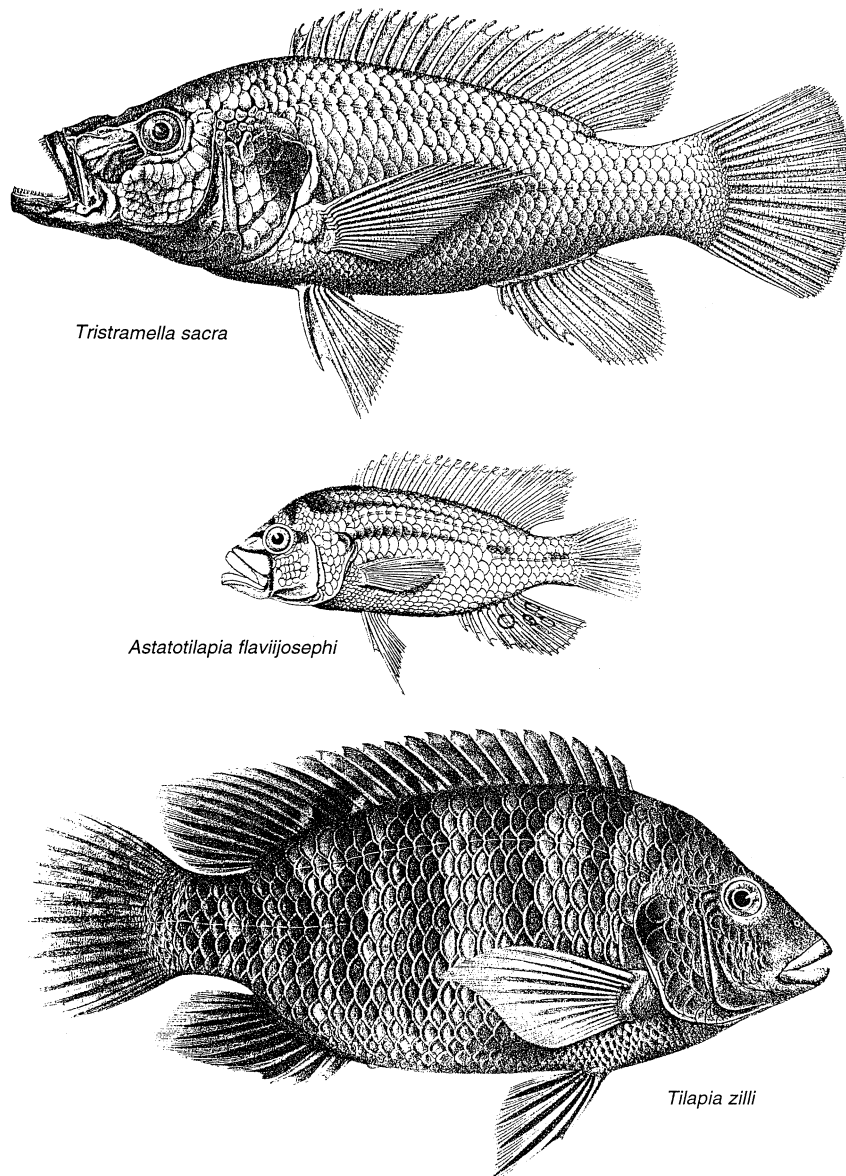


FIGURE 14 Cichlids of the Levant: *Tristramella sacra*, an endemic genus of African origin; *Astatotilapia flavijosephi*, an endemic species of African origin; and *Tilapia zilli*, an African species that may have reached the Levant via the Nile and Mediterranean Sea (from Lortet, 1883, *Etudes Zoologiques sur la Faune au Lac de Tiberiade*, Part I. *Arch. Mus. d'Histoire Nat.* 3, 99–194, Lyon, France).

Elsewhere in the southern Levant, there is little of this rich freshwater fauna. In all the springs and brooks of Israel, usually only two species of Ephemeroptera can be found, *Cloeon dipterum* and *Caenis macrura*, which are well-known for their broad ecological valency. Only a small number of freshwater invertebrates are found in every spring and brook. This is a monotonous fauna characterized by *Melanopsis*

*buccinoidea* (Mollusca), *Proasellus coaxalis*, *Echinogammarus syriacus* and *Eucyclops serrulatus* (Crustacea), and *Dina lineata* (Hirudinea). However, even this impoverished fauna does not advance much further than the oasis springs of En Avdat and of Qadesh Barnea, on the brink of the extreme desert. In most springs of Sinai, usually only insects or passively transported animals occur (Por, 1975).

## VII. BIODIVERSITY: THE HUMAN IMPACT

Since the dawn of the Pleistocene, the Levant has been inhabited by the evolving hominids, starting with the early *Homo erectus* of the Jordan Valley. It was in the Levant that the 10,000-year-old Neolithic revolution of the Old World, with its agriculture and animal domestication, occurred.

The history of the Levant is characterized by tides of highly sophisticated agricultural civilization and ebbs of nomadic destruction. Agricultural apogees saw the development of soil protection, terracing, aqueduct and irrigation works, and drainage of swamps, whereas nomadism brought erosion, overgrazing, destruction of irrigation works, and swamping. Nomads semiconsciously created the deserts, of which they are often called the sons. Throughout history there was a progressive destruction of the woodland habitats by ever-increasing deforestation.

However, the most profound effects of man on biodiversity of the Levant started at the beginning of the twentieth century with the tremendous increase in human populations, use of firearms, extent of cultivated areas, and use of pesticides. These changes have had a pronounced, and in many cases fatal, effect on wildlife in the Levant. The effects of various human activities on animal life in the Levant (mainly in Israel and mainly concerning vertebrates) have been studied in depth by Yom-Tov and Mendelssohn (1988).

### A. Hunting

The proliferation of firearms into the Levant by the end of the nineteenth century was followed by the overhunting and extermination of many game animals, mainly at the beginning of the twentieth century. Roe deer *Capreolus capreolus* and fallow deer *Dama dama mesopotamica* used to occur in the woods of Galilee and of Mt. Carmel; the last roe deer was shot in approximately 1912 and the last fallow deer probably at the beginning of the twentieth century. Oryx *Oryx leucoryx* still lived in Jordan at the beginning of the twentieth century; the last were shot before 1950. The last onagers *Equus hemionus hemippus* survived in the Syrian desert until the early 1930s. The bear *Ursus arctos syriacus*, which used to live near the Sea of Galilee and on the slopes of Mt. Hermon, was shot to extinction during World War I. The last cheetah *Acinonyx jubatus* in Jordan was shot in 1962. Two leopard subspecies existed in the Levant. The northern one, *Panthera pardus tulliana*,

used to be found in Turkey, Lebanon, and Syria and was not rare in the Galilee at the beginning of the twentieth century; the last individual was killed in 1965, and today this subspecies is extinct. The other subspecies, *P. p. nimr*, survives in the Judean desert and Negev. An ostrich (*Struthio camelus syriacus*) was seen in Jordan in 1932; now this species is extinct in the Levant. The Nile crocodile *Crocodylus niloticus* was found in swamps near Mt. Carmel as late as 1912 (Yom-Tov and Mendelssohn, 1988).

Hunting also caused the almost total elimination of the green and loggerhead turtles (*Chelonia mydas* and *Caretta caretta*) which used to nest along the sandy Mediterranean beaches. During 1920–1930 approximately 30,000 turtles were killed along Israel's seashores; in 1985 only 14 nests were found along its whole Mediterranean coast.

Today, effects of hunting on wildlife in the Levant vary in broad correlation to the enforcement of hunting laws (and hence to political borders). In Israel, hunting today has only minor effects. There are only approximately 5000 licensed hunters and the main game species are chuckar partridges, rock pigeons, doves, various wintering waterfowl, hare, porcupine, wild boar, and mountain gazelle where they damage agricultural crops. Many of these may be hunted only during well-defined seasons, and none are seriously affected by hunting. Indeed, increasingly efficient enforcement of hunting laws has enabled several species to recover: the leopard *Panthera pardus nimr*, of which approximately 20 individuals now dwell in the Judean Desert and Negev; the wolf *Canis lupus pallipes*; and the hyena *Hyaena hyaena*. All wild mammals, birds, and mollusks in Israel are completely protected by law, with the exception of designated pest species (rats, house mice, voles, fruit bats, house sparrows, etc.).

### B. Habitat Destruction

Most habitats of the Levant have been severely or completely destroyed by urban and agricultural development. Two examples are presented in the following sections (Yom-Tov and Mendelssohn, 1988).

#### 1. Coastal Sand Dunes

Coastal sand dunes form a narrow strip along the Mediterranean coast of the Levant, and until the beginning of the twentieth century were almost uninhabited by man. Today, however, the majority of the population of Israel lives on the coastal plain. The many urban settlements built on the coastal sands have caused loss of sand dune areas to such considerable extent that of

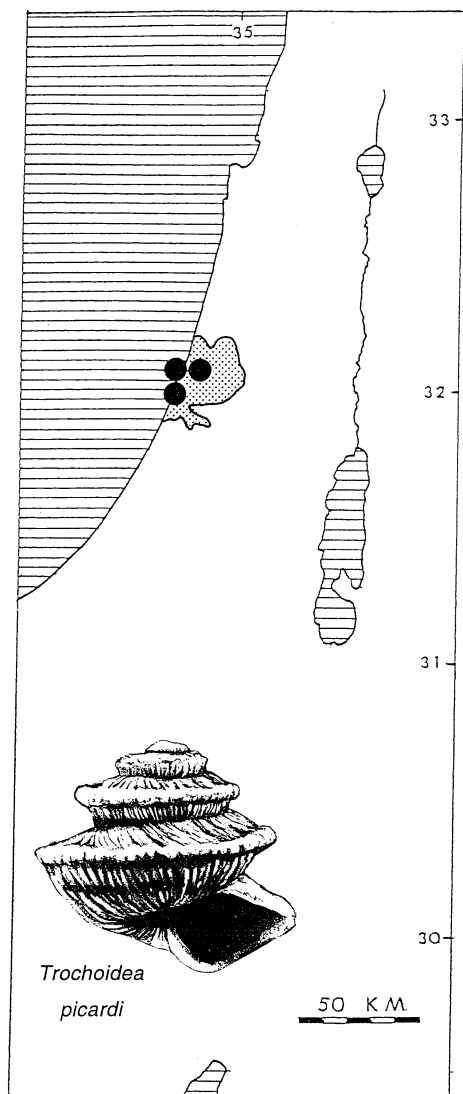


FIGURE 15 Extinction due to habitat destruction: total global distribution of the land snail *Trochoidea picardi* (museum records; each dot represents all samples collected within a  $5 \times 5$  km area). Stippled area, Tel Aviv metropolis, 1999 (adapted with permission from Heller, 1993).

110 km<sup>2</sup> of dunes which formerly occurred north of Nahal Soreq, only 3 km<sup>2</sup> have not been developed.

The land snail *Trochoidea picardi* was an endemic species whose entire small distribution area once ranged over a few kilometers of calcareous sand dune ("kurkar") along the coast of Israel. Today, the entire distribution range of *T. picardi* is buried beneath the motorways, skyscrapers, and residential areas of the Tel Aviv metropolitan area, and this species is almost surely extinct (Fig. 15). Concerning nonendemics, the coastal

sand dunes are the most eastern extension of Saharan sand-dwelling animals into the Mediterranean, temperate zone of the Levant: the Egyptian tortoise *Testudo kleinmanni*, snakes (*Cerastes vipera*, *Macroprotodon cucullatus*, and *Lytorhynchus diadema*), skinks (*S. scincus* and *Sphenops sepsoides*), geckos (*Stenodactylus stenodactylus* and *S. petrii*), lizards (*Acanthodactylus scutellatus* and *Agama savignii*), a monitor (*Varanus griseus*), a chameleon (*Chamaeleo chamaeleon musae*), rodents (*Gerbillus gerbillus*, *G. pyramidum*, and *G. andersoni allenbyi*, the endemic jerboa *Jaculus jaculus schlüeri*, and the endemic jird *Meriones sacramenti*), and a hedgehog, *Hemiechinus auritus*. The destruction of the coastal sand dunes (accompanied by predation by feral domestic cats and dogs) has reduced the distribution and abundance of these species.

## 2. Wetland Habitats

Wetland habitats that existed in the Levant at the beginning of the twentieth century included swamps, lakes, rivers, springs, and temporary (mainly winter rain) pools; man-made fish ponds were then barely existent. Today, however, water in the Levant is a scarce and much needed resource. All these habitats have been affected.

### a. Swamps

At the beginning of the twentieth century, the Hula swamp ranged more than 40 km<sup>2</sup> north of the (13 km<sup>2</sup>) Hula Lake (Fig. 16) and contained a rich diversity of (mainly widespread) animals. The swamp and lake were drained in the 1950s, but a small nature reserve (3 km<sup>2</sup>) was created that represents an impoverished version of the original swamp. Even so, 22 species endemic to the Levant have not been recorded from the Hula since its drainage (Table VI). Although the Hula nature reserve is today rich in bird life, eight bird species that once bred in the Hula Valley do not breed there anymore.

### b. Riverine Habitats

The rivers, streams, and wadis in the Levant drain, in general, either to the Mediterranean Sea or to the Rift Valley. Today, most Mediterranean-flowing rivers of the southern Levant are polluted, and consequently most fish have disappeared from them. Gray mullet (*Mugil cephalus* and *M. ramada*) and eel (*Anguilla anguilla*) fingerlings once used to spend their first years in these rivers (they reproduce in the sea) but now do not occur in most of them. Eight other, entirely freshwater, species are today very rare and the endemic *Acanthobrama telavivensis* now faces extinction. The soft-shelled turtle *Trionyx triunguis*, which once lived in many of the



FIGURE 16 Extinction due to habitat destruction: Lake Hula, now drained (photograph by Kluger).

Mediterranean-flowing rivers and nested on their shores, has disappeared from most of them. Only in Alexander River does a breeding population persist, but many of the eggs fail to hatch and the population consists only of very large specimens. The river otter (*Lutra lutra*), a former inhabitant of all these rivers, survives only in the Bezet, Keziv, and Na'aman, in which it is only occasionally seen.

In the Rift Valley-flowing streams the major threat to biodiversity is the extreme reduction in the amount of water that flows in them since most springs are piped by man and used for irrigation. This has reduced and even eliminated the populations of several animal species. The river otter, once common in each major tributary in the Jordan River system, is now rare. The blue-cheeked bee-eater *Merops supecilosus*, which bred along the Jordan River in the early 1930s, does not breed there any more (Yom-Tov and Mendelssohn, 1988).

### c. Temporary Winter Rain Pools

A multitude of temporary winter rain pools once existed in temperate regions of the Levant. Many existed near Arab villages, where their waters once served for watering livestock and for irrigation. Many of these village pools were man-made, built as reservoirs by erecting a dam across a wadi or hewn into rock at the foot of a water-collecting slope, and were probably ancient. These pools supported a rich and varied invertebrate fauna and several species of amphibians.

The number of winter rain pools decreased sharply during the 1960s and 1970s. Many of the formerly undrained valleys were turned into cultivated fields. Other pools were eliminated due to road construction and building. Continuous spraying of herbicides on roadsides and of insecticides on pools near human settlements destroyed this ecosystem or turned these pools into breeding sites for pesticide-resistant mosquito strains. One result of these operations is that the once abundant amphibians have become a threatened group. The spadefoot toad *Pelobates syriacus* (Fig. 17) breeds exclusively in rainwater pools and its tadpoles require a long (3 months) development period. Only large seasonal ponds of sufficient duration can support a viable population of spadefoot toads, and only 27 such ponds remain in Israel (including the Golan). Lack of suitable breeding localities has also caused the newt *Triturus vittatus* to become a threatened species, and even the formerly very common tree frog *Hyla arborea* and green toad *Bufo viridis* are becoming rare (Yom-Tov and Mendelssohn, 1988).

## C. Poisoning

More than 700 compounds are registered in Israel's Ministry of Agriculture for agricultural use, including insecticides, acaricides, nematocides, fungicides, herbicides, bactericides, molluskicides, rodenticides, insect attractants, bird and mammal repellents, fumigants,

TABLE VI

Aquatic Species Endemic to the Levant and Not Recorded from the Hula Since Its Drainage<sup>a</sup>

---

Sponges (Porifera)
<i>Ephydatia syriaca</i>
Turbellarians (Turbellaria)
<i>Dendrocoelum</i> sp.
<i>Dugesia biblica</i>
<i>Phagocata?</i> <i>armeniaca</i>
Crabs (Crustacea)
Copepods Copepoda)
<i>Attheyella trispinosa affinis</i>
Insects (Insecta)
Dragonflies (Odonata)
<i>Rhythermis semihyalina syriaca</i>
<i>Urothermis edwardsi hulae</i>
<i>Calopteryx hyalina</i>
Caddisflies (Trichoptera)
<i>Orthotrichia?</i> <i>melitta</i>
Beetles (Coleoptera)
<i>Herophydrus cleopatre</i>
<i>Herophydrus galileae</i>
<i>Hydrovatus</i> n.sp.
<i>Laccobius levantinus</i>
<i>Canthydrus ornatus</i>
Mosquitoes (Diptera)
<i>Anopheles pharoensis</i>
Bivalves (Bivalvia)
<i>Potomida littoralis semirugata</i>
<i>Unio terminalis terminalis</i>
Fish (Pisces)
<i>Tristramella simonis intermedia</i>
<i>Num qalilaeus</i>
<i>Mirogrex terraesanctae hulensis</i>
Amphibians (Amphibia)
<i>Discoglossus nigriventer</i>
Mammals (Mammalia)
<i>Arvicola terrestris hintoni</i>
Birds (Aves) breeding
<i>Aquila clanga</i>
<i>Aquila pomarina</i>
<i>Anhinga melanogaster chantrei</i>
<i>Fulica atra</i>
<i>Porzana pusilla</i>
<i>Chlidonias hybrida</i>
<i>Chlidonias niger</i>
<i>Sterna albifrons</i>

---

<sup>a</sup> Adapted from Dimentman *et al.* (1992).

plant growth regulators, and defoliant. Effects of pesticide residues on wildlife started shortly after DDT began to be used at the end of World War II. The most dramatic effect, however, was during the early 1950s, when thallium sulfate-coated grain was widely used to control rodents. These poison campaigns were not necessary. The main damage of these rodents was caused before 1950, when fields were plowed with the traditional shallow-plowing plows that did not disturb rodent burrows. When, after 1950, deep plowing was introduced, rodent burrows were destroyed, the rodents were exposed to predation, and the damage they caused decreased (pest control officers, however, attributed the decrease in rodent damage to their intense poison campaigns). In unplowed areas, as in alfalfa fields, damage caused by voles is still considerable and no rodenticides can prevent it. Since the 1960s, fluoracetamid has replaced thallium sulfate. The residues of both substances accumulate in bodies of secondary consumers such as raptors, bats, and carnivores.

Of 39 species of birds of prey that occurred in Israel before the use of pesticides, all but 2 were seriously affected. Species that once were common breeders became very rare breeders (the black kite *Milvus migrans*, griffon vulture *Gyps fulvus*, long-legged buzzard *Buteo ferox*, Bonelli's eagle *Hieraetus fasciatus*, Egyptian vulture *Neophron percnopterus*, kestrel *Falco tinnunculus*, lesser kestrel *Falco naumanni*, and lanner falcon *Falco biarmicus*). Rare breeders went extinct or their populations decreased drastically (the lapper-faced vulture *Torgos tracheliotus*, spotted eagle *Aquila clanga*, peregrine *Falco peregrinus brookei*, marsh harrier *Circus aeruginosus*, black eagle *Aquila verreauxi*, white-tailed eagle *Haliaeetus albicilla*, bearded vulture lammergeier, and *Gypaetus barbatus*). Thallium sulfate also affected the wintering raptors, whose populations decreased considerably; some species disappeared completely for many years (the sparrowhawk *Accipiter nisus* and merlin *Falco aesalon*). Insect-eating birds, such as lesser kestrel *Falco naumanni* and scops owl *Otus scops*, were perhaps more affected by DDT and other persistent insecticides. The short-toed eagle *Circaetus gallicus* remained unaffected, probably due to its specialized reptilian diet and to its absence in winter when most of the poison grain was used. The latter factor also saved the hobby *Falco subbuteo*, the breeding population of which returns to the Levant only in May.

Today, after prohibition of the use of DDT, other chlorinated hydrocarbons, and thallium sulfate, some species have made a comeback but the existing populations of breeding raptors are only a fraction of their former populations. Hundreds of pairs of griffin vulture



FIGURE 17 Extinction due to habitat destruction: The amphibian *Pelobates syriacus* is now a threatened species because wetland habitats are disappearing (photograph by Y. L. Werner).

once bred in Galilee and on Mt. Carmel, but only 20 breed today; only one pair of lapper-faced vultures bred in 1986 in the Negev, where approximately 25 pairs had bred before the widespread use of agricultural poisons began. Continuous breeding attempts by long-legged buzzards, Bonelli's eagle, and golden eagle *Aquila chrysaetus* often fail; only the kestrel seems to have almost fully recovered.

The removal of many raptors had secondary effects on the fauna. The population increase of the blackbird *Turdus merula*, the bulbul *Pycnonotus barbatus*, the palm dove *Streptopelia (Stigmatopelia) senegalensis*, the Syrian woodpecker *Dryobates syriacus*, and the jay *Garrulus glandarius* may be partly attributed to the decrease in the numbers of their predators (mainly the sparrowhawk *A. nisus*).

Secondary poisoning by insecticides also affected some insectivorous birds, particularly species that lived near fields and human settlements. Populations of the swallow *Hirundo rustica*, red-rumped swallow *Hirundo daurica*, white-throat *Sylvia communis*, rufous bushchat *Cercotrichas galactotes*, nubian shrike *Lanius nubicus*, spotted flycatcher *Muscicapa striata*, roller *Coracias garrulus*, bee-eater *Merops apiaster*, and Egyptian nightjar *Caprimulgus aegyptius* all decreased considerably.

Also, insectivorous bats (Microchiroptera) have been affected by chemicals. Twenty-eight species are known

in the southern Levant, and many caves used to be inhabited by thousands of roosting bats. The numbers of both bat-inhabited caves and individuals roosting in them have decreased drastically since the end of the 1950s and most species have become rare. This decline is attributed to two main factors. The first is fumigation of caves with ethylene dibromid and later with lindane (Gammexan). Fumigations were intended to control the fruit-eating bats *Rousettus aegyptiacus*, which were considered fruit pests and which often roost together with insectivorous bats. It was later found that the damage allegedly caused by fruit bats was marginal at best and had been grossly exaggerated by pest control officers. The population of one insectivorous bat, *Pipistrellus kuhli*, remained less affected than the others, probably because it roosts not in caves but in hollow trees and in wooden houses. The second factor affecting insectivorous bats is secondary poisoning: Most bat species occur in the cultivated regions and are highly vulnerable to secondary poisoning because noctuid moths are an important part of their diet. Larvae of noctuids such as *Spodoptera littoralis* are serious agricultural pests, and fields are regularly sprayed with insecticides to eliminate them. Hence, secondary poisoning may be another factor contributing to the decrease of bat populations in the Levant.

The remaining bat populations suffer now from an-

other threat. Increasing numbers of hikers visit various caves and disturb roosting and hibernating bats, thus causing wastage of fat reserves and desiccation of the hibernators.

Carnivores are another group which suffered from poisoning. In 1964, the Plant Protection Department of Israel decided that jackals (*Canis aureus*), one of the few mammals not protected at that time by the "Wild Animals Protection Law," were a nuisance. They were being blamed for damaging plastic sheets used to cover certain crops. A large antijackal campaign was started and tens of thousands of chicks injected with 1081 (fluoracetamid) were spread over the Mediterranean area of Israel in an effort to eradicate the jackals. Several mammal predators were affected, including the jackal, the wolf *Canis lupus*, the red fox *Vulpes vulpes*, the Egyptian mongoose *Herpestes ichneumon*, the jungle cat *Felis chaus*, and the African wild cat *F. sylvestrus*. Most species recovered within a few years and their numbers today are similar to those before the poisoning campaign. The jackal, however, is much slower to recover and its numbers in Israel are still low. The wolf, whose populations in the Negev have increased in recent years due to greater food availability at garbage dumps, is endangered in central and northern Israel, where individuals are larger than those in the Negev and occasionally prey on livestock. Although the wolf is legally protected, stock owners often retaliate by poisoning.

The caracal *Caracal caracal*, which was formerly known only in the Negev, has increased its distribution area northwards into the Galilee. This range extension coincided with the population decrease in many Mediterranean carnivores following the jackal poisoning operation, suggesting that range extension of the caracal was possible due to the absence of competitors. Apparently, the main competitor of the caracal is the jackal, which preys on hares, the main food item of the caracal. After the jackal poisoning campaign, hares increased considerably. Today, after most predator species have recovered from the effects of poisoning, caracals still occur in the Mediterranean region of Israel but seem to be rare.

#### D. Changes in Agricultural Practices

Current agricultural practices in the southern Levant are different from those of the past mainly in that today approximately half of the cultivated area is irrigated. As a result, areas which formerly were left fallow in summer are now cultivated in the dry season.

Some species have responded to the change in agricultural practices with an increase in both abundance

and distribution. Population size of the mountain gazelle *Gazella g. gazella* in central and northern Israel increased between 1948 and 1985 from less than 500 to 10,000. Availability of succulent, nutritious food and of water throughout the year in irrigated agriculture enables female gazelles to deliver their first fawns at the age of 1 year (2 years in nonagricultural areas) and average 1.8 young annually (only 1 in nonagricultural areas). Gazelles have become a pest in several areas because of their high numbers: The males damage fruit trees by rubbing their horns against the bark, and both sexes eat cotton, wheat, corn, and other crops. Increasing food and water availability in agriculture has also affected the dorcas gazelle *G. dorcas* in the southern Negev. Whereas populations near agricultural settlements increased between 1964 and 1985 by 9% annually, the rate of increase in other populations was only approximately 3%. Normally, dorcas gazelles in the southern Levant have only one fawn per year, in spring. Even if this fawn is lost, the dam does not become estrous until the normal breeding season in autumn. In recent years, increasingly more fawns are born in autumn so that either some females breed twice per year, as do mountain gazelles, or females become estrous again after losing their fawn. Autumn fawns are seen mainly near agricultural areas.

Availability of green food in summer and leftover grain in wheat and barley fields has enabled several seed-eating birds to increase their populations to such an extent that they have become agricultural pests: the collared dove *Streptopelia decaocto*, palm dove *Streptopelia senegalensis*, house sparrow *Passer domesticus*, and feral domestic pigeons *Columba livia domestica*.

The planting of exotic trees and ornamental plants in human settlements has made new areas available to several bird species. Typical woodland species, such as the syrian woodpecker *D. syriacus*, blackbird *Turdus merula*, great tit *Parus major*, and jay *G. glandarius*, once restricted mainly to woodlands of the Galilee, Mt. Carmel, and Judean hills, are now widespread and common in the gardens and parks that accompany many human settlements, sometimes even in the Negev and Jordan Valley. The Syrian woodpecker has become a pest in certain areas because it drills holes in irrigation pipes and in telephone cables.

Approximately 1500 plant species, mostly ornamental plants, were imported to Israel during the twentieth century. The orange-tufted sunbird *Nectarinia osea* is favorably affected by this enriched flowering vegetation of ornamental exotic plants in human settlements. When first discovered (in the nineteenth century) it occurred only in the lower Jordan Valley and near the

Dead Sea, its distribution coinciding with that of the mistletoe *Loranthus acaciae*. Today, however, owing to the presence of ornamental plants with nectar-bearing flowers, the sunbird is widespread and common in settlements throughout much of the southern Levant.

Another species benefiting from the spread of plantations and gardens is the Arabian bulbul *Pycnonotus barbatus capensis*. It feeds mainly on fruits, flowers, and leaves of introduced plants such as *Melia azedarach*, *Lantana sp.*, *Erythrina sp.*, and other fruit trees and flowers. Due to the ample available food it has increased and become a pest in vineyards and orchards.

Man-constructed fish ponds and water reservoirs present an alternative habitat for some wetland species. The night heron *Nycticorax nycticorax*, little egret *Egretta garzetta*, and to a smaller extent squacco heron *Ardeola ralloides* are the main beneficiaries, but glossy ibis *Plegadis falcinellus* also breed near some reservoirs. All four species were rare breeders or nonbreeders in the Levant until the late 1950s, when they started breeding near fish ponds, which provide food and nesting sites. Fish ponds also enabled the pied kingfisher *Ceryle rudis* to extend its breeding distribution from the Jordan Valley to the coastal plain. Dirt roads surrounding fish ponds and reservoirs are a favorite nesting site of the spur-winged plover *Hoplopterus spinosus*, which has already extended its breeding distribution into the northern Negev. Stilts *H. himantopus*, which once bred only in the Hula swamp area, are now common breeders along the coastal plain on banks of fish ponds and reservoirs. Also, the little-ringed plover *Charadrius dubius* and the Kentish plover *C. alexandrinus*, which formerly bred mainly on the Mediterranean seashore, now breed on reservoir banks. Another wader that breeds on banks, especially of salt ponds in the coastal plain and in fish ponds with brackish water in the Bet Shean valley, is the avocet *Recurvirostra avocetta*.

Among mammals, the coypu *Myocastor coypu* was introduced into the southern Levant for the fur industry during the 1950s. Some individuals escaped from breeding farms, and others were released when their farming was discontinued. Currently, coypu are common near any water body, from northern Israel to the Gaza Strip. The jungle cat *Felis chaus*, which suffered heavily from the draining of wetlands and from poisoning (aimed at jackals and voles) during the 1960s, has fully recovered and occurs mainly near fish ponds. Also, the Egyptian mongoose *Herpestes ichneumon* has expanded, in numbers and distribution, due to the existence of fish ponds. This increase has affected one of its prey species, the water snake *Natrix tessellata*, so that from a very abundant species (traps in fish ponds yielded 30–40 snakes

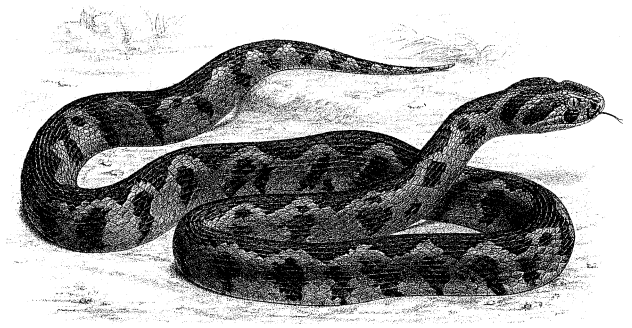


FIGURE 18 The viper *Vipera palaestinae* (from Tristram, 1884, Fauna and flora of Palestine, Committee of the Palestine Exploration Fund, London).

per pond in several days) it has become rather rare. However, the poisoning of jackals also reduced the number of mongooses. This was followed by an increase in the number of snake bites in Israel by the Palestine viper *Vipera palaestinae* (Fig. 18) from approximately 200 per year during 1960–1964 to 430 in 1967. Recovery of the mongoose populations was followed by a decrease of snake bite cases to the former level.

In the future, animal diversity in the Levant will be far from the great ecological and zoogeographical variety which makes it so unique.

### See Also the Following Articles

DESERT ECOSYSTEMS • MEDITERRANEAN-CLIMATE ECOSYSTEMS • NEAR EAST ECOSYSTEMS, PLANT DIVERSITY

### Bibliography

- Arad, Z. (1990). Resistance to desiccation and distribution patterns in bush-dwelling land snails. *J. Zool.* 221, 113.
- Ayal, Y., Broza, M., and Pener, M. (1999). Geographical distribution and habitat segregation of bushcrickets (Orthoptera: Tettigoniidae) in Israel. *Israel J. Zool.* 45, 49.
- Benyamini, D. (1988). The zoogeography of the butterflies (Lepidoptera, Rhopalocera) of Israel and nearby areas. In *The Zoogeography of Israel* (Y. Yom-Tov and E. Tchernov, Eds.). Junk, Dordrecht.
- Botosaneanu, L. (1992). *Trichoptera of the Levant*. Israel Academy of Science and Humanities, Jerusalem.
- Dimentman, Ch., Bromley, H. J., and Por, F. D. (1992). *Lake Hula*. Israel Academy of Sciences and Humanities, Jerusalem.
- Dumont, H. J. (1991). *Fauna Palaestina. Insecta 5. Odonata of the Levant*. Israel Academy of Sciences and Humanities, Jerusalem.
- Fishelson, L. (1985). *Orthoptera: Acridoidea*. Israel Academy of Sciences and Humanities, Jerusalem.
- Freidberg, A. (1988). The zoogeography of the Diptera of Israel. In *The Zoogeography of Israel* (Y. Yom-Tov and E. Tchernov, Eds.). Junk, Dordrecht.
- Freidberg, A., and Kugler, J. (1989). *Fauna Palaestina. Insecta 4—*



- Diptera: Tephritidae*. Israel Academy of Sciences and Humanities, Jerusalem.
- Heller, J. (1988). The biogeography of the land snails of Israel. In *The Zoogeography of Israel* (Y. Yom-Tov and E. Tchernov, Eds.). Junk, Dordrecht.
- Kadmon, R., and Heller, J. (1998). Modelling faunal responses to climatic gradients with GIS: Land snails as a case study. *J. Biogeogr.* 25, 527.
- Krupp, F. (1987). Freshwater ichthyogeography of the Levant. In *Proceedings of the Symposium on the Fauna and Zoogeography of the Middle East, Mainz 1985* (F. Krupp, W. Schneider, and R. Kinzelbach, Eds.). Ludwig Reichert Verlag, Weisbaden.
- Kugler, J. (1988). The zoogeography of social insects in Israel. In *The Zoogeography of Israel* (Y. Yom-Tov and E. Tchernov, Eds.). Junk, Dordrecht.
- Levy, G. (1985). *Fauna Palaestina. Arachnida 3. Araneae: Thomisidae*. Israel Academy of Sciences and Humanities, Jerusalem.
- Levy, G. (1991). On some new and uncommon spiders from Israel (Araneae). *Bull. Br. Arachnol. Soc.* 8, 227.
- Levy, G. (1998). *Fauna Palaestina. Arachnida 3. Araneae: Theridiidae*. Israel Academy of Sciences and Humanities, Jerusalem.
- Levy, G., and Amitai, P. (1980). *Scorpiones*. Israel Academy of Sciences and Humanities, Jerusalem.
- Levy, G., and Shulov, A. (1964). The Solifuga of Israel. *Israel J. Zool.* 13, 102.
- Martens, K., and Ortal, R., (1999). Diversity and zoogeography of inland-water Ostracoda (Crustacea) in Israel (Levant). *Israel J. Zool.* 45, 159.
- Por, F. D. (1975). An outline of the zoogeography of the Levant. *Zool. Scripta* 4, 5.
- Shkolnick, A. (1988). Physiological adaptations to the environment: The Israeli experience. In *The Zoogeography of Israel* (Y. Yom-Tov and E. Tchernov, Eds.). Junk, Dordrecht.
- Tchernov, E. (1988). The palaeobiogeographical history of the southern Levant. In *The zoogeography of Israel* (Y. Yom-Tov and E. Tchernov, Eds.). Junk, Dordrecht.
- Theodor, O. (1980). *Fauna Palaestina. Insecta 2—Diptera: Asilidae*. Israel Academy of Sciences and Humanities, Jerusalem.
- Vet, V., Polis G. A., and Sissom, D. (1998). Life in sandy deserts: The scorpion model. *J. Arid Environ.* 39, 609.
- Werner, Y. L. (1987). Ecological zoogeography of the Saharo-Arabian, Saharan and Arabian reptiles in the sand deserts of southern Israel. In *Proceedings of the Symposium on the Fauna and Zoogeography of the Middle East, Mainz 1985* (F. Krupp, W. Schneider, and R. Kinzelbach, Eds.). Ludwig Reichert Verlag, Weisbaden.
- Werner, Y. L. (1988). Herpetofaunal survey of Israel (1950–1985), with comments on Sinai and Jordan and on zoogeographical herpetogeny. In *The Zoogeography of Israel* (Y. Yom-Tov and E. Tchernov, Eds.). Junk, Dordrecht.
- Yom-Tov, Y. (1988). The zoogeography of the birds and mammals of Israel. In *The Zoogeography of Israel* (Y. Yom-Tov and E. Tchernov, Eds.). Junk, Dordrecht.
- Yom-Tov, Y., and Mendelsohn, H. (1988). Changes in the distribution and abundance of vertebrates in Israel during the 20th century. In *The Zoogeography of Israel* (Y. Yom-Tov and E. Tchernov, Eds.). Junk, Dordrecht.



# NEAR EAST ECOSYSTEMS, PLANT DIVERSITY

Avinoam Danin

*The Hebrew University of Jerusalem*

---

- I. Environmental Conditions
  - II. Flora
  - III. Vegetation
- 

## GLOSSARY

**batha** A biblical term for semishrub communities, mainly of seral vegetation of old field succession in the Mediterranean part of the Near East.

**contracted vegetation** Vegetation restricted to wadis that receive additional water supply.

**diffused vegetation** Vegetation occupying all slopes and most habitats.

**semisteppe batha** Semishrub communities developing on most soil types at the boundary of the Mediterranean zone of the Near East.

**wadi** Dry water course.

---

and Jordan and the extreme desert areas in their southern part. A long history of human pressure of cultivation and grazing by domestic animals led to strong stress on the existing flora and enabled the introduction of many alien species, many of which occupy habitats created by human activity. In this article, these factors are discussed and examples are provided.

## I. ENVIRONMENTAL CONDITIONS

Currently, the most important factors affecting the distribution of species are environmental conditions. The evolutionary history of the area reviewed influenced the composition of the flora, but the environment influences the composition and their ability to coexist and prosper. Before discussing the vegetation and its gradients of species diversity, the physical setup of the study area will be discussed.

### A. Topography

*THE HIGH RICHNESS OF PLANT SPECIES OF ISRAEL*, expressed as species to area and being 9.06 species/100 km<sup>2</sup> (with a similar number for Jordan), is related mainly to its position in a meeting zone of plant geographical regions, each with its own typical flora. This wealth is supported by the existence of many habitats needed to support these species. The wealth of habitats derives from the climatic transition between the relatively moist area in the northern part of Israel

The topography of Israel and Jordan can best be described as north–south topographic belts, which are influenced by the geomorphological features of the area. The Mediterranean coastal plain is narrow in the north of Israel (Fig. 1, 1) and becomes wide in the south (Fig. 1, 5). Low hills with gentle topography and wide valleys with deep soil constitute the foothills. The mountainous area reaches elevations of approximately 1000 m at the Judean Mountains (Fig. 1, 12) and the Negev Highlands

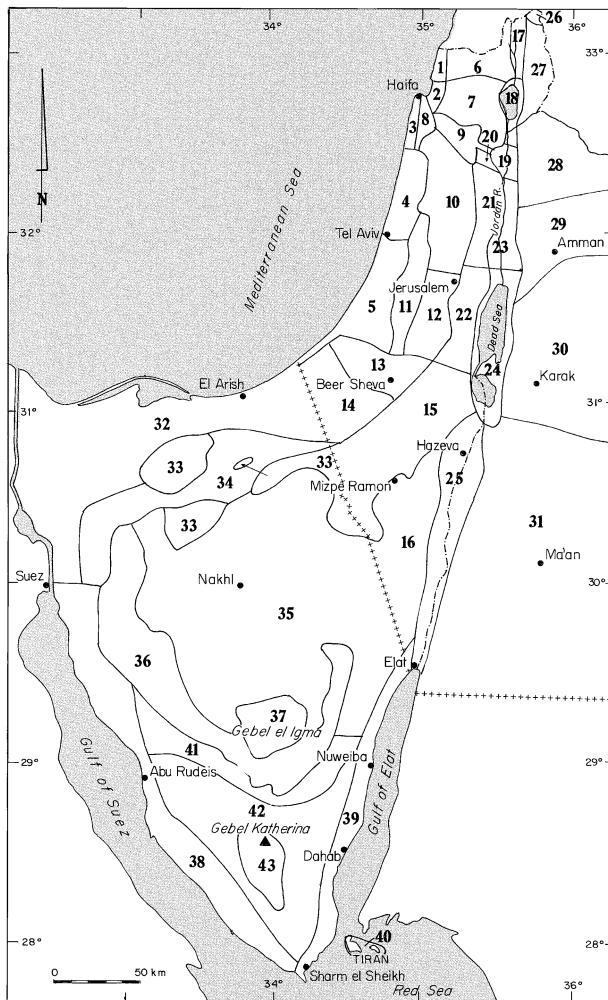


FIGURE 1 Geographical subdivision of the study area (based on Danin, 1999). Israel: 1, coastal Galilee; 2, Acco Plain; 3, coast of Carmel; 4, Sharon Plain; 5, Philistine Plain; 6, Upper Galilee; 7, Lower Galilee; 8, Mt. Carmel; 9, Esdraelon Plain; 10, Samaria; 11, Shefela; 12, Judean Mountains; 13, northern Negev; 14, western Negev; 15, Negev Highlands; 16, southern Negev; 17, Hula Plain; 18, Kinnroth Valley; 19, Beit Shean Valley; 20, Mt. Gilboa; 21, Samaritan Desert; 22, Judean Desert; 23, Lower Jordan Valley; 24, Dead Sea Valley; 25, Arava Valley; 26, Mt. Hermon; 27, Golan. Jordan: 28, Gilead; 29, Ammon; 30, Moav; 31, Edom. Sinai: 32, Mediterranean sands and salt marshes; 33, anticlines of northern Sinai; 34, gravelly plains covered with shifting sands; 35, gravelly plains of central Sinai; 36, Table mountains of central and western Sinai; 37, Gebel el Igma; 38, coastal plain of the Gulf of Suez; 39, coastal plain and foothills of the Gulf of Elat; 40, Tiran and Sinafir Islands; 41, sandstone belt; 42, Lower Sinai Massif; 43, Upper Sinai Massif.

(Fig. 1, 15) and 1200 m at the Galilee (Fig. 1; 7). The Jordan–Dead Sea–Arava rift valley is part of the Afro-Syrian Rift Valley and constitutes the lowest topographic element in the area. This valley is 400 m below

sea level in the deepest part of the Dead Sea (Fig. 1, 24). The terrain ascends to 200 m above sea level at the foot of Mount Hermon (Fig. 1, 26) in the north and at the water divide of the Arava Valley (Fig. 1, 25) to the south. A few transversal valleys, such as the Esdraelon Plain (Fig. 1, 9) running from east to west, dissect the mountainous area into large blocks. Steep escarpments, dissected by canyons, typify the areas east and west of the rift valley. There are a few internal lakes in this valley. Lake Hula between the Golan and Galilee was drained in the 1950s. The Kinneret, also known as the Sea of Galilee, is the largest freshwater natural reservoir of Israel. The Dead Sea is a highly salty water body. The large salt marshes at the Arava Valley are in fact a large underground lake. At the vicinity of these water bodies there are small or large springs. Mount Hermon, at the common border of Israel, Lebanon, and Syria, reaches an elevation of 2800 m and terminates sharply at the basalt-covered plateau of the Golan to the south. This plateau descends gently from an elevation of 1200 m toward the Yarmouch River and toward the Kinneret at 200 m below sea level. The Jordanian plateau starts south of the Yarmouch River (Fig. 1, 28). Its escarpments delimiting the rift valley are as steep as, and in many places steeper than, the escarpments in Israel. The Jordanian plateau has several peaks 1200–1600 m high. A few deep rivers dissect the plateau from east to west toward the rift valley. East of the water divide, which is near the western edge of the plateau, the landscape becomes flat and gradually descends toward Mesopotamia.

Sinai has a wide coastal area along the Mediterranean Sea (Fig. 1, 32), ascending gradually southwards. The northern Sinai sand belt reaches the anticlinal ridges of Gebel Maghara, Gebel Halal, and Gebel Yiallaq (Fig. 1, 33). The central Sinai gravel plains (Fig. 1, 35) are surrounded by crescent-like ranges of mountains or plateaus (Fig. 1, 36), dissected by a few valley passages such as those of Wadi Giddi, Wadi Mitlah, and Wadi Sidr. South of the gravelly plains there are two large plateaus ascending gradually in a north–south direction. The Gebel el Igma plateau (Fig. 1; 37) peaks at 1600 m and the Gebel et Tih plateau peaks at 1400 m. The erosion escarpments of the two plateaus are steep and that of the latter descends to the sandstone belt. The morphology of the sandstone belt (Fig. 1, 41) is highly diverse in different parts of Sinai. The Southern Sinai Massif (Fig. 1, 42 and 43), built of magmatic and metamorphic rocks, is highly dissected topographically. The mountain peaks reach 2500–2600 m. The escarpments toward the Gulf of Elat–Aqaba are not as steep as those of the rift valley in Israel and Jordan. The

coastal plain along this gulf is wide in the southern part of Sinai, whereas in many places the mountain slopes descend directly into the sea without any coastal plain. There is a wide coastal plain along the Gulf of Suez.

## B. Rock Types, Geomorphology, and Edaphic Conditions

The most common rock types of northern Israel and Jordan are sedimentary limestones, dolomites, chalks, and marls of the Cretaceous and the Tertiary. Terra Rossa soils occur in the mountainous areas on hard rocks, and Rendzinas occur on the soft ones. Basalt rocks typify much of the Golan plateau and northeast Galilee: Both are covered with basaltic brown or red Mediterranean soils. Tertiary and Pleistocene rocks and derived soils fill up the small and large valleys. Calcareous sandstone, locally known as kurkar, and sandy-loamy soils, known as hamra, typify the coastal plain near the coast and deep clay soils, known as grumusols or vertisols, far from the Mediterranean coast. Cretaceous and older sandstone typify the vicinity of the rift valley of the Jordan River, the Dead Sea, and Arava in Jordan. The edaphic conditions in northern Jordan and south to Amman are similar to those of the area west of the Jordan River.

The steppe and desert areas of Israel, Jordan, and Sinai develop on a much more diverse assemblage of rocks. The main contributors are limestone, chalk, marl, chert, sandstone, magmatic, metamorphic rocks, and gravel of alluvial origin or rocks weathered *in situ*. Loess is important eolian sediment, composed mainly of silt and clay particles, that influences much the soils of the transition zone of the Mediterranean territory and the steppelands. Large flatlands south of this transition zone occur in the northern Negev of Israel and east of Irbid, Jarash, and Amman in Jordan. The poor moisture regime of loessial soils with low quantities of rainfall makes the desert boundary prominent.

Dan *et al.* (1975) provide soil maps of Israel, and Al-Eisawi (1996) provides a map of Jordan; a detailed soil map of Sinai is not available. The specific influence of each substratum type on the vegetation is discussed in detail elsewhere (Danin, 1983). The relationships between rock type, soil, and vegetation in the desert areas of Israel and Sinai are similar to those in many areas of Jordan.

Geomorphological structures in Israel are small, and they are larger in Sinai and Jordan; however, these are smaller than those in countries such as Saudi Arabia in which huge areas of similar structures occur. This is the expression of relatively high geomorphological di-

versity. One may expect to find a high diversity of rock and soil types under a certain climatic regime. Hence, habitat diversity in such areas is expected to be high.

## C. Climate

### 1. Rainfall

The climate of Israel, Jordan, and Sinai is Mediterranean, characterized by a cold and rainy winter and a rainless and warm summer. Rainfall quantities vary in two directions: Rainfall decreases gradually from north to south—sharply from the water divide eastwards in Israel and gradually in Jordan from the water divide eastwards (Fig. 2). The north–south gradient is influenced mainly by the intensity and frequency of rain-

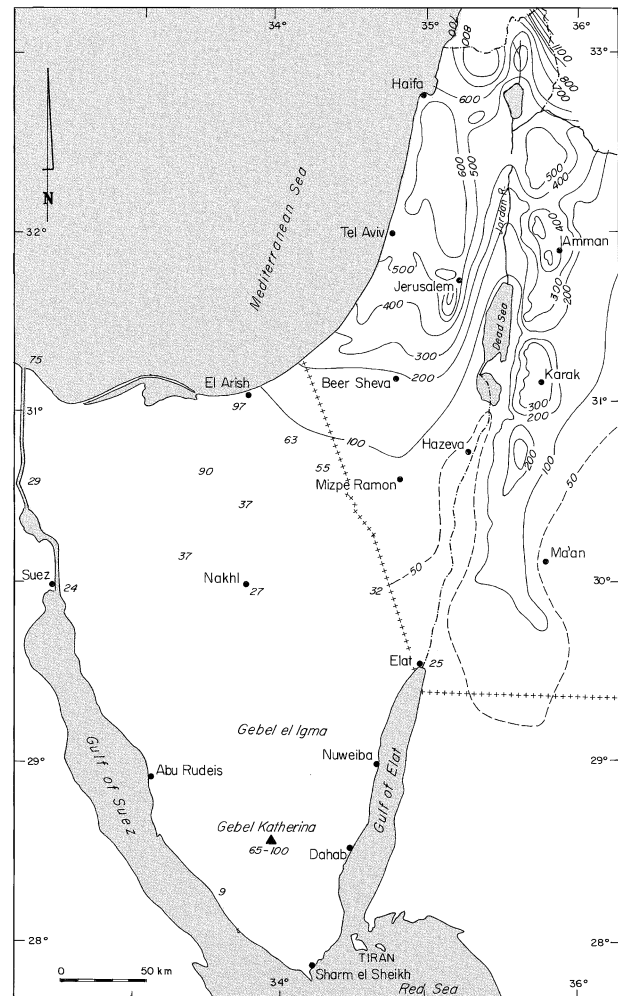


FIGURE 2 Mean annual rainfall map [based on Danin (1999) and Al-Eisawi (1996)].

contributing systems, whereas the west–east gradients are influenced by the topography. Thus, the ascent of air bodies from the Mediterranean Sea toward the mountains of Judea, Samaria, and Galilee results in cooling of the air and an increase in rainfall with increasing elevation. Air bodies descending toward the rift valley become warmer and drier; hence, the Samarian and Judean Deserts and the eastern Galilee occur at the rain shadow of the Judean, Samarian, and Galilee mountains, respectively. The ascent to the Jordanian plateau mirrors the rainfall map with maximum at the mountaintops. There is a gradual decrease in mean annual rainfall eastwards with increasing continentality far from the Mediterranean toward the desert. The increase in mean annual rainfall in Moav and Edom, located south of the west–east section of the 300-mm isohyet in Israel, is related to the orographic influence of the high elevation of this area. Snow may fall in winters of cold and wet years on the highest summits of mountains in Israel, Jordan, and Sinai and remain for a few days.

## 2. Dew and Fog

Dew and fog are important sources of humidity for the poikilohydric organisms growing on rock outcrops. The measurement of dew by Duvdevani's dew gage indicated a high similarity to events of efficient dew when lichens imbibed on stones and rocks at the Negev Highlands. The mean annual number of nights with more than 0.02 mm per night of dew has been  $191 \pm 22$  for the past 15 years at Avdat; of these, there are only  $124 \pm 28$  nights annually with dew amounts of 0.11–0.5 mm. Dew measurements in other parts of Israel, Jordan, and Sinai are not available. However, lithobiont communities near Avdat were used to extrapolate the regional distribution of dew. There are many similarities in weathering features found at the top of the northern Sinai anticlines (Fig. 1, 33) and on limestone outcrops at the area marked 11d in Jordan (Fig. 3). I assume that the dew and fog regime is similar in the latter two areas to that of the Avdat area where real measurements of dew and rainfall were carried out.

## 3. Temperature

Temperature regime in the study area is influenced by the altitudinal and latitudinal position of the site under review. Mean annual temperature in the desert areas varies from  $9^\circ$  to  $25^\circ\text{C}$ . The lowest temperatures prevail in the peaks of southern Sinai and Jordan near Shoubak, whereas the highest are those of the Dead Sea and Arava Valleys, where elevation reaches 400 m below sea level. Mean annual temperature maps of Israel are presented in the atlas of Israel and of Jordan in Al-Eisawi (1996).

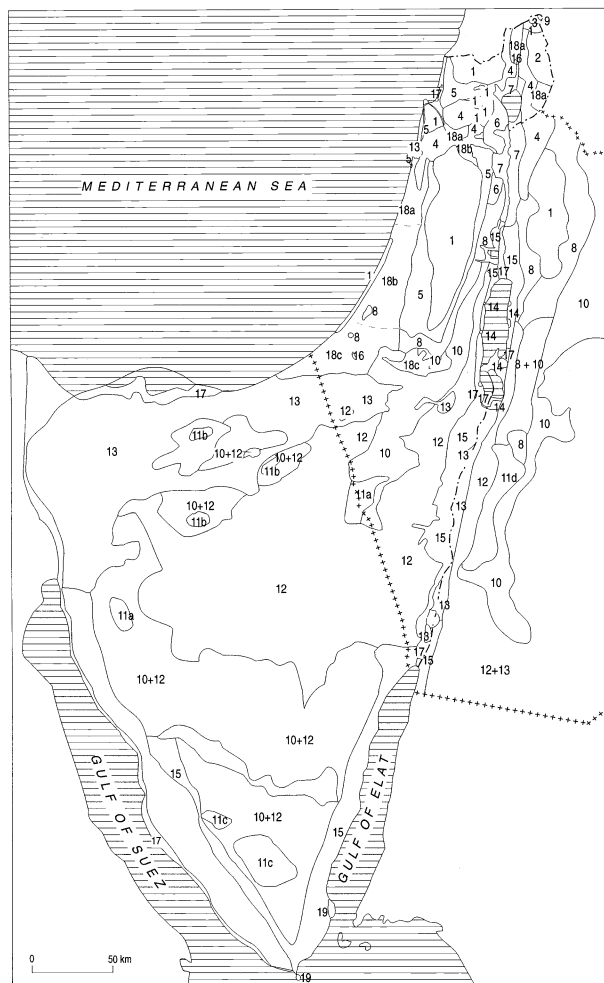


FIGURE 3 Vegetation map of Israel, Sinai, and western Jordan: 1, maquis and forests; 2, *Quercus calliprinos* woodlands on basalt; 3, montane forest of Mt. Hermon; 4, open forests of *Quercus ithaburensis*; 5, open forests of *Ceratonia siliqua* and *Pistacia lentiscus*; 6, *Ziziphus lotus* with herbaceous vegetation; 7, Mediterranean savannoid vegetation dominated by *Ziziphus spina-christi*; 8, Semisteppe batha; 9, tragacanth vegetation; 10, shrub–steppes; 11, shrub–steppes with trees of (a) *Pistacia atlantica*, (b) *Juniperus phoenicea*, (c) *Pistacia khinjuk* (Sinai), (d) *Quercus calliprinos* and trees of 11a, 11b, and 11c (Jordan); 12, desert vegetation; 13, sand vegetation; 14, oases with Sudanian trees; 15, desert savannoid vegetation; 16, swamps and reed thickets; 17, wet salinas; 18, synanthropic vegetation; 19, mangroves (reproduced with permission from Danin, 1999).

## II. FLORA

There are 2682 plant species in Israel, including approximately 200 species which occur in the Mt. Hermon area and are absent from the other districts of Israel. The number of species in Jordan is 2078 according to Al-Eisawi (1996). There are 2885 species in the combined lists of plants of Israel and Jordan (A. Danin,

TABLE I  
Species Richness of the Flora of Several Countries<sup>a</sup>

Country/State	Species	Area(km <sup>2</sup> )	Species/100 km <sup>2</sup>
Israel	2682	29,600	9.06
Sinai	889	61,100	1.45
Egypt (excluding Sinai)	2100	1,000,000	0.21
Saudi Arabia	2200	2,200,000	0.1
Syria	3060	185,000	1.65
Spain	5500	504,000	1.09
Tunisia	2250	164,000	1.37
Greece	4200	132,562	3.17
Italy	5600	301,100	1.86
Britain	1666	229,850	0.72
California	5057	411,000	1.23

<sup>a</sup> After Danin (1996) and Le Houérou (1995).

unpublished data). The high species richness of Israel, expressed as the parameter of species to area (Table I), is related mainly to the following factors:

1. Its position in a meeting zone between plant geographical regions, each with its own typical flora.
2. The existence of many habitats needed to support these species: The wealth of habitats derives from the climatic transition between the relatively moist area in the northern part of the two countries and the extreme desert areas in their southern part and also in Jordan's eastern part. Topography is also a factor in creating the warm climates of the rift valleys and the relatively cold climate of the mountainous areas. Similarly, other highlands and lowlands have local climatic influence which increases the habitat diversity of Israel and Jordan and increases the number of habitats which support high plant species diversity. The high number of rock types influences the development of many soil types in a small area, increasing the diversity of habitats available for plants.
3. A long history of human pressure of cultivation and grazing by domestic animals led to strong stress on the existing flora and enabled the introduction of many alien species, many of which occupy habitats created by human activity.

According to various studies of the flora of Israel, it may be divided into the following groups:

1. Mediterranean species, which are distributed around the Mediterranean Sea.
2. Irano-Turanian species, which also inhabit Asian

steppes of the Syrian Desert, Iran, Anatolia in Turkey, and the Gobi desert.

3. Saharo-Arabian species, which also grow in the Sahara, Sinai, and Arabian deserts.

4. Sudano-Zambesian species, typical of the subtropical savannas of Africa.

5. Euro-Siberian species, also known in countries with a wetter and cooler climate than that of Israel. These species grow mainly in wet habitats and along the Mediterranean coasts.

6. Biregional, triregional, and multiregional species that grow in more than one of the regions mentioned previously.

7. Alien species derived from remote countries; these plants propagate without human assistance.

### III. VEGETATION

Danin (1999) discusses the vegetation of the entire area in detail. The most prominent factors influencing the high species diversity of the study area are discussed in the following sections. The reader may combine the general information on the vegetation of Israel (Danin, 1999) with that of Sinai (Danin, 1983) and Jordan (Al-Eisawi, 1996).

#### A. Maquis and Forests

The principal spontaneous woodland areas of Israel are found in the mountains of Judea, Carmel, and Galilee and at the foot of Mt. Hermon; those of Jordan occur north of Amman (Fig. 3, 1). These forests or maquis are dominated by the sclerophyllous evergreen *Quercus calliprinos* and the deciduous *Pistacia palaestina* and grow on hard limestone with terra rossa soil. The companions of *Q. calliprinos* vary according to edaphic and climatic conditions. In the mesophytic aspect of the oak woodlands, found mainly in Upper Galilee, there are several trees, shrubs, vines, and geophytes that are not found in the xerophytic aspect of the southern Judean Mountains. In the driest maquis stands, *Rhamnus lycioides* subsp. *graecus* is the only arboreal companion of *Q. calliprinos*. Typical vines in these maquis are *Rubia tenuifolia*, *Lonicera etrusca*, *Asparagus aphyllus*, and *Ephedra foeminea*.

Marly chalk is a common rock type that has high moisture-holding capacity and is covered with a shallow light rendzina soil. The aeration of the rhizosphere, of trees and shrubs that penetrate into the soft rock, is poor and leads to nutritional stresses, which only specially adapted plants can withstand. Much of the nitrogen in these soils is in the form of ammonium ions, whereas

in the terra rossa it is in the nitrate ion form. The vegetation cover of the light rendzinas on marly chalk is poor when compared with that on terra rossa and includes very few annual species. The principal trees of these soils are *Arbutus andrachne* and *Pinus halepensis*.

In most of the area, cultivated plants have replaced the spontaneous trees. A few thousand years ago, people in the eastern Mediterranean countries started to clear the natural vegetation to create agricultural land. Trees such as olives (*Olea europaea*) and almonds (*Amygdalus communis*), which grow spontaneously in the area, have been domesticated. Timber derived from the forests and maquis was used for the construction of houses, for agricultural tools, and for fire fuel. For the past few millennia shepherds have burned large woodland areas to open paths for domestic animals, and the pasture quality has been improved through the replacement of trees and shrubs with palatable herbaceous plants.

As a result of thousands of years of deforestation and agricultural and urban development, large areas of Israel and Jordan look like mosaics of seral communities. Abandoned cultivated ground becomes populated for dozens of years by herbaceous and low lignified plants. Habitat diversity, and thus species diversity, is extremely high due to the removal of the shading trees and the opening of poor and temporary fertile soils to annual plants which could not grow in the closed woodlands. Frequent human-induced fires reopen areas, which become populated with shading trees and shrubs and ephemeral plants that increase species diversity. Some terra rossa soils that are leached and poor become fertilized during the years of crop cultivation. Once abandoned, a series of replacement annual plant communities occurs, from those using fertile ground, including tall plants, to those using poor soil and small in size. By sampling approximately five such stages, each a square of  $1 \times 1 \text{ m}^2$ , in the vicinity of Jerusalem, I found approximately 100 different species in  $8 \text{ m}^2$ . Local enrichment of the poor soils by gazelle droppings and at the vicinity of nests of harvesting ants creates an ecotone of soil fertility which enables the coexistence of many species.

Light rendzinas were often cultivated due to the high water-holding capacity of the substratum of soil plus rock. When abandoned, this soil may support the growth of a few annual species. After time passes and the aerated plowed soil becomes flattened, microbiotic crust, including cyanobacteria, lichens, and mosses, develops on the surface, thus influencing soil aeration and changing edaphic conditions. Such minute edaphic changes along the precipitation gradients of the moun-

tainous areas (Fig. 2) contribute to the high species diversity of the country.

### B. *Quercus calliprinos* Woodlands on Basalt

The basalt flows of the northern Golan are younger and differ in rock types from those of the rest of the Golan. The woodlands constituting these woodlands (Fig. 3, 2) are rich in their herbaceous companions. Contrary to the vegetation of the calcareous rocks, they are almost devoid of semishrub communities at the early succession stages because of destruction and abandonment. The main environmental factor influencing this is the high phosphorous content of the rock and soil which favors the development of ephemeral plants. Thus, a dense maquis of *Q. calliprinos* covers the gentle north-facing slope of the ancient volcanic cone of Har Odem, the Golan, at an elevation of 900–1000 m. Regarding the trees, the rich ephemeral vegetation includes approximately 20 species of *Trifolium*, many other Papilionaceae, and rich annual flora of other families. The phytomass is low, thus indicating some nutrient deficiency when compared to that of the vegetation of other basalt flows. Under such conditions, without one dominant species there is high species diversity.

Other basalt rocks derived from older basalt flows at lower elevation, down to 500 meters above sea level (m a.s.l.), support denser and taller ephemeral vegetation with different composition due to differences in soil fertility and climatic conditions, which follow elevation. The oak trees and their arboreal companions are found only occasionally and they have no impact as shadow creators in most of the Golan and eastern Galilee.

Ancient volcanoes provided much volcanic ash to large areas surrounding them. Because of a different pedogenetic nature, this substratum supports a rich flora, many components of which are not common in other kinds of soil in this area. Many of the volcanic ash layers have a porous and well-aerated texture. In contrast to the grumusols and protogrumusols of much of the area surrounding it, this substratum supports many narrow-range species.

### C. Montane Forest of Mt. Hermon

The montane forest (Fig. 3, 3) stretches at Mt. Hermon from 1300 to 1700 m a.s.l., where there are only scattered remnants of the woodlands which are believed to have been there before their destruction by human activity. The dominants are deciduous trees such as

*Quercus boissieri* and *Q. libani* and several species of *Crataegus*, *Amygdalus*, *Acer*, and *Prunus*. Their companions are mainly perennial grasses and many herbaceous species that differ from those of the rest of the Mediterranean territories of the Near East. These occur at lower elevation and presumably thrive on a different temperature and rainfall regime than that of the montane forest. There are no representative areas of this category in Jordan because there are no mountains of this elevation in the northern part of the country. However, in southern Jordan there are such high mountains, but due to their drier climate woodlands of this kind are absent. Nonetheless, a few representative shrubs of the genera *Astragalus* and *Cotoneaster*, common at Mt. Hermon, occur at high elevations in crevices of smooth-faced rocks in southern Sinai and in southwestern Jordan.

#### D. Open Forests of *Quercus ithaburensis*

The tabor oak is the most frequent tree in the open forests prevailing in large areas of Israel and Jordan (Fig. 3, 4). *Styrax officinalis* also appears with the tabor oak when developed on chalky ground of the Lower Galilee, Israel, and the Gilead in Jordan. *Pistacia atlantica* is the companion on basalt rocks of the Golan. Plenty of herbaceous plants cover the open space among the trees in all areas of this category. The few semishrubs found in this community when developing on chalky ground are mainly of *Majorana syriaca*. This is contrary to the dominance of *Sarcopoterium spinosum*, the typical component of seral communities on terra rossa. This type of forest, which once dominated the Sharon Plain, is restricted in the Sharon (Fig. 3, 4) to solitary, sporadic tabor oak trees which survived urbanization and agricultural development. There are large woodlands of the tabor oak in the Lower Galilee and in the Golan, lower than 500 m. Some of the largest indigenous trees in Israel are those of tabor oak at the Hula and Dan Valleys in the northern section of the rift valley in Israel. The tabor oak woodlands of the northwest of Jordan are well preserved. They cover large areas of rocky terrain west of Irbid, at the western escarpments of the Jordanian plateau, from sea level to 500 m a.s.l. The drier and more continental climate of the Jordanian woodlands allows for a different assemblage of companions to the oak.

#### E. Open Forests of *Ceratonia siliqua* and *Pistacia lentiscus*

Open woodlands dominated by carob trees and shrubs of *P. lentiscus* cover the lower altitudinal belt of the

main mountain ranges, 0–300 m on both sides of the central mountain range on terra rossa soils. This community inhabits rendzina soils at the foothills of the Judean mountains and sandy soils of the Sharon Plain in the littoral aspect of the community (Fig. 3, 5). Generally, the community is more drought and heat resistant than the communities dominated by *Q. calliprinos* and has a similar position in the aridity sequence of communities as those dominated by the tabor oak. Wild olive (*Olea europaea* var. *sylvestris*), which resembles the cultivated olive but has much smaller fruits, is an important companion of the carob in rocky sites at Mt. Carmel and Galilee. In Jordan, this category is almost missing; however, scattered carob trees occur in the open woodlands of *Quercus ithaburensis* of the Gilead, mainly in the transition zone from the belt dominated by the tabor oak to that of *Q. calliprinos* at elevations of 500–600 m. The main companion of the carob at the southern Judean foothills is *Rhamnus lycioides* subsp. *graecus*. The latter totally replaces all the other arboreal components in dry habitats.

The flora of this community is rich due to high habitat and microhabitat diversity. The hard limestone rocks have many crevices and soil pockets, which enable diverse microhabitats supporting different species to coexist that are isolated from each other by the rock body. Soft chalk rocks, covered by hard nari crust, also known as caliche, constitute additional diverse habitats. The nari layers have an upper hard crust layer of 5–10 mm which covers softer rock but is still consolidated chalk. Trees and shrubs, the roots of which penetrate into the rock, may establish themselves, thus avoiding the competition of the rich herbaceous flora which accompany them in the deep soil pockets. Thus, in many places the local boundary of the land dominated by shrubs and trees coincides with that of the nari rocks. Various kinds of shade are provided by evergreen trees and shrubs, each having a different architecture that influences the companion life. Opened for grazing by domestic and wild animals, the potential strong competition ordained by a few herbaceous plants is decreased, thus enabling many species to coexist.

#### F. *Ziziphus lotus* with Herbaceous Vegetation

The phosphorous-rich soils developing on basalt rocks of southeastern Galilee and the slopes near the Sea of Galilee and limestone and other sedimentary rocks down to the Samarian desert support grasslands dominated by Gramineae with large seeds. The most common grasses are wild wheat (*Triticum dicoccoides*), wild



barley (*Hordeum spontaneum*), and wild oats (*Avena sterilis*). This area (Fig. 3, 6) contains the highest genetic diversity of these species, which are believed to have been domesticated in the Near East thousands of years ago (Zohary and Hopf, 1993). A similar formation covers the west-facing slopes of the Gilead, Jordan, below the tabor oak belt. In drier and warmer sites or sites at which the soil is shallow, *Stipa capensis* is the dominant plant. The lignified and spiny shrub *Ziziphus lotus*, spread as green patches over the area, typifies most areas of these grasslands. Sites of relatively high soil fertility, common in this plant community, are the nests of the harvesting ant *Messor messor* and piles of dung of the male gazelles (*Gazella gazella*). This category is presented as a mosaic with that of the "savannoid Mediterranean vegetation" (Fig. 3, 7) in the steep topography at the vicinity of the rift valley north of Jericho to north of the Sea of Galilee.

### G. Mediterranean Savannoid Vegetation

Warm, stony/rocky slopes of the Galilee, Golan, Gilead, and Samaria descending to the rift valley (Fig. 3, 7) at sea level and below are covered with grasslands of Mediterranean annuals with large seeds, such as wild wheat, barley, and oats, and scattered *Ziziphus spinachristi* trees. The latter is a low, spiny tree with edible fruits that also grows in true savannas in Africa, where its companions are Sudanian perennial grasses. Therefore, the vegetation here is named "savannoid." Stands of this tree are also established along the transversal valleys on deep clay soils and along the Sharon Plain, where it grows on sandy-loamy soils together with the grass *Desmostachya bipinnata*. Local diversity of nutrients and derivative high species diversity are caused by harvesting ants and gazelles as discussed previously.

At the Bet-Shean Valley south of the Sea of Galilee there are considerable areas of slightly saline springs, the salts of which remain in the soil while water evaporation occurs. This results in a wide range of microhabitats which vary in soil salinity and the availability of spring water. The northernmost site of *Balanites aegyptiaca*, an African savanna tree, is in the vicinity of one of these springs. The Sudanian mistletoe *Plicosepalus acaciae* (= *Loranthus acaciae*) has its northernmost site on these *Balanites* trees.

### H. Semisteppe Batha

Semishrub communities of the *Ballotetea undulatae* at the boundary of the Mediterranean zone (Fig. 3, 8), where mean annual rainfall is 300–400 mm, are re-

garded as semisteppe bathas. There are several communities dominated by Mediterranean plants such as *Sarcopoterium spinosum*, which dominates bathas in more mesic parts of the country. Others, such as *Artemisia sieberi* and *Noaea mucronata*, dominate steppe areas in the Negev, Sinai, Jordan, and eastwards to Afghanistan. Many plants which play an important role in the seral communities in fallow fields at the center of the Mediterranean region grow here in what is regarded as their primary habitats. No anthropogenic disturbance assists or enables their growth here, and they occupy their typical habitats which are related to certain edaphic and climatic conditions. *Sarcopoterium spinosum*, which becomes ethiolant and dies under the shade of trees and shrubs in the course of plant succession in maquis and forests, has no such competitors in the semisteppe bathas. The semisteppe bathas extend further south and east of the regional boundary of the maquis in Jordan and in Israel. Because semishrubs are the dominant tallest growth form, the semisteppe bathas are rich in their flora. Many plants of mesic and xeric origin coexist here, contributing to the high species diversity. Phytogeographical analysis of each association in this vegetation indicates a higher diversity in origin when compared to associations of drier and moister conditions.

### I. Tragacanth Vegetation of Mt. Hermon

Semishrubs which look like spiny cushions constitute the most prominent formation of vegetation developed on the windward slopes of the peaks of Mt. Hermon above 1900 m (Fig. 3, 9). Many species of section *Tragacantha* in the genus *Astragalus* and of the genus *Acantholimon* are components of the cushion plants formation throughout the Middle East. The spiny cushion seems to have some biological advantages which make it adaptable to some of the components of the harsh conditions of this habitat: cold winter with high-velocity winds, precipitation mainly as snow, and dry summer. The short growth season and the harsh environment are the main factors enabling the success of very few annual species as companions of the cushion plants.

Snow accumulates on the slopes in the wind shadow of small local ridges or crests because of the low wind velocity. Consequently, snow may reach a depth of 10 m for a few months. Water drainage from the slopes of this area is through karst systems and not through wadis as in many areas of the country. A common feature of the karst topography is the occurrence of small valleys which are filled with fine-grained soil known as dolinas. The soil in these valleys is waterlogged for a long time,

thus increasing habitat diversity and the growth of many additional species compared to those of the rest of Israel.

Several species of tragacanth *Astragalus* occur in areas which are far from Mt. Hermon and the Anti Lebanon mountains. *Astragalus bethlehemiticus* is a typical companion of steppes and rock vegetation in the shrub–steppes and semisteppe bathas of the Negev Highlands and of southwestern Jordan; *A. echinus* is a representative of this group at the high elevations of southern Sinai, where it is confined to crevices of smooth-faced granite.

### J. Vegetation Patterns in the Dry Areas of Israel, Jordan, and Sinai

When dealing with the vegetation of drylands, which receive less than 200 mm mean annual rainfall, one has to consider general patterns of distribution of vegetation. The two principal patterns of distribution of vegetation in the dry areas of the Near East are related to dry watercourses, which are known as wadis from the Arabic language or arroyos in the southwestern United States. Wherever plants grow on slopes and depressions, the pattern is regarded as “diffused.” In extremely dry desert areas, where vegetation is restricted to wadis that receive additional water supply, the pattern is “contracted.” In an area in which the climatic conditions enable the development of diffused vegetation on most soil types, silty or clayey soils which are relatively dry soils in desert will locally support contracted vegetation. Plants demanding relatively high quantities of moisture may inhabit special habitats of wadis in zones of contracted vegetation. Such special habitats are derived from local contribution of runoff water from rocks to their crevices and soil pockets. In hard and fissured limestone, dolomite, granite, and metamorphic rocks, much of the rainfall is available to the semishrubs growing there because water infiltration is good and the soil in the rock fissures is leached. Accumulating deep in the soil and the weathered rock, this water is protected from direct evaporation by rocks and stones. Trees which grow on slopes of the Mediterranean zone that receives 500–700 mm of annual rainfall occur in desert areas with 100 mm or even less in proximity to outcrops of smooth-faced hard rocks. These rocks do not absorb much water and their crevices receive high amounts of water through runoff (Danin, 1999).

One of the ways in which plants are adapted to desert conditions is that they avoid the extremely dry season, which may last from 6 months to several years. Thus, annual plants may be seen in some habitats every few years. This depends on edaphic and climatic conditions;

for example, annual halophytes may germinate and grow to produce seeds only on salty soils and only in rainy years when the soil solution is diluted to the appropriate level. On the other hand, annuals which develop in crevices of smooth-faced rocks may have sufficient resources to be sustained almost every year. Thus, species diversity in the desert may be high in areas in which habitat diversity is high, especially if there are sufficiently large outcrops of smooth-faced hard rocks which contain many different microhabitats.

### K. Shrub–Steppes

The area of the Negev Highlands, the Judean Desert, Sinai, and southwestern Jordan, which receives 80–250 mm of rainfall annually, is covered with semishrubs at a diffused pattern, thus forming shrub–steppes (Fig. 3, 10). The most common dominants in these steppes are *Artemisia sieberi*, *Noaea mucronata*, and *Gymnocarpus decander*. The phytomass produced by annuals in the plant communities developing on stony or rocky shallow soils is always small when compared with that of fine-grained and deep soils. The latter types hold much of their water close to the soil surface, thus losing much of it through direct evaporation. The minute quantities of 8 ppm of salts carried by clouds and later by rain from the Mediterranean sea by climate systems remain in the soil and accumulate there. The soil may be too dry or too saline for the growth of annuals in regular or dry years. Nearly monospecific communities of semishrubs exist, each best adapted to the specific local saline conditions. The most common dominants under these conditions are *Reaumuria hirtella*, *R. negevensis*, *Salsola vermiculata*, *Bassia (Chenolea) arabica*, and *Atriplex glauca* on chalk- and marl-derived soils; *Anabasis syriaca* and *Haloxylon scoparium* are the shrubby dominants on loess-derived soils. However, in moist years there is much development of annuals, although in patches with high salinity there are monospecific patches of salt-resistant annuals. Showy geophytes such as species of *Tulipa*, *Iris*, *Ixiolirion*, *Ranunculus*, and *Anemone* may bloom in large quantities in the shrub–steppes in moist years.

Outcrops of smooth-faced hard limestone support plants that differ much from those on the other soil types. This vegetation is typically characterized by *Chiliodendron iphionoides*, *C. montanus*, *Globularia arabica*, *Stachys aegyptiaca*, *Polygala negevensis*, *Tanacetum sinai-cum*, and *Capparis aegyptia*. Isolated populations of dozens of Mediterranean relicts and many rare desert plants are found in this habitat in the Negev, the Judean Desert, Sinai, and Jordan. The semishrub *Sarcopoterium*

*spinosum* and the geophytes *Narcissus tazetta* and *Sternbergia clusiana* are examples of this phenomenon in the Negev (Danin, 1983). Several species that were discovered as endemic and new to science and many species that were not previously collected in the steppe areas of the Near East are confined to this habitat (Danin, 1999; Danin and Künne, 1996).

Shrubs of *Retama raetam* and *Achillea fragrantissima* are the dominants in wadis of the terrain of hard limestones. *Atriplex halimus* prevails in this habitat, in which the catchment area is built up from the salty soils on chalk, clay, or marl. At lower elevations, *Acacia raddiana*, *A. pachyceras*, *Tamarix nilotica*, and *T. aphylla* also occur. Many small springs exist in the limestone hills of western Sinai and the sandstone hills of southwestern Jordan. Most of these springs can be detected from afar by the date palm (*Phoenix dactylifera*), which is confined to sites with a high water table of fresh water. It is accompanied by *Nitraria retusa*, *Juncus arabicus*, *Phragmites australis*, and *Cressa cretica*. Canyons occur in many wadis and may have long-lasting water pools supplied by floods and support rich flora of hydrophytes such as *Zannichellia palustris* and *Potamogeton* spp. and green algae such as *Chara* spp.

### L. Shrub-Steppes with Trees

Shrub-steppes similar to those discussed previously cover most of the area of this category (Fig. 3, 11); however, prominent trees occur in special habitats sporadically. Approximately 1400 adult trees, most of which are *Pistacia atlantica*, accompanied by a few *Amygdalus ramonensis* and *Rhamnus disperma* trees and shrubs (Danin, 1983), typify subunit 11a in the Negev Highlands and eastern Sinai. The trees are found in affinity to outcrops of smooth-faced rocks. The high yields of runoff water from the rock outcrops to their crevices enable the successful germination and establishment of seedlings even outside of wadis. Sufficiently large outcrops of limestone support trees, some of which are several hundred years old. Wadis with such outcrops in their catchment area support large trees. Rare relicts, such as the vines of the Mediterranean maquis *Prasium majus* and *Ephedra foeminea*, and endemic plants such as *Origanum ramonense* and *Ferula negevensis* also occur in these rocks.

*Juniperus phoenicea* is the tree of subunit 11b. Of the three anticlinal ridges of northern Sinai, Gebel Halal is the richest in trees and accompanying rare plants. The junipers occur in crevices of smooth-faced limestone outcrops and in wadis. Rare Mediterranean relicts which occur in the rock crevices are *Ephedra foeminea*,

*Rubia tenuifolia*, and *Astomaea seselifolium*. A rare component of the rock vegetation is *Origanum isthmicum*, which is endemic to an area of Gebel Halal. Its closest relative, *O. jordanicum*, was discovered in Jordan near Petra (Danin and Künne, 1996). The juniper populations of Gebel Maghara and Gebel Yiallaq grow mainly in wadis.

Subunit 11c is richer both in nondesert trees and shrubs and in companions. The large outcrops of smooth granite and the high elevation of southern Sinai Massif result in a high number of habitats available for the survival of rare species. The typical trees of the rocky environment in 11c are *Pistacia khinjuk*, *Crataegus sinaicus*, and *Ficus pseudosycamoros*. The west-facing escarpments of Gebel Serbal are rich in *Moringa peregrina*, which also grows in the vicinity of Wadi Feiran oasis. The typical shrubs of the rocky habitat of 11c are *Rhamnus disperma*, *Rhus tripartita*, *Cotoneaster orbicularis*, *Periploca aphylla*, and *Sageretia thea*. Most of the endemic and rare species of Sinai occur in the crevices of rocks that also support trees.

The *Q. calliprinos* and *J. phoenicea* woodlands in Edom, Jordan, are marked 11d on the vegetation map. The climatically controlled belt of the arboreal Mediterranean vegetation terminates in the vicinity of Amman, approximately 120 km north of the Dana-Tafilea area. Large areas of shrub-steppe, semisteppe batha, and steppe-forest typify the western ridge of the Jordanian plateau between At-Tafilea and Petra. These are dominated by *Artemisia sieberi*, *Noaea mucronata*, and spiny species of *Astragalus*. The occasional arboreal components include *Pistacia atlantica*, *Crataegus aronia*, *J. phoenicea*, and *Q. calliprinos*. These formations develop on fissured limestone, basalt, and chalk rocks. The smooth-faced hard sandstone outcrops of the Dana-Petra area support the richest relict Mediterranean flora in the Near East. The trees and shrubs growing here are *Q. calliprinos*, *P. atlantica*, *P. palaestina*, *P. khinjuk*, *C. aronia*, *J. phoenicea*, *Amygdalus korschinskii*, *Ceratonia siliqua*, *Olea europaea*, *Arbutus andrachne*, *Rhamnus punctata*, *R. lycioides*, *R. disperma*, *Ficus carica*, *F. pseudosycamoros*, and *Sageretia thea*. Typical root parasites of the Mediterranean maquis, such as *Osyris alba* and *Thesium bergeri*, occur in the rock crevices, as do typical vines of the maquis—*Rubia tenuifolia*, *Ephedra foeminea*, *Hedera helix*, *Bryonia cretica*, and *Lonicera etrusca*, which are represented by a higher number of individuals than other Mediterranean vines in any other refugia in the Near East. The rich Mediterranean flora with many endemics may be regarded as existing in a successful refugium which has functioned for a long time in the evolutionary history of the region.

### M. Desert Vegetation

Sparsely vegetated shrub–steppes dominated by *Anabasis articulata* and *Zygophyllum dumosum* occur at the broad boundary area of the desert with the steppes zone (Fig. 3, 12 and 10). These typical semishrubs of the desert grow in this transition zone in a diffused pattern. Chalk and marl outcrops are populated with the xerohalophyte communities of *Suaeda asphaltica*, *Salsola tetrandra*, and *Haloxylon negevensis*. The nearly monospecific communities of semishrub xerohalophytes are accompanied by a diverse assemblage of herbaceous plants that grow on the leached soil only in rainy years.

The extreme desert areas may have no higher plants growing out of wadis, but the entire area may be covered by a highly diverse microbiotic crust. Dor and Danin (1996) studied the microbiotic succession and changes in floristic composition of the crust in the Dead Sea area.

The sequence of plant communities along the wadis of a more arid zone of this category is typified by a stretch of annuals which occur in rainy years. The nature of the annuals community is heavily influenced by the edaphic and climatic conditions of the site. A lower section of the wadi receives higher quantities of water and supports small and short-living semishrubs such as *Pulicaria incisa*, which may function locally as an annual. Further down, larger and long-living semishrubs such as *A. articulata* or *Z. dumosum* grow. A section dominated by shrubs such as *R. raetam*, *Ochradenus baccatus*, or *Lycium shawii* prevails further down the wadi system. In the lowest section of the wadi system, trees, mostly *Acacia* species or *Tamarix* species, may be found. The nature of plant communities and the sequence of their occurrence along the wadis are in close affinity to rock and soil types, which greatly influence the moisture, salinity, and nutrient regime in the wadi systems.

In similar gravel plains in the large desert area south of Ma'an to the Saudi Arabian border, *A. articulata* grows with trees of *A. pachyceras* at the fifth-order section of the wadi system. The rocky terrain at the escarpments of southwestern Jordan to the Arava rift valley supports diverse communities in a diffused pattern as occurs in Sinai and the Negev.

### N. Sand Vegetation

The main areas in which sand dunes or sand sheets occur in Israel (Fig. 3, 13) are the Mediterranean coastal plain, the western Negev, a few valleys of the northeastern Negev, and the Arava Valley. Different climatic regime prevails in each of these areas, sand texture differs,

and hence the vegetation and processes of its development differ accordingly. Much of this vegetation, patterns of species diversity, and the nature of the microbiotic crust of the sand in desert areas are discussed in other publications. Thus, some background concerning the sand vegetation of southern Jordan is presented here.

Weathering of nubian sandstones in the Jordanian plateau contributes mobile sand to the Arava Valley. The relatively high water table at the Arava enables the successful development of the tall shrub *Haloxylon persicum* at sites in which sand is sufficiently deep. However, most stands of this plant in Israel became intensively irrigated agricultural areas. Nonetheless, large areas of *Haloxylonetum persici* cover considerable areas of the Arava and the southeastern desert of Jordan (Al-Eisawi, 1996). Annuals accompany the shrubs of the sand sheaths in rainy years. The sandstone area of the Jordanian plateau is built of sandstone inslebergs which project from large flat valley and are filled with stable sand. Huge areas of sand sheets in the Wadi Rum area are dominated by *Haloxylon salicornicum*, *A. articulata*, and occasional patches of *H. persicum*. The sandstone hills support many relict Mediterranean species and there is particularly interesting vegetation at the vicinity of the contact zone between the sandstone and the Precambrian crystalline rocks.

### O. Oases with Sudanian Trees

The climate of the rift valley is much warmer than that of the hilly terrain surrounding it. Springs in which a high quantity of fresh water is available throughout the year (Fig. 3, 14) support thermophilous Sudanian trees such as date palms (*P. dactylifera*), *A. raddiana*, *Acacia tortilis*, *Calotropis procera*, *Moringa peregrina*, *Balanites aegyptiaca*, *Cordia sinensis*, *Maerua crassifolia*, *Dalbergia sisoo*, *Capparis decidua* (in Jordan), and *Z. spina-christi*. Rich annual flora may accompany sites which are not under the influence of the springwater but successfully grow in rock crevices, soil patches, and abandoned cultivated land. In many freshwater springs/oases near the Dead Sea in Israel and Jordan and in southern Sinai, *Adiantum capillus-veneris* grows on dripping water at shady places with the orchid *Epipactis veratrifolia*. A prominent species of the freshwater springs in the Jordanian desert, *Nerium oleander*, is absent from the desert springs of Israel and Sinai.

### P. Desert Savannoid Vegetation

A considerable part of the Arava and Dead Sea Valleys is covered by savannoid vegetation in which *Acacia*

trees are accompanied by desert semishrubs (Fig. 3, 15). The upper soil layers support in wadis the typical desert vegetation. *Acacia pachyceras* is the dominant tree of areas which are at a relatively high elevation, such as the upper tributaries of Nahal Paran in Israel, Wadi Jirafi in Sinai., and large areas of gravel plains in Jordan south of Ma'an. *Acacia raddiana*, which is less resistant to low temperatures, is the dominant tree at lower elevations, and *A. tortilis*, which has the highest demands for high temperatures, grows in the southern Arava or below sea level in the northern Arava and Dead Sea Valley. In a few places between Wadi Watir and Sharm el Sheikh in eastern Sinai, rare trees of *Capparis decidua* grow, which is an important component of the savanna vegetation of southern Egypt and Sudan.

### Q. Swamps and Reed Thickets

Only small nature reserves, such as that of the Hula Lake, remain (Fig. 3, 16) from the swamps which covered large areas at the beginning of the twentieth century. Freshwater springs still flow and their vegetation is composed of a small number of species that produce high quantities of phytomass. The most common hydrophytes are *Phragmites australis*, *Arundo donax*, *A. pliniana*, and *Typha domingensis*. Large areas of swamps in the Hula Valley supported *Cyperus papyrus*, which reached its northernmost station at this valley. Several rivers of the coastal plain that supported riparian vegetation in the past have become sewage canals and their polluted water has destroyed most of the past rivers' vegetation.

### R. Wet Salinas

Wet, salty soils occur in places in which springs of salty water occur or at sites in which the water table is near the surface and the evaporating water leaves salt in the upper soil layers (Fig. 3, 17). A salt crust may occur at the soil surface in arid regions and plants are restricted to small wadis in which leaching occurs. Most plants growing in the desert salt marshes are perennials that establish themselves in the rare event of leaching. The typical plants of desert salty soils are *Suaeda monoica*, *S. fruticosa*, *S. vermiculata*,

*Nitraria retusa*, *Seidlitzia rosmarinus*, and a few species of *Tamarix*.

### S. Synanthropic Vegetation

The category of synanthropic vegetation (Fig. 3, 18) is further divided in Israel into three subcategories according to the remnants of trees found in the intensively cultivated areas: In 18a, it is trees of *Quercus ithabur-ensis*, in 18b the tree is *Z. spina-christi*, and in 18c it is *A. raddiana* and *Z. spina-christi*. The synanthropic vegetation of Jordan and the current status of the synanthropic vegetation in Sinai need further investigation.

### T. Mangroves

The mangrove of Nabq, eastern Sinai (Fig. 3, 19), typically growing in muddy soils of the tidewater, constitutes the most northerly population of *Avicennia marina* on Earth at 28°10'N. This species was recorded even farther from the equator in South Australia at 37°S. The only population found in the Gulf of Suez shares its water, and possibly warmth, with the Gulf of Elat at Ras Mohammed.

### See Also the Following Articles

DESERT ECOSYSTEMS • MEDITERRANEAN-CLIMATE ECOSYSTEMS • NEAR EAST ECOSYSTEMS, ANIMAL DIVERSITY

### Bibliography

- Al-Eisawi, D. M. (1996). *Vegetation of Jordan*. UNESCO, Cairo.
- Danin, A. (1983). *Desert Vegetation of Israel and Sinai*. Cana, Jerusalem.
- Danin, A. (1996). Vegetation of Israel and Sinai. *Bot. Zhurn.* 81(11), 14–31.
- Danin, A. (1999). Desert rocks as plant refugia in the Near East. *Bot. Rev.* 65(2), 93–170.
- Danin, A., and Künne, I. (1996). A new species of *Origanum* (Labiatae) from Jordan: *O. jordanicum* Danin et Künne sp.n., and notes on the species of section *Campanulatalyx*. *Willdenowia* 25(2), 601–611.
- Dor, I., and Danin, A. (1996). Cyanobacterial desert crusts in the Dead Sea Valley, Israel. *Algolog. Stud.* 83, 197–206.
- Le Houérou, H. N. (1995). Bioclimatologie et biogéographie des steppes arides du Nord de l'Afrique; Diversité biologique, développement durable et désertisation, OPTIONS méditerranéennes, serie b (tudes et recherches No. 10). CIHEAM (Montpellier) and ACCT (Paris), France.



# NEST PARASITISM

Scott K. Robinson\* and Stephen I. Rothstein†

\*University of Illinois at Urbana-Champaign and †University of California at Santa Barbara

---

- I. Natural History
  - II. Coevolution between Brood Parasites and Their Hosts
  - III. Modeling Host–Parasite Coevolution
  - IV. Impacts on Host Population Dynamics
  - V. Ecology and Social Behavior of Brood Parasites
  - VI. Conspecific Brood Parasitism
  - VII. Conclusions and Research Needs
- 

## GLOSSARY

**acceptance** Hosts treat brood parasitic eggs as if they were their own eggs.

**coevolution** Cycles of adaptation and counteradaptation that occur between interacting lineages.

**coevolutionary arms race** The continuing bouts of coevolving defenses and counterdefenses that occur between hosts and parasites.

**conspecific brood parasites** Brood parasites that lay their eggs in the nests of other individuals of the same species.

**cowbirds** American blackbirds in the genus *Molothrus*, which contains important brood parasites.

**cuckoos** A family of birds of which approximately half the species (61 of 125) are obligate interspecific brood parasites.

**egg rejection** Host responses to parasitism that include ejection of parasitic eggs, abandonment of parasit-

ized nests (usually with renesting), or a new cup built over a parasitized clutch.

**evolutionary equilibrium hypothesis** The hypothesis that frequencies of acceptance of brood parasitism reflect an equilibrium between costs and benefits of host defenses against parasitism.

**evolutionary lag hypothesis** The hypothesis that hosts lack defenses against brood parasites because the defenses have not yet had time to evolve.

**generalist brood parasites** Species that parasitize many (up to more than 200) host species.

**gentes (singular gens)** Lineages of cuckoos in which individual females specialize on a single host and lay mimetic eggs.

**host** The individual that is parasitized.

**interspecific brood parasites** Brood parasites that lay their eggs in the nests of other species.

**mafia effect** Interspecific brood parasites that may destroy clutches from which parasitic eggs have been ejected by the host.

**mimicry** Brood parasitic eggs or nestlings that closely match those of the hosts.

**specialist brood parasites** Species that parasitize only one or a few host species.

---

**BROOD PARASITISM**, also called social parasitism, is the exploitation by one individual (the brood parasite) of the parental care of another (the host). Brood para-

sites can deposit eggs in the nests or broods of another individual of the same (conspecific brood parasitism) or of a different (interspecific brood parasitism) species. Hosts often raise young of the brood parasite, typically at the expense of their own young. This article presents a comprehensive overview of the natural history, evolution, and consequences of brood parasitism, with a special focus on birds, the taxon in which it has been best studied.

## I. NATURAL HISTORY

Brood parasites lay their eggs in the nests of other individuals, which then raise the parasitic young at the expense of part or all of their own brood. Brood parasitism was noted by Aristotle and even earlier (approximately 2000 B.C.) in India. Brood parasitism is best known in birds, approximately 1% of which are obligate interspecific brood parasites that only lay their eggs in the nests of other species. Interspecific brood parasitism, however, has also been documented in insects and fish. Conspecific brood parasitism, in which individuals facultatively lay eggs in nests of conspecific individuals, is more widespread.

Brood parasites have generated intense interest in the public and scientific communities. Brood parasites tend to be vilified in the media because of a human tendency to moralize about such things as killing baby birds (which parasitic birds often do) and because at least some may pose a conservation threat to some of their hosts. Among scientists, brood parasites have generated intense debate about the coevolutionary processes that may be responsible for the seemingly maladaptive acceptance of parasitic eggs by hosts.

### A. Adaptations of Obligate Avian Brood Parasites

Brood parasites search for host nests, synchronize their laying with that of the host, and often remove a host egg from nests they parasitize. Brood parasitic eggs often have unusually thick eggshells, are usually small relative to the size of the parasite (but large relative to the size of the host), often mimic the coloration of the hosts' eggs, and have rapidly developing embryos. All these traits increase fitness of the parasite by reducing competition with host nestmates and making parasitic eggs more difficult to detect and remove. Egg mimicry is most pronounced in specialist brood parasites. In the

well-studied common cuckoo, the species as a whole is a generalist, but each individual female is a specialist member of a "gens." Any one region has only one to several coexisting gentes but different arrays of gentes occur in different regions.

Brood parasitic nestlings also have a variety of mechanisms that enhance their ability to compete with host nestlings. Cuckoo nestlings have concave backs that they use to push out host eggs and nestlings, whereas several other brood parasites have specialized bill hooks that they use to kill nestmates. Some brood parasites apparently increase the amount of food hosts bring to them by mimicking the juvenile plumages or complex mouth markings (the "gapes") of host nestlings or by having large mouths and intense begging behavior.

Because brood parasites do not have to engage in costly breeding activities such as incubation and nestling feeding, they often have more time and energy to devote to egg production. Some generalist brood parasites lay 40 or more eggs in a season; estimates for one tropical brood parasite suggest that more than 100 eggs/year are typical. Other brood parasites, however, may be only slightly more fecund than their hosts.

### B. A Survey of Avian Brood Parasites

There are 90–95 species of obligate avian brood parasites in five unrelated families (Table 1).

#### 1. Cuculidae

The cuckoos are a diverse family that contains both parasitic and nonparasitic species. The Cuculinae has traditionally been recognized as an Old World subfamily, all of whose approximately 50 species are obligate

TABLE I  
Major Groups of Avian Brood Parasites

Taxon (family/subfamily)	No. of species
Old World cuckoos ( <i>Cuculinae</i> ) <sup>a</sup>	50
New World cuckoos ( <i>Neomorphinae</i> )	3
Honeyguides (Africa and Asia) ( <i>Indicatoridae</i> )	17?
Vidua finches (Africa) ( <i>Ploceidae</i> )	14
Cuckoo-finch (Africa) ( <i>Ploceidae</i> )	1
Cowbirds (New World) ( <i>Icteridae</i> )	5
Black-headed duck (South America) ( <i>Anatidae</i> )	1

<sup>a</sup> Recent work indicates that at least one genus in this subfamily belongs in another cuckoo subfamily (see text).

brood parasites. Most species are host specialists that are relatively uncommon and only a few have been well studied (the common cuckoo and the great spotted cuckoo). Another traditionally recognized subfamily, the Neomorphinae or New World ground cuckoos, has 11 species, 3 of which are relatively rare obligate parasites. The latter tend to parasitize hosts that build domed nests and are poorly known overall. Nestlings of one species, *Tapera naevia*, have pincer-like bills that they use to kill host nestlings, a case of convergence on honeyguides (Indicatorinae). Recently published DNA sequence data indicate that at least 1 genus (*Clamator*) within the Cuculinae is more closely related to the Phaenicophaeinae, a subfamily of New World cuckoos. None of the currently recognized members of the Phaenicophaeinae are obligate parasites but some parasitize conspecifics and occasionally other species. These new DNA data place obligate parasitism within three groups of cuckoos, two of which have both parasitic and non-parasitic species, and indicate that parasitism evolved separately up to three times in cuckoos or that some parasitic cuckoo lineages reverted to parental behavior.

## 2. Indicatoridae

The honeyguides, a family named for the habit of one species that guides people and the African honey badger or ratel to beehives, has 17 obligately parasitic species. The species of *Indicator* and two related genera parasitize cavity nesters, whereas 3 species of *Protodiscus* parasitize open-cup nesters. Most species are poorly known and the eggs of many species have never been described. Some honeyguides have raptorial hooks on their bills, which they use to kill host nestlings.

## 3. Viduine Finches

The approximately 16 species of obligate brood parasitic viduines occur in Africa and provide some of the best examples of nestling mimicry, particularly of the intricate gape patterns of their hosts. Compared with other brood parasites, viduines exert relatively low costs on their hosts, which can usually raise mixed broods. Most *Vidua* species specialize on single species of grassfinches (Estrildidae), to which they may be related. Unlike most brood parasites, which are insectivorous, hosts of viduines are granivorous. Some species of viduines incorporate the songs of their hosts into their own songs, which are then used by females to choose males that have been raised by the same species.

## 4. Cuckoo-Finch (*Anomalospiza imberbis*)

This African species is closely related to Viduine finches and parasitizes approximately 11 species of

warblers in grasslands. Recently reported DNA evidence suggests that parasitism evolved once in a single lineage that then gave rise to both *Anomalospiza* and the Viduines.

## 5. Icterinae

The five species of New World brood parasitic cowbirds include the most generalized of all brood parasites, the brown-headed and shiny cowbirds, both of which parasitize more than 200 host species. Another species, the tropical giant cowbird, parasitizes almost exclusively colonial American blackbirds. Of the two remaining species, the bronzed cowbird is a host generalist (>80 host species), whereas the screaming cowbird is one of the most specialized of all brood parasites. It mainly parasitizes a single species of blackbird, which was formerly thought to be a species of cowbird. Unlike most brood parasites, cowbirds are often more abundant than many of their hosts and can pose a significant threat to populations of rare, localized host species. Most cowbirds benefit greatly from changes in the landscape caused by humans, which has enabled several species to expand their geographic ranges and increase their population sizes. As a result, they are coming into contact with new hosts that have not had recent contacts with brood parasitism. Some of these new hosts suffer extremely high levels of parasitism. Cowbirds appear to be extremely fecund; the shiny cowbird may lay more than 100 eggs a year. The invasion of North America by this cowbird may pose an additional threat to species that are not currently being parasitized by the brown-headed cowbird. The brown-headed cowbird may be the most intensively studied North American bird.

## 6. Black-Headed Duck (*Heteronetta atricapilla*)

This species is the only obligately parasitic species with precocial young that feed themselves when they hatch. For this reason, this brood parasite has little effect on the nesting success of its hosts, which only have to incubate an extra egg or two. Because incubation is so similar among all bird species, *Heteronetta* successfully parasitizes a wide range of birds, such as gulls and ibises, as well as other ducks.

## 7. Conspecific Brood Parasitism

Recent studies have shown widespread conspecific brood parasitism, especially in waterfowl, gallinaceous birds, and in a small number of songbirds (generally species with rare nest sites or that breed colonially).



## 8. Other Brood Parasite Systems

### a. Fish

An African catfish is the only fish species known to be an obligate brood parasite. It parasitizes mouth-breeding cichlids and apparently consumes all of its host's young while in the mouth of its foster parent. Conspecific brood parasitism and occasional parasitism of other species occur in many nest-building fish. Some host defenses such as increased nest guarding may have developed in response to this parasitism but egg recognition has not been reported.

### b. Insects

A range of insects practice brood parasitism, especially members of the order Hymenoptera (bees, wasps, and ants). Some of these species are obligate brood parasites, such as certain ants that kill the queens of nonparasitic species and use the workforce of the entire colony to rear their own young. Other species are facultative parasites that victimize conspecifics or raise their own young. To our perception, some of these parasitic insects look remarkably different from their hosts so their acceptance by hosts seems inexplicable. However, olfaction is critical to these insects and it is likely that parasitic species have evolved chemical cues that mimic those of their hosts. Because both humans and birds perceive the world largely via vision and audition, and have relatively poor olfactory abilities, it is not surprising that parasitic birds have been subjects of much more research than have parasitic insects, even though the latter undoubtedly have many equally interesting examples of coevolved adaptations.

## II. COEVOLUTION BETWEEN BROOD PARASITES AND THEIR HOSTS

### A. Cuckoo-Host Systems

Cuckoo-host systems are much more highly coevolved than the other well-studied systems, those of cowbirds, and DNA divergence data show that they are also much older. Most cuckoo species use only one or a few host species. Some species; such as the Eurasian common cuckoo, might be called generalists when viewed over their entire ranges because they parasitize 50–100 or more host species. However, even these cuckoos are more properly viewed as specialists because they use only one to several host species in each region and individual females, or members of a gens, specialize on a single host species or several similar host species. In

regions where several cuckoo species coexist, they tend to show little or no overlap in host use, which reduces competition for hosts. Most cuckoos are much less abundant than their hosts and probably have little effect on host population dynamics because they affect only a small proportion of their host populations.

Some cuckoo-host systems may have reached an evolutionary equilibrium, whereas others show strong evidence of ongoing coevolution and provide some of the strongest examples in vertebrates of microevolution that has occurred during ongoing research.

## B. Adaptations of Cuckoos

### 1. Behavior

Cuckoos typically approach nests stealthily and may quickly drop their thick-shelled eggs onto host eggs to increase chances of breaking host eggs. Males of some cuckoos resemble hawks. This resemblance or simply host recognition of cuckoos incites host aggression and males in some species use this aggression to draw hosts away from nests. This male distraction is coordinated with and facilitates nest entry by the stealthy and drab-colored females in several cuckoo species. Nestlings of some cuckoos mimic the begging calls or plumage of host young. There is evidence that great spotted cuckoos revisit nests they have parasitized and destroy host eggs or nestlings if the parasitic egg has been removed by the host. Such a protection racket or "mafia effect" could select for acceptance of host eggs.

Cuckoos find nests by watching hosts from hidden perches and laying is synchronized with that of the hosts. Most cuckoos lay one egg per nest, except in a few species in which cuckoo nestlings do not kill those of the host. Cuckoos defend territories against other cuckoos and usually remove a host egg before laying one of their own.

### 2. Fecundity

Cuckoos lay 8–25 eggs per season, which is generally more than their hosts lay. Eggs are laid every other day.

### 3. Egg Adaptations

Cuckoo eggs usually mimic the coloration of their hosts' eggs. Cuckoos that parasitize many host species have individually specialized females, with each group of females or gens (plural gentes) mimicking the eggs of a single host species or a group of hosts with similar eggs. Egg mimicry is usually assumed to have arisen in response to host egg recognition. An alternative hypoth-

esis argues that egg mimicry evolved to keep a second cuckoo from identifying and removing a cuckoo egg from an already parasitized nest and then laying its own egg in the nest. However, it is not clear that more than one cuckoo attempts to parasitize the same host nest often enough for this to be a significant selective pressure. Most cuckoos parasitize hosts that are much smaller than themselves and have correspondingly small eggs for their body sizes. However, cuckoo eggs are usually slightly larger than those of the hosts. The few species that parasitize hosts larger than themselves do not have smaller eggs. The thick shells of cuckoo eggs may be adaptive because they reduce chances of breakage of cuckoo eggs during rapid laying, facilitate breakage of host eggs (thereby reducing competition with hosts), or make cuckoo eggs resistant to pecking by hosts. A counteradaptation to egg mimicry that some hosts have developed is variable egg types. In response, some cuckoo species have evolved an equivalent range of egg types; but hosts still seem to benefit because individual female cuckoos lay only one type and may not always parasitize a host female who lays the same type. The highly variable host eggs may also function in the context of conspecific parasitism.

#### 4. Nestling and Fledgling Adaptations

Nestlings in most cuckoo species have concave backs that they use to push host nestlings or eggs out of nests. Cuckoos that parasitize larger hosts do not show this eviction behavior but can compete successfully with host nestlings because they beg more loudly and incessantly, hatch earlier, and develop more rapidly. Mimicry of host begging calls occurs mainly, but not exclusively, in species that do not evict host nestlings.

### C. Host Defenses Against Cuckoos

#### 1. Defenses against Adults

Hosts often respond aggressively to cuckoos, but this may actually provide a cue to cuckoos about their stage of the nest cycle and the proximity of nests. Mobbing cuckoos has both genetic and learned components. One possible advantage of mobbing cuckoos is that it may trigger ejection of parasitic eggs.

#### 2. Egg Rejection

Hosts reject cuckoo eggs by abandoning parasitized nests or ejecting eggs directly. Small hosts are more likely to abandon (and re-nest), perhaps because ejection results in the incidental breakage of some of their own

eggs and is therefore costly or because they are simply too small to eject cuckoo eggs. Egg rejection is most prevalent in hosts that have a long coevolutionary history with cuckoos. These hosts, which often show fine abilities to discriminate among egg types, are species that have intrinsic characteristics that make them suitable for parasitism, such as animal food appropriate for the nestling parasites and nests accessible to adult cuckoos. Species unsuitable as hosts, such as ones that have specialized nestling diets (e.g., seeds in some finch species), have presumably never been parasitized intensely in the past and generally lack egg recognition altogether. Among suitable hosts, egg recognition is more prevalent in species that are currently rarely parasitized than in species that are currently common hosts. This trend suggests that cuckoos shift away from using suitable hosts with well-developed egg recognition. It is possible that cuckoos have a dynamic system of host usage in which they repeatedly switch from hosts with well-developed defenses only to switch back to these hosts after their defenses have declined and defenses in other hosts have increased. Such a system could result in never-ending cycles of host switches and coevolution, but it is also possible that hosts retain high levels of egg recognition for long periods in the absence of parasitism, which would force cuckoos to become increasingly more specialized. Long-term retention is indicated by high levels of egg recognition in New World magpies and shrikes, which are currently not parasitized by any brood parasites but are descended from Old World ancestors that are cuckoo hosts. Populations of some suitable hosts that have only recently come into contact with cuckoos have apparently undergone rapid increases in egg ejection and discrimination, apparently because they possessed some recognition even in the absence of parasitism. This possession could be due to retention of recognition from past bouts of parasitism or to gene flow from parasitized populations of the same species.

#### 3. Nestling Discrimination

Despite there being many species that have some degree of mimicry of host nestling appearance or begging calls, there are no known cases in which cuckoo hosts show outright rejection (i.e., removal) of nonmimetic nestlings. This lack of outright rejection appears to be true in all other systems of avian parasitism that show equivalent or even better mimicry of host nestlings. Therefore, the selective value of nestling mimicry may relate to ensuring that parasitic nestlings receive high-quality care from their hosts rather than avoiding outright rejection.

## D. Cowbird–Host Systems

The five brood parasitic cowbirds include one of the most specialized of all brood parasites and the two most generalized of all brood parasites. We have chosen to highlight cowbird–host interactions in this article because there is a vast literature on cowbird–host systems and because these systems are quite different from cuckoo–host systems.

### 1. Cowbird Adaptations

#### a. Egg Adaptations

As described previously, some cowbirds are extremely fecund, an apparent reflection of a trade-off between egg production and parental care. In captivity, brown-headed cowbirds can lay up to 77 eggs in the short temperate breeding season and field estimates of shiny cowbird productivity exceed 100 eggs during the much longer tropical breeding season. As in cuckoos, the eggs of the smaller cowbirds are small relative to the size of the cowbird but are larger, thicker shelled, rounder, and faster to develop than those of most of their hosts. Only the giant cowbird, which parasitizes hosts as large as itself, has normally proportioned eggs. The other smaller species usually parasitize smaller hosts. Egg mimicry has not been firmly established in any cowbirds, but cowbird egg shape and patterns have evolved, perhaps in response to diffuse selection by dominant hosts or groups of hosts. The two most generalized cowbirds, the shiny and brown-headed, have generalized spotted egg colorations that are similar to those of many passerines. Although cowbird egg coloration is not finely tuned to host egg coloration, as in cuckoos, it has undergone some shifts, presumably in response to host use, during the evolution of the cowbird lineage. Different cowbird species have generalized spotted eggs or immaculate unspotted ones; however, two species, the giant and shiny cowbirds, have both kinds of eggs.

#### b. Behavioral Adaptations of Adults

Similar to cuckoos, cowbirds search for nests primarily by observing the behavior of hosts. Cowbirds approach nests stealthily and lay eggs very early in the morning before most hosts appear at their nests during the egg-laying period. In contrast, giant cowbirds appear to circumvent the defenses of colonial hosts by visiting colonies as a group in which males appear to distract hosts while females stealthily enter nests. Screaming cowbirds also visit their communally breeding hosts in groups, but males do not appear to help distract the hosts. Cowbird eggs are usually synchronized properly with host laying but improperly timed eggs that appear

before the host eggs or after hosts have finished laying occur more commonly than in cuckoos. Although generalist cowbird species avoid parasitizing some species that are clearly unsuitable as hosts, such as doves, there is little evidence that they select the best host species in a community. Many cowbird eggs are seemingly wasted in nests of hosts that feed their nestlings inappropriate diets of seeds or fruit or that eject cowbird eggs. However, parasitism of the latter species may sometimes be adaptive for two reasons. First, some species that exercise egg recognition and ejection have low nest predation rates. Second, these species learn the appearance of their own eggs by imprinting on the first egg or eggs that they lay during their lives. Accordingly, naive hosts parasitized about the time they begin laying may come to learn cowbird eggs as their own eggs and may provide cowbirds with nests that have a relatively low likelihood of failing due to nest predation. Multiple parasitism (two or more cowbird eggs in a nest) occurs in approximately one-third of all nests parasitized by cowbirds, especially with some large hosts. Several female cowbirds sometimes lay in the same nest even though multiple parasitism usually reduces the success rate of cowbird eggs and many host nests receive no cowbird eggs. Female cowbirds are usually highly aggressive toward one another in habitats in which hosts occur (but not in which cowbirds feed), but there is still substantial overlap among home ranges. Cowbirds sometimes, but not always, remove host eggs by puncturing them. Egg removal is rarely done during the same nest visits on which females lay but may occur the afternoon before or later in the morning on which cowbird eggs are laid. There are many observations of cowbirds depredating unparasitized host nests and there is evidence that cowbirds do this regularly to stimulate re-nesting, which will increase the future availability of host nests to be parasitized. However, some studies have found no evidence for this depredation-re-nesting hypothesis and the behavior may occur only in special circumstances, if at all.

#### c. Nestling and Fledging Adaptations

In the specialized screaming cowbird, there is almost perfect mimicry of host plumage and vocalizations of its usual host species, the bay-winged cowbird (which is not closely related to the parasitic cowbirds despite its name). This mimicry does not seem to be essential for nestling care but instead seems to be an absolute requirement for receiving host care after the parasites fledge from host nests. Such perfect mimicry is lacking in the other, more generalized cowbirds. However, giant cowbird nestlings have white bills, like those of their

oropendola hosts. Bills turn the usual cowbird black after the fledglings become independent. Nestling cowbirds grow more rapidly than the nestlings of many, but not all, of their hosts and beg more loudly. They have relatively large mouths and show intraspecific variation in their gape colors, unlike the majority of nonparasitic nestlings. This gape variation may relate to host use in some unknown way. There is a recent report of a cowbird ejecting a host nestling, although the ejection could have been inadvertent.

## 2. Host Defenses against Cowbirds

### a. Preventing Access to Nests

Many hosts react aggressively to cowbirds and colonial nesting may provide protection for some hosts. Aggressive responses to cowbirds, however, may reveal nest locations to cowbirds and be counterproductive for small hosts, which may not be able to drive cowbirds away from nests. Some hosts sit on their nests when approached by a cowbird, but there are records of cowbirds pulling hosts off their nests. Some hosts reduce parasitism by nesting in cavities or in dense vegetation, but even cavity nests are often parasitized in many species.

### b. Egg Rejection

Cowbird host species can be divided into acceptors, which accept cowbird eggs, and rejecters, which abandon parasitized nests, build a new nest on top of the old nest, or, most commonly, eject cowbird eggs. Within acceptor species, nearly 100% of the individuals accept cowbird eggs. Rejecter species show nearly 100% rejection and there is little geographic variation in a species' responses to cowbird eggs. Birds that eject cowbird eggs either puncture them or grasp them and usually drop them at least several meters from the nest. Smaller hosts tend to puncture-eject, whereas larger hosts grasp-eject. Nest abandonment occurs most frequently when cowbird eggs are laid too early in a host's laying cycle and when hosts are too small to eject eggs (or to eject without breaking too many of their own eggs). Nest desertion appears to be triggered by encounters at the nest between hosts and cowbirds. In nearly all host species, detection of cowbird eggs in nests does not appear to play a role in eliciting desertion, even in hosts whose eggs are highly divergent from cowbird eggs. Some hosts may begin incubating early in the morning to prevent undetected cowbird parasitism. Most species that reject cowbird eggs occur in open habitats where they may have a long coevolutionary history with cowbirds, all of which forage in open areas in association with grazing ungulates. Hosts that eject cowbird eggs

recognize their own eggs and remove cowbird eggs even if these outnumber their own eggs, unless they have misimprinted on cowbird eggs laid in their very first nest.

## III. MODELING HOST-PARASITE COEVOLUTION

There is little doubt that the adaptations and counteradaptations described in Section II result from coevolution between parasites and their hosts. Parasitic counterdefenses to host defenses must correspond to the characteristics of hosts, as in mimicry of host eggs. Thus, parasites cannot have effective defenses against more than a few hosts at any one time because hosts vary in key characteristics such as egg coloration. The traditional view therefore hypothesizes that parasites specialize on increasingly fewer hosts as a parasite-host system becomes older and more hosts evolve defenses. This view is supported by studies showing that (i) recently derived brood parasites tend to be more generalized than older species (e.g., cowbirds versus cuckoos); (ii) the costs of parasitism are often much higher than the costs of potential host defenses, which suggests that there is a continuing evolutionary arms race; (iii) some systems in which hosts and parasites have only recently come in contact show either an absence or low levels of host defenses; (iv) host defenses have evolved since recent contact with a brood parasite; (v) parasites have evolved egg mimicry in response to host defenses; and (vi) cuckoo-host systems are dynamic as evidenced by host shifts in parasitic cuckoos. In accord with these generalizations, most Old World passerines in Africa and Eurasia, which are exposed to numerous species of an ancient group (the cuckoos), possess egg recognition. However, most New World passerines, which are exposed to a smaller and relatively younger group of parasites (the cowbirds) lack egg recognition.

The question then arises, why do so many hosts accept easily distinguishable parasitic eggs? Two hypotheses have been proposed: "evolutionary lag" and "evolutionary equilibrium." There is a burgeoning literature by proponents of each hypothesis. The evolutionary lag hypothesis is that these acceptors have not yet had time to evolve antiparasite adaptations. Acceptance of parasitic eggs by some hosts in heavily parasitized, unproductive populations may also result from gene flow from populations in areas where cowbirds are absent and reproduction is high. The lag hypothesis is clearly applicable to situations in which parasites have

begun to use host species with no recent history of exposure to parasitism but may even apply to some host species with long histories of parasitism because egg recognition/rejection is a totally new feature that may be difficult to evolve. In the evolutionary equilibrium hypothesis, hosts accept parasitic eggs because of the high costs of rejection. In this scenario, parasitism is costly, but trying to reject parasitism would be even more costly, i.e., acceptors are making the best of a bad situation.

Distinguishing between lag and equilibrium has proven to be difficult. Testing either hypothesis requires measurements of the costs of rejection in species that accept, which has not yet been done effectively. Sometimes, likely costs for acceptors can be estimated from similar species that eject. For example, the number of their own eggs that small ejecting hosts break when they eject cowbird or cuckoo eggs can be extrapolated to acceptor hosts of similar size. These ejection costs, generally less than 0.5 host eggs per ejection, do not seem to be sufficiently high to make acceptance a more adaptive option because acceptance usually results in the loss of all the host's young for cuckoo hosts and for small cowbird hosts. There is evidence that rejecter hosts mistakenly eject their own eggs on occasion, but it is not clear if this cost is sufficient to outweigh the demonstrable costs of egg acceptance. Equilibrium may be most likely in species for which parasitism is relatively infrequent and in which parasitic eggs resemble those of the host, which greatly increases the probability of mistakes. Nevertheless, distinguishing conclusively between lag and equilibrium will require new experimental and phylogenetic studies.

#### IV. IMPACTS ON HOST POPULATION DYNAMICS

Brood parasitism is undoubtedly almost always costly for individual hosts, but the extent to which brood parasites threaten host populations is much less clear. Most brood parasites have little impact on host populations, presumably because coevolutionary processes reduce the frequency of parasitism and because most parasites are much less abundant than their hosts. In Europe, for example, less than 5% of most host nests are parasitized by the common cuckoo. One brood parasite, the giant cowbird, has even been hypothesized to benefit its hosts because its nestlings remove parasitic botfly eggs from its host nestmates.

Nevertheless, brood parasites can pose threats to host populations, especially generalist brood parasites and those that have undergone recent range expansions and population increases. In a few cases, cuckoos have expanded their ranges as a result of human-caused increases of early successional habitat. In areas recently colonized by cuckoos in Japan, levels of parasitism often exceed 40%, although in most of these areas the incidence of host defenses appears to be increasing rapidly, probably because these host populations already had some level of defense before being parasitized. Such rapid development of host defenses, however, appears to be absent from hosts of the brown-headed and shiny cowbirds. Furthermore, range expansions are much more pervasive in these two generalist cowbirds than in any cuckoo species. Cowbirds have undergone enormous population increases as the result of human activities, especially those associated with cattle. As a result, many host populations with no recent history of parasitism are being exposed to massive levels of brood parasitism. For some species, especially forest-nesting ones, widespread parasitism may be occurring for the first time in the history of the species' lineage. Parasitism levels of many forest species nesting in the midwestern United States, for example, exceed 80% with most nests receiving multiple cowbird eggs. Most newly exposed host lineages lack effective defenses. Several recent studies have linked large-scale population declines of songbirds with increasing levels of cowbird parasitism. Concerns about cowbird parasitism have stimulated an enormous amount of research since the late 1980s and two major symposia have been published recently on this subject.

This increased focus on cowbirds has resulted in pressure to trap and kill cowbirds to reduce their impacts on hosts. However, host species that have declined have experienced massive loss of habitat and increased rates of nest predation in addition to increased rates of cowbird parasitism. For example, most passerine species that breed in riparian habitat in the southwestern United States have declined in the past century but so too has their habitat. Dams, water diversions, overgrazing, urbanization, and exotic plants have resulted in the degradation or loss of more than 90% of the riparian habitat present a century ago in the Southwest. Even if cowbirds are not the primary cause of some or all declines, they may now be exposing some reduced host populations to additional stresses that threaten the populations with extinction. Importantly, a cowbird population can be stable or grow even as it pushes a rare host

to extinction because individual female cowbirds do not specialize on single host species.

Nevertheless, it is not clear whether cowbird parasitism threatens more than a few host species. Species that are most at risk are those with small geographic ranges that are wholly included within areas that contain abundant cowbird foraging habitat (pastures, feedlots, mowed grass, and bare soil) and that cannot be rescued by emigrants from more productive populations. This list includes several endangered species that are brown-headed cowbird hosts (Kirtland's warbler, black-capped vireo, least Bell's vireo, and southwestern willow flycatcher) and many species restricted to islands in the Caribbean that have recently been invaded by shiny cowbirds (yellow-shouldered blackbird and Puerto Rican vireo).

Most species in North America have very large geographic breeding ranges that include regions in which cowbirds are rare and restricted to areas near human habitations. These refugia from cowbird parasitism tend to occur in large, unfragmented habitats that may act as "sources" of surplus host young that can recolonize populations in more fragmented habitats in which levels of parasitization (and nest predation) are often very high. In these population "sinks" in fragmented habitats, levels of reproduction may be too low to compensate for adult mortality; that is, such populations can only be sustained by immigration from source habitats. Evidence for this source-sink scenario derives from well-documented sink populations in fragmented midwestern U.S. forests. These populations are nevertheless relatively stable, probably due to immigration from populations in large, unfragmented forests in the region in which levels of both parasitism and nest predation are very low and reproductive success is sufficiently high for populations to act as sources. Such large-scale source-sink population dynamics can slow the evolution of host defenses because most young are being produced in areas in which cowbird parasitism is rare.

### A. Cowbird Management

Many cowbird control programs have been initiated as a result of concern regarding several endangered songbird species. Controlling cowbirds, which is easily done because their social nature attracts them into traps that contain cowbirds that function as decoys or "bait," has become a multimillion dollar a year business in the American Southwest and there are also active programs elsewhere in North America and in the Caribbean. Local control efforts, in conjunction with habitat restoration,

may have prevented the extinction of several endangered host species and races, and two endangered hosts have increased in population since cowbird control began. However, control programs for two other endangered hosts resulted in no increases in the sizes of host breeding populations, even though all control programs have resulted in increases in host reproductive output. Unfortunately, even when host populations increase, cowbird control must be done every year because cowbird removal has little or no year-to-year effect on the numbers of cowbirds that occur in an area due to high dispersal by cowbirds. Therefore, most workers view cowbird control as a temporary, stop-gap measure, although a U.S. Fish and Wildlife recovery plan for the least Bell's vireo in California advocates cowbird control in "perpetuity." Larger scale control programs, such as killing cowbirds by the millions at winter roosts, have been suggested by some workers but do not seem justified given that regional breeding season control programs are effective in eliminating nearly all cowbirds from the ranges of endangered hosts. Such large-scale killing programs may also raise important ethical issues and some workers have argued that an undue emphasis on cowbirds may detract from more productive and more long-lasting management actions such as habitat restoration.

## V. ECOLOGY AND SOCIAL BEHAVIOR OF BROOD PARASITES

The lack of parental care in brood parasites potentially sets them apart from most other birds in their foraging ecology, mating systems, spacing behavior, and vocal development.

### A. Foraging Ecology

Most brood parasites have unusual diets and foraging behavior. Honeyguides eat wax, many cuckoos eat hairy and toxic caterpillars, and cowbirds prefer to forage in short grass close to ungulates such as cattle, horses, or bison. In each of these cases, either the diet of the brood parasite would be difficult for nestlings to digest (wax and hairy caterpillars) or the foraging habitat may be so ephemeral (the proximity of ungulates) that it may not be available for an entire nesting cycle. It is not known, however, if these unusual foraging ecologies were precursors of brood parasitism or if they were made possible by the evolution of brood parasitism.

## B. Mating Systems

Mating system theory predicts that birds freed from the needs of parental care should be promiscuous. Nevertheless, avian brood parasites show a remarkable array of mating systems, including monogamy and resource-based polygyny. This variation in mating behavior has been attributed to the diverse ways in which parasites gain access to nests (some of which require cooperation between several individuals), to the great variety of foraging ecologies of brood parasites, and to such unusual features of individual host-parasite systems as strongly male-biased sex ratios in cowbirds. In general, there have been few detailed studies of the mating systems of brood parasites. The mating system of the brown-headed cowbird has received the most study among parasitic birds and has been described as promiscuous, polygynous, or monogamous. Promiscuity has been proposed because it is a common sight to see two or more males associating with a single female. However, studies of color-marked cowbirds that enabled researchers to determine which males and females mated together and associated with one another the most have indicated monogamy. A recent DNA fingerprinting study confirmed that monogamy prevails and even found that cowbirds have fewer matings outside of their pair bonds than do most nonparasitic songbirds. Field observations show that the mate faithfulness in cowbirds is due both to males guarding females from the advances of other males and to females being reluctant to mate with males other than their usual consort. Because males outnumber females, many males do not acquire a mate. A single study demonstrated promiscuity in an area in which extremely high cowbird abundances may have made it difficult for the birds to maintain pair bonds.

In one Asian honeyguide, males defend beehives and only allow access to wax if the females mate with them. In some viduine weavers and cuckoos, males display from prominent perches and are chosen by females presumably on the basis of mate quality, which may be indicated by song elements or plumage.

## C. Spacing Behavior

Many brood parasites defend breeding areas that are rich in hosts. At least some cowbirds and many cuckoos defend home ranges against conspecifics. However, the home ranges are technically not true territories because several males and females often occupy the same area.

In areas of home range overlap, some host nests are often parasitized by several female cowbirds. Cuckoos may occupy areas that are more mutually exclusive and thus this is closer to classic territoriality. In areas of low abundance, cowbirds may occupy mutually exclusive areas and thus may appear to be territorial.

Cowbirds often breed and feed on a daily basis in different areas that can be separated by as much as 15 km. After searching for nests in the morning, cowbirds commute to feeding sites in pastures, plowed fields, and in other habitats with short mowed grass. The uncoupling of breeding and feeding areas made possible by brood parasitism allows cowbirds to select a wide variety of breeding habitats, even if there are only a few foraging sites in a region. Some cuckoos also appear to have very large home ranges. Home range size is related to a bird's body size and trophic level, with predators needing especially large areas. Although cowbirds and cuckoos are not predators of other birds, adult breeding birds are a key resource for them, and they have home ranges similar in size to those of raptors that feed on adult birds.

## D. Vocal Behavior

Brood parasites have provided some of the clearest examples of genetically hard-wired vocal behavior, but recent studies have also shown a key role for subsequent learning in the modification of vocalizations. Because cowbirds are raised by different species, it is not surprising that at least one of their songs, the perched song, is genetically programmed. Cowbirds, however, learn to modify this song in response to female preferences and interactions with other males and develop individual repertoires of approximately five perched song types. A second song type, the flight whistle, is almost totally learned. It occurs as discrete spatial dialects generally tens of kilometers in diameter, which are so variable that trained observers may fail to recognize as cowbird vocalizations whistles from dialects they have not yet experienced. These dialects are true examples of culture and show that considerable amounts of biodiversity within a species can be related to learned/cultural differences rather than to genetic differences. Unlike the majority of songbirds that complete their vocal development by the time they are 1 year old, male cowbirds in some regions do not master local versions of whistles and perched songs until they are 2 years old.

Vocal learning occurs among one other parasitic group. Male viduine weavers incorporate elements of

their host's song into their repertoire, which presumably enables females to choose a male raised by the same host species. As such, viduine weavers show a high potential for sympatric speciation by cultural learning. Such assortative mating maintains coevolutionary adaptations such as the gape mimicry that is specific to just one host. Assortative mating is essential because gape appearance is due to paternal and maternal genes. In contrast, common cuckoo and cowbird males do not incorporate host songs into their repertoires, perhaps because egg appearance, which is essential for acceptance by their hosts, is determined solely by the maternal genotype.

## VI. CONSPECIFIC BROOD PARASITISM

Female birds sometimes lay their eggs in the nests of conspecifics. This behavior has been documented in more than 100 bird species and has also been documented in some insects and fish. Among birds, conspecific brood parasitism (CBP) is most prevalent among precocial species, such as ducks and gallinaceous birds, and among altricial species that breed colonially and/or use specialized nest sites, such as cavities. CBP is rare or absent in most songbirds, even some that are colonial, and it has been well studied in swallows.

In many cases, CBP occurs when individuals lose their nests during the laying period or are unable to obtain a territory or nest site. In these situations, individuals may be making the best of a bad situation. In some species, CBP may be an alternative reproductive strategy that increases reproductive success by spreading the risk among several nests, reducing competition within a female's own nest, and exploiting the parental care of others. Recent theoretical work suggests that CBP is a necessary precursor to the evolution of interspecific brood parasitism. Conspecific brood parasitism is especially prominent in New World cuckoos of the genus *Coccyzus*, which tend their own nests and also parasitize conspecifics and congeners. In at least some species, such as the ostrich and some ducks, individuals may benefit from CBP because extra eggs or young dilute losses to predators.

Nest defenses against CBP parallel those used to combat interspecific brood parasitism. Hosts guard their nests, especially at high population densities in barn swallows, which may exert a significant cost of nesting colonially. Parasitic females of another swallow species, the cliff swallow, actually carry their eggs between nests.

The rejection of parasitic eggs laid before the host has begun laying is a common adaptation against CBP but does not necessarily involve egg recognition. True egg recognition, in which a bird discriminates among egg types, is a rare response to CBP because it is difficult to evolve because parasitic and host eggs are similar in appearance, given that they are from the same species. Also, CBP is less deleterious to hosts than is interspecific parasitism because it does not involve adaptations for killing host nestlings or asymmetries in size and incubation period that accomplish the same end. However, egg recognition in response to CBP does occur in some birds such as ploceid weaver finches of Africa and Asia, which have extensive variation among the eggs of conspecific females. Indeed, it has been hypothesized that the extreme intraspecific egg variation of these weavers is an evolutionary response to CBP. Some species with typical amounts of variation in egg appearance may lay smaller clutches than they could potentially feed to leave room for parasitic eggs. If hosts cannot recognize parasitic eggs, then laying a smaller clutch might avoid starvation of most or all nestlings when parasitism does occur.

CBP may actually affect population sizes. In species with frequent CBP, fitness may be reduced at high population densities, which may cause populations to be cyclic. Nests of wood ducks that are placed too close together are subject to extreme levels of CBP, which can reduce population fitness.

## VII. CONCLUSIONS AND RESEARCH NEEDS

Brood parasitism raises fascinating questions about coevolution and conservation. Studies of brood parasitism provide some of the strongest evidence of microevolution yet documented in vertebrates. The extreme mobility of cowbirds and their potential threats to host populations illustrate the importance of landscape-level processes in conservation biology. Cowbird song development has become a model system showing that both genetic and environmental factors are important to the development of behavior in general. Despite these and other lessons learned from studies of brood parasitism, there are many unanswered questions that are vital to our understanding of this subject.

We still have much to learn about topics such as (i) the frequency of recognition errors in hosts with



rejection behavior, (ii) reasons for the general lack of parasite nestling recognition by hosts, (iii) the frequency of “mafia”-like behaviors and nest predation by brood parasites, (iv) how brood parasites choose from available hosts, (v) whether or not cowbirds pose a significant threat to populations of widespread host species, and (vi) the genetic mating systems of brood parasitism. The recent invasion of the United States by a new generalist brood parasite, the shiny cowbird, also offers an excellent opportunity to study its interactions with other cowbirds and its new hosts. Finally, most host–parasite systems remain very poorly known, especially those of the New World cuckoos, many Old World cuckoos, and some old honeyguides. Some members of the two latter groups are so poorly known that there is no direct evidence that they are parasitic. However, they are assumed to be so because closely related species are known to be parasitic.

### See Also the Following Articles

BIRDS, BIODIVERSITY OF • COEVOLUTION • PARASITISM • POPULATION DYNAMICS • SPECIES COEXISTENCE

### Bibliography

- Davies, N. B., Bourne, A. F. G., and de L. Brooke, M. (1989). Cuckoos and parasitic ants: Interspecific brood parasitism as an evolutionary arms race. *Trends Ecol. Evol.* 8, 2–4.
- Johnsgard, P. A. (1997). *The Avian Brood Parasites: Deception at the Nest*. Oxford Univ. Press, New York.
- Ortega, C. P. (1998). *Cowbirds and Other Brood Parasites*. Univ. of Arizona Press, Tucson.
- Payne, R. B. (1997). Avian brood parasitism. In *Host–Parasite Coevolution: General Principles and Avian Models* (D. H. Clayton and J. Moore, Eds.), pp. 338–369. Oxford Univ. Press, New York.
- Rothstein, S. I., and Robinson, S. K. (1998). *Parasitic Birds and Their Hosts: Studies in Coevolution, Oxford Ornithology Series*. Oxford Univ. Press, Oxford.



# NITROGEN, NITROGEN CYCLE

Sandy L. Tartowski and Robert W. Howarth  
*Cornell University*

---

- I. An Overview of the Nitrogen Cycle and Its Ecological Importance
  - II. Human Impacts on the Nitrogen Cycle
  - III. Effects of Nitrogen on Biodiversity
  - IV. Effects of Biodiversity on Nitrogen Dynamics
  - V. Considerations for the Future
- 

## GLOSSARY

**anoxia** The absence of molecular oxygen.

**biogeochemistry** The discipline which studies biotic controls on the chemistry of the environment and geochemical controls on the structure and function of ecosystems.

**denitrification** The reduction of nitrate or nitrite to gaseous nitrogen products, mainly  $N_2$  and  $N_2O$ , by bacteria.

**deposition** The delivery of material inputs to the earth's surface from the atmosphere.

**eutrophic** High in production or species typical of high-production environments.

**mesotrophic** Intermediate in production or species typical of intermediate-production environments.

**mineralization** The conversion of an element from an organically bound form to an inorganic form.

**nitrification** The oxidation of ammonia to nitrite and nitrate by bacteria.

**nitrogen fixation** The reduction of atmospheric  $N_2$  to ammonia or other organic or inorganic compounds by bacteria, lightning, or photochemistry.

**nitrogen saturation** Nitrogen supply in excess of the capacity of the ecosystem to retain nitrogen.

**nutrient limitation** The addition of a nutrient or nutrients to an ecosystem causes an increase in net primary production.

**oligotrophic** Low in production or species typical of low-production environments.

**opportunistic species** Species typical of transient, unstable, unpredictable, frequently disturbed, or periodically extreme environments, usually having strong dispersal abilities and faster growth, smaller size, and shorter life spans than potential competitors.

**primary production** Synthesis of organic matter from carbon dioxide by photosynthesis.

---

**NITROGEN IS AN ESSENTIAL ELEMENT FOR LIFE** and it frequently controls productivity and community structure in both terrestrial and aquatic ecosystems. During the past few decades, human activity has changed the nitrogen cycle more than any other element cycle, resulting in many deleterious environmental changes, including decreased biotic diversity.

## I. AN OVERVIEW OF THE NITROGEN CYCLE AND ITS ECOLOGICAL IMPORTANCE

Nitrogen (N) is an essential element for all life on Earth, a required component of DNA, all proteins (including

enzymes), chlorophyll, and many other critical constituents of biological structure and function. More than 99% of nitrogen on Earth is present as molecular  $N_2$ , with most of this in the atmosphere and some dissolved as a gas in the oceans. However, most organisms cannot use the relatively inert  $N_2$  molecule. Paradoxically, molecular  $N_2$  is not a thermodynamically stable compound in the presence of oxygen, but these two gases coexist and comprise virtually all the atmosphere. The nitrogen in  $N_2$  becomes biologically available only when the N–N bond is broken and nitrogen is combined with other elements to form compounds such as ammonium ( $NH_4^+$ ) and nitrate ( $NO_3^-$ ) in a process called nitrogen fixation. Prior to the massive alteration of the global nitrogen cycle by human activity, small amounts of nitrogen were fixed by lightning and by volcanic activity, but the vast majority of nitrogen fixed on Earth each year (more than 95%) was fixed by bacteria. Not all bacteria are capable of fixing nitrogen, although the ability is fairly widespread among diverse groups of bacteria including cyanobacteria, nonoxygenic photosynthetic bacteria (such as purple sulfur bacteria), and many classes of heterotrophic bacteria. No other organisms are capable of nitrogen fixation, although many plants, lichens, mosses, and some animals have bacterial symbionts which fix nitrogen. These plants with symbiotic nitrogen-fixing bacteria include many species of trees (such as alders in the temperate zone and many species in moist tropical forests) and many agricultural crops, including alfalfa, peas, peanuts, soybeans, and rice. Plants with symbiotic nitrogen-fixing bacteria are often loosely called “nitrogen fixers,” although it is always the bacteria that are actually fixing nitrogen.

For most of the history of life on Earth, the demand by plants for biologically available forms of nitrogen such as nitrate and ammonium has been greater than the rate of supply by nitrogen fixation. Consequently, nitrogen limits the primary productivity in many ecosystems, and when more biologically available nitrogen is added to these systems plant production increases. This relative lack of availability of nitrogen globally has also been a major factor shaping the evolution of life and structuring the communities in many different types of ecosystems, both terrestrial and aquatic.

The general patterns which relate productivity and diversity in ecosystems are well established. In aquatic ecosystems, species diversity declines with increasing productivity, and diversity is greatest in systems of very low productivity, such as oligotrophic lakes and areas of low productivity in the oceans in which colimitation by multiple nutrients (including nitrogen) may be common. In terrestrial ecosystems, the greatest diversity of

plant species usually occurs in sites of intermediate fertility and productivity. In sites of very low fertility, the resources can support only a few viable populations able to tolerate the severe conditions. In very high fertility sites, competition intensifies and may shift from competition for nutrients to the more asymmetrical competition for light or space, resulting in dominance by a few species. For both aquatic and terrestrial systems, when nitrogen is limiting production, it is also influencing diversity.

The importance of nitrogen as a factor limiting productivity and regulating community structure varies among the earth's ecosystems, with nitrogen being of paramount importance in some systems and not limiting at all in others. Although there are exceptions, among terrestrial ecosystems nitrogen availability generally controls production in temperate forests, boreal forests, arctic and alpine tundra systems, temperate and tropical grasslands, and agroecosystems. On the other hand, phosphorus or other elements—and not nitrogen—tend to limit production in most tropical forests if any element is limiting. Among aquatic ecosystems, nitrogen limits production in most estuaries and coastal seas of the temperate zone and perhaps in some tropical seas, although these are sometimes phosphorus limited. Nitrogen with phosphorus as a colimiting element may regulate production in much of the oceans away from shore as well. However, phosphorus is generally more important in regulating production in freshwater lakes, at least in moderately or highly productive lakes in the temperate zone.

These patterns of nutrient limitation are the result of many processes which regulate the availability of nitrogen in comparison to other elements essential for plant growth. For terrestrial ecosystems, one important aspect is the change in phosphorus biogeochemistry as soils develop over geological time. In a geologically young soil, the initial amount of phosphorus is often high, but as the soil is weathered this phosphorus is exported from the system or stored in mineral forms which make the phosphorus biologically unavailable. In the temperate zone, soils are often young due to glaciation within approximately the past 20,000 years. On the other hand, many tropical soils are extremely old and highly weathered and therefore have very low availabilities of phosphorus. This is one factor which tends to make tropical forests phosphorus limited and temperate forests nitrogen limited.

Differences in nitrogen processes among systems are also important in determining which element is most limiting to primary productivity. Nitrogen fixation is one critically important process, and it is much more

prevalent and occurs at higher rates in tropical forests than in temperate forests. This contributes to the greater tendency for nitrogen limitation in temperate forests. An interesting paradox is that nitrogen fixation does not alleviate nitrogen shortages in temperate forests and thereby alleviate nitrogen limitation. Why should any ecosystem be nitrogen limited if nitrogen fixation could provide sufficient inputs? Why is nitrogen fixation in temperate forests so much less than that in tropical forests? The reasons for the relatively low rates of nitrogen input to nitrogen-limited temperate forests remain poorly known and are an active area of research. A similar paradox occurs in aquatic ecosystems. One of the reasons that freshwater lakes tend to be limited by phosphorus is that planktonic cyanobacteria in these systems can often fix sufficient nitrogen to alleviate any potential nitrogen limitation. In contrast, nitrogen fixation by planktonic cyanobacteria rarely occurs in even strongly nitrogen-limited estuaries. Recent research suggests that the relative lack of nitrogen fixation in estuaries is due to an interaction of two factors: Trace metals that are required for the process of nitrogen fixation (molybdenum and iron) are less available in estuaries than in lakes, slowing the potential growth rate of nitrogen-fixing organisms, and this slow growth rate makes cyanobacteria highly susceptible to grazing mortality by zooplankton and benthic animals.

Nitrogen fixation is only one process affecting nitrogen availability in an ecosystem. Also important are the processes which regulate the loss of nitrogen from the system and the input of biologically available forms of nitrogen from neighboring ecosystems and from the atmosphere. A brief summary of the nitrogen cycle illustrates the complexity. The nitrogen that is fixed accumulates over time in both the living and nonliving components of ecosystems and is transported and transformed by numerous processes (Fig. 1). Ammonium and nitrate are taken up by plants as well as by microorganisms and converted to organic nitrogen compounds. This organic nitrogen is cycled through food webs but also accumulates in soils, water, and sediments as organisms die, excrete matter, or drop tissues such as the fall of leaves from trees. Much of the nitrogen in this nonliving organic matter is not readily available for use by organisms, but microbial decomposition of the organic matter releases ammonium, which can then be oxidized to nitrate by bacteria. Both nitrate and ammonium are highly soluble in water, but the negatively charged nitrate ion is more easily leached from soil and transported to aquatic ecosystems. In ecosystems where oxygen concentrations are low (such as in the sediments of wetlands, lakes, and estuaries), the nitro-

gen in nitrate can be converted back to molecular  $N_2$  (or the slightly oxidized gas  $N_2O$ ) by bacteria in a process called denitrification. Also, in soils and waters where the pH is sufficiently high to favor the dissociation of ammonium to ammonia ( $NH_3$ ), nitrogen can be lost to the atmosphere as ammonia gas. Fires can also volatilize both ammonia and oxidized gases of nitrogen (such as  $NO$  and  $N_2O$ ) to the atmosphere. Most of these forms of nitrogen are redeposited from the atmosphere onto the earth's surface in precipitation or as dry deposition.

## II. HUMAN IMPACTS ON THE NITROGEN CYCLE

Human activity has altered the nitrogen cycle globally more than that of any other element cycle, and much of the change has occurred during the past 50 years. The increase of carbon dioxide and other greenhouse gases in the atmosphere is, appropriately, a cause of widespread alarm, but the rate at which humans have altered nitrogen availability on land is even greater. As recently as 30–40 years ago, the major source of fixed nitrogen was still natural biological nitrogen fixation, but today more nitrogen is fixed through human activities than by all nitrogen fixation in natural terrestrial ecosystems on Earth (Fig. 2). The greatest change has been the widespread use of inorganic nitrogen fertilizer, and the production of such fertilizer today accounts for more than half of the total amount of nitrogen fixed by all human activities. The process for making nitrogen fertilizer from atmospheric  $N_2$  was invented in the early twentieth century, but such fertilizer was not widely used until the 1950s. The rate of use has increased steadily since then, with only a modest interruption caused by the collapse of the former Soviet Union and the disruption of agriculture in Russia and eastern Europe in the early 1990s. Despite that interruption, half of the inorganic nitrogen fertilizer that has ever been used on the planet has been used during the past 15 years.

The synthesis of inorganic fertilizer accounts for approximately 60% of the total amount of nitrogen fixed globally through human activities, with the rest coming primarily from the production of nitrogen-fixing crops in agriculture and from the combustion of fossil fuels. Fossil fuel combustion not only releases nitrogen from geological storage in the fuel but also catalyzes the reaction of  $O_2$  and  $N_2$  in the air drawn into internal combustion engines. This nitrogen is released as oxi-

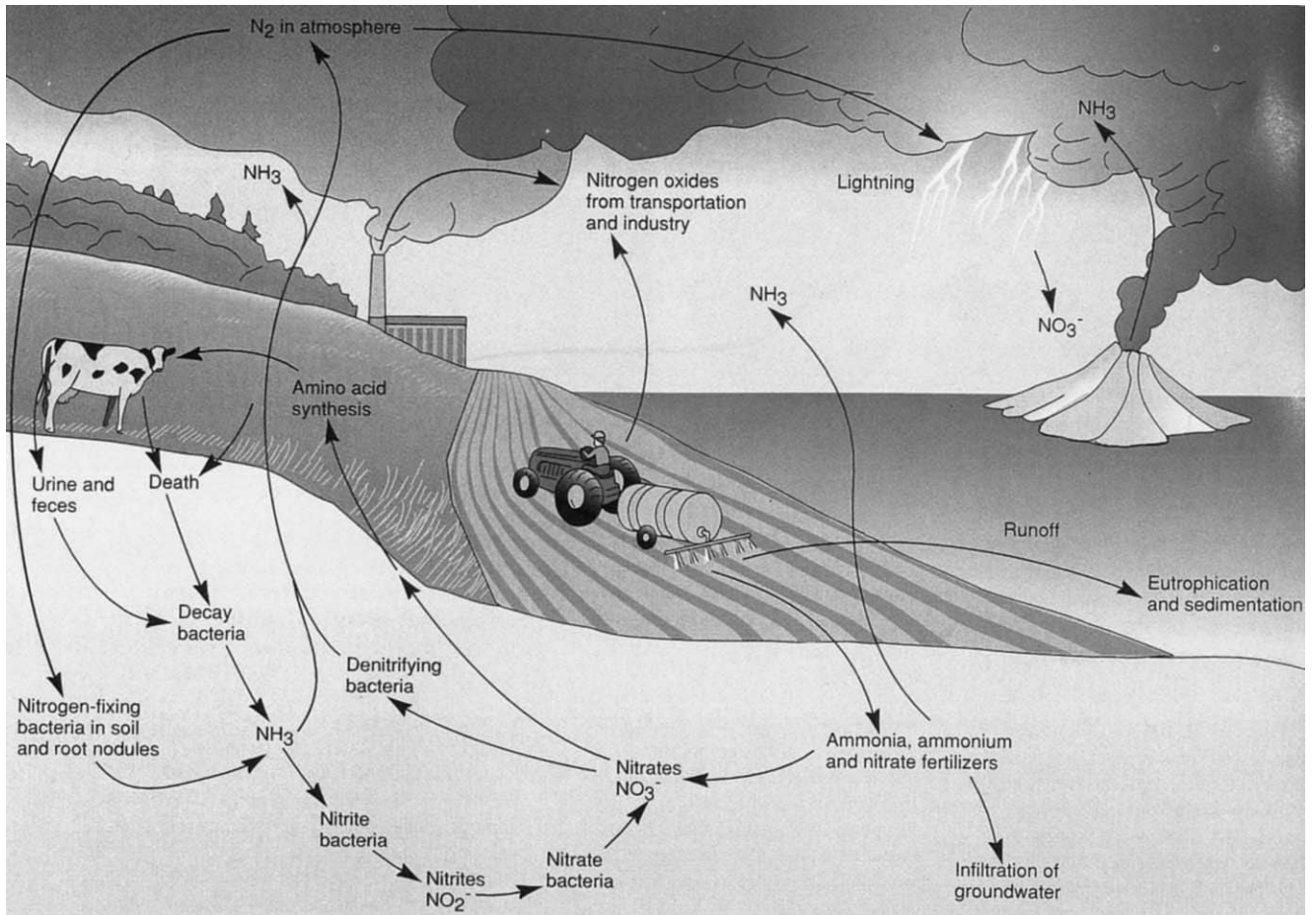


FIGURE 1 Simplified diagram of the nitrogen cycle (reproduced with permission from Cunningham and Saigo, 1995).

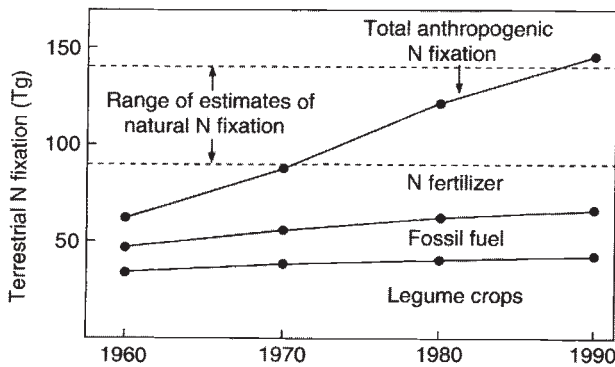


FIGURE 2 Anthropogenic fixation of nitrogen compared to the range of estimates of natural biological fixation on land (reproduced with permission from Vitousek et al., 1997).

dized nitrogen gases, generically referred to as  $NO_x$ , which are subsequently deposited onto the earth's surface in precipitation or dry deposition. Overall, human fixation of nitrogen (including production of fertilizer, combustion of fossil fuel, and production of nitrogen-fixing agricultural crops) increased globally approximately two- or threefold between 1960 and 1990 and continues to increase. By the mid-1990s, human activities made fixed nitrogen available at a rate of approximately  $140 \text{ Tg year}^{-1}$ , or more than  $0.8 \text{ g N m}^{-2} \text{ year}^{-1}$  when averaged over the entire land surface of the earth.

Nitrogen from human sources can travel for fairly long distances in the environment. For instance, much of the nitrogen deposited from the atmosphere in the northeastern United States comes from the combustion of fossil fuels in the Midwest, and the majority of nitrogen flowing down the Mississippi River and causing eutrophication in the Gulf of Mexico originates from agricultural sources in Illinois, Iowa, Indiana, Minne-

sota, and Wisconsin. Of the nitrogen fertilizer applied to agricultural fields, some leaches into groundwater and surface water and affects downstream ecosystems; on average in the United States, approximately 20% of nitrogen fertilizer is thus exported from agroecosystems, although this percentage ranges from only 3 to 80% depending on the soil type, climate, and agricultural practices. Nitrogen from agricultural sources also gets volatilized into the atmosphere and thereby redistributed onto nonagricultural lands, including forests. Some of this nitrogen is volatilized directly from the agricultural fields, as both ammonia and  $\text{NO}_x$ , with the flux of  $\text{NO}_x$  being particularly important in tropical areas. Also of importance is the volatilization of ammonia to the atmosphere from animal wastes. In developed countries, approximately half of the nitrogen used as fertilizer is exported from the field in crop harvest, on average, with most of these crops being fed to livestock and poultry. Much of the nitrogen in these feedstocks ends up in the animal wastes, with a high percentage being volatilized. The total flux of such volatilization in the United States is almost as large as the leaching of nitrogen directly from fertilized fields.

Reactive nitrogen compounds are usually transported for only a few hundred kilometers to at most approximately 1000 km through the atmosphere, and rivers and ocean currents transport biologically available forms of nitrogen on the same sort of spatial scales. As a result, the alteration of the nitrogen cycle is not

uniform over the earth, and the greatest changes are concentrated in the areas of greatest population density and greatest agricultural production. Generally, the largest changes have occurred in the northern temperate zone. The tropics have seen less change, at least to date. Most terrestrial ecosystems and most coastal marine ecosystems in the temperate zone are nitrogen limited, and therefore the acceleration of nitrogen cycling in the northern temperate zone has had a great impact. Current deposition of nitrogen from the atmosphere onto the landmasses of the Northern Hemisphere temperate zone is almost seven times more than the preindustrial deposition, and in some regions, such as the northeastern United States, Western Europe, and east Asia, the increase has been far greater. The natural rate of biological nitrogen fixation in European watersheds of the North Sea and in the northeastern United States is approximately  $0.5 \text{ g N m}^{-2} \text{ year}^{-1}$  or less and human activity has increased nitrogen input to the watersheds of the North Sea to  $7.5 \text{ g N m}^{-2} \text{ year}^{-1}$  and in the northeastern United States to  $4.1 \text{ g N m}^{-2} \text{ year}^{-1}$ . The export of nitrogen in major rivers to coastal oceans is closely related to the increased inputs of nitrogen to the basins from human activities (Fig. 3). Downstream transport of nitrogen to estuaries and coastal oceans has increased up to 20-fold in some areas such as the North Sea.

The rate at which human activity has accelerated nitrogen cycling varies widely among regions of the

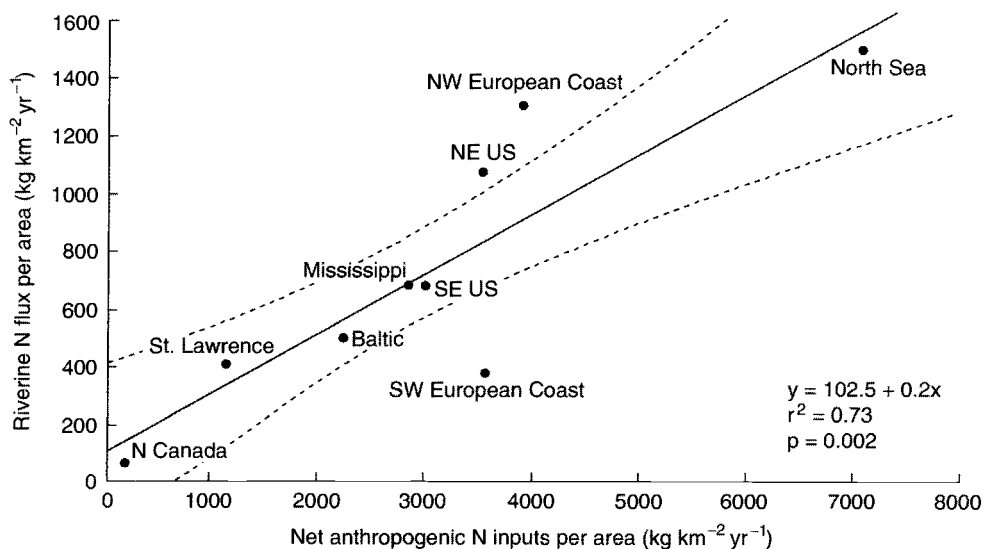


FIGURE 3 Export of nitrogen from rivers emptying into the North Atlantic Ocean is strongly correlated with the net nitrogen inputs to the watersheds caused by human activities, especially agriculture and combustion of fossil fuels (reproduced with permission from Howarth *et al.*, 1996).

world. Globally, the use of inorganic nitrogen fertilizer continues to increase, but in the United States and most other developed countries fertilizer use has changed little since 1980. In China, on the other hand, the use of inorganic nitrogen fertilizer has more than doubled since 1980. By 1995, China was using almost one-third of all inorganic nitrogen fertilizer used globally, and the rate of use there continues to increase. The acceleration of nitrogen use in developing countries has allowed major increases in food production and has greatly reduced starvation, but the environmental consequences are becoming increasingly apparent.

### III. EFFECTS OF NITROGEN ON BIODIVERSITY

Human acceleration of the nitrogen cycle has tremendous implications for biodiversity at all scales, from the genome to the globe. Reduced genetic diversity within species exposed to high levels of nitrogen deposition has been documented in tree species, soil mycorrhizae, and *Rhizobium spp.* in beans. Changes in species composition and reductions in species diversity are obvious and widespread. Declining ecosystem and landscape diversity are becoming evident. In return, changes in biodiversity influence nitrogen dynamics and processes, creating complex nonlinear interactions and feedbacks between nitrogen cycling and biodiversity. Nitrogen enrichment can reduce biodiversity by several mechanisms including acidifying soil and water as well as increasing the growth and dominance of a few particularly responsive species in nitrogen-limited ecosystems.

#### A. Effects of Eutrophication on Biodiversity

The substantial increase in nitrogen inputs to nitrogen-limited ecosystems as a result of human activities alters species composition (which species are present) and decreases species diversity (the number of species and the evenness of their relative abundance). A few fast-growing opportunistic species, with greater ability to take advantage of the increased availability of nitrogen to produce additional biomass, usually become dominant in fertilized ecosystems, while the species characteristic of less fertile environments (oligotrophic species) decline and disappear. Often, a relatively diverse flora, including rare endemic plants, characteristic of an oligotrophic ecosystem is replaced by a more productive, but less diverse, flora dominated by a few common

weedy species, normally found in mesotrophic or eutrophic sites. Usually, invasive pest plants respond strongly to increased availability of nitrogen, and generally eutrophication facilitates the spread of introduced and native pest plants. As plant species are lost from an ecosystem, additional animal species associated with those plants may be lost as well, further decreasing species diversity. The conversion of whole ecosystems from oligotrophic to eutrophic and the disappearance of entire types of ecosystems encompass the extinction of large groups of species, not just a few particularly vulnerable endangered species, and reduces biodiversity on the local, landscape, and regional scale.

The decrease in species diversity caused by the eutrophication of terrestrial and aquatic ecosystems has been observed consistently across a wide variety of nitrogen-limited ecosystems. However, the speed and magnitude of the effects of increased inputs of nitrogen on biodiversity depend on the characteristics and history of the affected ecosystem. The immobilization and redistribution of nitrogen by human activity can cause local areas of nitrogen depletion, even though the amount of fixed nitrogen has increased overall. Semiarid and arid lands and unfertilized agricultural and overgrazed systems are particularly susceptible to the loss of nitrogen-containing topsoil. In some cases, nitrogen depletion can threaten biodiversity much as nitrogen addition reduces biodiversity. In very unproductive ecosystems, the addition of nitrogen may cause an initial, usually transient, increase in species diversity by allowing the invasion of native species from more fertile sites in the same region or exotic species introduced from other regions. Even when species diversity remains elevated, often the oligotrophic species are lost from the ecosystem, contributing to a regional decline in species diversity. The effects of nitrogen enrichment may be delayed in systems which are severely nitrogen limited due to previous depletion of nitrogen by fire or human activities, such as forest clearing or timber harvesting.

Enrichment of grasslands with nitrogen causes productivity increases in a few dominant grasses and severe declines in species diversity, as shown by experiments in North America, Australia, and Europe (Fig. 4). In one long-term study of fertilizer addition to a grassland in England, species diversity was reduced by more than fivefold. Oligotrophic, calcareous grasslands in Europe, with high biodiversity and many endemic species, are particularly sensitive to species loss due to eutrophication and most sites have already experienced changes in species composition. Fertilization and increased productivity decrease species diversity by allowing fast-growing, tall plants to shade competitors and sometimes

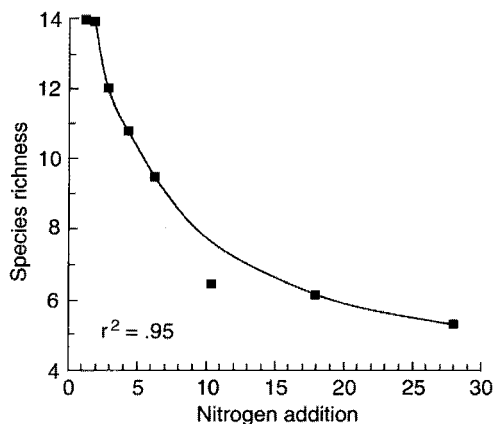


FIGURE 4 Average number of plant species present after 4 years of nitrogen additions ( $\text{g}/\text{m}^2/\text{year}$ ) to grassland plots in Cedar Creek, MN (reproduced with permission from Huston, 1997; Tilman, 1996).

by increasing production of litter which prevents the establishment of seedlings. Increased availability of nitrogen suppresses nitrogen-fixing plant species and consequently many of the forbs lost from eutrophied grasslands and heaths are nitrogen-fixing species. The mechanisms by which nitrogen enrichment reduces biodiversity may be quite specific to the circumstances of each ecosystem. For example, when nitrogen was added to fertile wet and infertile dry sedge meadows in Colorado, diversity declined in the wet meadow but not in the dry meadow. In the infertile dry meadow, nitrogen additions released rare species from nitrogen limitation, increasing the evenness of species abundances, but in the more fertile wet meadow the already dominant species increased in abundance, suppressing subordinate species. Nitrogen inputs affect not only plant diversity but also the diversity of the animal community. It has recently been suggested that nitrogen deposition near San Francisco may be a contributing factor leading to the decline of a threatened, endemic butterfly.

In many ecosystems, moderate levels of grazing—within the range typical of the evolutionary history of the system and plants—tend to increase plant species diversity. Grazing can interact with nitrogen availability to affect diversity, and ungrazed or lightly grazed grasslands are more susceptible to decreases in biodiversity caused by nitrogen enrichment because heavy grazing usually suppresses the dominant grass species stimulated by nitrogen. Nitrogen enrichment usually reduces the spatial heterogeneity of nitrogen and therefore reduces the potential for different species to coexist using microsites with different availabilities of nitrogen. Graz-

ing redistributes nitrogen and usually increases the spatial heterogeneity and temporal variability of nitrogen availability, contributing to the maintenance of biodiversity. However, when fertilization increases productivity sufficiently to support greatly increased grazing, grazing-intolerant plant species may be lost from the ecosystem.

Diversity in forests is also affected by nitrogen additions, as has been particularly well studied in parts of Europe in which nitrogen deposition from the atmosphere is high. In both deciduous (the Netherlands and France) and coniferous forest (Sweden and Finland), nitrogen enrichment has caused shifts in the species composition of the understory, favoring opportunistic species with high nitrogen requirements and reducing biodiversity of lichens, mosses, and vascular plants. Increased nitrogen deposition has caused weakened tree trunks, decreased frost hardiness, and increased pest outbreaks in some forests. After many years of high nitrogen deposition, some forests of Western Europe have become “nitrogen saturated,” unable to store additional nitrogen in soil or vegetation. In nitrogen-saturated ecosystems, continuing nitrogen inputs are balanced by comparable nitrogen outputs, accelerating both the eutrophication of downstream waters and the loss of other essential elements from the forest.

In the Netherlands, the high population density, with accompanying industry and vehicles, is combined with intensive livestock production to generate perhaps the highest rates of nitrogen deposition in the world. More than 35% of oligotrophic, diverse heathlands in the Netherlands have suffered nitrogen accumulation, invasion by grasses, and reduced diversity. In dry heathlands, this nitrogen fertilization has stimulated outbreaks of the heather beetle, accelerating the conversion to grassland. Some heaths have been completely converted to much less diverse grasslands and forest, reducing the landscape and regional diversity as well as the diversity within the heathlands. The beginnings of similar eutrophication of heathlands have been observed in Norway and in the United Kingdom.

Estuaries and other coastal waters receive a variety of insults from human activities. Perhaps the greatest pollution problem, however, is the eutrophication that results from increased nitrogen inputs. Eutrophication of coastal marine ecosystems frequently results in hypoxia (the depletion of dissolved oxygen) or anoxia (the absence of dissolved oxygen), both of which result not only in fish kills and the loss of other biotic resources but also in sharp decreases in biotic diversity. Hypoxia has severe effects on the diversity and abundance of benthic species, shifting dominance from large, long-



lived, less mobile species to smaller, mobile, opportunistic, short-lived species. Frequent hypoxia may prevent successional development beyond the early colonizing community. The loss of burrowing benthic organisms that irrigate and oxygenate the sediments may strongly influence biogeochemical processes such as phosphorus adsorption and nitrification and denitrification, which can feed back to influence diversity. Increased nitrogen inputs have caused increased hypoxia and anoxia in Chesapeake Bay, the Baltic Sea, the Black Sea, Long Island Sound, Florida Bay and the Florida Keys, the northern Adriatic Sea, and many other areas globally. The hypoxic “dead zone” in the northern Gulf of Mexico increased from 9500 km<sup>2</sup> when first reported in 1991 to 20,000 km<sup>2</sup> in 1999 and is the direct result of nitrogen inputs coming down the Mississippi River.

Not all the effects of eutrophication on biodiversity are due to hypoxia and anoxia. Nutrient enrichment also leads directly to changes in the composition of the phytoplankton community, and these can cascade up the food web generally leading to lowered diversity at each trophic level. Eutrophication of coastal waters is often accompanied by a decrease in silica availability, due both to increased sedimentation of silica within estuaries and to trapping of silica in upstream, eutrophic fresh waters. Since silica is required by diatoms, but not by other types of phytoplankton, eutrophication often results in a relative loss of diatoms from the community. A 4-fold increase in the relative inputs of nitrogen compared to silica along the German coast during the past few decades was accompanied by a 10-fold decrease in diatoms and a comparable increase in flagellates, which are less conducive to supporting food webs leading to commercially valuable fisheries. Decreased silica availability and the concomitant loss of diatoms may also be responsible, at least in part, for the increased frequency of occurrence and duration of blooms of harmful algae that seem to accompany coastal eutrophication. Harmful algal blooms (“red tides,” “brown tides,” and “green tides”) can kill animals, such as sea lions, far up the food chain. The resulting loss of important top predators such as sea otters can cascade back down the food chain, altering species composition and reducing biodiversity.

Coastal systems that are both high in biodiversity and severely affected by eutrophication are seagrass beds and coral reefs. Temperate seagrass beds are usually nitrogen limited, whereas tropical seagrass beds and coral reefs are often phosphorus limited. However, even these tropical systems probably become nitrogen limited once eutrophication begins, and therefore additional inputs of nitrogen can cause immense harm. For

the seagrasses, eutrophication leads to an immediate loss in diversity and can lead to a complete loss of the grasses due to shading by the increased biomass of phytoplankton in the water column, shading by mats of opportunistic species of macroalgae, or an accumulation of toxic decomposition products in the sediments. Eutrophic shallow estuaries and lagoons with large accumulations of macroalgae may experience frequent episodic oxygen depletion, further reducing biodiversity. Seagrasses provide food and shelter for a rich and diverse fauna, and reduced seagrass depth distribution or replacement by macroalgal blooms can result in marked declines in the abundance and biodiversity of the associated fauna.

Coral reefs occur in the oligotrophic shallow waters of the tropics, support extraordinary biodiversity, and are extremely sensitive to damage from eutrophication. The biodiversity of coral reefs declines dramatically with eutrophication from nitrogen inputs, especially when species-poor turf algae or macroalgae communities overgrow and replace corals and coralline algae. Nutrient enrichment disrupts the coral-zooxanthellae symbiosis, inhibiting calcification and contributing to increased coral “bleaching” (loss of zooxanthellae). Increased phytoplankton biomass and production increase turbidity and sedimentation. The decreased quality and quantity of light penetrating to the corals slow growth and reduce the maximum depth at which the corals can survive, reducing the available habitat. Algal blooms can cause hypoxia, decreasing the biodiversity of reef organisms through mortality and reduced habitat quality. Also, eutrophication can damage reefs indirectly by increasing predation on corals, facilitating the expansion of opportunistic filter feeders, and reducing the recruitment of corals.

## B. Effects of Acidification by Nitrogen on Biodiversity

A variety of oxidized nitrogen compounds (NO<sub>x</sub>) are released into the atmosphere in the exhaust of vehicles, electric power plants, and other fossil fuel combustion. Oxides of nitrogen are also released from soils and burning vegetation. In the presence of sunlight, NO<sub>x</sub> catalyzes the formation of photochemical (or brown) smog and reacts with oxygen and hydrocarbons from automobile exhausts to form ozone, an air pollutant damaging to many plant species. In the atmosphere, NO<sub>x</sub> reacts to form nitric acid, one of two major components of acid rain (the other is sulfuric acid).

Approximately 70% of global ammonia (NH<sub>3</sub>) emissions are caused by humans, mostly by volatilization

from fertilized fields, animal wastes, and forest burning. In the atmosphere,  $\text{NH}_3$  neutralizes some of the acidity in aerosols, cloud water, and precipitation. However, once deposited on the earth's surface, ammonium ( $\text{NH}_4^+$ ) is taken up by plants or converted to nitrite ( $\text{NO}_2^-$ ) and then nitrate ( $\text{NO}_3^-$ ) by bacteria. These processes of biological uptake and nitrification release hydrogen ions, acidifying the soil and downstream aquatic systems. Other processes, such as increased biological nitrogen fixation or increased rates of decomposition, which increase the amount of ammonium in the soil, also contribute to acidification.

As nitrogen inputs to ecosystems increase and the capacity of the ecosystems to retain nitrogen is lessened, ecosystems begin to release increasing amounts of  $\text{NO}_3^-$ . When the negatively charged ion leaves the ecosystem, it carries with it positively charged base cations, such as  $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$ , and  $\text{K}^+$ . These elements are essential plant nutrients and their removal can deplete soil fertility and cause serious nutrient imbalances which are detrimental to plants. As the other soil cations are depleted, toxic  $\text{Al}^{3+}$  is mobilized. Soils with low capacity to neutralize, or buffer, the increased acidity from nitrogen and acid deposition are especially sensitive to acidification. Increased acidification may be the major impact of increased nitrogen deposition in ecosystems which are not nitrogen limited, such as many tropical forests and most temperate lakes, or are already nitrogen saturated, such as some temperate forests.

Until recently, most of the impacts of humans on nitrogen dynamics were focused in the temperate zone, but industrial and agricultural practices are changing rapidly in the tropics, where most of the earth's biodiversity resides and for which our ecological knowledge is rudimentary. By 2020, approximately two-thirds of the global application of industrial nitrogen fertilizer and energy-related nitrogen inputs will occur in the tropics and subtropics. The majority of tropical forests are probably not nitrogen limited and nitrogen enrichment will likely have little direct effect on plant production but instead will substantially increase the output of dissolved and gaseous nitrogen from these ecosystems. The acid soils of the tropics may be particularly susceptible to further acidification, depletion of base cations, decreased availability of plant nutrients, and release of toxic  $\text{Al}^{3+}$ , all possibly reducing productivity and biodiversity.

The effects of acidification on sensitive terrestrial and aquatic ecosystems are clear and severe. Increasing acidity of thousands of lakes and streams and the loss of fish and amphibian populations have been documented in Scandinavia and eastern North America. Ex-

perimental acidification of an entire Canadian lake eliminated shrimp, minnow, and crayfish species, all important food sources for the lake trout, which subsequently ceased reproduction. Increased nitrogen deposition has been implicated in the dieback of poorly buffered high-elevation forests in Europe and damage to high-elevation red spruce and fir forests in the northeastern United States. Acid deposition causes direct damage to trees by leaching base cations from leaves. Exposure to acid fog and cloud water, usually much more acidic than rain, increases sensitivity to frost damage. Acidification of forest soils causes nutrient imbalances, cation loss, aluminum toxicity, reduced growth rates, and increased mortality in trees. The diversity of lichens is greatly reduced by acidic deposition and the species composition of vascular plants, soil animals, and microbes shifts toward a less diverse community of acid-tolerant species. Acidification of grasslands and heathlands is widespread in Europe and results in decreased plant diversity when acid-tolerant species replace the species typical of less acid soils.

Terrestrial and aquatic ecosystems subjected to increased inputs of nitrogen often experience both eutrophication and acidification. The combined effects on biodiversity may be independent, additive, or synergistic depending on the characteristics of the ecosystem, but interactions between eutrophication and acidification are common. For example, nitrogen saturation increases the loss of base cations from the soil, accelerating acidification, while acidification strips base cations from plants, increasing the severity of the nutrient imbalances caused by eutrophication.

#### IV. EFFECTS OF BIODIVERSITY ON NITROGEN DYNAMICS

The global biodiversity crisis has prompted increased concern for the possible effects of extinction of native species and introduction of invasive species on ecosystem function and the provision of ecosystem services essential to humans. The effects of biodiversity on nitrogen dynamics have not been studied as thoroughly as the converse effects of nitrogen on biodiversity, but there are numerous examples of significant effects of particular species or functional groups of species on ecosystem processes, including nitrogen dynamics. There are far fewer examples of the effects of species diversity (the number and relative abundance of species), independent of the species composition (specific identity of the species present), on ecosystem processes such as primary production and biogeochemistry.

At larger scales, the effects of ecosystem and landscape diversity on nitrogen dynamics vary with the specific ecosystems and their spatial relationships. Ecosystems which capture or accumulate nitrogen, such as riparian zones bordering farm fields or wetlands upstream of estuaries, can be very important in controlling the transport and downstream impact of nitrogen. Declines in biodiversity associated with the fragmentation and homogenization of landscapes, land clearing, and development may increase the magnitude and variability of exports of nitrogen from the disturbed or converted ecosystems.

Particular species sometimes play unique or dominant roles in ecosystems and the introduction or elimination of these key species can dramatically alter the structure and function of particular ecosystems. For example, the introduction of a nitrogen-fixing tree to a Hawaiian ecosystem which contained no other nitrogen-fixing trees rapidly increased the availability of nitrogen and led to other changes in species composition of the vegetation. Burrowing animals, such as pocket gophers or prairie dogs, increase the rate of mineralization of nitrogen and increase the heterogeneity and availability of nitrogen. In many grasslands, grazers accelerate the rate of nitrogen cycling and influence the distribution and availability of nitrogen as well as the species composition of the plant community. In the boreal forest, in contrast, selective browsing by moose on deciduous trees containing more readily mineralized nitrogen slows nitrogen cycling and speeds succession toward coniferous forest. Addition or removal of whole groups of species that are performing the same functional role ("functional group") has effects similar to removing or adding a single species which is performing a unique function (a functional group with one member). For example, functional groups including nitrogen fixers, leaf-chewing insects in streams, and early spring ephemeral plants of the forest understory all have significant effects on nitrogen dynamics. To date, most of the investigations of the effects of diversity on ecosystem processes have focused on the diversity within groups of primary producers, usually herbaceous plants, but the effects of diversity at other trophic or organizational levels (consumers, predators, microbes, soil invertebrates, or whole ecosystems) may differ from the effects on primary producers. The influence of species or functional group diversity on ecosystem processes may differ for each functional group.

Undoubtedly, the idiosyncratic traits of individual species can influence nitrogen dynamics, but species also share many similar characteristics, compete for some of the same limited resources, and carry out some

of the same ecosystem functions. If species are very similar, then redundancy is high and the removal of a single species would have little impact on ecosystem function. Alternatively, if species are very different (idiosyncratic) and there is little niche overlap (rather, niche differentiation), then the removal of a particular species would have a greater impact on ecosystem function. As the number of species increases, the probability of including an extremely effective dominant species in the more diverse community increases. This simple "sampling effect" is a result of changing species composition which occurs simultaneously with increasing species diversity.

Spatial and temporal variability in resource availability and environmental conditions are characteristic of most ecosystems; therefore, there are many ways in which species might differ in their use of resources and the environment. One species may be capable of symbiotic nitrogen fixation, whereas another can use organic nitrogen and another prefers  $\text{NO}_3^-$  or  $\text{NH}_4^+$ . One species may be better adapted to drier environments or another may flower earlier in the growing season. Theoretically, a greater number of species, with different traits, will make more complete use of the available resources, increasing productivity beyond the maximum of any less diverse mixture of species. Complementarity of species in the use of nitrogen as a limiting resource and facilitation of one species by another as occurs when a nitrogen fixer increases the availability of nitrogen to other species can cause productivity of diverse mixtures of species to be greater than the productivity of any of the single species.

Species diversity is often correlated with other factors, such as productivity, biomass, predation, or resource availability, that complicate the assessment of the effects of species diversity separate from other factors. Only a few studies have examined the effect of species diversity, independent of species composition, on nitrogen cycling. These initial results suggest that greater species diversity among the primary producers more completely uses the available nitrogen and increases plant uptake of nitrogen. In a Minnesota grassland, the  $\text{NO}_3^-$  concentration in the soil of the rooting zone (Fig. 5) and below the rooting zone decreased with increasing diversity, indicating that less  $\text{NO}_3^-$  was lost via leaching and more nitrogen was retained within the ecosystem. In these same Minnesota grasslands, the nitrogen concentration in plants and the total nitrogen content of the plant biomass increased with the number of functional groups as well, and the composition and diversity of the functional groups had approximately equal influences on ecosystem processes. In California

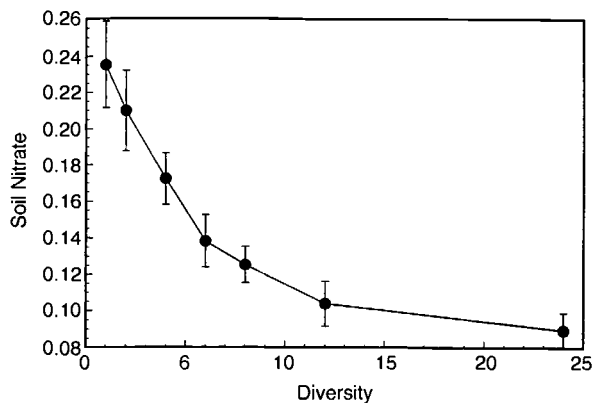


FIGURE 5 Effect of plant species diversity (number of species) in grassland plots on the concentration of nitrate in the soil (mg N/kg) containing most of the plant roots (0–20 cm) (reproduced with permission from Tilman, 1999).

serpentine grasslands, temporal partitioning of resources and facilitation by nitrogen fixers allowed more complete use of the greater available nitrogen, but the identity of the functional groups was more important than the number of functional groups. Hypothetically, increased diversity and resource exploitation increase nitrogen retention, thereby creating a positive feedback loop of increased fertility and productivity. With more complete capture or immobilization of a limiting resource such as nitrogen, more diverse communities may be less susceptible to invasion.

Variability of  $\text{NO}_3^-$  in the rooting zone decreases as the number of plant species increases, but this is partly due to a simple statistical artifact—a “portfolio effect” in which the aggregate variance of several species (like the diversified portfolio) is less than the variance of individual species. More diverse communities may show less variability in ecosystem function under the varying conditions typical in the natural world. Thus, species diversity may provide “insurance” that regardless of the varying environmental conditions, ecosystem functions will be carried out by some species suited to the prevailing conditions.

In general, it appears that the number of species or functional groups may influence nitrogen dynamics, but these are probably less influential than the particular identities of the species or functional groups, especially within the context of the many other factors which strongly influence nitrogen dynamics, including climate, geology, hydrology, type and history of disturbance, etc. Species and functional group diversity may be most important at the very low diversity characteristic of intensively managed and exploited ecosystems

typical of modern agriculture or very degraded ecosystems. It appears that the upper limit of the number of plant species for species diversity, independent of species composition, to have an impact on nitrogen cycling may be quite low (1–12) in most ecosystems. Of course, there may be some ecosystems in which nitrogen cycling cannot be maintained by a few species—perhaps extremely variable and heterogeneous ecosystems such as semiarid grasslands.

## V. CONSIDERATIONS FOR THE FUTURE

Unfortunately, there is no obvious reason to expect human impact on nitrogen cycling or other human impacts on biodiversity to decrease. Rather, we can expect modification of the nitrogen cycle and declining biodiversity to continue, within the context of all the other modifications which we humans are making to our environment, including climate change, air and water pollution, and overexploitation of natural resources. Many indirect effects of alterations of the nitrogen cycle on biodiversity are possible. For example, human activities have increased the release of  $\text{N}_2\text{O}$ , a greenhouse gas which contributes to climate change. Climate change has potentially major effects on biodiversity. In addition,  $\text{N}_2\text{O}$  which reaches the stratosphere participates in the destruction of the stratospheric ozone shield, allowing more ultraviolet light to penetrate to the earth's surface, potentially decreasing biodiversity. The variety of important ecological interactions which involve nitrogen, the complexity of controls on nitrogen dynamics, and the huge scale of human alterations to the biogeochemistry of nitrogen imply that many more interactions and feedbacks between nitrogen and biodiversity remain to be discovered as the domination of the environment by humans intensifies. The interactions between nitrogen and biodiversity are multiple and complex, but it is clear that humans cause both the changes in the biogeochemistry of nitrogen that usually reduce biodiversity and the changes in biodiversity that alter the biogeochemistry of nitrogen.

### See Also the Following Articles

ACID RAIN AND DEPOSITION • ATMOSPHERIC GASES • BACTERIAL BIODIVERSITY • BIOGEOCHEMICAL CYCLES • CARBON CYCLE • ESTUARINE ECOSYSTEMS • EUTROPHICATION AND OLIGOTROPHICATION • SOIL CONSERVATION

## Bibliography

- Bobbink, R., Hornung, M., and Roelofs, J. G. M. (1998). The effects of air-borne nitrogen pollutants on species diversity in natural and semi-natural European vegetation. *J. Ecol.* 86, 717.
- Charles, D. F. (Ed.) (1991). *Acidic Deposition and Aquatic Ecosystems: Regional Case Studies*. Springer-Verlag, New York.
- Cunningham, W. P., and Saigo, B. W. (1995). *Environmental Science: A Global Concern*, 3rd ed. Brown, Dubuque, IA.
- Hooper, D. U., and Vitousek, P. M. (1998). Effects of plant composition and diversity on nutrient cycling. *Ecol. Monogr.* 68, 121.
- Howarth, R. W. (Ed.) (1996). *Nitrogen Cycling in the North Atlantic Ocean and Its Watersheds*. Kluwer, Dordrecht.
- Howarth, R. W. (1998). An assessment of human influences on fluxes of nitrogen from the terrestrial landscape to the estuaries and continental shelves of the North Atlantic Ocean. *Nutrient Cycling Agroecosyst.* 52, 213.
- Howarth, R. W., Billen, G., Swaney, D., Townsend, A., Jaworski, N., Lajtha, K., Downing, J. A., Elmgren, R., Caracao, N., Jorden, T., Berendse, F., Freney, J., Kudeyarov, V., Murdoch, P., and Zhao-Liang, Z. (1996). Regional nitrogen budgets and riverine N & P fluxes for the drainages to the North Atlantic Ocean: Natural and human influences. *Biogeochemistry* 35, 75–139.
- Huston, M. A. (1997). Hidden treatments in ecological experiments: Re-evaluating the ecosystem function of biodiversity. *Oecologia* 110, 449–460.
- Jørgensen, B. B., and Richardson, K. (1996). *Eutrophication in Coastal Marine Systems*. American Geophysical Union, Washington, D.C.
- Kennedy, I. R. (1992). *Acid Soil and Acid Rain*, 2nd ed. Wiley, New York.
- Mooney, H. A., Cushman, J. H., Medina, E., Sala, O. E., and Schulze, E.-D. (Eds.) (1996). *Functional Roles of Biodiversity: A Global Perspective*. Wiley, New York.
- Naeem, S., Thompson, L. J., Lawler, S. P., Lawton, J. H., and Woodfin, R. M. (1994). Declining biodiversity can alter the performance of ecosystems. *Nature London* 368, 734.
- National Research Council (1996). *Understanding Marine Biodiversity*. National Academy Press, Washington, D.C.
- National Research Council (2000). *Nutrient Over-Enrichment in Coastal Waters*. National Academy Press, Washington, D.C.
- Tilman, D. G. (1999). The ecological consequences of changes in biodiversity: A search for general principles. *Ecology* 80, 1455–1474.
- Townsend, A. R. (Ed.) (1999). *New Perspectives on Nitrogen Cycling in the Temperate and Tropical Americas*. Kluwer, Dordrecht.
- Vitousek, P. M., Walker, L. R., Whiteaker, L. D., Mueller-Dombois, D., and Matson, P. A. (1987). Biological invasion by *Myrica faya* alters ecosystem development in Hawaii. *Science* 238, 802.
- Vitousek, P. M., Aber, J. D., Howarth, R. H., Likens, G. E., Matson, P. A., Schindler, D. W., Schlesinger, W. H., and Tilman, D. G. (1997). Human alteration of the global nitrogen cycle: Source and consequences. *Ecol. Appl.* 7, 737–750.



# NOMENCLATURE, SYSTEMS OF

David L. Hawksworth

*International Committee on Nomenclature*

---

- I. Purpose
  - II. Origins
  - III. Hierarchical Systems
  - IV. Principles
  - V. *Codes* of Nomenclature
  - VI. The *Draft BioCode*
  - VII. Special Cases
  - VIII. Other Systems
  - IX. Why Names Change
  - X. Naming Newly Discovered Species
  - XI. Determining Current Names
  - XII. Indices of Names
- 

## GLOSSARY

**available** Of a scientific name of an organism, one that must be taken into account when deciding the correct (q.v.) name to be used for it.

**binomen, binomial** The name of a species, consisting of the name of the genus in which it is placed followed by a word peculiar to that species.

**Code** In nomenclature, one of the sets of internationally agreed rules governing the scientific names of organisms.

**conserved** Of a scientific name, one that the appropriate international body has decided should continue to be used for an organism in cases where a strict application of the *Code* (q.v.) would mean it had to be replaced.

**correct** Of a scientific name of an organism, one that

conforms to the appropriate international *Code* (q.v.) for the position in which it is placed.

**priority** Of a scientific name, its date of valid publication; the first published name is generally the one to be used, subject to the provisions of the pertinent *Code* (q.v.).

**synonym** One of two or more scientific names applied to the same organism.

**taxon** Taxonomic entity or group of any rank, including all subordinately ranked taxa placed within it.

**type** Of an organism, the specimen, culture, or other element on which a scientific name of an organism is based and that fixes its application; the name-bearing type.

**valid** Of a scientific name, one that conforms to the conditions of valid publication proscribed in the relevant *Code* (q.v.).

---

**SYSTEMS OF NOMENCLATURE** are the procedures developed to provide the scientific names of organisms. This article concentrates on the consensual systems that operate in different groups of organisms through a series of published, internationally adopted *Codes*. It also discusses why scientific names change and the procedures involved in the naming of newly discovered species, and provides hints for ascertaining the current name of an organism. The article is concerned only with the nomenclature of organisms, that is, bionomenclature, and not with systems developed for naming

parts of the genome or communities of plants or animals.

## I. PURPOSE

The purpose of systems of nomenclature is to provide an unambiguous mechanism that enables biologists and all others who work with organisms to communicate by a scientific name so as to avoid misunderstandings and confusion. Scientific names are latinized and given to organisms of all kinds, whether living or fossil, and are in effect a universal currency. The same species may have different common or colloquial names in a variety of current languages, but can bear only a single correct scientific name.

Nomenclature is not the same as taxonomy, although the two are often confused. Taxonomy, a component of systematics (or biosystematics), is concerned with drawing up classifications, with deciding, for example, to what genus a species belongs, whether two species are really the same, or whether one species should be divided into two. Nomenclature is the process of determining what scientific names or labels should be applied to the units that taxonomic research considers to merit independent names. Nomenclature is subservient to taxonomy; it is not science itself, but the method by which scientific results are made available for general use.

The application of naming systems is complicated because new research may show that a particular taxonomy does not reflect the phylogenetic relationships of the organisms concerned. Further, the same species may have been described more than once by different scientists, and placed in a number of genera according to different taxonomic opinions. The net result is that there are many more scientific names than known species in the scientific literature. The full extent of the problem is unclear, but in the case of nonfossil botanical groups there are about 1.7 million species names available for 406,000 accepted species. It is the challenge of nomenclatural systems to provide clear procedures of how to handle a morass of names that may have been applied over the centuries to the same organism.

## II. ORIGINS

From the earliest times, humans needed to communicate taxonomic information: to tell each other which animals or plants are a threat, which can be eaten, which make tools or dyes, or which can help cure

particular ailments. Indigenous peoples today confront the same problems, and often have sophisticated naming systems for organisms of importance to them; interestingly, the species concepts they employ are often not too divergent from those a taxonomist would adopt. The basic need to name organisms was recognized even in the Bible, where one of Adam's first tasks was to name the animals.

Problems arose when different peoples needed to communicate with each other, and even the use of Latin as a language of scholars failed to solve the problem. Descriptive phrases written in Latin came to be used, but these "polynomials" were cumbersome and difficult to remember. For instance, the tea plant was referred to as *Evonymus affinis arbor orientalis nucifera, flore roseo* by Leonard Plukenet in 1696. By the early eighteenth century the situation was becoming impossible, and it was Carl Linnaeus (1707–1778) in Sweden who developed a solution that remains the basis of our current nomenclatural systems. Linnaeus at first followed several earlier authors in grouping his species into categories that had a single name, in effect a genus. However, in 1737 he started to use a single second name to denote the species in the indexes in his books. This shorthand practice entered the main texts of some of the works he was associated with in 1751, and he then went on to produce accounts of all organisms known throughout the world using this approach; the species names were printed in the margins offset from the main text. Instead of Plukenet's polynomial, Linnaeus simply used the binomial *Thea sinensis*. Such a radical change was not well received by all of his contemporaries, who regarded him as autocratic, but his *Species Plantarum* (1753) and *Systema Naturae* (10th edition, 1758) were destined to become the starting points of modern botanical and zoological nomenclature, respectively.

## III. HIERARCHICAL SYSTEMS

Nomenclatural systems are based on a series of ranks. Each species is referred to a genus, a genus to a family, a family to an order, and so on. The principal ranks available are, in descending order: domain, kingdom, phylum (or division), order, class, family, tribe, genus, section, species, and subspecies. Additional ranks can be intercalated using the prefixes super- and sub- (e.g., superfamily above family, subfamily below family). In practice, all the possible ranks are rarely used.

Though the number of available ranks may seem large, the current system cannot adequately satisfy those who wish to recognize each branching point in a clado-

gram, and various proposals for modified systems have been made, but are not yet accepted in any of the current *Codes*.

Some of the ranks above genus have standardized suffixes that enable them to be recognized (Table I), but there are exceptions for some well-known names approved by the relevant international nomenclatural bodies; for example, the grass family *Gramineae* (not *Graminaceae*). There are no rank-indicating endings used above phylum in botany, none over order in bacteriology, nor any over superfamily in zoology.

The names of families are based on an included genus, such as *Nostocaceae* on the genus *Nostoc*. Order names are also based on generic names in botanical groups, but not necessarily so in zoology, where the characters of the order tend to be employed. At ranks above order, there are no strict rules, but the names usually reflect some key distinguishing feature.

The only rank below species that is formally recog-

nized in bacteriology and zoology is that of subspecies. In zoology the subspecies name follows that of the species without any indication of rank, for example, *Mus musculus domesticus*, whereas under the other main *Codes* the rank is made clear by the insertion of “subsp.” or “ssp.”, as in *Silene vulgaris* ssp. *maritima*. The insertion of a rank term in those cases is essential to make clear which is referred to: “var.” denotes variety and “f.” denotes a form. In botany, additional terms are used in particular groups, most notably “cultivar” (“cv.”) for stable propagated cultivated plants and “special form” (“f. sp.”) for pathogenic fungi attacking different host plants that cannot be separated morphologically.

Scientific names in the rank of genus and above always start with a capital letter. Generic, specific, and formal ranks below species (see Table I) are always printed in italic script. The practice at other ranks varies and is a matter of editorial style as this is not ruled on in any of the existing *Codes*. While in zoology it is not

TABLE I  
Hierarchical Ranks and Suffixes Indicating Them

Rank	Bacteriology	Botany	Virology	Zoology
Domain				
Kingdom				
Subkingdom		-mycetia (fungi)		
Phylum		-mycota (fungi) -phyta (algae)		
Subphylum		-mycotina (fungi) -phytina (algae)		
Class		-mycetes (fungi) -opsida (ferns) -phyceae (algae)		
Subclass		-idae (ferns) -mycetidae (fungi) -phycidae (algae)		
Order	-ales	-ales	-virales	
Suborder	-ineae	-ineae		
Superfamily				-oidea
Family	-aceae	-aceae	-viridae	-idae
Subfamily	-oideae	-oideae		-inae
Tribe	-eae	-eae		-ini
Subtribe	-inae	-inae		
Genus			-virus	
Subgenus				
Section				
Species				
Subspecies				
Variety				
Form				



usual to italicize any other ranks, in botany scientific names at all ranks are italicized in the most recent *Code* (see Section V) and this practice is increasingly being adopted in botanical publications.

#### IV. PRINCIPLES

All *Codes* share a common objective: the scientific name of an organism must be unique and stable in a particular taxonomy. Nomenclature can be thought of as a tool used when it is necessary to decide on the name to apply to a particular organism or other hierarchical rank; it must not constrain taxonomy. To achieve this objective, each name needs to be formally introduced by agreed upon procedures. There must also be a means of ensuring that the application of every name is fixed so that it can always be recognized. In cases where more than one name has been proposed for the same organism, the basis on which to make a choice must be unambiguous.

Names of genera and species can be formed from any source, but are treated as if they were Latin. Ideally they should be informative, giving some clue to their features, but they are often based on the names of their discoverers or the places where they were found. This is not always the case; they can even be derived from acronyms or be anagrams. Today, the etymology of names that are introduced is generally expounded when they are first introduced.

A consequence of treating species names as if Latin is that their endings must agree with the gender of the genus in which they are placed. The ending of the name of the same species may thus change if it is transferred from a genus that has a female gender to one that is male.

The name of the person (or persons) who first introduced a scientific name is often indicated after the name itself, sometimes with the date of the publication. This device is essentially a much abbreviated reference to a particular publication, and while often interpreted as recognizing the contribution of that person, it can also be viewed as attributing responsibility for the person that burdened us with it—especially as many names prove to be unnecessarily coined. If an author's name appears without brackets, it was introduced just like that. If the name is in brackets, the name was originally in another genus or rank.

The citation *Simulium albellum* Rubtsov tells us that Rubtsov published that species name in the genus *Simulium*; *Obuchovia albella* (Rubtsov) makes clear that Rubtsov introduced the species name of this fly but did not refer it to the genus *Obuchovia*. In bacteriology and

botany, the name of the person making the change in generic placement or rank follows any name placed in brackets. For example, the citation of the Fool's Watercress, *Apium nodiflorum* (L.) Lagasca, indicates that "L." coined the name (actually as *Sium nodiflorum* L.) but that Lagasca first placed it in the genus *Apium*. Names such as *Simulium albellum* and *Sium nodiflorum*, which give the name being used in a different rank or combined in other genera, are termed "basionyms"; names using such basionyms in a different rank or position are "combinations."

The abbreviation of the names of authors, such as "L." for Linnaeus in this case, is commonplace, especially in botany where an internationally recognized list of abbreviations to be used for particular authors was issued in 1992; initials are inserted where there is more than one author with the same surname. An "ex" will sometimes appear inserted between the names of two authors; this means that the one before it first used the name but did not meet some requirements of validity or availability (see the following), whereas the second author did.

The way in which names are published is also specified in the different *Codes*. These relate to the publication itself (effective publication), which must be available to the scientific community, and the matter that must be associated with a name when it is introduced (valid publication), for example, a description and details of the material on which the name was based (the name-bearing type, discussed next). Only if the specified criteria are met does a name formally enter the body of scientific literature and become available for use.

The name-bearing type irrevocably fixes the application of a name. This is usually a specimen permanently preserved in a museum or other public institution, but in some cases it can also be a microscopic preparation, drawing, photograph, or microbial culture. The original material used and specified by the author is a "holotype"; a "lectotype" is material from the original author's collection selected to act as the name-bearing type by a later researcher; a "neotype" is a later collection chosen to serve as the name-bearing type where no original material remains. Types are not necessarily representative of the range of a species or subspecies, but are simply a nomenclatural device to fix the usage of names. They are consequently a very important reference point in biological communication and merit careful preservation and handling as they will remain of relevance for centuries.

A key feature in all *Codes* is the priority by date accorded to the names to be used for an organism. The

name to be adopted is generally the first published in the same rank or group of ranks, depending on the *Code*.

Surprisingly, with the exception of bacteriology and some groups of cultivated plants, there is no obligation on a person introducing a new scientific name to register it with some international authority. Proposals to introduce such a system for botanical groups were accepted in principle in 1993 and were to be introduced for the year 2000, but the 1999 International Botanical Congress reversed that decision.

Conscious of the fact that the application of a rigid system of rules could lead to unwelcome and confusing name changes, the *Codes* have developed a variety of mechanisms for overcoming this. These involve cases being published for open comment and consideration by international committees who debate them and express an opinion or recommend a particular course of action. Names can be “conserved” against ones that otherwise would have to be used, or “rejected,” in which event they cannot be taken up. In some instances, whole publications can be “suppressed” from usage.

An *Approved List* of names is published in bacteriology. *Virus Taxonomy* approaches this, and there is an *Official Lists and Indices of Names and Works in Zoology*. There is no exact equivalent in botany, although some of the appendices to that *Code* approach them, at least in part; however, proposals to develop protected lists of botanical *Names in Current Use* are currently being debated.

Names other than the one to be used that have been proposed for the same biological unit are termed “synonyms.” These can be of two types: “homotypic” (“objective” or “nomenclatural” synonyms) and “heterotypic” (“subjective” or “taxonomic”). Homotypic synonyms are based on the same name-bearing type, whereas heterotypic ones are based on different types.

## V. CODES OF NOMENCLATURE

The five current *Codes* differ in their detailed provisions. All are complex documents approved by particular internationally mandated bodies. Each has its own history of development and special characteristics. Some of their distinctive features, differences, and procedures are summarized here.

### A. Bacteriological

This *Code* arose as an offshoot of the botanical *Code* in 1939 in large part due to the need to accept living cultures as name-bearing types. In consequence, there

are many similarities between the two *Codes*, although they have diverged in some respects. Most strikingly, and as a consequence of the large numbers of bacterial names of uncertain application, a start date of 1980 was established and linked to an *Approved List of Bacterial Names*; names published prior to 1980 and not on the list do not exist for nomenclatural purposes, that is, they are not treated as validly published and so are no longer available for use. At a stroke, the number of species names that had to be considered by bacteriologists was reduced from around 32,000 to 3000.

New names of bacteria now have to be published in or be registered by the *International Journal of Systematic Bacteriology*, and will then be added to the next edition of the *Approved List*; a continuously updated on-line system is available through the internet.

The realization that many bacteria, and especially archaea, can be identified on the basis of molecular sequence data but not grown in culture is challenging the requirement for living cultures to be designated as name-bearing types. In addition, many cyanobacteria have name-bearing types that are dried specimens or microscopic preparations.

While the only rank below species to be formally recognized in this *Code* is that of subspecies, “pathovar” (“pv.”) is employed for plant pathogens essentially distinguished only by the ability to cause diseases in particular plants; for instance, *Xanthomonas campestris* pv. *durantae* (on *Duranta repens*) and pv. *erythrinae* (on *Erythrina indica*). Regulations regarding pathovar names and lists of those proposed are prepared by a committee of the International Society for Plant Pathology. Strain or serological type names or numbers are commonly used in bacteria of medical importance, but these have no standing under the *Code*.

The *Code* operates under the aegis of the International Committee on Systematic Bacteriology (ICSB), which is a part of the International Union of Microbiological Societies (IUMS). The Committee has a Judicial Commission that reports back to the ICSB and has wide powers to propose changes to the *Code*, which have to be ratified by IUMS congresses.

### B. Botanical

The botanical *Code* covers all groups traditionally studied by botanists even if they are no longer classified in the plant kingdom (*Plantae*), namely, cyanobacteria (*Bacteria*), fungi (*Fungi*), slime-moulds (*Protozoa*), and various algal groups (*Chromista* or *Straminopila*).

Botanical nomenclature is considered as starting from the publication of Linnaeus' *Species Plantarum* in

1753, but the first internationally agreed *Code* dates from the *Lois de La Nomenclature Botanique* prepared by Alphonse de Candolle in 1867 and adopted by an International Botanical Congress in Paris that year. Later starting dates, all linked to particular major publications by authors other than Linnaeus are used for some groups, notably mosses (1801), certain groups of algae (1848–1900), and fossils (1820). Fungal names formerly dated from 1801 or 1821 depending on the group, but that practice was discontinued in 1981. Names of fungi accepted in the previous starting point works remain “sanctioned” for continued use even if earlier competing names exist.

In addition to the insertion of terms to denote ranks below species and differences in the practice of how describing authors are cited, several features of this *Code* are unique to it. New scientific names published after 1935 must, with a few special exceptions, have a description or diagnosis (i.e., a statement of how the organism differs from others) in Latin. The desirability of retaining this practice is questionable and regularly debated now that English has occupied the role held by Latin in the eighteenth century.

The botanical *Code* recognizes the priority of names only within the particular rank under consideration. This means that even if a plant was recognized as a subspecies long before a species name was coined, the species name is nevertheless the one to be used. In addition, this *Code* rules as “illegitimate” names that have been introduced unnecessarily when another should have been adopted by the author. Also “illegitimate” are names spelled in exactly the same way; these are termed “homonyms” and only the oldest is generally available for use; for example, *Erica hibernica* (Hook & Arnott) Syme 1866 is illegitimate because of the existence of *E. hibernica* Utnet 1839, which represents a different species and is based on a separate name-bearing type. That Syme’s name was based on *E. mediterranea* var. *hibernica* Hook & Arnott 1835 does not affect the situation, as that name has priority from 1835 only in the rank of variety and not of species.

Name-bearing types have had to be cited in the *Code* since 1958, and from 1990 the institution where they are preserved must also be cited. Living type material is not permitted, but dating from 1993 freeze-dried (lyophilized) material or specimens preserved in liquid nitrogen are acceptable as they are in a metabolically inactive state.

The botanical *Code* has special provisions for hybrids, fungi with pleomorphic life cycles, and fossils, all of which are considered separately in Section VII below, and also appendices of conserved and rejected

names, and suppressed publications. The provisions of the *Code* are now debated at each six-yearly International Botanical Congress, after which a new edition is published. Any changes proposed have to be published in the journal *Taxon* and are balloted first by mail and then at the Congress itself, where a 60% majority is normally required to effect any change. The Congress establishes a series of permanent Committees that are charged with considering and making recommendations on proposals to reject or conserve names in different groups; those also have to be published in *Taxon* and ratified by a subsequent Congress.

### C. Cultivated Plant

The cultivated plant *Code* split from the botanical one in 1952. It is primarily concerned with the naming of cultivated varieties or “cultivars,” which are propagated horticulturally by any method that retains their characteristics in a stable way. It is complementary to the botanical *Code* and does not compete with it. Species names and those in other formal recognized ranks of cultivated plants remain subject to the botanical *Code*.

Cultivar names, with a few established exceptions, are not latinized and since 1958 have been required to be in a modern language; they are placed within single quotation marks and never printed in italic type, for example, *Rubus nitidoides* ‘Merton Early,’ in which ‘Merton Early’ is the cultivar name and *R. nitidoides* is the botanical species name. The abbreviation “cv.” can be inserted before the cultivar name but is generally omitted. It is also not uncommon for the species epithet to be dropped as well in garden catalogs (e.g., *Rubus* ‘Merton Early’). Similar cultivars can be grouped into cultivar-groups; this is achieved by placing a group name in brackets before the cultivar name and without quotation marks, for instance, *Rosa* (Hybrid Tea Group) ‘Richmond’; if the particular cultivar is not being cited, the brackets are omitted (*Rosa* Hybrid Tea Group).

In the case of orchids, groupings of cultivars based on parentage are called “grexes” (or more correctly “greges”); their application follows separate provisions in *The Handbook of Orchid Nomenclature and Registration* (1993).

Graft-chimeras, in which living tissues of two species are made to grow together by grafting, can also be named under this *Code*. These are indicated by the use of a “+” sign in a manner equivalent to the use of “x” in the names of hybrids (see Section VII,C). The form *Crataegus monogyna* + *Mespilus germanica* indicates a graft with tissues of both of those plants. A latinized

name can be used instead provided that the “+” sign is retained and the resultant name is not identical to a name of a hybrid; in this example, the name +*Crataegomespilus dardarii* has been coined. Cultivar names formed like this can be added after the name of the graft-chimera.

Names of cultivars have to be established through publication and not by garden labels or confidential trade lists. There must also be a reference point for their application, a “standard” dried specimen preserved in a herbarium, or in some cases an illustration. There is then a system of 45 International Registration Authorities, each dealing with particular groups of cultivated plants, that register newly proposed names and make lists of these available.

The cultivated plant *Code* operates under the auspices of the International Commission for the Nomenclature of Cultivated Plants.

## D. Virological

The naming of viruses was controversial for many years. Two different systems were operating in parallel: a traditional system recognizing families, genera, and species with latinized names, and a divergent one employing the categories of group, subgroup, and type (or virus) and in a modern language. The former held sway among those working with viruses in animals, and the latter among plant pathologists. In the case of plant viruses, a scientific name indicative of the structure of the virus was generally preceded by an indication of the host and (or) the symptoms caused; for instance, the “Desmodium mosaic potyvirus” and the “Desmodium yellow mottle potyvirus.”

The situation started to change rapidly in 1991 when the International Committee on the Taxonomy of Viruses decided to use the traditional approach (the Committee operates within the International Union of Microbiological Societies). New guidelines were drawn up and adopted by the International Congress of Virology in 1993. Proposals to modify the rules relating to virus names may be aired in *Archives of Virology*, which also includes reports of meetings of the Committee and periodically a survey of all accepted virus taxa.

Virus names can be recognized by their endings even down to the rank of genus; generic names have the suffix *-virus* (see Table I).

## E. Zoological

Just as the botanical *Code* covers all groups traditionally studied by botanists, so the zoological *Code* embraces all that have been considered by zoologists. Although

the two *Codes* both arose in the mid-nineteenth century, their provisions have diverged in some significant ways through subsequent modifications. The first zoological *Code* appeared in 1843, and revisions were considered at each International Congress of Zoology until these were discontinued after 1972. New editions are now prepared more irregularly. The starting point for zoological nomenclature is the tenth edition of Linnaeus' *Systema Naturae* published in 1758; there are no exceptions for particular groups.

The zoological and botanical *Codes* rule that they are independent. This means, for example, that the same generic name can be used for disparate organisms provided they belong to groups treated under different *Codes*; a classic case is *Drosophila*, a genus of fungi and one of flies. The occurrence of such cross-*Code* homonyms can cause confusions in database searching.

The zoological *Code* is distinctive in operating a principle of coordination of ranks. Three series of names are recognized as coordinate: family-, genus-, and species-groups. Each group includes names of several ranks, for example, species includes superspecies, species, and subspecies; these are treated as equivalent for nomenclatural purposes, such as assigning priority of publication and author citations. No rules relating to names above the family-group level are included.

The description or diagnosis accompanying the publication of new scientific names in zoology can be in any language that uses words, although English, French, German, Italian, or Latin are recommended. Surprisingly, the designation of name-bearing types, while recommended, has not been mandatory although this situation is expected to change in the 1999 edition of this *Code*. Species names in which the generic name and specific epithet are identical, for example, *Troglodytes troglodytes*, are termed “tautonyms”; these are permitted in zoology but not in botany.

There are also differences when a species is moved from one genus to another. In zoology, the earliest name takes precedence and is to be taken up in the new genus even if an identical name based on a different name-bearing type already exists in the genus; in this way, “secondary homonyms” can be created.

Terminological differences between the zoological and botanical *Codes* are a particular source of confusion to biologists; the same term can have different meanings and the same concept different terms (Table II). Proposals to harmonize terminology are now being made.

The International Commission on Zoological Nomenclature has wide powers to set aside provisions of the *Code* in particular cases where to follow the rules would lead to changes in well-established names. Appli-

TABLE II  
Equivalent Terms Used in the Current Codes of Nomenclature

Draft BioCode	Bacteriology	Botany	Cultivated Plants	Zoology
<i>Publication and dates of names</i>				
published	effectively published	effectively published	published	published
registerable	effectively published			
date	date	date (or priority)	date	priority
priority	priority	priority		priority
precedence	priority	priority	precedence	precedence
earlier	senior	earlier	earlier	earlier
later	junior	later	later	junior
<i>Nomenclatural status</i>				
established	validly published	validly published	established	available
registration	validation	registration	registration	
acceptable	legitimate	legitimate	acceptable	potentially valid
<i>Taxonomic status</i>				
accepted	correct	correct	accepted	valid
<i>Types of names</i>				
name-bearing type	nomenclatural type	nomenclatural type	name-bearing type	name-bearing type
nominal taxon	name and type	name and type		nominal taxon
<i>Synonymy</i>				
homotypic	objective	nomenclatural	homotypic	objective
heterotypic	subjective	taxonomic	heterotypic	subjective
replacement name		avowed substitute		explicit replacement
<i>Setting aside the rules</i>				
conserved	conserved	conserved	conserved	conserved
rejected	rejected	rejected	rejected	conditionally suppressed
suppressed	rejected	explicitly rejected		suppressed

cations have to be published in the *Bulletin of Zoological Nomenclature*, and after a period to permit zoologists to publish Comments in the *Bulletin*, the Commission expresses and publishes an Opinion that is binding. In the absence of International Congresses of Zoology, the Commission prepares new editions of the *Code* largely by correspondence, but subject to ratification at a General Assembly of the International Union of Biological Sciences, under whose aegis it operates.

## VI. THE DRAFT BIOC CODE

The need for a more unified approach to biological nomenclature has long been recognized, and this has been of particular concern to the International Union of Biological Sciences. Concerted action to address the problem emerged from a symposium held during the International Congress of Systematic and Evolutionary Biology in 1985. Pressure for harmonization between the various *Codes* arises from the unification of biology as a discipline and the consequent difficulties in teaching nomenclature, the mismatch between the *Codes* and

the kingdoms of life now recognized, the needs of users for increased stability in names, the needs of systematists to spend less time on nomenclature, and the difficulties faced by developing nations. In addition, the existing *Codes* were finding that they had to confront similar issues, for example, in relation to electronic publication, living name-bearing types, organisms that could be treated under different *Codes* depending on whether they were interpreted as animals or plants, protected lists of names, and the registration of newly proposed names.

The move toward harmonization is operating along three fronts: harmonization of terminology (see Table II), bringing similar provisions into all *Codes* where new problems are being confronted, and developing a unified *Code* that would apply to all kinds of organisms. In 1995 the IUBS and the International Union of Microbiological Societies, the two bodies that between them oversee all five *Codes*, jointly established an International Committee on Bionomenclature to promote these initiatives. This Committee, which consists of representatives of the bodies responsible for the five organismal *Codes*, issued a *Draft BioCode: The Prospective Interna-*

*tional Rules for the Scientific Names of Organisms* for discussion in 1996.

The *Draft BioCode*, subject to approval by the appropriately mandated bodies, is planned to operate for names introduced after a date to be agreed on and for groups where lists of protected names are in existence. It was recognized early on that a single retrospective *Code* would be too disruptive of names in use. The existing *Codes* would consequently continue to apply to names introduced before that date, but be expected to adopt a common terminology and similar approaches to new problems. Among the provisions of the *Draft BioCode* are the adoption of coordinate status for names, a requirement for descriptions and diagnoses to be in either English or Latin, provision for electronic publication, registration of newly proposed names, unacceptability of homonyms, and the conservation or rejection of names to promote stability. The starting point for the new *Code* will be lists of protected names that cannot be displaced by unlisted names; a more open and more stable system will result, which at the same time will reduce the time inputs required from systematists.

Debates on the latest version of the *Draft BioCode* took place in 1998–1999, and final implementation decisions by the international bodies will not be made for some years; at the 2005 International Botanical Congress in the case of botany.

## VII. SPECIAL CASES

Particular provisions or problems in the *Codes* relate to the special cases of ambireginal organisms, fossils, hybrids, lichens, and pleomorphic fungi.

### A. Ambireginal Organisms

Ambireginal organisms are ones that can be treated under different *Codes* depending on whether they are considered to be plants or animals, algae or bacteria, etc. The problem is particularly acute in the case of unicellular and often flagellate organisms now placed in the kingdoms *Protoctista* or *Protozoa*, and especially where related species may or may not have chloroplasts and so have traditionally been treated as plants or animals, respectively. Because different rules apply when a “plant” is found to be an “animal,” name changes in both the generic and species names may result. For example, the euglenoid name *Entosiphon* B. Stein 1878 is acceptable under the zoological *Code*, but when treated as a plant it has to be replaced by *Entosiphonomo-*

*nas* Larsen & Patterson 1991, as there is an earlier flowering plant genus *Entosiphon* Beddome 1864.

The resultant frustrations and confusions have led to demands for a separate *Code* for protoctistan groups, but the preferred route forward is to have parallel provisions in the existing *Codes*. The problem would not arise under the *Draft BioCode* as no decisions as to whether an organism is a “plant” or an “animal” would be needed.

In the case of slime moulds (*Myxomycota*), because these organisms have been traditionally treated under the botanical *Code* as fungi, that practice is continued to prevent the disruption in names that a switch to the zoological *Code* would entail. The *Cyanobacteria*, traditionally treated under the botanical *Code* as blue-green algae, are pragmatically best left there rather than transferred to the bacterial *Code* in order to avoid unfortunate and unnecessary name changes.

### B. Fossils

Special provisions for fossils are made only under the botanical *Code*. Pieces of trunks, stems, roots, and leaves, and seeds, pollen, and spores can all be given latinized binomials. These are considered to be “form-” or “organ-” genera and species in the absence of intact plants. Even when such pieces are later found connected together, as in the case of the *Archaeopteris* leaves and *Callixylon* wood, the two systems are often maintained because of the time spans involved. In particular, it is not known whether a *Callixylon* leaf always had an *Archaeopteris* leaf system.

Where names of living and fossil plants compete, the living always has precedence. The names of fossils thought to be close but not necessarily identical with a modern genus are often given the suffix “-ites,” as in *Ginkgoites* for fossil plants and *Ginkgo* for species now living.

### C. Hybrids

Only the botanical *Code* has special arrangements for the naming of hybrids. The zoological *Code* excludes them from consideration.

Botanical hybrids can be referred to by indicating their parentage and the use of the multiplication symbol “x” to make a “hybrid formula.” Alternatively, and particularly if the hybrid is regularly encountered in nature, it can be given a separate scientific name prefixed by the same symbol. For example, the hybrid between *Potentilla anglica* and *P. erecta* can be indicated as *Potentilla anglica* x *P. erecta* or alternatively as *Potentilla*

*xsuberecta*. The same principle applies to generic names, for instance, *xAgroelymus* is a hybrid between species belonging to the genera *Agropyron* and *Elymus*; the name *xAgroelymus* is called a “condensed formula.”

In order to be established, the separate names of hybrids have to conform to other parts of the *Code* that apply to their rank. In the case of condensed formulae for generic names, the parent genera have to be indicated.

## D. Lichens

Although generally appearing as discrete individuals, lichens are mutualistic symbioses composed of two or sometimes more different kinds of organisms. A fungal partner combines with a green alga or cyanobacterium, or both, to form characteristic shapes for the particular association. The different partners, when isolated into pure culture, fail to produce the same morphological structures. The classification of both the fungi and algae or cyanobacteria involved is based on their own characteristics and not those of the combined association; both have independent scientific names.

Before the dual nature of lichens was recognized in 1867, and for many years afterward, names were given to the composite lichen structure. The botanical *Code* rules that for nomenclatural purposes the names given to lichens are to be considered as belonging to the fungal partner. Pedantically, this means that lichens do not have names, and that it is impossible to “name a lichen.”

## E. Pleomorphic Fungi

Fungi that have a life cycle with separate sporulating stages, known as pleomorphic fungi, present particular problems in naming. In many cases the different stages were named separately in different genera and were never thought to be connected until fresh research was carried out, for example, by germinating single sexual spores. In many cases the names given to the sexual (“teleomorph”) and asexual (“anamorph”) stage are both well established and mycologists have been reluctant to simply apply the earliest name regardless of the stage it represents. This resistance is understandable, as it is often a particular sporulating form that is associated with a plant disease or the production of toxins.

In basidiomycete and ascomycete fungi, with the exception of lichen-forming species, the correct name for a fungus with a pleomorphic life cycle that covers all its stages is that which produces sexual spores, that is, the teleomorph. The anamorph (or anamorphs) can also have independent names so that just those stages

can be referred to; the name-bearing types of the names of anamorphs must represent only the asexual (“mitosporic”) sporing stage, and the sexual stage must not be mentioned in its establishing description.

For example, when *Penicillium brefeldianum* B. O. Dodge 1933 was introduced, the description and type covered both the teleomorph and anamorph; that name is therefore applicable to the teleomorph, even though the generic name *Penicillium* is based on an asexual species. In this case, the teleomorph belongs to the sexual genus *Eupenicillium* and is correctly named *E. brefeldianum* (B. O. Dodge) Stolk & Scott 1967. The anamorphic name *Penicillium dodgei* Pitt 1980, with only asexual structures mentioned in its description and represented on the type, was introduced for those wishing to refer just to the asexual *Penicillium* stage.

Now that molecular techniques enable the position of asexual fungi to be determined with confidence in overall fungal classifications, even in the absence of a sexual stage, doubt is being expressed as to the desirability of maintaining this special provision for dual nomenclature in the botanical *Code*.

## VIII. OTHER SYSTEMS

This article focuses on the nomenclatural systems used in and operating under the mandate of internationally mandated committees established by the scientific community. But, in addition to infraspecific systems such as those for special forms and pathovars that are specifically excluded from these *Codes*, alternative nomenclatural systems have been proposed, although none has found general approval.

Concern over the adequacy of the currently accepted systems has been voiced most strongly by those involved in the determination of phylogenetic relationships using cladistic analyses. If formal scientific hierarchical names were to be given to equivalent branching points, considerable numbers of new ranks would be needed and there would be an exponential expansion in the numbers of such names. A further complication is that phylogenetic analyses may show that different families or orders generally regarded as distinct have some recent ancestors in common (i.e., are monophyletic) and arguably could therefore be united. A pragmatic approach to this problem has been suggested, namely, continuing the current taxonomic practice while recognizing that all higher category names may not be equivalent, even to the extent that some may be nested within units bearing names of the same rank. In a system proposed by K. E. Kinman in 1994 to take

the best from cladistic analyses and include them in a practical system, markers are used to indicate where orders or higher ranks include organisms belonging to different evolutionary lines (i.e., are paraphyletic) while retaining established names.

A more radical nomenclatural system is the New Biological Nomenclature, which arose in 1971 out of the frustration felt by Belgian zoologists with the current system of *Codes*. Names being introduced have to be agreed on by a group of specialists and are presented in Esperanto and not latinized. Generic names under this system relate more to common perceptions than those of scientists and can embrace several traditionally accepted genera. For example, the New Biological Nomenclature Esperanto generic name *Delfeno* embraces 17 genera of cetaceans; *Delfeno diverskolora* is used for the Common Dolphin (i.e., *Delphinus delphis*) and *Delfeno ordotipa* for the Bottle-nosed Dolphin (i.e., *Tursiops truncatus*). The system has not gained general acceptance, but does have an Association devoted to its promotion; detailed rules have been issued, the latest edition in 1991, and 522 names had been "officialized" by the end of 1990.

A novel approach to the nomenclature of fossils of all groups, a Paleontological Data Handling Code, was proposed by N. F. Hughes in 1989. Instead of first matching a newly discovered fossil to what has already been described, each specimen is recorded separately with particular reference to its stratigraphic position; only then are comparisons made with what has already been described in that particular position in the fossil record.

Various systems of coding using letters and/or numbers rather than latinized names have been proposed, the first by P. Hartig in 1871. A particularly elaborate system was that published by G. Tornier in 1898, which took the first letters of words of the higher ranks to which a species belongs, followed by a number indicating the particular species; where more than one name started with the same first letter, subscripts were employed. Numerical codes regained popularity in the 1960s in response to the need to compress data in the computing systems in use at the time; one such code extended to 36 digits. With the advances in current technology, the interest in alphanumeric systems has waned. Humans find it easier to remember words than numbers.

## IX. WHY NAMES CHANGE

Changing scientific names are a constant cause of irritation and frustration to all who use them. That this

occurs at all is sometimes interpreted as a criticism of the discipline, and while that may be justified in some instances, in most cases it is in the long-term interests of users.

New research can show that a certain species is actually a member of a different genus than the one in which it was placed, and it is therefore transferred. Because classifications are predictive of the properties of an organism, accurate placement will benefit users searching for species with particular attributes or wishing to determine the risks they may pose. Detailed studies may also show that one species is a complex of several that merit independent recognition, or that two species currently treated as distinct are really the same; again, these are categories of information that are relevant to users.

Name changes that arise from new taxonomic research, or "taxonomic changes," are therefore to be welcomed, but that is not so for "nomenclatural changes." Nomenclatural changes are ones that arise from the provisions of the *Codes* and not from any new scientific data. Two categories can be distinguished. First, some changes are due to a failure to apply the relevant *Code*, for example, the discovery of an earlier name for the same species, that a name did not meet all the requirements for establishment, or that a name was superfluous under the botanical *Code* when published. Second, some changes arise from rules of the *Code* itself that are retroactive, for example, in relation to starting point dates or new provisions relating to issues ranging from recommendations on spelling to reworkings of particular sections.

Problems leading to name changes also arise from the application of the system of name-bearing types. Species names are sometimes found to have come to be applied to a species different from that represented by its type. This category of changes is especially unfortunate as it can mean that a name that has been used in one sense must be applied in a different one.

The responsibility for deciding when enough data are at hand to justify a change in taxonomy will always lie with the taxonomist. Provided that the taxonomist is self-critical in approach and conscious of the need to be conservative in the interests of the image of the subject, most taxonomic changes will prove to be in the users' interests. That is not the case with nomenclatural changes where all *Codes*, and particularly the *Draft Bio-Code*, are increasingly addressing this fundamental issue. The powers now given to the botanical and zoological *Code* Committees mean that, provided a sound case is made, rules may be set aside to protect familiar names; this can even extend to the designation of different



name-bearing types in order to maintain the current usage of names.

## X. NAMING NEWLY DISCOVERED SPECIES

As only about 1.75 million of the estimated 13.7 million species on Earth have yet been given scientific names, the task of naming the balance is one of the major challenges facing biology as we enter the twenty-first century. However, it is always easier to describe an organism as new to science than to establish if it really has never been described before, albeit in a different genus or even in another family. The repeated unnecessary description of species that are not new is a major cause of the inflation of taxonomic literature. In the case of the fungi, one analysis showed that each accepted species had been given on average 6.5 names! Once introduced into the body of scientific knowledge, names can never be expunged and will have to be taken note of by all future workers on the group concerned. Before deciding to introduce a new name, careful checking is a prerequisite.

This checking proceeds in a series of steps. First, see if the prospective species matches any in the same (or allied) genera in any checklists for the geographical area from which it comes. Then proceed to any regional or world monographs of the group if they exist. If these steps yield no matches, check names in the world indices of names (see Section XI), investigating possible names in allied genera and under names no longer in use. This will entail making lists of possible names and their places of publication, looking up original descriptions, and locating and making comparisons with name-bearing types for likely candidates. If a specialist in the group concerned can be located, that can be a most valuable shortcut as the checking process can be exorbitantly time-consuming if dealing with, for example, a genus of over 1000 described species.

When novelty has been confirmed in the checking phase, the procedures detailed in the appropriate *Code* for introducing (and in some cases registering) new names must be followed. These include clear statements of why the species is different, a full description (which may have to be in Latin or English), illustrations (photographs as well as drawings are recommended), and a name-bearing type deposited in a secure public institution (ideally with duplicates deposited elsewhere for security). Publication should ideally be in a journal that will be readily accessible to other specialists in that

group and geographical region. Much can be learned from recent publications of new species in the same group by other authors.

In some instances, notably certain bacteria, internationally developed minimum standards for descriptions are available and should be followed. In some cases codes of practice have been developed, for example, there is one for fungi by the International Commission on the Taxonomy of Fungi.

## XI. DETERMINING CURRENT NAMES

When taxonomic work is undertaken, a number of different names are often found to have been given to the same species. The steps to be followed in determining which is to be used can be considered as a series of sieves, a "nomenclatural filter." Although there are differences in the detailed requirements at each step, and some will depend on the rank, the sequence in which these need to be followed is the same.

**Step 1: Publication** If a name has not met the requirements for effective publication, generally being printed in a book or journal available to other researchers, it need not be considered further.

**Step 2: Establishment** All of the following requirements must be met: species description (and in some cases illustrations), clarity concerning the rank (or coordinate group of ranks in zoology) and the intention to introduce a new scientific name, registration where appropriate, citation of the name-bearing type, and in some cases its place of preservation. If any requirement is not fulfilled, those names also need not be considered further.

**Step 3: Typification** The most critical step and the one that requires scientific judgment is checking the status of the type and examining the specimen, microscopic preparation, illustration, and living culture as appropriate. In some cases, no name-bearing type may have been designated and it may be necessary to select a lectotype or neotype to fix the application of the name. Names whose types are different from the taxon to be named can be excluded at this step.

**Step 4: Acceptability** Does the name contravene any special provisions, such as being superfluous under the botanical *Code*? Is the spelling identical to another name of the same rank (or coordinate group of ranks in zoology)?

**Step 5: Precedence** What is the date of establishment of each name and which is the earliest in the same rank (in botany) or family-, genus-, or species-group (zoology)?

**Step 6: Accepted Name** The name to be used under the *Code* will be the one with precedence, that is, the earliest acceptable name to be established. This name may have to be used in a new combination. Under the botanical *Code*, a new combination cannot be made if the resultant binomial would be spelled exactly like an already existing name (i.e., a “homonym”). The situation in zoology differs in that a new homonym can be made if the basionym takes precedence by date; the already existing homonym becomes a “secondary homonym” and can no longer be used. Names applied to the same suprageneric rank or organisms other than the accepted one are “synonyms.” Synonyms are of two main types: “homotypic” when based on the same basionym or name-bearing type, and “heterotypic” when based on different basionyms or name-bearing types. Homotypic synonyms will always appear together, as their applications are irrevocably linked by a common type, whereas heterotypic synonyms may be treated differently depending on the disparate views of taxonomists on the particular types. If the steps potentially result in the disruption of a well-known name, the procedures for the conservation and rejection of names are available as a safeguard (see earlier discussion).

This daunting series of steps, once performed, leads to an accepted name. The advantage of the internationally agreed on *Codes* is that whoever follows the steps faithfully will come to the same answer regarding the name to be used. That name will remain stable unless new nomenclatural or taxonomic information comes to light (see earlier).

Faced with such complexity, users of names wishing to check the currently accepted name for an organism need shortcuts. Recent world monographs and checklists are the first port of call, and then any regional or national treatments. If these are not available, and there is no taxonomist specializing in the group that can be consulted, it will be necessary to check the main indices of names and supporting bibliographic databases or abstract publications (e.g., *Biological Abstracts*, *Kew Record*, or *Bibliography of Systematic Mycology*). Online checking will improve markedly when the SPECIES-2000 system is fully operational (see the next section). Checking should be viewed as a necessary evil. If research is published using an unfamiliar obsolete name, it may never be retrieved from

the hundreds of thousands of scientific articles that appear each year.

## XII. INDICES OF NAMES

There is a series of major reference works of the scientific names of organisms that focus on different major groups and that vary with respect to the information captured. There are too many to list here, but attention should be drawn to those that are currently active: *Index Kewensis*, *Index of Fungi*, *Zoological Record*, and the *Approved Lists of Bacterial Names*. These are increasingly available as on-line systems through the internet or as CDs.

Unfortunately there is still no one-stop shop for information on names, though there is now an international initiative, SPECIES-2000, that is linking key databases around the world that have not just lists of names but ones that are taxonomically up-to-date for particular families, orders, and so on. International funding is currently being sought to complete the linking of existing data sets, and also to establish additional databases where no taxonomically vetted ones exist. For zoology, the BIOSIS *Taxonomic Reference File* is particularly valuable in the interim.

### See Also the Following Articles

SPECIES, CONCEPT(S) OF • SUBSPECIES, SEMISPECIES • SYSTEMATICS (OVERVIEW) • TAXONOMY, METHODS OF

## Bibliography

- Greuter, W., Barrie, F. R., Burdet, H. M., Chaloner, W. G., Demoulin, V., Hawksworth, D. L., Jørgensen, P. M., Nicolson, D. H., Silva, P. C., Trehane, P., and McNeill, J. (eds.). (1994). *International Code of Botanical Nomenclature (Tokyo Code) adopted by the Fifteenth International Botanical Congress, Yokohama, Japan, August–September 1993*. Regnum Vegetabile No. 131. Koeltz Scientific Books, Königstein.
- Greuter, W., Hawksworth, D. L., McNeill, J., Mayo, M. A., Tindall, B. J., Trehane, P., and Tubbs, P. (1997). Draft BioCode (1997): The prospective international rules for the scientific names of organisms. *Taxon* 47, 127–150.
- Hawksworth, D. L. (ed.). (1991). *Improving the Stability of Names: Needs and Options*. Regnum Vegetabile No. 123. Koeltz Scientific Books, Königstein.
- Hawksworth, D. L. (1994). *A Draft Glossary of Terms used in Bionomenclature*. IUBS Monograph No. 9. International Union of Biological Sciences, Paris.
- Hawksworth, D. L. (ed.). (1997). *The New Bionomenclature: The BioCode Debate*. Biology International, Special Issue No. 34. International Union of Biological Sciences, Paris.

- International Commission on Zoological Nomenclature (1999). *International Code of Zoological Nomenclature*, 4th ed. International Trust for Zoological Nomenclature, London.
- Jeffrey, C. (1989). *Biological Nomenclature*, 3rd ed. Edward Arnold, London.
- Lapage, S. P., Sneath, P. H. A., Lessel, E. F., Skerman, V. B. D., Seeliger, H. P. R., and Clark, W. A. (eds.). (1992). *International Code of Nomenclature of Bacteria (1990 Revision)*. American Society for Microbiology, Washington, D.C.
- Melville, R. V. (1995). *Towards Stability in the Names of Animals*. International Trust for Zoological Nomenclature, London.
- Murphy, F. A., Fauquet, C. M., Bishop, D. H. L., Ghabrial, S. A., Jarvis, A. W., Martelli, G. P., Mayo, M. A., and Summers, M. D. (1995). *Virus Taxonomy: The Classification and Nomenclature of Viruses*. Archives of Virology, Supplement No. 10. Springer-Verlag, Vienna.
- Nicolson, D. H. (1991). A history of botanical nomenclature. *Ann. Missouri Botanical Garden* 78, 33–56.
- Ride, W. D. L. (1988). Towards a unified system of biological nomenclature. In *Prospects in Systematics* (D. L. Hawksworth, ed.), pp. 332–353. Systematics Association Special Volume No. 36. Clarendon Press, Oxford, United Kingdom.
- Ride, W. D. L., and Younés, T. (eds.). (1986). *Biological Nomenclature Today*. IUBS Monograph No. 2. IRL Press, Oxford, United Kingdom.
- Stalleu, F. A. (1971). *Linnaeus and the Linneans*. A. Oosthoek, Utrecht, Netherlands.
- Trehane, P., Brickell, C. D., Baum, B. R., Hettterscheid, W. L. A., Leslie, A. C., McNeill, J., Sponberg, S. A., and Vrugtman, F. (eds.). (1995). *The International Code of Nomenclature for Cultivated Plants—1995*. Regnum Vegetabile No. 133. Quarterjack Publishing, Wimborne, United Kingdom.



# NORTH AMERICA, PATTERNS OF BIODIVERSITY IN

Martin J. Lechowicz  
*McGill University*

---

- I. Continental Diversity
  - II. Terrestrial Bioregions of North America
  - III. Diversity in Major Groups of Organisms
  - IV. Changes in North American Biodiversity
  - V. Conclusion
- 

tral America with strong affinities among the regional floras and faunas but different from the tropical biota of Africa and Southeast Asia (the Paleotropics).  
**physiography** The landforms that give shape and character to the continental landscape, for example, the configuration of mountains and drainage basins.

---

## GLOSSARY

**Beringia** A region at the northwestern corner of North America separated from Asia only by the shallow waters of the Bering Strait. When ocean levels drop during glaciation, Asia and North America are connected by land in this region.

**bioregion** A landscape subunit at the scale of the continent that is set apart by its coherent geological and biotic history.

**continental shelf** The continental margin, which is submerged to depths to approximately 180 m at current sea levels and is bounded by the abrupt drop to the abyssal depths of the oceans.

**ecoregion** A landscape unit within a bioregion in which a distinct assemblage of organisms interact ecologically, usually within a spatial context defined by drainage systems, mountain ranges, or similar natural boundaries.

**Holarctic** A biotic realm in the temperate, boreal, and arctic regions of the Northern Hemisphere with strong affinities among the regional floras and faunas.

**Neotropics** A biotic realm in tropical South and Cen-

**THE DIVERSITY OF THE FLORA AND FAUNA OF NORTH AMERICA** has been the focus of study by naturalists and scientists for the past few hundred years. Studies of species diversity have proceeded even as parts of North America have undergone dramatic changes driven by activities associated with settlement, mining, forestry, agriculture, and industrial development. Large areas of the midcontinent have less than 5% of their primeval landscape intact. This article attempts to identify the patterns of species diversity that existed in North America 500–1000 years ago, before the arrival of the Europeans. The focus is on the diversity and distribution of native or indigenous species, as opposed to species that people have introduced from other places. Current or historical records from relatively undisturbed habitats provide the basis for estimates of species diversity from region to region. This article is restricted to the diversity of the terrestrial and freshwater biota of mainland North America, including offshore islands but excluding the rich marine biota associated with the continental shelf and coastal estuaries. Although the continental shelf is quite properly considered part of

North America, its biota is determined not only by continental but also oceanic influences of considerable complexity. The evaluation and summary of North American marine diversity is a separate and important task, but it is beyond the scope of this article.

## I. CONTINENTAL DIVERSITY

Mainland North America, with an area of 24,258,000 km<sup>2</sup>, accounts for 16% of the land surface on Earth and stretches over 65° of latitude. The ancient core of North America is well over 3 billion years old, but parts of the continent are much younger. The continent is bounded on the west by a series of mountain ranges along the coast of the Pacific Ocean. A more extensive coastal plain and less rugged mountains characterize the Atlantic coast of North America. The Arctic Ocean lies at the northern limit of the continent, which includes the extensive Arctic Archipelago. Rugged mountains mark the commonly accepted limit of the continent in southern Mexico, but in fact distinguishing the boundary in this mountainous and geologically complex region where North America grades into Central America is arbitrary. The continental interior is an extensive plain with relatively little topographic relief, although highlands, escarpments, and erosional features occur regionally. Major drainage systems flow both north and south from the interior and also through the Great Lakes basin to the east. The Mississippi River system flowing south into the Gulf of Mexico is among the largest drainage basins in the world. The Caribbean Sea lies southeast of the continent. North America has a land connection to South America through the Isthmus of Panama, a link that formed only approximately 3 million years ago (Ma). In the past few million years, as ocean levels have fluctuated during glacial cycles, there has also been an intermittent land connection to Asia through Beringia. People arrived in North America by this route, and perhaps others, only very recently—sometime in the past 15,000–35,000 years. With the exception of Antarctica, the Americas were the last of the continents to be colonized by people. Although recently much influenced by people, the biota of North America evolved for hundreds of millions of years without a human presence. The current diversity of the continental biota is founded on this long period of geological and evolutionary history.

Two biological processes interact over millions of years in the evolutionary history of the continental biota: speciation and extinction. Populations, isolated from one another by changing landforms or climates, diverge

to form new species. Species go extinct in the face of changing environments or by chance events in small populations. The number of species in a continental biota at any point in time reflects in part an ongoing balance between speciation and extinction on the continent as a whole. The evolutionary history of a continent is intimately interwoven with its geological history, especially through the influence of continental drift. In the 400 million years since life appeared on land, interchange among different continental biota has been possible whenever continents have drifted into contact with one another. Species that have evolved separately on other continents have had opportunities to disperse and colonize North America. This occasional mixing of continental biota together with *in situ* evolution in North America has created the total pool of species on the continent today.

Although this pool of species defines the current magnitude of diversity in North America, the regional patterns of diversity are set by the different distributions of individual species within the continent. Some regions have more species, and others have fewer. The range of a species within a continent reflects both environmental requirements and the ability to disperse to suitable environments. Species differ in their rates of dispersal and establishment in the face of changing environments and in their ability to cross barriers such as mountains, bodies of water, or inhospitable habitat. The physiography of the continent is thus an important influence on the current patterns of species diversity within North America. Species distributions also depend on climatic patterns within the continent, which arise from both the local influence of physiographic features and the global patterns of atmospheric circulation from season to season. In the end, the complement of species found in a region depends on the ability of species to disperse to the region, establish there, and survive and reproduce there. Ecological processes acting on relatively short timescales (years to millennia) are the proximate controls on species distribution, but the availability of the species and the distribution of suitable habitats on the continent ultimately have arisen in geological and evolutionary processes acting over much longer timescales (millions of years). It is this interplay of processes on ecological and evolutionary timescales that effects the natural patterns of diversity in North America.

## II. TERRESTRIAL BIOREGIONS OF NORTH AMERICA

Mainland North America and its offshore islands can be divided into six bioregions: (i) the Canadian Shield

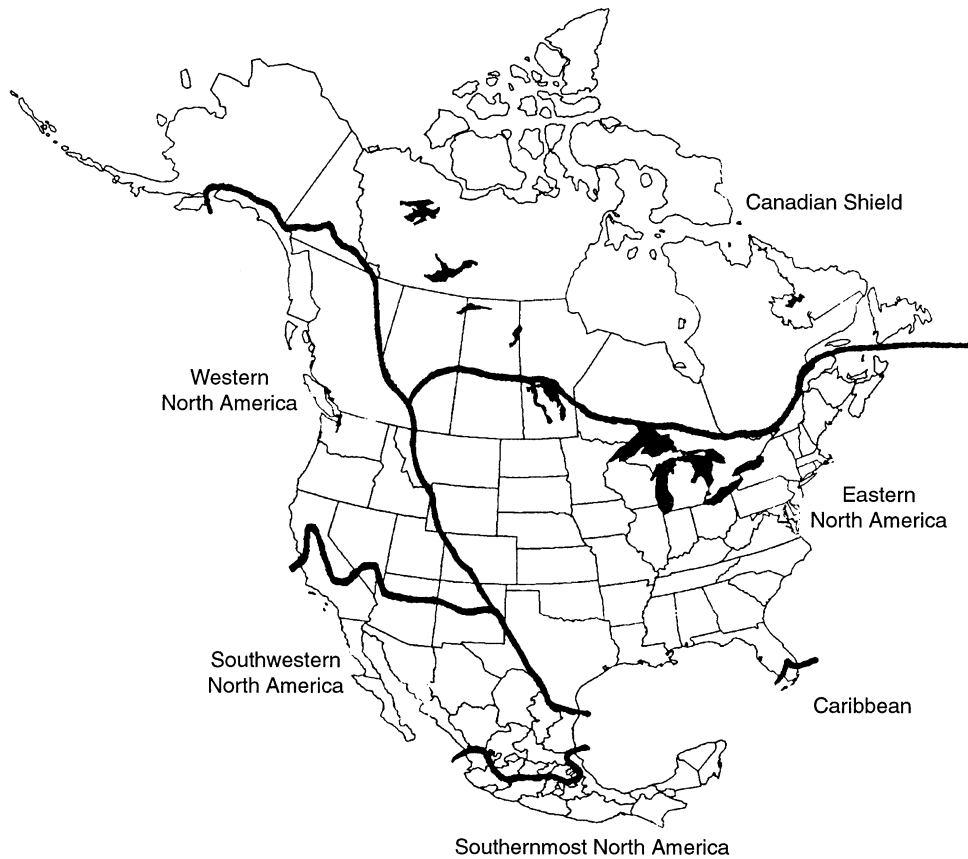


FIGURE 1 Schematic diagram of the six terrestrial bioregions of North America.

in the north, (ii) eastern North America, (iii) western North America, (iv) southwestern North America, (v) a northern extension of the Central American biota in southern Mexico, and (vi) outliers of the Caribbean biota on the southern tip of the Florida peninsula. Figure 1 illustrates and depicts in a general way the boundaries of these six terrestrial bioregions in North America. There is, of course, substantial physiographic and biotic diversity within each bioregion, and the boundaries between adjacent bioregions are not precisely identifiable. Fairly distinct ecoregions can be recognized at smaller spatial scales within each bioregion, and these are important for understanding and conserving biodiversity (Abell *et al.*, 1999; Ricketts *et al.*, 1999). Larger drainage systems, such as that of the Mississippi River and its tributaries or the Great Lakes basin, subsume many terrestrial ecoregions. The geography of the terrestrial versus freshwater biota of North America differs at the scale of ecoregions and to some degree at the scale of the continental bioregions. Similarly, the recognition of bioregion boundaries at the current con-

tinental margins is arbitrary. Continental watersheds are directly linked ecologically to the marine diversity of the continental shelf, and this shelf typically shares a geological history with the adjacent continent. Despite these uncertainties in the boundaries and internal structure of the six bioregions, each does have a reasonable coherence in its geological history, current physiography, and current biota. These bioregions therefore provide a useful working framework in which to summarize the general patterns of terrestrial and freshwater species diversity within North America.

The first three of these bioregions are part of the Holarctic, a zonal assemblage of flora and fauna that has coherence across the whole of the middle and higher latitudes of the Northern Hemisphere. Species may differ among North America, Europe, and Asia, but many families and genera of plants and animals are represented throughout the Holarctic. Higher latitudes within the Holarctic realm, including the Canadian Shield, have especially strong biotic affinities. High-latitude sites in North America, Europe, and Asia are

in close proximity around the polar region and share common climatic regimes; biota can disperse and establish among these high-latitude regions more easily than they can move between the more distant and climatically disparate latitudinal extremes within North America. The biota at midlatitudes in the Holarctic, on the other hand, are more strongly influenced by longitudinal climatic gradients and longitudinal contrasts in geologic history and physiography. Hence, we distinguish eastern and western bioregions in the midlatitudes of North America, whereas the Canadian Shield bioregion stretches across the top of the entire continent. Some biogeographers, recognizing similar longitudinal differences at midlatitudes on a global scale, divide the Holarctic into the Nearctic and Palearctic regions. North America comprises the Nearctic.

The last three North American bioregions have strong affinities with the Neotropics, the biotic realm to which South America belongs. The biota of equatorial regions do not have the global coherence that characterizes the Holarctic. Tropical regions of Africa and Southeast Asia, which have a flora and fauna that is distinct from that in the Neotropics, are distinguished as the Palearctic. The Caribbean and southernmost North American bioregions are unambiguously Neotropical, basically representing the northern extent of floras and faunas centered on the continent of South America. The southwestern North American bioregion, on the other hand, is a mix of Neotropical and Holarctic biota and is one of the most species-rich regions in the world.

### A. Canadian Shield

This bioregion takes its name from the Canadian Shield, which is the ancient core, or craton, of the North American continent. The bioregion and the craton are not geographically identical. The craton extends as basement rock well under the central and eastern parts of the continent, but this does not affect the regional development of the modern biota. The southern boundary of this bioregion is marked by the shift in surface geology from the ancient, crystalline bedrocks of the craton to younger, sedimentary bedrocks. The bioregion actually extends beyond the craton in northern Alaska and the Yukon, even though the bedrock in these western regions is not nearly as ancient. Two major biomes occur with the Canadian Shield bioregion: the boreal forest and the arctic tundra.

Tundra occupies the coasts and islands of the Arctic Ocean and higher elevations in interior Alaska and the Yukon. These barren lands with a short, cold growing season are relatively poor in plant and animal species.

From one ecoregion to another within the Canadian Shield bioregion, there are only 100–750 plant species in tundra. Some tundra ecoregions have as many as 50 mammal and 160 bird species, and from 0 to 75 butterfly species, but very few have any reptile or amphibian species. There are usually only approximately 5–25 fish species in tundra. Tundra diversity decreases seasonally as birds migrate to more southern bioregions for the winter. The most diverse tundra biota occurs at high elevations in the Wrangell–St. Elias Range in southern Alaska and the adjacent Yukon. The least diverse tundra occurs along the coasts of the heavily glaciated Axel Heiberg and Ellesmere Islands, both of which extend north of 80°N latitude. There are only approximately 110 species of vascular plants in this northernmost part of North America.

The boreal forest, which occupies the more southern parts of the Canadian Shield bioregion, generally has greater biodiversity than the tundra. Different boreal ecoregions have between 250 and 1250 vascular plant species that include 5–30 tree species. There are never more than 10 coniferous tree species in any ecoregion, even though conifers dominate much of the boreal landscape. Ecoregions in the boreal forest usually have approximately 20–50 mammal species, 105–185 bird species, 35–95 butterfly species, and 20–60 fish species. Amphibians are less diverse, usually 1–5 species in an ecoregion and occasionally more than 10; there may be 1–3 reptile species, but often there are none in a given ecoregion. The boreal forests of central and eastern Canada are the most diverse overall, and those of Newfoundland are the least diverse.

### B. Eastern North America

The Appalachian Orogeny essentially created the character of this bioregion. These mountains on the eastern margin of North America arose approximately 300 Ma in the collision of all the ancient continents to form the supercontinent Pangea. The Appalachians are in fact a part of the African continent left behind as the mid-Atlantic Rift developed beginning approximately 240 Ma, gradually breaking up Pangea, creating the Atlantic Ocean and moving Europe, Africa, and the Americas to their current geographic positions. As the young and rugged Appalachians eroded, huge deposits of sediment accumulated in the continental interior forming the bedrock of much of the interior plains that comprise the central part of this bioregion. Their sedimentary deposits also account for an extensive continental shelf, rich in marine biota today and also an important corridor for plant and animal migrations during and after glacial

peaks when ocean levels are low. From approximately 170 to 70 Ma much of the more western interior of this bioregion lay beneath shallow continental seas in which marine biota built up extensive reef systems and sediments. The Canadian Shield forms the basement rock of essentially the entire bioregion, but it is these much younger mountains and sedimentary features that influence the current patterns of biodiversity.

The eastern parts of this bioregion are generally forested with some combination of conifer and broadleaf trees. The western parts are more often grasslands, except for gallery forests along watercourses. The transition from forest to grasslands, which is more a mosaic dependent on local topography and drainage than a smooth gradient, generally occurs to one side or the other of the Mississippi River. Savannas, grasslands with a scattering of trees, are common in the transitional ecoregions. Many of these transitional ecoregions, such as the Edwards Plateau in Texas, have exceptionally high biodiversity. There are approximately 650–3350 species of plants in the different ecoregions of eastern North America, 30–85 mammals, 160–270 birds, 5–65 amphibians, 5–85 reptiles, 75–230 butterflies, 40–230 fish species, 2–260 land snails, 5–125 mussel species, and 1–65 crayfish. The mixed forests of the southeastern United States comprise the ecoregion most rich in plant species within eastern North America, and the Pine Barrens of the Atlantic Coastal Plain are the least rich. The western short grasslands are richest in mammal species, and the Florida Pine Scrub is the least rich. These western short grasslands also have the most butterfly species in this bioregion, whereas the lowland forests of the Gulf of St. Lawrence have the least. The grasslands of the western coast of the Gulf of Mexico are most rich in bird species, and the Nebraska Sand Hills are least rich. Very high numbers of both amphibian and land snail species occur in the forested Blue Ridge and southern Appalachian Mountains. The Tennessee and Cumberland River systems, which drain the unglaciated Appalachian–Blue Ridge region, are the most rich in fish, mussel, and crayfish species of any region in the bioregion; the glaciated, northern parts of the bioregion are most poor in freshwater biota.

### C. Western North America

Western North America is younger and more rugged than the northern and eastern parts of the continent, with some coastal mountains having been created only in the past 1 million years. All of western North America was created as the Farallon Plate, which has been squeezed between the Pacific and North American

Plates, gradually subducted under the North American Plate. The subduction of the Farallon Plate caused a series of major episodes of mountain building in western North America: the Sonoran Orogeny approximately 245 Ma, the Nevadan Orogeny approximately 150 Ma, the Sevier Orogeny approximately 120–90 Ma, the Laramide Orogeny approximately 70–60 Ma, and the Basin and Range Orogeny approximately 15 Ma. The continued subduction of the Juan de Fuca Plate, a remnant of the now broken up Farallon Plate, accounts for the current geologic activity in the Pacific Northwest. All this mountain building associated with the expansion of the North American continent has created and defined the essential character of this bioregion. The landscape is physiographically diverse, with a high degree of regional isolation leading to many distinct ecoregions and a biota rich in local endemics.

The mountains that dominate the landscape in the bioregion are predominantly forested, but various grasslands, shrublands, and deserts occur in basins and on plateaus amid the mountains. Between ecoregions, there is considerable variation in species diversity. There are approximately 450–2550 species of plants in the different ecoregions of western North America, 30–110 mammals, 140–240 birds, 1–20 amphibians, 0–60 reptiles, 5–225 butterflies, 0–50 land snails, 5–60 fishes, 1–5 mussels, and 0–5 crayfish. Plant species diversity is lowest on the isolated Queen Charlotte Islands and high in the Sierra Nevada, in the Great Basin, and on the Colorado Plateau. The Colorado Plateau and the Sierra Nevadas are notably rich in bird species. The Colorado Plateau and the Colorado Rocky Mountains are rich in butterfly species, whereas the Queen Charlotte Islands are very poor in butterfly species. The forested coastal mountains have the most amphibians in the bioregion, and the Colorado Plateau and the Great Basin have the most reptile species. The Siskiyou Mountains, the Sierra Nevada, and the north Central Rocky Mountains are all notably rich in species of land snails. Although low in terrestrial biodiversity within the bioregion, the watersheds of the mountains along the northern Pacific Coast are the most rich in aquatic species; the isolated and arid watershed of south-central Oregon is the least rich. The Colorado Plateau shrublands are the most species-rich ecoregion overall, whereas the Queen Charlotte Islands and the adjacent northern Pacific Coast are the most species poor.

### D. Southwestern North America

Northern and central Mexico and immediately adjacent parts of the American Southwest form a distinct biore-



gion within North America, which like western North America has rugged and relatively young terrain. The Sierra Madre Occidental and Sierra Madre Oriental running down the western and eastern coasts converge on a belt of high mountains of volcanic origin in southern Mexico, which form a natural boundary at the southern limits of this bioregion. These mountainous regions surround an interior plateau that grades north into increasingly arid deserts. The mountain ranges stand in contrast to the low relief of the coastal plains along the Pacific Ocean and the Gulf of Mexico and to the Baja California Peninsula. Ecoregions in southwestern North America are especially rich in bird, mammal, reptile, and butterfly species. Among the ecoregions, bird species number 90–280, mammals 50–110, reptiles 10–105, fishes 10–50, and there are at least 130–260 butterflies. Numbers of plant species are also high—1500–4000 species among the ecoregions—but there are as many as 8000 species along the Sierra Madre Oriental. Not surprisingly given the aridity of this bioregion in general, the diversity of amphibians is low—only 5–20 species in most ecoregions. There are insufficient data to designate an ecoregion that is the most rich in biodiversity overall, but the more diverse ecoregions are almost certainly in Mexico as opposed to the southwestern United States.

### E. Southernmost North America

The tropical highlands of the Sierra Madre del Sur and the Sierra Madre de Chiapas and the tropical lowlands on the Pacific and Gulf Coasts of Mexico comprise this bioregion, which is located at the southern limits of North America. This part of North America has a complex and poorly understood geological history involving the convergence of five plates and various crustal fragments of uncertain continental affinity. It appears certain, however, that the rugged mountains of Central America and the land connection between North and South America have developed only in the past 2–5 Ma. For the preceding 70 Ma or more, North and South America had been separated by an ocean channel deep and wide enough to severely limit biotic exchange between the two continents. Once Central America arose as a bridge between the two continents, its biota was enriched by migrations in both directions, but especially by Neotropical flora and fauna from South America that were well-suited to its tropical climate. Most ecoregions in this part of southern Mexico have on the order of 8000 plant species, although this decreases to only approximately 2000 on the Yucatan Peninsula where water quickly drains deep underground in the karst land-

scape. Some ecoregions have as many as 125 amphibian species, but most have only 10–30. Many ecoregions have as many as 120–160 reptile species, 100–300 bird species, and 80–160 mammal species. The overall diversity of this bioregion is high and increases toward the south as the influence of the South American biota strengthens. Defining the southern limit of North America at the Isthmus of Panama would greatly increase the biodiversity of the continent as a whole.

### F. Caribbean

The Caribbean Plate has a poorly understood geological history. Most of the plate is submerged in a shallow sea dotted with island arcs that are closest to North America at the Florida Peninsula. The southern tip of Florida, the only part of North America that is located in the Caribbean bioregion, has a biota that is a mix of continental and Caribbean elements. There are approximately 1000 species of plants, 25 species of mammals, 175–200 species of birds, 15–20 species of amphibians, 45–50 species of reptiles, 130–135 species of butterflies, and 60–65 species of land snails. Despite the addition of Caribbean species, the biota is not exceptionally rich compared to adjacent parts of the continent.

## III. DIVERSITY IN MAJOR GROUPS OF ORGANISMS

An alternative perspective on the general levels and patterns of biodiversity in North America is to focus on the distribution of species in particular taxonomic or functional groups of organisms. Although our knowledge of the biodiversity of North America is better than that for many other parts of the world, it is far from complete and we can only adopt this perspective for the better studied organisms in North America. A few groups of vertebrate animals are fairly well studied, but the diversity of even these is poorly known in parts of Mexico. Some groups of insects and vascular plants are fairly well studied throughout the continent, but the sheer numbers of species in these groups undoubtedly indicate that many have yet to be discovered and properly mapped. Well-founded summaries of diversity in microbial groups, many invertebrate animals, fungi, and nonvascular plants currently do not exist for North America as a whole. Despite our incomplete knowledge of North American diversity in individual groups of organisms, the numbers of species and distributional patterns in the better known groups yield insights and provide guidance for conservation efforts.

## A. Plants

There are approximately 1320 species of mosses in North America north of Mexico and approximately 1200 species known from Mexico. The liverworts and hornworts in North America north of Mexico number approximately 650 species, with about 800 species known from Mexico. All these numbers are expected to increase as the continental bryoflora is better explored. There is no comprehensive North American flora for these plant groups.

The pteridophytes (ferns, horsetails, club mosses, spike mosses, and quillworts) are among the most primitive vascular plants represented in the North American flora. There are approximately 440 species of pteridophytes in North America north of Mexico, more than 75% of which are ferns. Mexico has on the order of 1000 pteridophytes species, most of which are ferns of moist habitats and Neotropical affinity. For example, there are only approximately 125 pteridophyte species in the Chihuahuan Desert of northern Mexico compared to almost 700 species in the southern state of Oaxaca. The moist, tropical forests of southern Mexico harbor the greatest diversity of pteridophytes in North America.

The gymnosperms, an ancient group of plants that includes the pines, spruces, and other conifers, are commercially important and especially well-known in North America. There are approximately 90 conifer species in North America north of Mexico; numbers of conifers in Mexico are higher but uncertain. The well-studied pines of Mexico offer one specific point of comparison to the rest of North America. There are approximately 35 pine species in North America north of Mexico, and only about one-third of these have ranges extending into Mexico. There are another 60 species of pine in Mexico! On the other hand, there are 11 spruce species in North America, but only 2 occur in Mexico. The frequently isolated, arid mountain habitats of Mexico clearly have provided a center for the evolutionary diversification of pines in North America. North of Mexico the greatest concentration of conifer species is found in the Klamath–Siskiyou ecoregion on the mountainous coast at the California–Oregon border; approximately 30 species of conifer are found there. Conifer diversity is generally greater in the western mountains than in eastern or northern North America.

Trees in general, angiosperm as well as gymnosperm species, are well studied in North America. There are approximately 650 native species of trees in North America north of Mexico, approximately 85% of which

are angiosperms; the arboreal flora of Mexico is not known with certainty but there are at least as many, and probably substantially more, tree species as there are in the rest of the continent. The oaks, which are well studied in North America including Mexico, provide a useful point of comparison. As many as 150 species of oaks are known from Mexico, approximately 30 of which are also found farther north. There are only about 60 additional oak species in the rest of North America. The Mexican oaks predominate in arid habitats, with the mountains of central and eastern Mexico being a center of diversity. The center of tree diversity north of Mexico is in the southeastern United States, where as many as 190 tree species can be found in a given ecoregion. With the exception of Mexico, the diversity of hardwood species declines steadily to the west and north of this concentration of species in the southeastern United States.

The angiosperms, both herbaceous and woody, are the most recently evolved and most species rich of the vascular plant groups in North America and indeed throughout the world. The angiosperms comprise the largest part by far of any regional flora. They are also the least completely identified and mapped in terms of their North American patterns of biodiversity. There is no comprehensive floristic inventory for North America, only a series of regional descriptions. There may be as many as 17,000–20,000 native plant species in North America north of Mexico. At midlatitudes, the highest diversity of angiosperm species in an ecoregion is in the forests of the southeastern United States, where there are approximately 3000 native angiosperms; species diversity decreases to the west and north on the continent but increases dramatically in Mexico. The state of California, which includes many diverse ecoregions, has more than 5000 native plant species. The whole of Canada has about only 3125 native angiosperm species, and the Canadian Arctic Archipelago has only 330. Mexico, on the other hand, has approximately 19,000 angiosperm species; the state of Oaxaca alone has about 8000. Mexico has approximately 2600 of the 5500 species in the Asteraceae (sunflower family); there are only 400 native species of Asteraceae in Canada. The Fabaceae (bean family) has approximately 1700 species in Mexico, with more than half occurring only in Mexico, compared to only about 130 in Canada. The 900 species of Cactaceae (cacti) in Mexico represent more than half the species in the world. The greatest angiosperm diversity in North America is clearly found at the southern end of the continent.

## B. Butterflies

There is no reliable estimate of the total number of butterfly species in North America. Comprehensive surveys of western and southwestern North America report only approximately 530 species of butterflies, although this is believed to be from the part of the continent north of Mexico where butterfly diversity is greatest. Surveys of remnant prairie in the midcontinent yielded only approximately 90 species of butterflies, but the prairie ecosystems have been much disrupted in the past 200 years. There are estimated to be approximately 2200 butterfly species in Mexico, and it is certain that butterfly diversity within North America is greatest in Mexico. The swallowtails (Papilionidae), which are among the more conspicuous and better known butterflies, provide evidence of the relative richness of the Mexican butterfly fauna. There are almost 60 swallowtail species in Mexico, about twice as many as in the more northern parts of the continent.

## C. Land Mammals

There are approximately 4500 mammal species in the world. Canada has approximately 140 species of mammals, whereas the United States has about 350 and Mexico about 440–450. There is no good composite estimate for the continent as a whole. North of Mexico, the Colorado Plateau harbors the greatest diversity of mammal species but the Chihuahuan Desert, which extends into Mexico, is equally rich in mammal species. The species richness of mammals in North America decreases steadily to the north, and at midlatitudes it is relatively low in the east compared to the west.

## D. Birds

There are approximately 9000 bird species in the world. Canada has approximately 425 species of birds, the United States approximately 650 species, and Mexico approximately 960–1050 species. There is no good composite estimate for the number of bird species on the continent as a whole. The geographic pattern of species richness for birds in North America is generally similar to that of mammals: richest in Mexico and the adjacent southwestern United States and steadily decreasing to the north.

## E. Amphibians

The amphibians in North America primarily include frogs and salamanders. There are approximately 4000

amphibian species in the world. Canada has approximately 40 species of amphibians, none of which do not also occur farther south on the continent. Mexico, in contrast, has approximately 285 amphibian species, approximately 170 of which do not occur elsewhere in North America. The United States has approximately 230 amphibian species. There is no good estimate of the number of amphibian species for the continent as a whole. With 55–70 amphibian species, the Appalachian Mountains comprise the region north of Mexico that is most rich in amphibian species, but this is only one-fourth of the number of species in the tropical highlands of central and southern Mexico.

## F. Reptiles

The reptiles in North America include turtles, lizards, snakes, and a negligible number of crocodylians. There are about 6000 reptile species in the world. Canada has only about 40 species of reptiles, none of which do not also occur farther south on the continent. The United States has approximately 280 species of reptiles. Mexico has about 690–720 reptile species, approximately 370 of which do not occur elsewhere in North America. The Chihuahuan Desert is the part of the continent most rich in reptiles.

## G. Freshwater Fishes

North America has approximately 1200 species of freshwater fish. There is, however, substantial variation in the diversity of fishes in glaciated and unglaciated parts of the continent. The glaciated regions are notably poor in fish species. Canada, despite its very large land area rich in lakes and river systems, has only approximately 180 native fish species. The extensive Hudson Bay Basin, which was entirely glaciated 18,000 years ago, has only approximately 100 fish species and most of these occur in the southern headwaters of the drainage system. The Canadian Arctic Archipelago has fewer than 10 fish species. In contrast, the largely unglaciated Mississippi River Basin has approximately 375 fish species. The southeastern United States has 485 freshwater fish species, with the greatest diversity in the Appalachian and adjacent Interior Plateau. Despite the aridity of much of Mexico, about 380–500 freshwater fish species occur there, with notably high numbers of species in the Panuco and Papaloapan River systems along the southwestern coast of the Gulf of Mexico. The arid regions of southwestern North America had wetter climates and more extensive wetland systems as recently as 7500 years ago, but the fish fauna now is restricted

to remnant bodies of permanent water. Although the habitats available for fishes in southwestern North America are currently limited, there are many remnant species with a high degree of endemism.

#### IV. CHANGES IN NORTH AMERICAN BIODIVERSITY

There is a long history of patterns of biodiversity in North America. The native species that occupy North America today belong to groups of organisms that have developed over the long history of life on this continent. Some of the earliest life on land is represented by fossil plants more than 400 million years old that are found along the Acadian coast of North America. The continent has shared in the evolutionary diversification of life on land. Ancestors of conifers in the family Pinaceae grew in what is now North America 135 Ma. Angiosperm fossils from North America date to approximately 120 Ma, near the origins of this now dominant group of plants. Dinosaurs that dominated the animal life of North America beginning approximately 200 Ma are ancestors to the modern birds; the mammals rose to dominance beginning 65 Ma with the mass extinction of dinosaurs. Primitive insects are among the earliest fossils of terrestrial life in North America, but the rapid increases in species numbers of groups such as the butterflies and beetles that have high diversity in North America today also began only 65 Ma. Patterns of continental drift and climate change during the 400 million years that life has been on land have facilitated the exchange of biota from different continents. The strong affinity in the flora and fauna of the Holarctic regions of North America, Europe, and Asia has its origins in the period 170–80 Ma, when these continents were still in close proximity to one another. Similarly, the richness of the Central American flora and fauna in southernmost North America stems in part from biotic exchanges initiated in the past 2 or 3 Ma when the Americas were reconnected by land after 60–70 million years of relative isolation from one another. Some contemporary patterns of diversity in North America that we might seek to attribute to current environmental conditions in fact originated in the ancient history of life on the continent.

The influence of past events on contemporary patterns of biodiversity has been most marked during the past 1 million years, during which climatic changes associated with glacial cycles have repeatedly disrupted the North American biota. Species ranges in North

America have shifted dramatically from glacial to interglacial, and current distributions of species reflect patterns of dispersion from the most recent glacial refugia. The relative poverty of species in the more northern parts of North America in part stems from the failure of some species to disperse back into these glaciated regions in the 5000–14,000 years since the ice melted. Major climatic changes from glacial to interglacial periods also have influenced the biodiversity of unglaciated regions of the continent. For example, during the most recent glacial maximum parts of western and southwestern North America had a wetter climate. Large pluvial lakes and extensive river systems existed that are now mostly reduced to playas and dry channels. Today, desert pupfish occur only in isolated springs and pools, whereas once they were widespread in extensive lake and river systems throughout the Southwest. Once widespread palms now exist only as remnant populations in isolated canyons. In another cycle of changing climate, such remnant populations may again become common, but now we value them as rare species that enhance the biodiversity of North America. The rich and sometimes peculiar patterns of biodiversity in North America are built on a dynamic history of changing environments and changing species distributions.

In terms of the most recent history of the North American biota, the arrival of humans coincident with the end of the most recent glacial cycle is especially noteworthy. As the interglacial was beginning, hundreds of the larger animals in North America went extinct. It is difficult to account for these extinctions by changing climate alone, and it appears that human hunting pressures were a significant factor. In a brief period, the mammoths, mastadons, camels, horses, sabre-tooth tigers, giant ground sloths, and many other large mammals that undoubtedly were major elements in the functioning of North American ecosystems simply disappeared from the continent. The patterns of biodiversity at the time of European settlement not many thousands of years later must reflect this fairly recent mass extinction as well as the continued influence of the few million aboriginal people who already lived in North America at that time. Whatever patterns of continental diversity we identify, we should not assume that they occurred independent of human influence.

Human influence on patterns of diversity in North America has increased greatly in the past 500 years. The European colonists consciously or inadvertently brought many new plant and animal species to this continent. Indeed, people continue to introduce species from outside North America, some of which establish

and spread to natural ecosystems well beyond the confines of urban or agricultural lands. Approximately 25% of the current flora of Canada consists of alien species; the United States has approximately 3700 alien plant species, on the order of 20% of the flora. Approximately 6% of the fishes of Canada are not native species; the United States has about 75 fish species native outside of North America and approximately 200 that have been introduced outside their native North American ranges. The fishes currently dominating the ecosystems of the Great Lakes are entirely the outcome of introduction and management of nonnative species by people. Inadvertently introduced zebra mussels are now altering ecosystem function in the Great Lakes Basin and ousting native mussel species. Introduced starlings and sparrows dominate the avifauna in many settled landscapes. The landscapes and habitats of large parts of the continent have been drastically altered from pre-Columbian times. Almost all rivers in the United States have had their flow altered by dams or their drainage basin hydrology has been manipulated. Terrestrial habitats reasonably unaltered from pre-Columbian times now occupy only 48% of North America north of Mexico, and these are largely concentrated in remote mountainous or northern regions. Remnants of natural prairie and forest habitats represent only a few percent of many ecoregions in eastern North America. In western North America natural habitats are essentially obliterated in the grasslands of central California and in the interior grasslands of the Pacific Northwest. We are faced with the challenge of identifying natural patterns in the biodiversity of our native flora and fauna while the continental biota is increasingly disrupted by the introduction of alien species and the alteration of natural landscapes.

## V. CONCLUSION

From the preceding discussion, assembled from a variety of sources, one point is most clear: We have really only begun to know the patterns of biodiversity in North America. The biodiversity of no group of organisms in North America is completely characterized. Some of the groups of organisms discussed in this review are reasonably well studied, but all are incompletely known. Remote parts of arctic North America are little explored and may yield new species despite the general tendency of species diversity to decrease at higher latitudes. There are unusual habitats in the high arctic that are unlike anywhere else on Earth, such as the mineral-rich hot springs on Axel Heiberg Island

that emerge through hundreds of meters of permafrost. Such habitats, which are being studied as analogs for Martian polar environments, may well harbor previously unknown microbial life. The more heavily settled and biologically explored midcontinent seems well-known, but every year new localities and range extensions, and even new species, are recorded. The remote and rugged parts of southern North America undoubtedly harbor unknown species even in groups that have been well studied on the continent as a whole. Mexico is recognized as one of the top five countries in the world in terms of overall levels of biodiversity. By far the most species-rich part of North America, Mexico is ironically the least well studied.

In addition to our incomplete knowledge of species diversity in well-studied groups of organisms in North America, we must recognize that there are many other groups of organisms on the continent whose distribution and abundance are scarcely known. There are almost 91,000 species of insects described from North America north of Mexico, but this number is certainly only a small fraction of the total insect species on the continent. Almost all the insect groups are too little studied to assemble any sort of meaningful summary of their patterns of continental biodiversity. Similarly, there may be as many as 120,000 species of microfungi in the United States alone, only one-fourth of which have been described and far fewer have been studied sufficiently to map and inventory them for the continent. The 5000–10,000 species of macrofungi (mushrooms) estimated to occur in the United States are only marginally better known. It is too early to write the definitive treatise on patterns of biodiversity in North America. There is much exciting and worthwhile work yet to do.

## See Also the Following Articles

CENTRAL AMERICA, ECOSYSTEMS OF • SOUTH AMERICA, ECOSYSTEMS OF

## Bibliography

- Abell, R., Olson, D. M., Dinerstein, E., Hurley, P. T., Diggs, J. T., Eichbaum, W., Walters, S., Wettengel, W., Allnutt, T., Loucks, C. J., and Hedao, P. (1999). *Freshwater Ecoregions of North America. A Conservation Assessment*. Island Press, Washington, D.C.
- Boyce, S. G., and Martin, W. H. (Eds.) (1993). *Biodiversity of the Southeastern United States: Lowland Terrestrial Communities*. Wiley, New York.
- Brown, D. E. (Ed.) (1994). *Biotic Communities: Southwestern United States and Northwestern Mexico*. Univ. of Utah Press, Salt Lake City.

- Brown, D. E., Reichenbacher, F., and Franson, S. E. (1998). *A Classification System of North American Biotic Communities*. Univ. of Utah Press, Salt Lake City.
- Hackney, C. T., Adams, S. M., and Martin, W. H. (Eds.) (1993). *Biodiversity of the Southeastern United States: Aquatic Communities*. Wiley, New York.
- Harper, K. T., St. Clair, L. L., Thorne, K. H., and Hess, W. W. (Eds.) (1994). *Natural History of the Colorado Plateau and Great Basin*. Univ. Press of Colorado, Niwot.
- LaRoe, E. T., Farris, G. S., Puckett, C. E., Doran, P. D., and Mac, M. J. (Eds.) (1995). *Our Living Resources: A Report to the Nation on the Distribution, Abundance, and Health of U.S. Plants, Animals, and Ecosystems*. U.S. Department of the Interior, National Biological Service, Washington, D.C.
- Lawford, R., Alaback, P., and Fuentes, E. R. (Eds.) (1993). *High Latitude Rain Forests and Associated Ecosystems of the West Coast of the Americas: Climate, Hydrology, Ecology, and Conservation*. Springer-Verlag, New York.
- Martin, W. H., Boyce, S. G., and Echternacht, A. C. (Eds.) (1993). *Biodiversity of the Southeastern United States: Upland Terrestrial Communities*. Wiley, New York.
- Ramamoorthy, T. P., Bye, R., Lot, A., and Fa, J. (Eds.) (1993). *Biological Diversity of Mexico: Origins and Distribution*. Oxford Univ. Press, New York.
- Ricketts, T. H., Dinerstein, E., Olson, D. M., Loucks, C. J., Eichbaum, W., DellaSala, D., Kavanagh, K., Hedao, P., Hurley, P. T., Carney, K. M., Abell, R., and Walters, S. (1999). *Terrestrial Ecoregions of North America. A Conservation Assessment*. Island Press, Washington, D.C.
- Rzedowski, J. (1994). *Vegetación de México*. Editorial Limusa, México.
- Schoenherr, A. A. (1992). *A Natural History of California*. Univ. of California Press, Berkeley.
- Stehli, F. G., and Webb, S. D. (Eds.) (1985). *The Great American Biotic Interchange*. Plenum, New York.
- West, F. H. (Ed.) (1996). *American Beginnings: The Prehistory and Palaeoecology of Beringia*. Univ. of Chicago Press, Chicago.





# NUCLEIC ACID BIODIVERSITY

Tamara L. Horton and Laura F. Landweber  
*Princeton University*

---

- I. Gene Scrambling
  - II. RNA Editing
- 

## GLOSSARY

- DNA** Linear polymer of nucleotides encoding information for a cell.
- eukaryote** An organism whose cell or cells have a membrane-bound nucleus.
- mitochondrion** A eukaryotic cellular organelle which is used for cellular respiration and energy production.
- nucleus** Compartment of a cell in which DNA is stored on chromosomes.
- protein** A three-dimensional macromolecule constructed of amino acids which is formed based on an RNA sequence.
- protist** A member of a diverse collection of eukaryotes, defined only by their exclusion from the groups plants, animals, and fungi.
- RNA** A linear polymer of nucleotides which is transcribed from DNA.
- 

quence of nucleotides in a genome is often just a starting point for the construction of DNA genes and RNA transcripts, which undergo many alterations before organisms actually use the information to fashion their building materials.

Two fascinating modes of nucleic acid sequence modification are gene scrambling and RNA editing. Gene scrambling is the rearrangement of DNA segments between a transcriptionally active copy and an archived germline copy. Some genes are broken into more than 50 unordered fragments along a germline chromosome and then unshuffled during formation of an active somatic chromosome. While gene scrambling assembles existing DNA information into a new order, RNA editing of transcripts creates completely novel sequences which are not even found in the genome. Processing of edited RNAs alters the transcript length and nucleotide identity, and the transformations can render the expressed RNA form unrecognizably different from the DNA molecule of its origin. Although gene scrambling and RNA editing complicate our analysis of genomes, understanding the methods by which organisms achieve these genetic revisions gives us an appreciation for the diversity of ways by which nucleic acids can store and recombine information.

*DNA IS OFTEN DESCRIBED* as a “blueprint for life,” implying that knowledge of the primary sequence of nucleic acids in a genome can give biologists a complete picture of the organism built from these plans. However, neither life nor DNA is that simple. In reality, the se-

## I. GENE SCRAMBLING

Gene scrambling occurs when coding segments of DNA are mixed in a randomly or nonrandomly shuffled order along a chromosome. Before the stored information can



be expressed by the organism, the pieces of the gene must be cut apart and reassembled in the proper sequential arrangement. These stunning acrobatics are performed in the genomes of spirotrichous ciliates, protists which have many unique and mysterious features. However, the fundamental lessons learned from spirotrichs about the flexibility of nucleic acid storage mechanisms are universally important.

### A. Ciliates and Nuclear Dualism

Ciliates are unicellular protists closely related to the “eukaryotic crown taxa,” meaning that on most phylogenetic trees they diverge as one lineage near the neighboring cluster of plants, animals, and fungi. The ciliates are a diverse monophyletic group, with certain ciliates estimated to be as evolutionarily distant from one another as corn from rats. All ciliates share two features: a coating of cilia on their cell surfaces and two types of nuclei within single cells.

The two nuclei types in each ciliate cytoplasm are different sizes; they are called the micronucleus and the macronucleus. The tiny germline micronucleus is transcriptionally inert and functions solely in sexual exchange. In contrast, the large somatic macronucleus is responsible for gene expression, but its contents are only transmitted to clonal offspring. Ciliates reproduce asexually by fission but are capable of exchanging genetic information in a sexual manner independent of reproduction. Conjugation between ciliates leads to an exchange of haploid micronuclei, which fuse to form a zygotic nucleus (Fig. 1). The biparentally created zygotic nuclei in each mating partner form new micronuclei and macronuclei as the old macronuclei are destroyed.

### B. DNA Processing during Macronuclear Formation

The micronuclear and macronuclear genomes do not have the same chromosomal structure. As the new macronucleus is formed, diploid micronuclear chromosomes are polytenized and reproducibly broken at certain sites, portions of DNA are deleted, chromosomes are rejoined, and new chromosome ends are healed by telomerase action. DNA in the new, shortened chromosomes is differentially amplified so that copy numbers of individual chromosomes in *Tetrahymena thermophila* range from 45 to 9000. Spirotrichous ciliates tend to have higher copy numbers of chromosomes than the oligohymenophoreans such as *Tetrahymena*, and the

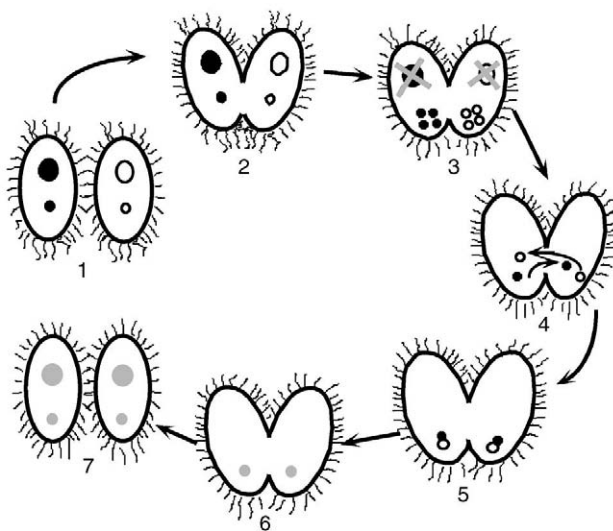


FIGURE 1 Ciliate sex. The micronuclei of conjugating ciliates undergo meiosis, exchange, and fusion to form new genetic combinations. Between steps 1 and 2, the ciliates conjugate. In the transition from step 2 to step 3, the micronuclei have undergone meiosis to form haploid micronuclei, while the old macronuclei have been destroyed. In step 4, the haploid micronuclei are exchanged, and in step 5 they fuse. By step 6, two unique diploid micronuclei are formed with genetic material from both parents. At step 7, a new macronucleus is formed from each new micronucleus.

spirotrich *Oxytricha nova* has 100,000 copies of a particular chromosome in the macronucleus. Disparities in DNA copy number relate to levels of gene expression because the most highly amplified DNAs encode the highly expressed ribosomal RNAs, and the quantities of mRNA synthesis and cognate protein secretion from *Euplotes raikovi* pheromone genes directly correlate with the genes' individual macronuclear copy numbers.

Coding sequences in micronuclear genes are called MDSs for macronuclear destined sequences (or segments), whereas IESs are internal eliminated (or excised) sequences. MDSs and IESs are superficially analogous to exons and introns, respectively, although the latter two terms only refer to a particular type of RNA processing and do not apply to this unrelated restructuring of DNA. The amount of DNA eliminated during macronuclear formation varies extensively across diverse ciliates: the oligohymenophorean *T. thermophila* eliminates approximately 15% of its micronuclear genome, whereas spirotrichous ciliates such as *Stylonychia* and *Oxytricha* eliminate up to 98% of their micronuclear genomes to create macronuclear chromosomes that bear single genes and very little noncoding sequence.

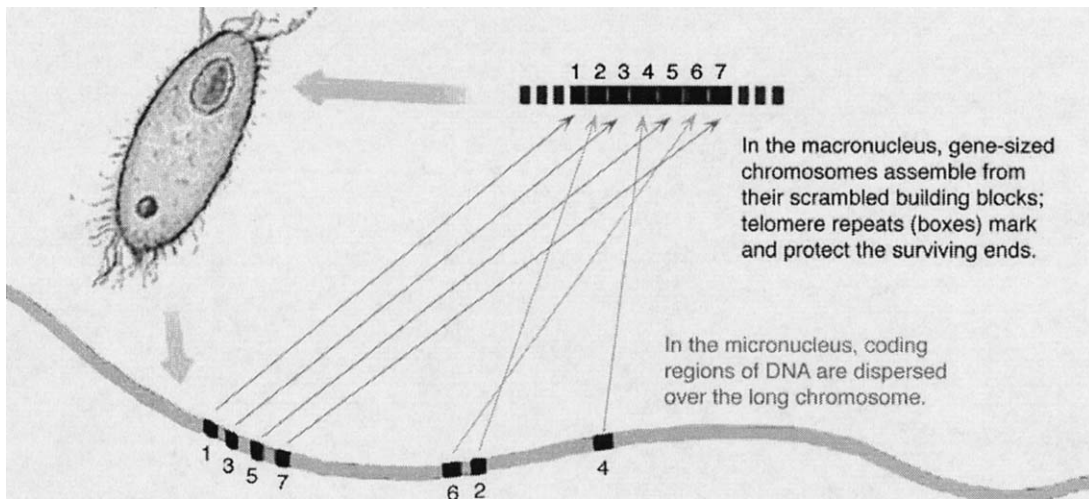


FIGURE 2 Overview of gene unscrambling. Dispersed coding MDSs 1–7 reassemble during macronuclear development to form the functional gene copy (top), complete with telomere addition to mark and protect both ends of the gene.

### C. Scrambled Genes and the Unscrambling Process

The drastic genome rearrangements of spirotrichous ciliates are not confined to the extreme quantity of DNA deletions. The protein-coding MDSs in *Oxytricha* and *Stylonychia* species are sometimes disordered relative to their final position in the macronuclear copy. For example, Prescott and colleagues found that in *O. nova*, the micronuclear copy of three genes (actin I,  $\alpha$  telomere binding protein, and DNA polymerase  $\alpha$ ) must be reordered while intervening DNA sequences are removed in order to construct functional macronuclear genes. In *O. nova*'s micronuclear genome, the MDSs destined to construct the gene for  $\alpha$  telomere binding protein ( $\alpha$ -TP) are arranged in the cryptic order 1-3-5-7-9-11-2-4-6-8-10-12-13-14 relative to their conventional order in the macronucleus of 1-2-3-4-5-6-7-8-9-10-11-12-13-14. Most impressively, the gene encoding DNA polymerase  $\alpha$  (DNA pol  $\alpha$ ) in *O. trifallax* and *S. lemnae* is apparently scrambled in 48 or more pieces in their germline nuclei (Fig. 2).

Homologous recombination at MDS boundaries probably helps guide the unscrambling process. A segment of DNA sequence at the junction between a particular MDS and its downstream IES usually matches the sequence at the junction of the next MDS and its upstream IES, leading to the correct ligation of the two MDSs over a distance. However, the presence of shared repeat regions as short as an average of 4 base pairs (bp) for nonscrambled MDSs and 9 bp for scrambled

MDSs suggests that although these recombination guides may be necessary, they are certainly not sufficient to guide accurate assembly of the genes. Hence, it is more likely that the repeats satisfy a structural requirement for MDS joining rather than perform any role in substrate recognition. Otherwise, incorrectly spliced products of promiscuous recombination would dominate genomes since 2–4 bp repeats occur many thousands of times throughout the micronucleus. This incorrect hybridization could be a driving force in the production of newly scrambled patterns in evolution. Nonetheless, if this sort of ambiguous unscrambling actually occurs during macronuclear development, then only unscrambled molecules which contain both 5' and 3' telomere addition sequences are selectively retained in the macronucleus, ensuring that most haphazardly ordered genes would be lost.

The pattern of MDSs in the micronuclear genome provides a strong clue to both the ciliates' orchestration of the unscrambling process and the mutational events which led to the scrambling in each gene sequence. For example, in the previously mentioned gene encoding *O. nova*'s  $\alpha$ -TP, the micronuclear order of MDSs (1-3-5-7-9-1-2-4-6-8-10-12-13-14) predicts a spiral mechanism in the unscrambling path to link odd and even segments in order (Fig. 3).

In contrast, DNA polymerase  $\alpha$  has at least 44 MDSs in *O. nova* and 51 in *O. trifallax*, scrambled in a nonrandom order with an inversion in the middle; some MDSs are located at least several kilobases away from the main gene in an unmapped fragment. A hairpin structure

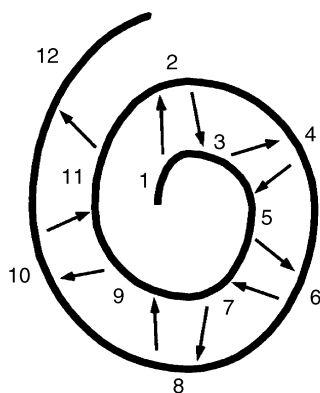


FIGURE 3 Spiral model for unscrambling in  $\alpha$ -TP (adapted from Mitcham *et al.*, 1992).

could resolve MDSs during unscrambling (Fig. 4). Comparison of the *O. nova* sequence with that of *O. trifallax* allows precise predictions for the origin of new scrambled segments (Figs. 4 and 5). Micronuclear junction sequences promote pairing between each MDS and the noncoding IES on the opposite side of a hairpin (Fig. 4). Any mutation which leads to an increased stabilization of this hairpin during macronuclear chromosome formation would also increase the chance of homologous recombination between correct segments

within the micronuclear genome. Consequently, selection would favor the appearance of more scrambled MDS segments in such a nonrandomly scrambled gene since each additional MDS adds more paired junction sequences to stabilize the hairpin necessary for unscrambling the already existing MDSs. This explanation is consistent with the additional MDSs in the DNA pol  $\alpha$  gene in *O. trifallax*. The arrangement of MDSs 2, 6, and 10 in *O. nova* could have given rise to the arrangement of eight new MDSs in *O. trifallax* (Fig. 4) by multiple crossovers in the germline micronucleus. Thus, the appearance of an inversion leads to the introduction of new MDSs in a nonrandomly scrambled pattern.

Gene scrambling in ciliates may have evolved as a product of an increased capacity for homologous recombination. The reason why it has been detected in only a restricted group of ciliates might reflect their increased levels of recombination, which generate both the scrambled arrangements and the subsequent process of unscrambling them. Although it is difficult to propose an adaptive argument for the presence and maintenance of such a complicated gene-decoding procedure, the forces that have led to ciliate nuclear dualism might be at the root of its origin. For example, Yao and Doerder independently proposed that the different roles required of the micronucleus and macronucleus,

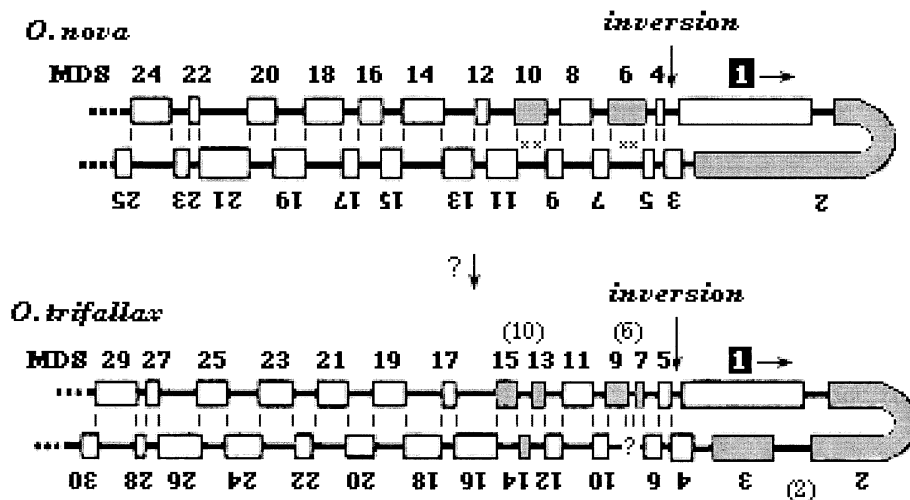


FIGURE 4 Model for scrambling of DNA pol  $\alpha$ . Vertical lines indicate recombination junctions between scrambled MDSs, guided by direct repeats. MDS 1 contains the start of the gene. MDS 10 in *O. nova* can also give rise to three new MDSs (13–15) in *O. trifallax*, one scrambled on the inverted strand, by two spontaneous intramolecular recombination events (x's) in the folded orientation shown. *Oxytricha nova* MDS 6 can give rise to *O. trifallax* MDSs 7–9 (MDS 8, shaded, is only 6 bp and was not identified in Hoffman and Prescott, 1997). *Oxytricha trifallax* nonscrambled MDSs 2 and 3 could be generated by the insertion of an IES in *O. nova* MDS 2 (similar to a model suggested by M. DuBois as cited in Hoffman and Prescott, 1997).

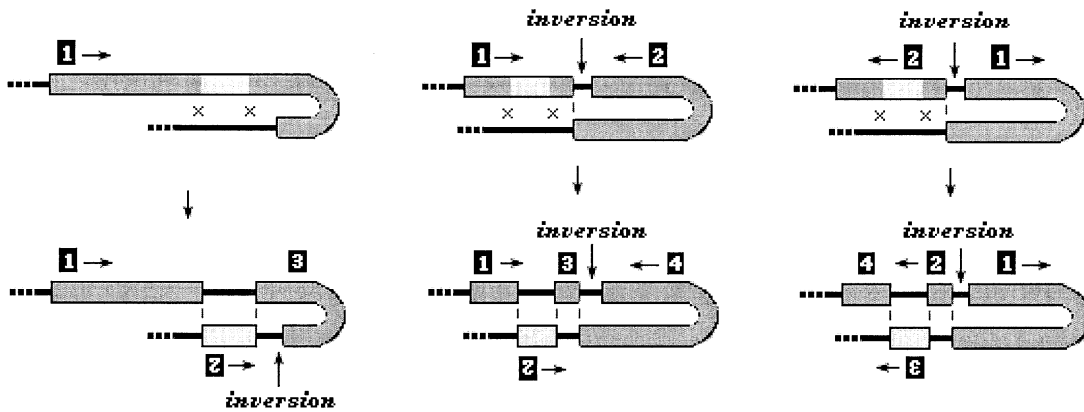


FIGURE 5 Proposed model for the origin of a scrambled gene. (Left) Birth of a scrambled gene from a non-scrambled gene by a double recombination with an IES or any noncoding DNA (new MDS order 1-3-2 with an inversion between MDSs 3 and 2). (Middle) Generation of a scrambled gene with a nonrandom MDS order from a non-scrambled gene with an inversion between two MDSs. (Right) Creation of new scrambled MDSs in a scrambled gene containing an inversion. Inversions may dramatically increase the production of scrambled MDSs by stabilizing the folded conformation that allows reciprocal recombination across the inversion.

as well as intragenomic competition, may have led to disruptive selection acting on their genetic organization and chromatin structure. The accommodation of both types of nuclei within a single cytoplasm may promote or at least permit profound differences to arise that distinguish active genes from transmitted genes. Likewise, the gene scrambling present in certain spirotrichous ciliates may be a profound exaggeration of this solution to problems presented by ciliate life.

## II. RNA EDITING

“RNA editing” is the alteration of RNA sequences by base modifications, substitutions, insertions, and deletions. The process of rewriting RNA transcripts by editing produces major effects, even adding more than half of the nucleotides in some mitochondrial transcripts of kinetoplastid protozoa. In contrast, the impact of editing can still be large even in cases in which the physical extent of editing is small. For instance, in human apolipoprotein B transcripts, replacement of a single cytidine (C) by uridine (U) results in the conversion of a glutamine codon to a stop codon; the early termination of apo B translation shortens the resultant polypeptide by one-half and removes a functional domain. Since the initial discovery of RNA editing as extensive U insertions and deletions in trypanosome mitochondrial mRNAs, many additional and apparently unrelated examples of editing have been found in organisms ranging from Ebola virus to humans. Substitution/modification

editing exists in certain nuclear and organellar RNAs among a diverse set of eukaryotes; however, insertion/deletion editing has only been found in mitochondrial RNAs of two protist groups (Table I).

Upon first inspection, the use of RNA editing in gene expression seems inefficient. Why not encode genes in their final edited form rather than require an additional revision step? However, organisms exploit the ability to edit RNA in amazingly clever ways. RNA editing allows the persistence of “deleterious” mutations in DNA genomes: New point mutations and frameshifts may persist if they are repaired at the RNA level, and DNA copies of genes and control sequences in crowded genomes may overlap but generate two or more discrete RNA sequences through editing. RNA editing also offers an array of posttranscriptional modes of genetic regulation through the formation of start and stop codons, intron splice sites, and open reading frames. Editing even permits single genes to produce multiple peptides, allowing combinatorial protein diversity.

The following sections review several models which explain our current understanding of the molecular and evolutionary basis of numerous forms of RNA editing.

### A. Kinetoplastids: Cryptogenes to Proteins by U Insertion/Deletion

Kinetoplastids are a group of unicellular protists which include the pernicious trypanosome and leishmania parasites responsible for deadly human diseases such as African sleeping sickness and Chagas disease. Trans-

TABLE I  
RNA Editing in Eukaryotes and Viruses

Organism	Genome	Class of RNA	Form of editing
Kinetoplastids	Mitochondria	mRNA	U insertion, U deletion
Myxomycetes	Mitochondria	mRNA, rRNA, tRNA	C insertion, U insertion, mixed dinucleotide insertion, C → U
Plants	Mitochondria	mRNA, rRNA, tRNA	C → U, U → C
Plants	Chloroplast	mRNA	C → U, U → C
Humans, rodents	Nucleus	mRNA	C → U
Humans, rodents, fish	Nucleus	mRNA	A → I
Humans, rodents	Nucleus	mRNA	U → C
Humans	Nucleus	mRNA	U → A
<i>Drosophila</i>	Nucleus	mRNA	A → I
Marsupials	Mitochondria	tRNA	C → U
Monotremes	Mitochondria	tRNA	U → A, U → C, A → C
Land snails	Mitochondria	tRNA	C → A, U → A, G → A
<i>Loligo</i> (squid)	Mitochondria	tRNA	G → A
Chicken	Mitochondria	tRNA	G → A
Acanthamoeba	Mitochondria	tRNA	U → A, U → G, A → G
<i>Spizellomyces</i> (fungus)	Mitochondria	tRNA	A → G, U → G, U → A, C → A
Hepatitis delta virus	Viral	mRNA	A → I
Paramyxoviruses	Viral	mRNA	G insertion (by polymerase stutter)
Ebola virus	Viral	mRNA	A insertion (by polymerase stutter)

lation of mitochondrial mRNAs of kinetoplastids is impossible without massive RNA editing of the transcripts by insertion and deletion of uridine. This editing drastically rewrites the coding regions of the transcripts by introducing and fixing frameshifts as well as creating stop and start codons. In some kinetoplastid genes, editing creates more than 90% of the amino acid codons. Since the DNA copies of many genes are barely recognizable as sources of their cognate mRNAs, they have been named “cryptogenes.”

Although all kinetoplastids display specific, reproducible editing patterns in certain mitochondrial mRNAs, comparison of the editing patterns in a particular gene transcript across a variety of species reveals a trend in the extent of RNA editing within each species. The earlier diverging kinetoplastids exhibit more editing within the cytochrome c oxidase III transcript than their later diverging relatives. This implies that this type of RNA editing is ancient within this lineage, and that its use has decreased over evolutionary time. However, the lack of U insertion/deletion editing in any other type of organism suggests that it arose specifically within the kinetoplastid lineage. Cavalier-Smith proposed that

glycosomes, special organelles found only in kinetoplastids, may provide a clue to the origin of editing in these protists. Glycosomes permit an efficient anaerobic lifestyle, during which RNA editing may have become fixed by drift in the seldom-used mitochondria of the early kinetoplastids.

Guide RNAs (gRNAs) provide the specificity of the kinetoplastid editing mechanism. gRNAs are tiny RNA transcripts which guide the editing machinery through base pairing with the mRNAs in edited regions. The gRNAs are complementary to small segments of the fully edited mRNAs and thus serve as minitemplates to guide the addition and deletion of uridines from the preedited mRNA as they form a locally double-stranded RNA helix. For instance, an unpaired A in a gRNA signals for an editing event to insert a U into the mRNA. Complete editing of a gene requires a set of many overlapping gRNAs. The affiliated set of overlapping gRNAs sequentially edit the mRNA from its 3' end to its 5' end. During editing, a complex of proteins called the “editosome” catalyzes the insertion and deletion of uridines until the length of each paired gRNA/mRNA region is maximized. The extent of G · U wobble pairs

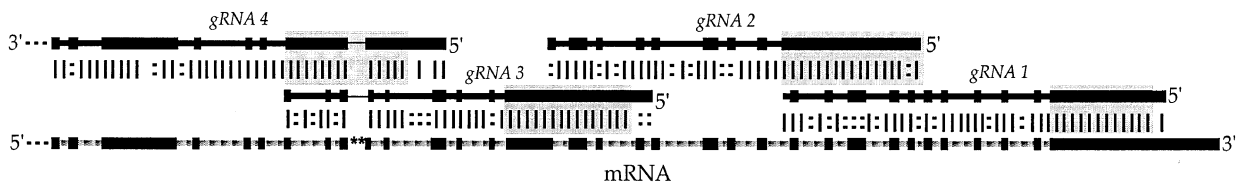


FIGURE 6 Editing of a gene region by four overlapping gRNAs. Thick lines in the mRNA are encoded in the mitochondrial DNA. Thin shaded lines are inserted U's; the two asterisks are deleted U's (Maslov and Simpson, 1992). Thin lines in the gRNAs are guide nucleotides (A or G) that pair with inserted U's. Vertical lines indicate Watson-Crick base pairs; colons indicate G:U wobble base pairs, illustrating formation of well-paired "anchors" between the 5' ends of gRNAs and the corresponding region of the mRNA.

and conventional A-U and G-C pairs of each gRNA/mRNA pair controls the cascade of editing. Each subsequent gRNA binds the mRNA more stably than the last by possessing more Watson-Crick base pairs (A-U and G-C) in the so-called "anchor region" on one end of the gRNA. These Watson-Crick base pairs displace the less stable G·U base pairs on the opposite end of the previous gRNA to dislodge the upstream gRNA and lead to an overall 3' to 5' directionality of editing (Fig. 6).

gRNAs are encoded in the kinetoplastid mitochondrial genome. Kinetoplastids are named after the unusual structure of their mitochondrial genomes, which consist of a "kinetoplast," a network of DNA inside a single, large mitochondrion located at the base of the flagellum. In one kinetoplastid subdivision, the trypanosomatids, the kinetoplast consists of a few identical maxicircles of 20–40 kb intertwined with thousands of heterogeneous minicircles of 0.5–3 kb. The maxicircles encode the mitochondrial genes and cryptogenes for respiratory proteins, ribosomal proteins, and rRNA. The only known role of the minicircles is to encode gRNAs. Bodonids, the other group of kinetoplastids, share the presence of a DNA network with the trypanosomatids but have a different structure for their minicircle homologs. In *Bodo caudatus*, the "minicircles" are noncatenated 1.4-kb circles. *Trypanoplasma borreli* has minicircle-like structures of 170 and 200 kb, with tandem 1-kb repeats encoding most gRNAs.

How did the gRNAs and minicircles come to exist as they do today? A product of mitochondrial DNA recombination provides a plausible explanation. Lunt and Hyman (1997) identified a minicircle of DNA excised from the major mitochondrial genome circle in a nematode. Similar intramolecular recombination within a kinetoplastid maxicircle might have led to the formation of minicircles encoding tiny portions of mRNAs while other unrecombined maxicircles still contained complete functional copies of the mRNA. If the minicircle copy of a gene fragment acquired a promoter that allowed transcription of an antisense

RNA, such a gene-encoding complementary RNA could have given rise to a proto-gRNA. If the protein equipment responsible for catalyzing RNA insertions and deletions simultaneously arose or was recruited from another function to serve in RNA editing, mutations in the mitochondrial mRNA could have been repaired by editing. gRNA-containing minicircles would be selectively retained in the genome, ensuring their survival.

## B. Myxomycetes: The Slimes, They Are A-changin'!

*Physarum polycephalum* is a myxomycete, or plasmodial slime mold. It takes on many shapes and sizes throughout its life, morphing from microscopic amoeba to a multinucleate syncytium which can be as large as several feet across, and then forming millimeter-scale delicate, mushroom-like fruiting bodies. *Physarum*'s mitochondrial transcripts for almost all messenger and structural RNAs require several types of RNA editing to create functional products. Editing positions are sprinkled regularly throughout the entire length of each edited transcript, spaced an average of 27 bases apart, but never closer than nine nucleotides. The majority of editing events are insertions of cytosine, but there are also a small number of specific and reproducible insertions of uridines and dinucleotides into sequences. Although there appears to be no consensus sequence that defines the hundreds of insertion sites, they usually follow a purine/pyrimidine dinucleotide and show significant preference for third codon position. Visomirski-Robic and Gott (1997) found that under conditions of stalled RNA polymerization, sites only 14–22 nucleotides from the 3' end of the RNA have correct insertion editing, suggesting that the editing reactions closely follow transcription and they may even be coupled. The tight association of transcription and editing yields a high editing efficiency; unedited or partially edited transcripts are rarely detected. *Physarum* insertional editing must occur within a narrow window of possibility:

CUAUUUUUAGUCUGCAUCUAGCUGGUGUUUCUUCUAUGUUAGGUGCUAUCA  
 AUUUCAUUUGUACCAUUAAAAUAUGCGUCUUAAAGGAUUAACAGGAGAAC  
 GUUUAUUUUUAUUGUUUGGGCUGUAUUAGUAACUGUGAUUUUAUUAUJAC  
 UUUACUGCCUGUCUUAGCAGGUGCUAUCACUAUGUUAAUUAACUGAUCGUA  
 AUUUUAAUACAUCUUUUUUUGAUGCAACCGGUGGUGGAGAUCUUAUUUUAU  
 AUCAACAUUUUGUUUUGGUUUUUUGGCCAUCCAGAAGUUUACAUUUUAAUUU  
 UACCUGGUUUUGGUAUCGUUUCUAUUUAUUUAAGCCCUAUGCUAAUAAAAG  
CUAUUUUUGGUUAUUUAGGUAUGGUGUAUGCUAUGUUGUCUAUUGGUAUCU  
 UGGGUUUUAUAGUGUGGGCUCAUCAUAUGUAUACUGUAGGAUUGGAUGUGG  
 AUACUCGCGCUUAUUUCACCGCUGCUACUAUGAUCAUUGCUGUGCCAACCG

FIGURE 7 RNA editing in *Physarum polycephalum*. Partial *P. polycephalum* mRNA sequence for gene encoding cytochrome c oxidase subunit I, with underlined text indicating inserted nucleotides and boxed text indicating C → U substitution events (from Gott *et al.*, 1993).

If editing is restricted *in vitro* by low concentrations of “the required nucleotide” CTP, and CTP is subsequently restored, then sites that were “missed” by the editing process never get edited.

Myxomycete insertional editing seems distinct from the uridine insertion/deletion editing found in the kinetoplasts. This is supported by the dissimilar pattern of edited sites, the different identity of nucleotides involved, and *Physarum*'s apparent lack of gRNA-like template molecules. Although these differences set myxomycete editing apart, another remarkable feature of myxomycete editing makes it unique among all other editing systems. Myxomycetes are the only organisms known to combine multiple types of editing within the same transcript. For example, *P. polycephalum*'s 1.5-kb transcript encoding cytochrome c oxidase subunit I (coI) undergoes 59 C insertions, a single U insertion, three dinucleotide insertions, and, astoundingly, four sites of C → U base conversion (Fig. 7). C insertion has been separated from C → U conversion in isolated mitochondria, implying separate mechanisms or components. We have found that insertional and base conversion editing have distinct evolutionary histories. The coI gene of *Stemonitis flavogenita*, another myxomycete, shares all three types of insertional editing with *P. polycephalum* but lacks the C → U conversion editing in this transcript. Even the three types of insertional editing did not all evolve simultaneously because U insertional editing is found in all myxomycetes to date, whereas C insertion and dinucleotide insertion are found only in some slime molds.

The sites of C → U editing in *P. polycephalum* are all in first or second codon position, similar to the

case in plant organelle editing. Also similar is that the process of editing restores the conserved peptide sequence in this region so that the unedited *S. flavogenita* transcript codes for the same amino acids at these positions as does the edited *Physarum* transcript. In contrast, the role of the inserted nucleotides, usually in third codon positions, may be primarily to restore the correct reading frame since they rarely change crucial coding information.

### C. Plant Organelles: C → U and U → C Editing Restores Conserved Amino Acids

Plant mitochondria employ rampant RNA editing. In nearly every mitochondrial mRNA, RNA editing converts many cytidines to uridines, and some mitochondrial mRNAs have uridine to cytidine conversions as well. Several plant chloroplast RNAs also show similar editing patterns. Although the mechanism for plant organellar editing is not fully understood, when CTP residues of *in vitro* transcripts were labeled on their  $\alpha$ -phosphate and on their cytosine base, the labels were retained after editing. This means that the C → U changes occur through a deamination of cytidine rather than a base or nucleotide substitution. Plants with different nuclear genotypes exhibit different degrees of editing of transcripts from identical mitochondrial genomes, indicating that at least part of the editing machinery is nuclearly encoded. Although no consensus sequence for editing site recognition has been determined, recent analysis (Bock *et al.*, 1997; Williams *et al.*, 1998) of natural mutants and creation of mutant sequences indicate that the specificity for editing sites in both organelles lies in

the local upstream primary sequence rather than in the predicted folded conformation of an edited region of RNA.

RNA editing is present in both the mitochondrial and the chloroplast genomes of all land plants, but it is absent in green algae and the liverwort *Marchantia polymorpha*. This distribution suggests that the plant editing systems share common components and may have arisen simultaneously in both organelles or have been transferred from one to the other. However, although the editing seems to appear suddenly in the land plant lineage, the degree of mitochondrial or chloroplast editing does not correlate with phylogenetic position, indicating multiple evolutionary losses and gains of editing at particular gene sites.

Plants which possess the ability to perform base conversion may exploit the conversions as a mechanism for repairing sequences disturbed by mutational drift. The observation that most of the edited sites in plant genes are in first or second codon positions supports this hypothesis because editing produces nonsynonymous changes in amino acids. Furthermore, Malek *et al.* (1996) found that the level of mitochondrial editing in plants correlates with G/C content. Indeed, Lu and associates (1998) determined that editing “reverses” all nonsynonymous U → C substitutions at the RNA level in *coI* genes from eight gymnosperm species, eliminating almost all variation in the predicted protein sequences. Editing can regulate translation or enzyme activity; it creates stop and start codons and proteins of varying function. However, the pressure toward loss of editing appears to be greater than any selective advantage it confers since Shield and Wolfe’s (1997) comparison of edited sites over many plant species reveals a high rate of mutation of the genes toward the edited (uridine) nucleotide.

#### D. Mitochondrial tRNA Editing: Acceptor Stems and Anticodons

Mitochondrial genomes of some organisms seem pressed to economize space. “Junk” DNA is held to a minimum, and genes immediately about their neighbors. Some tRNA genes overlap their 5′ and 3′ extremities. Base pairing of a tRNA’s 5′ and 3′ ends is essential to form an acceptor stem for aminoacylation. In six described cases of tRNA editing, mitochondria repair incompletely paired tRNA acceptor stems by exchanging mismatched bases on one side of the acceptor stem for ones that complement those bases present on the opposing stem. Thus, editing and the presence of an

internal acceptor stem template allow the genome to repair mutations that may occur in response to selection for either nucleotide compositional bias or sequence compression.

In land snails, squids, and chickens, the overlapping ends of certain tRNAs disturb the crucial base pairing of the acceptor stems. RNA editing activity restores complementarity to these tRNAs by replacing mismatched 3′ guanosine, uridine, and cytidine nucleotides with adenosines. In a platypus tRNA, there are three exchanged nucleotides (U → A, U → C, and A → C) in the 3′ half of the stem to complement the 5′ half of the acceptor stem. There is also editing of mitochondrial tRNA acceptor stems in the amoeboid protist *Acanthamoeba* and the fungus *Spizellomyces*, using the second half of the stem as guides. However, the editing mechanism in these two organisms may be unrelated to the tRNA editing described in animals since it occurs on the 5′ half of the stem. Editing of the *Acanthamoeba* and *Spizellomyces* tRNAs exchanges uridines, adenosines, and cytidines in the first three 5′ nucleotides for purines complementary to a corresponding pyrimidine on the 3′ template side of the stem.

In opossums and other marsupials, the DNA encoding mitochondrial tRNA<sup>Asp</sup> has the wrong anticodon. Although the rest of the tRNA has a canonical tRNA<sup>Asp</sup> sequence, the anticodon, GCC, will pair with glycine codons rather than aspartic acid codons. About half of the tRNA<sup>Asp</sup> transcripts are not edited and consequently are aminoacylated with glycine. The remainder are edited by a C → U change that restores the GUC aspartic acid anticodon. These are charged correctly with aspartic acid since this anticodon is used as a determinant of tRNA charging. Surprisingly, the standard tRNA<sup>Gly</sup> (with a UCC anticodon) is present in the genome as well, and it could theoretically recognize all four GGN glycine codons by wobble. However, a mutation just outside of the anticodon region of this tRNA<sup>Gly</sup> renders it incapable of recognizing the two glycine codons which the mutated tRNA<sup>Asp</sup> can bind.

Boerner and Pääbo (1996) suggest that marsupial mitochondrial editing became fixed through two mutational steps. First, a mutation occurred in the anticodon for tRNA<sup>Asp</sup>, transforming it into a functional tRNA<sup>Gly</sup>. Genomes with this mutation were still viable since the altered tRNA could regain its necessary function as tRNA<sup>Asp</sup> through editing. A subsequent mutation in the original tRNA<sup>Gly</sup> was not lost because of redundancy in tRNA<sup>Gly</sup> activity: A mutation outside the anticodon of the original tRNA<sup>Gly</sup> eventually left it unable to recognize the two glycine codons that the newly mutated



tRNA<sup>Asp/Gly</sup> could translate. After this second mutation in the original tRNA<sup>Gly</sup>, back-mutation of the tRNA<sup>Gly/Asp</sup> anticodon (to form simply a tRNA<sup>Asp</sup> anticodon) would be deleterious since the mutant tRNA<sup>Gly/Asp</sup> serves double duty by pairing with both aspartic acid and glycine codons. Thus, RNA editing permitted a deleterious change in a genome and then became fixed when a second mutation made editing a requirement for expression of mitochondrial proteins.

### E. Human Apo B and NF1: Shorter Peptides by Editing

Apolipoprotein B is present in two different forms in both humans and rodents. The long form, apo B100, is part of very low-density lipoprotein particles which have a role in cholesterol metabolism, whereas the short form, apo B48, contributes to chylomicrons that transport dietary lipids. The two forms of the apo B protein are actually encoded by a single gene which undergoes tissue-specific C → U RNA editing at nucleotide position 6666. The editing converts an encoded glutamine codon into a stop codon to create the short form of the peptide.

A complex of proteins edits the apo B transcript. The main catalytic peptide is apo B RNA editing cytidine deaminase subunit 1 (APOBEC-1). APOBEC-1 expression varies throughout development, and apo B editing levels vary correspondingly. Specificity of the cytidine deamination is determined by the primary sequence of the apo B RNA transcript, particularly an 11-base "mooring sequence," located just four nucleotides downstream of the edited site. In addition to this required mooring sequence three other "efficiency" elements are found in the 140-base region surrounding the edited cytidine. These upstream and downstream elements increase the effectiveness of the editing reaction.

APOBEC-1 is part of a family of cytidine deaminases. The family is divided into two groups of larger and smaller deaminases with various structural features in common. APOBEC-1 is categorized as a member of the group of larger deaminases, as is *Escherichia coli* cytidine deaminase (ECCDA). The two enzymes are approximately the same size, form homodimers, and share structural features such as the carboxy-terminal core domain, which is absent in the smaller, homotetramer-forming deaminases. ECCDA catalyzes the deamination of single cytidine nucleosides as part of bacterial biosynthetic pathways. Comparison of the APOBEC-1 primary sequence with the ECCDA crystal structure predicts the presence of an additional hollow space

within the APOBEC-1 structure. Navaratnam and associates (1998) noted that this cavity is the correct size and shape to accommodate an RNA transcript, which suggests how this type of editing might occur.

Cytidine deamination may even play a role in human tumorigenesis. For example, an imperfect APOBEC-1 mooring sequence and efficiency elements are present within the coding region of the neurofibromatosis type I (NF1) tumor suppressor gene. Despite slight differences in position and sequence context, normal individuals exhibit very low levels of editing. In comparison, tumor tissues from NF1-affected individuals showed more than eight times the normal quantity of edited transcript. The NF1 gene encodes the neurofibromin protein, which is a putative homolog of yeast proteins in the *ras* signal transduction pathways. The proposed GTPase activating domain of the neurofibromin lies just downstream of the NF1 editing site, and indeed C → U RNA editing at the site transforms an arginine codon into a premature translation stop upstream of the domain. Thus, the editing of the NF1 transcript most likely cripples the tumor-suppressing activity of neurofibromin.

### F. A → I Deamination: One Gene, Many Peptides

Deamination of another type alters additional RNA sequences of humans and rodents. Removal of adenosine's amino group yields inosine, a nucleotide that the translation machinery reads as guanosine. Several such A → I transitions cause predicted amino acid changes in glutamate and serotonin receptor subunits in the human and rodent brain. Teleost fish also share one of these A → I editing sites in their glutamate receptors. These editing events adjust the calcium channel permeability and the speed of the desensitization response in glutamate receptors. In the serotonin receptors, protein loops encoded by the edited region of the RNA interact with G proteins to turn on signaling pathways. The editing, along with alternative splicing, allows the exquisite fine-tuning of neural responses by increasing combinatorial protein diversity available from a single transcript. Posttranscriptional regulation demands the expenditure of less time and energy than transcription of multiple gene sequences.

A family of enzymes called "ADARs" (adenosine deaminase acting on RNA) are the candidates for A → I editing activity based on *in vitro* experiments. At least three different ADAR enzymes are expressed differentially in human tissues, and their specific targets have not been definitively resolved *in vivo*, although they

display discrete editing abilities at various mRNA sites *in vitro*. The double-stranded RNA regions required for ADAR action are elegantly provided by pairing of the exonic sequences with intronic sequences in the pre-mRNAs. Neither mature mRNAs whose introns have been removed by splicing nor pre-mRNAs with mutations disturbing intron/exon complementarity act as substrates for editing in transfected cell lines.

Mice genetically disabled for A → I editing at a single site demonstrate the importance of RNA editing in glutamate receptors. Brusa *et al.* (1995) removed an intron region crucial for creating a double-stranded editing substrate from the mouse *Glur-B* gene. Heterozygotes for this editing-disabled allele displayed severe epileptic seizures and premature death within 3 weeks of birth. The decrease in edited mRNA levels led to a fivefold increase in the calcium flow into their neurons and serious damage to their nervous systems.

Vertebrates are not alone in using the A → I editing activity to generate diversity in nervous system components. Recent work (Petschek *et al.*, 1996; Smith *et al.*, 1996) reveals that two *Drosophila* mRNAs have apparent A → G substitutions between their genomic and RNA copies. The *4f-rnp* gene encoding a putative RNA-binding protein is expressed in several alternatively spliced variants during fruit fly development. The adult form, expressed heavily in the central nervous system, has probable deamination editing of 263 adenosine sites. Another *Drosophila* gene, which encodes a subunit of a calcium channel protein, has seven putative A → I editing sites, five of which alter codon identity, in addition to many alternatively spliced forms.

Scott (1995) suggests that ADARs might have evolved for an antiviral function, destabilizing RNA genomes by modification and subsequent unwinding. Interestingly, a devious viral genome coopted this function and uses host adenosine deaminases to its own benefit. Hepatitis delta virus (HDV), a single-stranded RNA subvirus of hepatitis B, has only one known gene product, the delta antigen. The short form of delta antigen stimulates replication of the genome, whereas the long form of the antigen suppresses replication and is necessary for packaging HDV. The short and long forms of the delta antigen differ only by a single A → I base change of the negative strand, or antigenome. The base conversion replaces a stop codon at the end of the shorter peptide's mRNA with a tryptophan codon to allow read through of the longer peptide. *In vitro*, ADARI is capable of inducing this A → I change on the antigenome of the HDV virus, suggesting that either this or a related enzyme edits HDV *in vivo*. In exquisite control of the reaction, the long-form delta antigen acts

as a repressor of editing in a negative feedback loop. The repression prevents accumulation of unduly high levels of long-edited antigen, which would lower the level of viral replication.

### G. Paramyxoviruses and Ebola Virus: Polymerase Stutter Unites Reading Frames

Two other types of viruses also capitalize on RNA editing processes as a method of increasing information storage while under pressure for genome size constraint. The RNA genomes of the paramyxoviruses have several overlapping genes. Some of these genes are activated for expression by ribosomal choice, whereas others become available by cotranscriptional RNA editing. The polymerase may slip as it travels through a purine run, adding from one to six extra guanosines, depending on the sequence present in the particular virus species. The stuttering polymerase thereby fuses coding regions together to make extended versions of genes.

Many viruses use an apparently similar mechanism to add poly A tails to mRNA transcripts. When the RNA polymerase reaches the U-rich regions, it adds additional nontemplated adenines to create the tails. In paramyxoviruses, RNA editing activity may be related to a process that corrects genome length, maintaining length in multiples of six nucleotides. The substrate for the RNA polymerase is the RNA genome complexed with the capsid in hexamer-length segments. Hausmann and colleagues (1996) note that if the genome is not a multiple of six, the polymerase inserts or deletes guanosines or adenines from the same region in which the RNA editing occurs.

Another RNA virus, the infamous Ebola virus, uses a comparable mechanism to insert adenines into a sequence to unite two reading frames. In this case, the sequence of the site where adenine addition occurs appears similar to viral poly A addition sites. In contrast to the paramyxovirus editing, the additions occur even when the RNA is produced by a *nonnative* polymerase and thus must be intrinsic to the template sequence or structure. T7 transcription of the Ebola mRNA *in vitro* results in edited product, although in a lower quantity than when transcribed by Ebola's RNA polymerase.

Thus, once considered a molecular anomaly unique to protists, RNA editing is now recognized as a vital part of gene expression in a wide distribution of eukaryotes and their viruses. Organisms can alter RNA sequences through many different mechanisms by recruiting enzymes such as deaminases, nucleases, ligases,

and special polymerases to specific RNA editing tasks. Editing clearly arose multiple times in evolutionary history in the various forms of both insertional and substitutional editing, all of which profoundly affect the expression of RNAs, the products they form, and the biological processes in which they participate. RNA editing is significant in cancers, cholesterol regulation, and neural function in vertebrates as well as in the *de novo* creation of coding sequences from obscured mitochondrial genes.

Furthermore, the impact of editing on the genomes which employ it is astounding. Genomes that use editing may become increasingly lenient toward persistence of deleterious mutations. Large amounts of genetic information and control devices may be confined to small spaces. Lastly, RNA editing, like gene scrambling, establishes a device for generating combinatorial sequence diversity.

### Acknowledgments

We thank Laura A. Katz and Andrew L. Goodman for helpful discussions. TLH is supported by a National Defense Science and Engineering Graduate Fellowship.

### See Also the Following Articles

DIVERSITY, MOLECULAR LEVEL • GENES, DESCRIPTION OF • GENETIC DIVERSITY

### Bibliography

- Arts, J. G., and Benne, R. (1996). Mechanism and evolution of RNA editing in kinetoplastida. *Biochim. Biophys. Acta* 1307, 39–54.
- Bass, B. L. (1993). RNA editing: New uses for old players in the RNA world. 418 In *The RNA World* (R. F. Gesteland and J. F. Atkins, Eds.), pp. 383–418. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Börner, G. V., Yokobori, S., Mörl, M., Dörner, M., and Pääbo, S. (1997). RNA editing in metazoan mitochondria: Staying fit without sex. *FEBS Lett.* 409, 320–324.
- Chan, L. (1995). Apolipoprotein B messenger RNA editing: An update. *Biochimie* 77, 75–78.
- Covello, P. S., and Gray, M. W. (1993). On the evolution of RNA editing. *Trends Genet.* 9 (8), 265–268.
- Grosjean, H., and Benne, R. (Eds.) (1998). *Modification and Editing of RNA*. ASM Press, Washington, DC.
- Landweber, L. F., and Gilbert, W. (1993). RNA editing as a source of genetic variation. *Nature* 363(6425), 179–82.
- Landweber, L. F., Kuo, T.-C., and Curtis, E. A. (2000). Evolution and assembly of an extremely scrambled gene. *Proc. Natl. Acad. Sci. USA* 97, 3298–3303.
- Prescott, D. M. (1994). The DNA of ciliated protozoa. *Microbiol. Rev.* 58(2), 233–267.
- Prescott, D. M. (1999). The evolutionary scrambling and developmental unscrambling of germline genes in hypotrichous ciliates. *Nucleic Acids Res.* 27(5), 1243–1250.
- Schuster, W., Hiesel, R., and Brennicke, A. (1993). RNA editing in plant mitochondria. *Cell Biol.* 4, 279–284.
- Seeburg, P. H., Higuchi, M., and Sprengel, R. (1998). RNA editing of brain glutamate receptor channels: Mechanism and physiology. *Brain Res. Rev.* 26, 217–229.
- Smith, H. C., Gott, J. M., and Hanson, M. R. (1997). A guide to RNA editing. *RNA* 3, 1105–1123.



# OCEAN ECOSYSTEMS

Richard T. Barber  
*Duke University*

---

- I. History of the Ecosystem Concept
  - II. Utility of the Concept
  - III. Partitioning the Ocean into Natural Functioning Units
  - IV. Characteristics of Ocean Ecosystems
  - V. Ocean Ecosystems and Global Change
- 

## GLOSSARY

**convective mixing** Vertical mixing produced by the increasing density of a fluid in the upper layer, especially during winter in temperate and polar regions.

**euphotic zone** The surface layer of the ocean in which there is adequate light for net positive photosynthesis.

**nutrients** Dissolved mineral salts necessary for primary productivity and phytoplankton growth: Macronutrients are phosphate, nitrate, and silicate; micronutrients are iron, zinc, manganese, and other trace metals.

**phytoplankton** Photosynthetic, usually single-celled, plants that drift with ocean currents.

**pycnocline** The layer in which density changes most rapidly with depth and separates the surface mixed layer from the deep ocean waters.

**Southern Ocean** The circumpolar ocean in the Southern Hemisphere between the Subtropical Front and the continent of Antarctica.

**stratification** The formation of distinct layers with different densities; stratification inhibits mixing.

**subpolar** Pertaining to the regions between the polar

and temperate zones, but for the oceans the boundaries are the Subtropical Front and the Polar Front.

**subtropical** Pertaining to the regions which, under the influence of the trade winds, are permanently stratified.

**upwelling** Upward vertical movement of water through the bottom of the surface mixed layer produced by a divergence at the surface.

**zooplankton** Animals that float or drift with ocean currents: Microzooplankton are protozoan plankton that graze on small phytoplankton; mesozooplankton are crustaceans that graze on larger phytoplankton such as diatoms.

---

**AN ECOSYSTEM IS A NATURAL UNIT** in which physical and biological interactions cause the organized flow of energy, mass, and information to have characteristic trophic, mass, and information cycles, and a successional pattern. Ecosystems have some degree of internal homogeneity, objectively definable boundaries, and a predictable pattern of temporal development. Ocean ecosystems are those ecosystems that exist in the open ocean more or less independent of solid substrates.

## I. HISTORY OF THE ECOSYSTEM CONCEPT

Natural scientists have long recognized that beyond populations or communities there is a higher level of

organization, but it has been difficult for physical and biological scientists to reach consensus about its nature. The origins of the ecosystem concept are indistinct and, indeed, involved polemics between some of the principals. Ernst Haeckel (1838–1919) coined the word “ecology” and he envisioned ecology as a science that studies the environment as a stage within which selection pressures and adaptations affect the evolution of species. Haeckel brought the physical environment to the forefront, but he was interested primarily in evolution, which acts on species or genetically interacting populations. Victor Hensen (1835–1924) coined the word “plankton” to describe the organisms that drift in the ocean following the paths dictated by currents and mixing. Hensen was interested in the integrated working of natural systems, the subject today called “ecosystem science.” He was particularly interested in the functional behavior of ocean systems from the smallest phytoplankton up through the food web and leading finally to fish, birds, and marine mammals. Hensen and Haeckel had a long professional dispute which seems ironic today because the eventual synthesis of their ideas by others led indirectly to the ecosystem concept.

The linguistic route to the modern word “ecosystem” was tortuous. From 1877 to 1939 the words “biocoenose,” “microcosm,” “naturkomplex,” “holocen,” and “biosystem” were all proposed as names for this idea. In 1935 an English scientist, Tansley, introduced the neologism “ecosystem” and this term rapidly eclipsed competing names, but ecosystem science was not a well-organized endeavor until 1959, when Eugene Odum and Howard Odum published a groundbreaking textbook, *Fundamentals of Ecology*. There were many treatises on ecology before the Odums’ book, but their book stressed the principle that the ecosystem (not the population or community) is the fundamental unit of ecology and indeed of biology. The 1959 text, which develops this concept in each of its chapters, has remained in print for many years. Although Eugene Odum’s definition of ecosystem (given at the beginning of this article) is widely accepted, the principle of the ecosystem as the fundamental unit of ecology is controversial. Populations, communities, adaptation, and evolution have all been explosively successful research areas. Ecosystem research, because of its inherent complexity and requirement for interdisciplinary work, has not enjoyed the success of the more reductionist areas of ecological research. The study of ecosystems is usually supported by encouraging words, but our institutions and agencies, which are organized along disciplinary lines, have trouble coping with a science composed of

equal parts of physics, chemistry, and biology. Ecosystem studies are placed with the life sciences, in which there is little institutional or agency enthusiasm for the complex and expensive atmospheric and physical oceanographic work required by the study of ocean ecosystems.

## II. UTILITY OF THE CONCEPT

The ecosystem concept was originally developed for terrestrial, intertidal, and benthic habitats, in which ecological succession on timescales of 10–100 years is an obvious and salient characteristic of the ecosystem. Succession on this timescale is difficult or impossible to detect in the fluid medium of pelagic ocean ecosystems. In the ocean, winds, mixing, and currents appear to reset the successional time clock to time zero each annual cycle or, perhaps, with each storm or passage of a front or large eddy. It has been shown that there are low-frequency (over the timescale of 10–100 years) changes in species abundance and community structure. These biological changes are sometimes so far reaching that they are called “regime shifts,” the term used by John Steele in 1998, but they are not functionally analogous to succession, which is (i) orderly and directional, (ii) involving biological modification of the physical environment, and (iii) resulting in a more stable climax community (Odum, 1969). Succession in ocean ecosystems is dramatically evident on the day to month timescale after spring stratification in temperate and subpolar waters or after an episode of upwelling. This short timescale succession, which appears to be cyclic, is not unidirectional and definitely not orderly.

If long-term directional succession does not appear to occur in pelagic ocean ecosystems and succession is a central tenet of ecosystem theory, isn’t the very existence of ocean ecosystems in doubt? The answer is that ocean ecosystems, as Steele (1985) noted, are clearly different from terrestrial, aquatic, intertidal, or benthic ecosystems, but ocean ecosystems meet most of the requirements of the definition set forth by Eugene Odum. The ocean ecosystems described here all have a characteristic and distinct trophic structure, characteristic and distinct material cycles, some degree of internal homogeneity and commonality, and definable hydrographic boundaries.

There is considerable heuristic power in the ecosystem concept because understanding gained in one ocean ecosystem can be used to predict the response of another ecosystem of the same kind that is geographically distinct from it. This predictive power is perhaps the

greatest benefit of ecosystem theory and provides evidence that each distinct kind of ocean ecosystem has characteristics that can be generalized and used in prediction.

Pomeroy and Alberts (1988) emphasized that the concept involves emergence of new properties. One consequence of the hierarchic organization of ecosystems is that as components (both biotic and abiotic) are integrated into a larger functional unit, new properties emerge that cannot be detected by study of the component populations or processes, no matter how thorough the reductionist study. This aspect of ecosystem theory makes the concept useful, even necessary, for predictive understanding of ocean ecosystems.

Determining specific emergent properties of a system as large as ocean ecosystems is difficult. Ocean ecosystem spatial domains of thousands of kilometers are difficult to sample adequately with ships; however, with satellites that measure wind, ocean temperature, ocean currents (from sea surface topography), and phytoplankton biomass, a new era has begun in understanding ocean ecosystems. Are there appropriate benefits to justify the large societal investment in remote sensing and data handling required to achieve this new level of understanding? One benefit deals with fisheries management. The historic approach to management of these fisheries has involved analysis of local populations and environment. It has become clear that ecosystem properties, especially physical conditions, operating on scales much larger than the range of the exploited population, may be responsible for changes in reproductive success and adult abundance. In this context, a valuable societal payoff of understanding ocean ecosystems is improved ability to predict local variations in living resources as shown by Sherman and others in 1986. Odum repeatedly emphasized that ecosystem science and economics are parallel disciplines and expressed regret that they are not perceived as such, particularly by the economic community. Investors and political leaders making policy decisions on resource development need an understanding of the probability, frequency, and intensity of natural variability for realistic economic decision making. In this context, economics and the ecosystem concept are closely related: By understanding the variability of ocean ecosystems, decision makers will also understand that variation in marine resources is a normal and inevitable characteristic that must be accommodated in economic plans. This kind of resource-related benefit, although important, is not the only societal benefit that will accrue.

Odum (1977) said, "It is the properties of the large-scale integrated systems that hold solutions to most of

the long-range problems of society." Although few in the scientific community recognized the wisdom of this comment two decades ago, the validity of Odum's prediction has now been well demonstrated. For example, carbon dioxide modification of the planet's heat budget, and therefore climate, is a phenomenon that can be understood only if the emergent properties of large-scale biogeochemical systems are understood.

### III. PARTITIONING THE OCEAN INTO NATURAL FUNCTIONING UNITS

#### A. Central Problem

The central problem for the lower trophic level of ocean ecosystems is obtaining light (energy) and inorganic nutrients (mass). Odum's definition requires that for a functioning system to be a distinct ecosystem it must possess characteristic trophic structure and material cycles. That is, how one kind of ocean ecosystem captures light and passes that energy on in the form of primary productivity, secondary productivity, and so forth is different from how another kind of ocean ecosystem processes and transfers its energy. Likewise, how mass (C, N, P, and Si), initially in the form of inorganic compounds, is taken up and transferred through the food web and eventually released back to the environment is different and very poorly understood.

What controls the supply of light and nutrients to an ocean ecosystem? Sverdrup in 1955 was the first oceanographer to note that the spatial and temporal patterns of physical processes, particularly the seasonal patterns of mixing, stratification, and upwelling as well as the seasonal pattern of irradiance, control the patterns of biological organization. The division of ocean ecosystems into six distinct types is based fundamentally on the pioneering work of Sverdrup and that of others in the intervening years.

#### B. Biome Concept

Longhurst (1998) presented a scholarly discussion of the attempt to partition the oceans into natural functional provinces. He described with considerable elan a regional ecology of the oceans which is clearly distinct from partitioning the ocean into ecosystems. The difference can be illustrated using one of the best described ocean ecosystems, the coastal upwelling ecosystem. It is well accepted that there is one coastal upwelling

ecosystem and that it is replicated at five coastal regions in the world ocean: the California Current off California and Mexico, the Peru Current off Peru and Chile, the Benguela Current, the Canary Current, and the Arabian Sea off Oman and Yemen. Longhurst recognized the same regions but counted them as five distinct ocean "provinces." The ecosystem concept emphasizes that there are geographic ocean regions that have much in common; ecological geography emphasizes the discreteness or autonomy of various ocean provinces. Both concepts are useful.

Longhurst (1998) proposed that the following information is required to define the functional nature of an ocean region:

*Latitude*, which determines if wind stress induces mixing or wind-driven advection

*Depth of water* because stratification may be broken down in shallow seas by tidal mixing

*Proximity to coastline*, which determines the effects of terrestrial runoff, river discharge, and release of nutrients (especially Fe) from sediments

*Seasonal irradiance*, which forces photosynthesis, stratification, and freezing or melting of ice

*Winds*, which force mixing or upwelling of subsurface waters and their nutrients up to the euphotic zone

*Precipitation*, which may induce strong stratification by making the surface layer less salty

*Nutricline depth*, which modifies the vertical flux of nutrients by wind mixing and upwelling

*Strength of the vertical nutrient gradient*, which determines the magnitude of the upward nutrient flux

*External source of iron*, because insufficient iron may limit uptake of macronutrients, phytoplankton growth, or primary productivity

Using these criteria, Longhurst partitioned the ocean into four major biomes: the westerlies biome, in which convective mixing and stratification are forced largely by the strong seasonal progression of winds, irradiance, and heat flux; the trades biome, in which upwelling and mixing are forced on the ocean-basin scale by both local and remote wind forcing; the polar biome, in which there is no significant thermal stratification and mixed layer depth is constrained by a surface low-salinity layer which forms each summer as the marginal ice melts; and the coastal biome, in which coastal processes such as tides or currents break down stratification and force mixing.

Longhurst's (1998) partitioning of the ocean into four biomes and 55 provinces is a significant accom-

plishment. Eventually, his ideas will be merged with the ecosystem concept to produce an internally consistent hierarchy of biomes, ecosystems, and provinces.

### C. Ecosystem Concept

Using the formal Odum definition of ecosystem, six more or less well-defined ocean ecosystems can be delineated (Table I), but note that the boundaries of these ecosystems are usually oceanographic, not geographic, features. [See Tomczak and Godfrey (1994) for a description of these oceanographic features.] Although it is possible to describe approximately where these ecosystems exist, the actual domain is determined by dynamic processes such as fronts where two kinds of ocean water converge. The approach is to define the characteristic trophic structure and material cycles by applying the same analysis to each system. Each of the six systems has a different combination of stratification, nutrient supply (Table I), primary productivity (Table II), and biotic characteristics (Table III). This analysis indicates that there are six distinct ocean systems that meet Odum's criteria:

1. A low-latitude gyre ecosystem is present in each of the five great low-latitude gyres—North Atlantic, South Atlantic, North Pacific, South Pacific, and South Indian Oceans—as well as in the warm pool of the western Pacific Ocean, the equatorial Indian Ocean, and in large marginal seas such as the Mediterranean Sea and the Gulf of Mexico. This ecosystem is also present in the Western Boundary Current regions between the western edge of each of the five great gyres and the coastal waters of the adjacent continent.

2. The Southern Ocean ecosystem occupies the circumpolar area between the continent of Antarctica and the Subtropical Front at approximately 40°S latitude; the Southern Ocean ecosystem has a subantarctic region from the Subtropical Front south to the Polar Front and an Antarctic region from the Polar Front to the Antarctic continent.

3. The equatorial upwelling ecosystem occupies an equatorial band from 5°N to 5°S and from South America westward to 180° in the eastern and central Pacific Ocean. This region is often called the "cold tongue" because upwelling keeps these tropical waters surprisingly cool. In the Atlantic the 5°N to 5°S band of equatorial upwelling reaches from Africa across to South America. In the Indian Ocean equatorial region there is no manifestation of this ecosystem because the upwelling, if present, does not extend upwards into the euphotic zone.

TABLE I  
Summary of Heat Flux, Stratification, Nutrient, and Light Characteristics of Ocean Ecosystems

Ocean ecosystem <sup>a</sup>	Heat flux	Stratification		Nutrient		Light	
		Strength	Duration	Level	Source	Level	Pattern
Low-latitude gyre	Neutral; negative in Western Boundary Current	Strong	Permanent	Low ( $\ll K_s$ ) <sup>b</sup>	Eddy diffusion, very weak convection	High ( $\gg E_k$ ) <sup>c</sup>	Continuous
Southern Ocean	Negative (ocean loses heat)	Very weak, except strong when ice melts in summer	Seasonal	High ( $>K_s$ ), except $\text{Si(OH)}_4$	Mixing and upwelling	Moderate in summer; low rest of year	Strongly seasonal
Equatorial upwelling	Positive (ocean gains heat)	Strong stratification following vertical transport	Permanent	High ( $>K_s$ ), except $\text{Si(OH)}_4$	Upwelling and mixing	High ( $\gg E_k$ )	Continuous
Subarctic gyre	Seasonally positive and negative	Moderate stratification following winter mixing	Seasonal convective mixing	High ( $>K_s$ in winter; $\approx K_s$ rest of year)	Convective mixing and eddies	Low in winter ( $\ll E_k$ ); moderate rest of year ( $\approx E_k$ )	Seasonal
Eastern Boundary Current	Positive	Medium	Permanent	Medium ( $>K_s$ )	Upwelling and lateral advection	Moderate ( $>E_k$ ), except in winter	Seasonal
Coastal upwelling	Positive	Strong stratification following vertical transport	Continuous	High ( $\gg K_s$ )	Upwelling	High ( $>E_k$ )	Weakly seasonal

<sup>a</sup> See text for the location and boundaries of these ecosystems.

<sup>b</sup>  $K_s$  is the nutrient concentration at which nutrient uptake occurs at one-half the maximal rate; at concentrations  $<K_s$ , uptake is limited.

<sup>c</sup>  $E_k$  is the light level at which the photosynthetic rate is light saturated; at light levels  $<E_k$ , photosynthesis is light limited.



TABLE II  
Summary of Size, Primary Productivity, Export, and Limiting Factors of Ocean Ecosystems

Ocean ecosystem	Size <sup>a</sup>		Primary productivity amount and pattern (mmol C m <sup>-2</sup> day <sup>-1</sup> ) <sup>b</sup>	Export ratio <sup>c</sup>	Factors limiting primary productivity	Factors limiting fish yield
	Area (m <sup>2</sup> × 10 <sup>12</sup> )	%				
Low-latitude gyre	164	52	35 continuously	Low; weak seasonality	Macronutrients	Low primary productivity, small size of organisms and long food web
Southern Ocean	77	25	120 summer mean; ≈35 in spring and fall; ≈0 in winter	High in summer; very low in winter	Iron and light	Short duration of summer bloom
Equatorial upwelling	22	7	90 continuously	Low continuously	Iron and silicate	Low export productivity
Subarctic gyre	22	7	150 spring bloom mean; ≈50 rest of year	Episodically high; moderate rest of year	Depth of winter mixing	Short duration of spring bloom
Eastern Boundary Current	21	7	150 summer mean; 75 rest of year; grades into gyre ≈35	High in summer; moderate rest of year	Unclear, but iron is a factor, and light is also in winter	Unclear
Coastal upwelling	6	2	300 continuously close to coast; grades into Eastern Boundary Current ≈150	High, but spatially variable	Iron	Small size of ecosystem

<sup>a</sup> Size was calculated for the world ocean exclusive of noncoastal upwelling, continental shelf regions; total area was 312 × 10<sup>12</sup> m<sup>2</sup>.

<sup>b</sup> Based mainly on recent measurements made by the author in the Joint Global Ocean Flux Study.

<sup>c</sup> Export ratio is the relative proportion of primary productivity that is exported vertically, horizontally, or to higher trophic levels. The maximum observed export ratios relative to total primary productivity are approximately 0.50.

TABLE III  
Summary of Biotic Characteristics, the Zooplankton Component of Ocean Ecosystems<sup>a</sup>

Ecosystem	Endemic spp.	Species richness	Species evenness	Variability of biomass	Size of organisms	Importance of protozoan micrograzers
Low-latitude gyre	Low	High	High	Low	Small	Dominant
Southern Ocean	High	Low	Very low	Very high (>5×)	Large	Moderate to low
Equatorial upwelling	Moderate	High	High	Low	Small	Dominant
Subarctic gyre	High	Low	Low	High (>2×)	Moderate	Moderate to low
Eastern Boundary Current	Moderate	Moderate	Moderate	Low (<2×)	Moderate	Dominant to low
Coastal upwelling	Low	High	Moderate	Moderate (≈2×)	Large	Low

<sup>a</sup> Adapted from McGowan (1974).

4. A subarctic gyre ecosystem is present in both the North Atlantic and the North Pacific in the region north of the Subtropical Front at approximately 40°N. In the Atlantic, the subarctic gyre ecosystem is bounded on the south by the North Atlantic Current and Subtropical Front, on the west by the Labrador Current, on the north by the East Greenland Current, and on the east by Europe. The North Atlantic subarctic gyre is a complex, but very well-delineated, ocean feature. The North Pacific subpolar gyre is bounded on the south at approximately 40°N by the North Pacific Current and the Subtropical Front, on the west by Siberia, on the north by the Aleutian Islands and Alaska, and on the east by Canada.

5. An Eastern Boundary Current (EBC) ecosystem is represented in each of the four great EBCs: the California Current off the west coast of the United States and Mexico, the Peru Current off Peru and Chile, the Canary Current off Northwest Africa, and the Benguela Current off Namibia and South Africa. These locations are characterized by a broad and weak equatorward-flowing current that usually extends offshore 400–600 km. Each of these ecosystems receives upwelled water from a coastal upwelling ecosystem and, in turn, exports water to the adjacent low-latitude ecosystem.

6. There are four great coastal upwelling ecosystems: Peru Current, Benguela Current, Canary Current, and California Current. This ecosystem is narrow but long; it extends for great distances along the west coasts of South America, North America, northwest Africa, and South Africa. This ecosystem is also present in the northwestern Arabian Sea, along the coasts of Oman and Yemen, where evidence of upwelled water is present up to 600 km off the Omani coast.

## IV. CHARACTERISTICS OF OCEAN ECOSYSTEMS

### A. Shared Characteristics

The following characteristics of ocean ecosystems are based on a 1974 synthesis by McGowan for the oceanic Pacific Ocean:

*They are few in number (only six).* This small number of ocean ecosystems contrasts dramatically with the much larger number of well-defined terrestrial, intertidal, and coastal habitats that occur in much smaller areas. The reason, of course, is that the fluid medium of the open ocean is fundamentally partitioned by the dominant physical processes that occur in an oceanic area, and there are only a few (six) of these overriding physical patterns. In contrast, terrestrial, intertidal, or coastal ecosystems are fundamentally partitioned by spatial or geographic features and there is a much larger number of unique geographic features associated with the solid surface of the earth's crust. The same fundamental concept pervades all aspects of the comparison of oceanic biodiversity with that of terrestrial, intertidal, or coastal biodiversity. The fluid medium of the ocean is relatively homogeneous, and the boundaries that do exist are dynamic processes such as the presence or absence of winter convective mixing. It is obvious that dynamic processes of this nature form physical boundaries that permit much more biological exchange than the physical barrier associated with the solid earth.

*They are large relative to terrestrial, intertidal, or coastal ecosystems.* The largest ocean ecosystem, the low-latitude gyre ecosystem, occupies 52% of the open

ocean area of the world ocean. This ecosystem's size is a simple reflection of the phenomenon that the earth has vast surface area occupied by the great circulation features called gyres which are slow-turning circular ocean current patterns set in motion ultimately by wind and the rotation of the earth. These gyres, together with the tropical warm pool regions and certain of the marginal seas, occupy a large portion of the earth's surface. Because of the size of the low-latitude gyre ecosystem, it plays an important role in the global heat budget and in exchanges with the atmosphere, and it is the single most important ecosystem for understanding future global change. Ironically and unfortunately, it is one of the least understood ecosystems on Earth. Its size and remoteness from societal activity, especially direct economic activity, have led both scientific and political policymakers to assign the large low-latitude gyre ecosystem a low priority for scientific effort.

On the other hand, the smallest ocean ecosystem, the coastal upwelling ecosystem, has received thorough, comprehensive, and multinational study during the past three decades. This ecosystem is small relative to other ocean systems; it has very strong physical, chemical, and biological processes; it has huge economic importance and considerable geological importance because much of the earth's oil is assumed to have been formed in past upwelling ecosystems. Coastal upwelling ecosystems do have great societal importance, and the global effort to study and understand these important systems reflects well on the wisdom of scientific and political decision makers.

*They are geologically old (with the exception of coastal upwelling).* Oceanic cores suggest that the defining oceanographic features have been in place for the past 200,000 years and perhaps for the past 1 million years. The coastal upwelling ecosystem, however, may have disappeared entirely during periods of lower sea level such as during the last Glacial Maximum approximately 18,000 years ago.

*They respond to climate but not to weather.* Subtle and pervasive changes in mixed layer depth, the depth of convective mixing, or the strength of stratification cause profound changes in ocean ecosystems. On the other hand, strong storms or the passage of a violent hurricane cause no change that is detectable in the ocean ecosystem 2 weeks after the event. El Niño events cause profound changes in equatorial upwelling, EBC, and coastal upwelling ecosystems, but when El Niño ends the lower trophic levels of these ecosystems (phytoplankton, bacteria, microzooplankton, and mesozooplankton) return to their pre-El Niño condition

within 1 month. Higher trophic levels, of course, require several years to recover.

*They have considerable internal homogeneity, i.e., they tend to be monotonous.* The key word is "internal." There are clear changes in almost all biological properties when physical or oceanographic boundaries are crossed; however, if oceanographic boundaries are not crossed, biological properties will remain surprisingly similar over distances of thousands of kilometers.

*They are relatively undisturbed by anthropogenic processes compared to other ecosystems.* The largest human disturbance thus far has been ruthless overfishing that removes top predators. The removal of these long-lived and slow-growing carnivores appears to set off a trophic cascade that changes the ecosystem even down to the level of nutrient regeneration by bacteria and protozoa. A second and more threatening change is that as the earth's heat budget changes the processes of precipitation, winter mixing, stratification, and depth of the nutricline all change in such a way that there is less transport of nutrients into the euphotic zone. This change, already apparent in the North Pacific Ocean, will reduce productivity and, hence, the yield of food resources.

*The basic organization (Fig. 1) is like that of other ecosystems.* It appears that the basic assembly rules for all ecosystems are very similar. There has to be a large number of primary producers, somewhat fewer herbivores, still fewer carnivores, and many fewer top predators. There have to be efficient recyclers to return a portion of the nutrients back to the euphotic zone. In addition, the ecological "rules" regarding diversity, species composition, and adaptation are expressed similarly in terrestrial and ocean ecosystems.

## B. Distinguishing Characteristics

There are qualitative differences in the basic biological processes. The most dramatic illustration of this is the observation that processes limiting productivity, biomass, export, and yield are different in almost every one of these six ecosystems. Although our understanding of limiting factors may change, currently it appears that the low-latitude gyres are limited by fixed nitrogen, especially nitrate and ammonia; the Southern Ocean appears to be limited by iron and light; the equatorial upwelling ecosystem appears to be limited by iron and silicon; coastal upwelling ecosystem rates appear not to be limited by any macronutrient, but the space where optimal coastal upwelling occurs is highly constrained and iron supply is involved; the EBC ecosystem appears to be limited by light and iron in winter and by iron

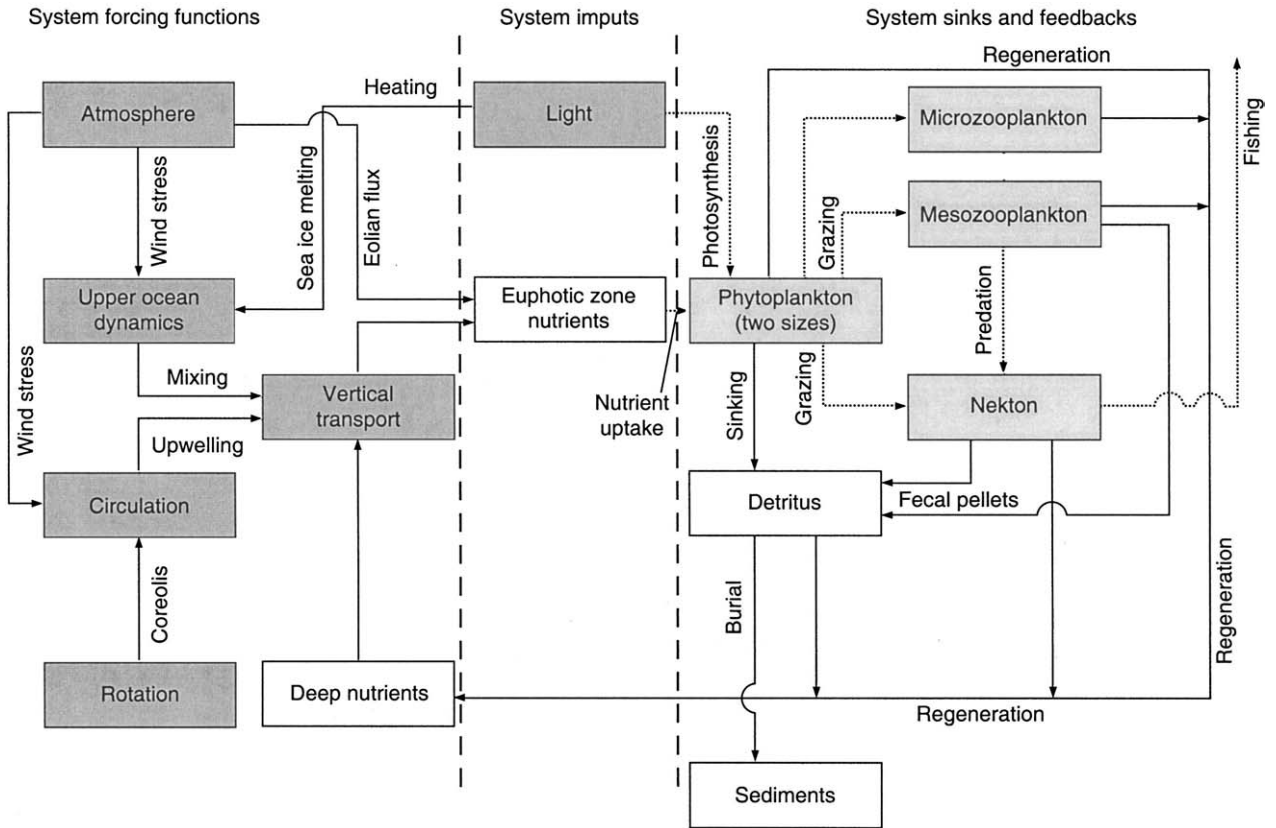


FIGURE 1 Generalized organization of ocean ecosystems. Physical forcing and input functions are shown in dark gray, chemical and geologic input functions are white, and biological components are in light gray. Physical transfers of energy, momentum, or mass are shown by arrows with solid lines; biological transfers of mass and energy between living components are shown by arrows with dotted lines. The export process at the extreme right, labeled fishing, is a proxy for all processes that remove biomass from the euphotic zone of the represented ecosystem.

alone in the summer; and the subarctic gyre ecosystem appears to be limited by the depth of the winter mixed layer or convection.

There are quantitative differences in basic processes such as primary productivity among different kinds of ocean ecosystems. One of the important milestones in the study and understanding of these differences was the work of John Ryther, who in 1969 published a work that provided a quantitative explanation of why fish yields vary by approximately 200-fold from the richest ocean ecosystems to the poorest. Variations in productivity, of course, are well-known from terrestrial ecosystems, but on land either aridity in deserts or freezing in polar regions is responsible for the low productivity of the poorest regions. Understanding why the benign low-latitude gyre ecosystem was so poor in fish production was much more difficult than understanding why productivity was low in deserts and polar regions. Part of the explanation was proposed in 1955 by Sverdrup, who said simply that the physical supply of nutrients to the euphotic zone is the reason for low fish yields in stratified ocean ecosystems. Ryther's contribution amplified the physical explanation by considering the nature of the ecological processes that lead to fish production.

First, Ryther estimated that approximately half the fish caught in the world are caught in coastal upwelling ecosystems, the smallest of the ocean ecosystems. Why? To begin, Sverdrup was correct: The physical processes of upwelling bring abundant nutrients to the surface layer, so primary productivity is very high in upwelling ecosystems. However, much more is involved. The phytoplankton that thrive in the rich coastal upwelling ecosystems are very large—so large that some are eaten directly by fish. This means that in coastal upwelling ecosystems the food chain leading to fish is very short. Ryther estimated that half the fish diet was phytoplankton and half was zooplankton. On average, then, the length of the food chain leading to fish had 1.5 transfers: large phytoplankton to fish or large phytoplankton to zooplankton to fish. At each ecological transfer, a large portion of the energy of the food is used to support the organism and this portion cannot be passed up the food chain. The shortness of the food chain is a factor that multiplies the high nutrient effect. Ryther also noted that in the small and food-rich coastal upwelling ecosystem the fish and zooplankton did not have to work as hard to get food, so the efficiency of transfer was increased relative to that of a poor environment such as the low-latitude gyre. Ryther proposed that fish yields were high in the coastal upwelling ecosystem because of high nutrients, high primary productivity, large size

of the primary producers, short food chains with few transfers, and increased efficiency of transfer. These effects multiply each other, leading to very high yields of fish. For the same reason, the abundance of seabirds and marine mammals is also very high in the coastal upwelling ecosystem.

The same arguments in reverse explain the low fish yields of the low-latitude gyre ecosystem. The other four ocean ecosystems range between these two extremes. The order of fish yield per unit area can be approximately estimated, from highest to lowest, as (1) coastal upwelling ecosystem, (2) subarctic gyres ecosystem, (3) EBC ecosystem, (4) Southern Ocean ecosystem, (5) equatorial upwelling ecosystem, and (6) low-latitude gyre ecosystem.

## V. OCEAN ECOSYSTEMS AND GLOBAL CHANGE

Human intervention in the material cycles and trophic structure of ocean ecosystems may already have caused some changes. The interventions are so varied and play out at so many different scales that it is difficult to know how to describe their impact on future climate change. One approach is to focus on model studies of how anthropogenic ocean warming will affect ocean ecosystems. Considerable effort has gone into investigating how increased atmospheric carbon dioxide will affect the physical ocean-atmosphere system. Other effort has gone into determining how the ocean's carbon dioxide system will behave in an anthropogenically warmed ocean. Today's atmosphere-ocean climate models are fully coupled to an ocean carbon model, including a full carbon system and the biological transfers of the pelagic food web. Results of these model runs are dramatic and, as expected, they predict (i) a warming of global mean sea surface temperature by 2.5°C, (ii) a slight decrease in light because of increased cloud cover, (iii) a large increase in vertical stratification because of increased precipitation which lowers salinity, and (iv) increased heating of the surface layer. The increase in stratification and decrease in convective mixing are predicted to cause a 37% reduction in the supply of nutrients to the euphotic layer. All physical processes that transport nutrients from the deep ocean reservoir to the surface layer are affected: upwelling, wind and tidal mixing, and convective mixing.

Of the three classes of impacts—warming, light, and increased stratification—increased stratification will

have the major impact on ocean ecosystems. The changes predicted vary in intensity from one ocean ecosystem to another, but they are most severe in productive ecosystems such as the subantarctic gyre, equatorial upwelling, and coastal upwelling ecosystems. Very little change in nutrient supply is predicted for the world ocean's five great low-latitude gyres. The predicted change is an expansion of the size of the world ocean's most oligotrophic ecosystems: Approximately half of the ocean area will have decreased nutrient supply; the low-latitude gyres, the poorest half, will have no change; and a very small area (less than 5%) in the high Arctic and Antarctic will have an increase in nutrient flux to the surface.

The El Niño phenomenon has provided evidence of the biological consequences of reducing an ocean ecosystem nutrient supply. New populations that became established under the low nutrient conditions are healthy, highly diverse communities, but they are dramatically different. Another consequence of reduced nutrients is that the biological pump sequesters less carbon dioxide. This change will feed back into the global carbon system to accelerate the increase in concentration of atmospheric carbon dioxide.

A reduction of new nutrients by 37% will have a strong impact on both the quantity and the kind of fish present. Because the increase will occur in what are now rich fishing regions, the coastal upwelling and subarctic gyre, the societal impact will be larger than the biological impact. A change of 37% in new nutrients is significant, but it is not enough to destroy ocean ecosystems. The current gradient in nutrient flux from the low-latitude gyre ecosystem to a coastal upwelling ecosystem region is more than a factor of 10. The predicted changes will reduce the yield from the world's rich fishing banks, but the expanded low-latitude gyre ecosystems should maintain their ecological integrity.

## See Also the Following Articles

CLIMATE CHANGE AND ECOLOGY, SYNERGISM OF •  
ECOSYSTEM, CONCEPT OF • MARINE ECOSYSTEMS •  
PELAGIC ECOSYSTEMS • PLANKTON, STATUS AND ROLE OF

## Bibliography

- Barber, R. T. (1988). Ocean basin ecosystems. In *Concepts of Ecosystem Ecology* (L. R. Pomeroy and J. J. Alberts, Eds.), pp. 171–193. Springer-Verlag, New York.
- Barber, R. T., and Smith, R. L. (1981). Coastal upwelling ecosystems. In *Analysis of Marine Ecosystems* (A. Longhurst, Ed.), pp. 31–68. Academic Press, San Diego.
- Karl, D. M. (1999). A sea of change: Biogeochemical variability in the North Pacific Subtropical Gyre. *Ecosystems* 2, 181–214.
- Longhurst, A. (1998). *Ecological Geography of the Sea*. Academic Press, San Diego.
- McGowan, J. A. (1974). The nature of oceanic ecosystems. In *The Biology of the Oceanic Pacific. Proceedings of the 33rd Annual Biology Colloquium* (C. Miller, Ed.), pp. 9–28. Oregon State Univ. Press, Corvallis.
- Odum, E. P. (1969). The strategy of ecosystem development. *Science* 164, 262–270.
- Odum, E. P. (1977). The emergence of ecology as a new integrative discipline. *Science* 195, 1289–1293.
- Odum, E. P., and Odum, H. T. (1959). *Fundamentals of Ecology*. Saunders, Philadelphia.
- Pomeroy, L. R., and Alberts, J. J. (Eds.) (1988). *Ecological Studies* 67, *Concepts of Ecosystem Ecology, A Comparative View*. Springer-Verlag, New York.
- Ryther, J. H. (1969). Photosynthesis and fish production in the sea. *Science* 166, 72–76.
- Sherman, K., and Alexander, L. M. (Eds.) (1986). *Variability and Management of Large Marine Ecosystems*, AAAS Selected Symposium No. 99. Westview, Boulder, CO.
- Steele, J. H. (1985). A comparison of terrestrial and marine ecological systems. *Nature* 313, 355–358.
- Steele, J. H. (1998). From carbon flux to regime shift. *Fisheries Oceanogr.* 7, 176–181.
- Sverdrup, H. U. (1955). The place of physical oceanography in oceanographic research. *J. Mar. Res.* 14, 287.
- Tomczak, M., and Godfrey, J. S. (1994). *Regional Oceanography: An Introduction*. Pergamon, Oxford.





# ORIGIN OF LIFE, THEORIES OF

Susanne Brakmann  
Evotec Biosystems AG

---

- I. Historical Overview
  - II. Theoretical Concepts
- 

## GLOSSARY

**autocatalysis** Self-acceleration of certain chemical reactions that yield catalytically active products. The reaction rate of autocatalytic processes increases exponentially.

**hypercycle** Cyclic sequence of self-supporting reactions between primitive prebiotic biomolecules, nucleic acids, and proteins. Hypercyclic organization of reacting systems was postulated to explain the spontaneous emergence of replicating systems. As a result of the hypercyclic reaction principle, functional characteristics of replicating molecules (i.e., phenotypes) are connected to their hereditary characteristics or their genotypes. Consequently, both phenotypes and appertaining genotypes are evaluated and selected in feedback loops.

**natural selection** Preferential reproduction and survival of certain species under given environmental conditions. Growth advantages of selected individuals reflect the efficiency of their reproduction pathways.

**quasispecies** Hierarchically ordered population of mutants that results from erroneous copying of a genotypic ancestor. Whole mutant distributions behave like single species. They are subjected to natural selection. The average sequence is defined as a con-

sensus sequence that represents the “center of gravity” of the distribution. The consensus sequence is synonymous to a wild type.

**self-organization** Spontaneous formation of complex structural aggregates from biological macromolecules such as proteins, nucleic acids, or lipids. The assembly pathways are usually determined by physicochemical properties of the constituent molecules.

---

**LIFE IS A STATE OF MATTER** that results from controlled cooperation of highly complex structural assemblies. Living matter shows more than the sum of properties of the isolate entities since none of them could survive separately. Many general features distinguish living from non-living matter, predominantly self-reproduction, functional self-organization, and evolution by natural selection. The origin of life is a consequence of prebiotic evolution of organic material. The process of prebiotic evolution is the object of various and contradictory hypotheses, none of which has been proven experimentally.

## I. HISTORICAL OVERVIEW

### A. What Is Life?

Currently, life appears on Earth with an immense variety of structures and functions. Despite this diversity,



some invariant features exist that are common to living systems and distinguish them from non-living matter.

All life takes shape with defined individuals. Cells are the smallest individual entities capable of performing all processes for living and reproducing. They represent the elementary structural units of life. Cells do not spontaneously originate from suitable mixtures of non-living matter. Louis Pasteur (1822–1895) was the first to disprove the idea of abiogenesis by showing that each cell results from cell division. Rudolf Virchow (1821–1902) confirmed this, noting that fusion of cellular precursors (germ cells) is the only alternative for generating new cells. Two general types of cellular organization are known: the prokaryotic type, which lacks a nucleus, and the more highly developed eukaryotic type, which harbors a nucleus. Basically, both cell types exhibit the same structure and functional organization suggesting a common origin. Simple organisms such as bacteria consist of single cells which are either prokaryotic or eukaryotic, whereas complex organisms, such as plants and animals, are highly organized multicellular assemblies of the eukaryotic type.

The different forms of life do not originate from nothing as Charles Darwin (1809–1882) discerned in his epoch-making work. He, and independently Alfred Russel Wallace (1823–1913), postulated that, instead, a continuous process of successive adaptation to environmental conditions led to the origin and development of species. Darwin also realized that evolutionary progress results from mutagenous reproduction and natural selection. Today, this theory can be explained with twentieth-century insights into structural and chemical characteristics of living cells.

Each single cell contains an extraordinarily efficient machinery for respiration and production of energy; for absorption and assimilation of nourishment; for synthesis, storage, utilization, and transport of new products; and for production of more or less identical copies of itself. The cellular machinery is instructed by hereditary information that is carried by large molecules which are composed of nucleic acid and known as genes. They comprise the architect's plan of all cellular components and also include structural elements regulating the formation of the different characteristics of an organism. During the process of reproduction, genes also reproduce, thereby passing all instructions on to the next generation. Proteins represent the executive that is coded by genetic information. They cooperate closely with nucleic acids in order to fulfill the diverse functional and structural tasks that maintain the cell machinery. Many proteins are enzymes that act as catalysts for the reproduction of nucleic acids and for the synthe-

sis of essential organic compounds, such as carbohydrates, amino acids, lipids, and hormones. Other proteins care for the cellular metabolism and the production of energy. The teamwork of these two molecular species, polynucleotides and polypeptides, predominantly characterizes the chemistry of living matter as we know it.

Generally, high-energy macromolecular compounds play a major role in living organisms. In addition to nucleic acids and proteins, biomolecules such as lipids, polysaccharides, and phosphates are found. They serve for demarcation of individual cells, for their division into different subcompartments, for their structural stability, and for energy transfer processes. However, the splendid variety of life on Earth is mainly due to the heteropolymeric nature of proteins. They consist of long chains that are built from 20 constituents, chemically defined as amino acids. Similarly, nucleic acids encoding the natural proteins are polymeric chains based on 4 different monomers. The building principle realized with both molecular species allows a variability that is far beyond the imagination—even with chain lengths of 20 monomers, more than  $10^{12}$  different nucleic acids (with 4 different building blocks) or  $10^{26}$  alternative proteins (with 20 different building blocks) can exist.

Structural variability is not the only feature that characterizes the molecules of life. Nucleic acids are also capable of serving as templates for their own replication, thereby guaranteeing the reproduction of living organisms. Self-reproduction is an inherent property of polynucleotides that is based on physical interactions between their complementary constituents. This phenomenon is known as base pairing and allows the synthesis of a unique complementary copy of each molecule. Replication of this copy then yields a copy of the original chain. The complete procedure very much resembles the positive–negative printing used in photography. Occasionally, the copying process is imperfect, leading to mutations in gene replication. A mutation alters the genetic information and therefore the instructions for one or more characteristics. Continuous replication and mutation can then produce genetic variability. Following expression (i.e., translation of nucleic acids into proteins), some mutations will produce altered characteristics that are favorable for the organism. This organism will reproduce preferentially over those with the unchanged gene. By far most mutations are deleterious, leading to reduced replication or even cell death. Only beneficial mutations which occur accidentally result in organisms with better adaptation to their environment. In this way, organisms evolve to-

ward higher fitness and often toward greater complexity. In summary, complex organisms evolved through time according to Darwin's theory because of replication, mutagenous replication, natural selection, and replication of resulting mutants. During the process of Darwinian evolution, information is created continually.

Another view of "life" is presented by the discipline of thermodynamics that distinguishes between open and closed systems. Corresponding to the second law of thermodynamics, no processes can occur in closed systems that increase the net order of the system or that decrease their entropy. In this perspective, any living organism is an open system that exchanges energy and matter with its surrounding, thereby establishing and maintaining a highly ordered structural assembly at the expense of a larger decrease of order of the universe outside. Living organisms participate in ecosystems and depend on resources found in their environment. In particular, most organisms are dependent on the flow of sunlight, which is absorbed by plants and utilized by them to synthesize high-energy molecules from simpler ones. These products serve as a source of food and energy for other links of the food chain (e.g., animals). By excretion and degradation of dead organic compounds, all material will be recycled without altering the organisms' state of life, which is characterized as being far from thermodynamic equilibrium. A state of thermodynamic equilibrium would be identical to a state of maximum entropy and thus correspond to cell death.

All organisms on Earth are closely related and the fundamental pattern of their life is essentially identical for all of them. The similarity among all living organisms probably implies that life is descended from a single instance of origin. This incident is central to many scientific and philosophical problems and the object of various different and contradictory hypotheses which will be described in the following section.

## B. Hypotheses and Theories on the Origin of Life

The traditional position of theology and some philosophy views the origin of life as the result of a supernatural event which is permanently beyond the descriptive powers of chemistry and physics. In its most general form, this view is not necessarily contradictory to contemporary scientific knowledge about prebiotic evolution, although the biblical descriptions of creation given in the first two chapters of Genesis, taken literally and

not metaphorically, are inconsistent with modern knowledge.

Until the mid-seventeenth century, the prevailing opinion was that God created man together with higher animals and plants, but that simple forms of life such as worms and insects arise steadily from mud, waste, and putrefied matter during short periods of time. The physiologist William Harvey (1578–1657), who studied reproduction and development of deer, was the first to challenge this view by postulating that every animal comes from an egg (*"omnia viva ex ovo"*) a long time before Karl-Ernst von Baer (1792–1876) discovered the existence of human egg cells by microscopy. An Italian scientist, Francesco Redi (1626–1698), found Harvey's idea to be true, at least for insects; he found that maggots in meat arise from fly eggs. Later, Lazzaro Spallanzani (1729–1799) discovered that spermatozoa were necessary for the reproduction of mammals. Before Pasteur, Spallanzani also showed that living matter (*"infusories"*) does not originate from boiled fluids kept in closed containers. Although Redi's and Spallanzani's findings definitely proved that insects and larger animals develop from eggs, it remained obvious to a large majority that at least microorganisms, because of their ubiquity, are generated continually from inorganic material. The debate of whether life is spontaneously generated from non-living matter or not culminated in the famous controversy between Louis Pasteur and Félix-Archimède Pouchet (1800–1872) which Pasteur won triumphantly. He showed that even microorganisms in fluids come from germs floating in the air, and he also demonstrated that nutrient solutions could be guarded against these creatures by suitable sterilization such as filtration or boiling. However, contemporary scientists were not satisfied by Pasteur's experiments because a delicate question remained: If living organisms do not arise from non-living matter, how had life come about in the first place?

In the late nineteenth century, another hypothesis was initiated by the Swedish chemist Svante Arrhenius (1859–1927). He strongly believed that the whole universe is replenished with living germs, a phenomenon that he called *"panspermia."* He suggested that microorganisms and spores of cosmic origin spread from solar system to solar system, and thus they arrived on Earth. Although Arrhenius' view avoids rather than solves the problem of the origin of life, and despite the extreme unlikelihood of microorganisms surviving the interstellar effects of cold, vacuum, and radiation, a few twentieth-century members of the scientific community returned to the idea of panspermia. Among these scientists are astronomer Fred Hoyle (1915–) and mo-

lecular biologist Francis Crick (1916–), who are convinced that the time span between the origin of Earth and the appearance of first cellular organisms on this planet was too short for life to have occurred spontaneously.

Darwin's theory of "natural selection as motive power for evolution" resulted in a new view on the phenomenon of life that is still valid. Although Darwin did not commit himself on the origin of life, contemporary scientists such as Thomas Huxley (1825–1895) extended his idea, asserting that life could be generated from inorganic chemicals. Pursuing this opinion, Alexander Oparin (1894–1980) was the most influential advocate of the successive origin of cellular organisms from non-living matter. He suspected this transition was preceded by a series of regular and progressive chemical reactions under the physical and chemical conditions on early Earth. Together with John Scott Haldane (1860–1936), Oparin recognized that the abiological production of organic molecules in the current oxidizing atmosphere of Earth is highly unlikely. Instead, both suggested that the beginning of life occurred in primordial hot waters under more reducing (i.e., hydrogen-rich) conditions. Furthermore, Oparin postulated the existence of pre-cellular coacervates—globular units with membrane-like surface structures—that may have high concentrations of certain chemical compounds. Coacervates indeed form spontaneously from colloidal aqueous solutions of two or more macromolecular compounds.

However, many fundamental problems on the transition from non-living to living matter remained unsolved. The central question concerned the role of the second law of thermodynamics, which defines the equilibrium in an isolated system as a state of maximum entropy that appears to contradict the origin and existence of highly ordered living organisms. Erwin Schrödinger (1887–1961) gave a decisive answer to this question, stating that "living matter evades the decay to equilibrium" or death by steadily compensating for the production of entropy. In any organism, this is achieved by feeding it free energy or energy-rich matter which is used by the cellular machinery to drive essential chemical reactions. Schrödinger and others also realized that living organisms can thermodynamically be described as open systems, but they could not explain the general physical conditions for self-ordering processes. These were perceived by Ilja Prigogine (1917–) and Paul Glansdorff (1904–1999), who worked on a thermodynamic theory of irreversible processes. According to Prigogine, selection and evolution cannot occur in equilibrated or nearly equilibrated reaction systems,

even if the right types of substances are present. Instead, certain combinations of autocatalytic reactions with transport processes may lead to peculiar spatial distributions of reaction partners, called "dissipative structures." These ordered structures are of importance for the formation of functional order in the evolution of life, especially for early morphogenesis. However, the first steps of self-organization probably involved little organization in physical space but extensive functional ordering of a tremendously complex variety of chemical compounds. Manfred Eigen (1927–) explained the process of ordering among molecules by augmenting the Prigogine–Glansdorff principle with phenomenological considerations on the behavior of self-replicating molecules: A certain quantity approaches a maximal value in any open system that is replicating autocatalytically with sufficient fidelity, and thereby continually consuming energy and matter. This quantity is called "information" and is closely related to the "negative entropy" postulated by Schrödinger. In addition to setting the stage for a molecular interpretation of biological information, Eigen developed the mathematical models for describing "selection." According to Eigen's theory, selection is the fundamental natural principle that brings order into any random arrangement of autocatalytically replicating species. With selection, information is generated successively, leading to a steady optimization of species, which can either be organisms or molecules.

The mathematical models developed by Eigen support a detailed hypothesis of the origin of life which comprises multiple, successive steps for the transition from inorganic to living matter. However, it should be mentioned that some scientists have theories on the emergence of life that differ from Eigen's theory. Among these is Stuart Kauffman (1939–), who believes that natural selection is important but not the sole ordering principle of the biological world. Instead, he considers spontaneous self-organization to represent the predominant source of natural order. Kauffman demonstrated that sets of inter-related autocatalytic reactions can undergo a transition to a newly ordered (i.e., self-organized) state as soon as their connectivity reaches a certain threshold value. Furthermore, Kauffman emphasizes that the phenomenon of autocatalysis, which plays the central role in his theory, is not limited to nucleic acids. Therefore, he concludes that even genes were not necessary for the origin of life. In contrast to Kauffman, Eigen distinguishes "random" autocatalytic or self-replicating activity which is observed for a variety of molecular species from the "inherently" self-replicating nucleic acids. Inherent capability for self-replica-

tion, in turn, represents the molecular basis for natural selection according to Eigen's theory.

Well-defined experiments were invented in order to simulate the principles that were postulated for molecular evolution. With certain experimental set-ups, replication and selection can be performed in a test tube. Similarly, the chemical conditions on primordial Earth can be mimicked in the laboratory. Several scientists attempted to verify experimentally the twentieth-century ideas on biogenesis. Their experiments are discussed in the following section.

## C. Experiments

### 1. Production of Building Blocks and Polymers

The ideas of Oparin and Haldane inspired the first simulations of prebiotic conditions on early Earth. Under the auspices of Harold Clayton Urey (1893–1981), graduate student Stanley Miller (1930–) attempted to simulate the primordial chemical processes in a famous experiment. He continually exposed a mixture of methane, ammonia, hydrogen, and water vapor above an "ocean" of boiling water to sparks of a corona discharge in a closed apparatus (Fig. 1). The gaseous reaction

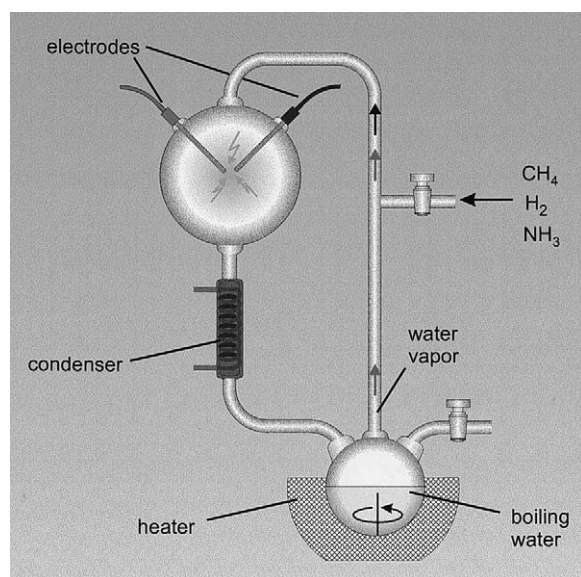


FIGURE 1 Simulation of prebiotic chemistry by Miller and Urey. Methane, ammonia, and hydrogen are led into the apparatus via stopcock 1 (top), driven into the reaction vessel by water vapor, and exposed to sparks of a corona discharge. The organic gaseous reaction products are condensed and dissolved in boiling water. Samples for analysis can be taken through stopcock 2 (bottom).

products were condensed in a cooler, dissolved in boiling water, and recirculated to the gaseous atmosphere. After several days of exposure to sparking, the aqueous solution changed color. Subsequent analysis indicated that several amino and hydroxy acids had been produced by this simple procedure. Because on early Earth probably much more energy was available in ultraviolet light than in lightning discharges, astrophysicist Carl Sagan (1934–1996) and colleagues altered Miller's experimental conditions. They synthesized amino acids by applying long-wavelength ultraviolet irradiation to a gaseous mixture of methane, ammonia, water, and hydrogen sulfide. Similar to Miller's experiment, these attempts also revealed that amino acids, particularly the biologically relevant ones, form readily under simulated primitive conditions. Also, it became obvious that reducing conditions were necessary for prebiological formation of amino acids since no yield was observed with oxidizing atmosphere. Further modification of these types of experiments showed that alkaline conditions led to the spontaneous synthesis of a variety of sugar molecules, including the five-carbon sugars fundamental for the formation of nucleic acids and six-carbon sugars such as fructose and glucose, which represent common metabolites and structural building blocks in contemporary organisms. As was shown by Juan Oró (1923–) and several other investigators, nucleotide bases and porphyrins can also be synthesized from simple mixtures of ammonia, water, and hydrogen cyanide. Therefore, it seems probable that most—if not all—essential building blocks of proteins, carbohydrates, and nucleic acids can be produced under quite general primitive reducing conditions.

The formation of polymers, long-chain molecules made of repeating units of the building blocks, would have to be the next major step during chemical evolution. However, polymerization of monomers was not achieved in the simple experiments described previously. These reactions involve the loss of one water molecule during the formation of each two-unit product and therefore are facilitated under dehydrating conditions. By dry heating of amino acid mixtures, Sidney Fox (1912–) accomplished the condensation reaction in the laboratory to yield polyamino acids. He observed that these molecules were not random polymers and they also exhibited some distinct enzyme-like catalytic activities; therefore, he called them proteinoids. In addition to this work, various experimental approaches concerned the catalysis of polymerization reactions by mineral surfaces, which may also protect growing chains from degradation by water molecules, as well as combinations of heating under dry and wet conditions mim-

icking the environment of submarine hydrothermal vents.

Some scientists cast considerable doubt on any contribution of prebiotic syntheses of organic compounds to the origin of life. Predominant among these is Günther Wächtershäuser (1938–), who imagines an origin without reducing atmosphere and without the primordial soup. According to Wächtershäuser, prebiotic chemistry began on the surface of minerals in the lithosphere. If these minerals were positively charged, they would form ionic bonds with negatively charged chemical groups of phosphate, carbonyl, or sulfide. As a result, a layer of negatively charged molecules would form, coating the minerals in two dimensions. Once a diversity of molecules accumulated, inorganic reactions involving the minerals could supply reducing power for the synthesis of more complex organic molecules from various possible interactions within the two-dimensional layer.

The most important implication of Wächtershäuser's theory is that the first forms of life could have been autotrophic instead of heterotrophic. Autotroph organisms get their carbon from carbon dioxide, whereas heterotrophs get it from organic molecules such as glucose. The fact that the oldest microfossils appear to be photoautotroph blue-green bacteria supports the autotroph hypothesis. Nevertheless, these ideas have not been adequately tested.

## 2. Evolution Experiments

The onset of self-replication—following the formation of polymeric nucleic acids—represents a major difficulty for the explanation of the origin of life. Experimental attempts by Leslie Orgel (1927–) and coworkers revealed that, under prebiotic conditions, it is possible to add monomeric units, called nucleotides, to complement a small template nucleic acid molecule and to combine the adhering nucleotides to the correct negative copy of the positive original. This is merely the first step of a self-replicating process that must be continued by a separation of both strands, positive and negative, and by repeated addition and concatenation of monomers. Without proteins supporting the polymerization of the building blocks, identical copies of the original nucleic acid molecule, the positive strand, could not be synthesized. However, because the simultaneous emergence of self-replication of nucleic acids and specific catalysis of their duplication by proteins seems very unlikely, Orgel, together with Francis Crick and Carl Woese (1928–), supposed that the first and primitive self-replicating systems involved nucleic acids

only. The discovery of catalytically active nucleic acids, or ribozymes, by Thomas Cech (1947–) and Sidney Altman (1939–) decisively supports this hypothesis. Much contemporary work therefore focuses on searching for ribozymes catalyzing the replication of nucleic acids.

Experimental studies on selection among replicating polynucleotides were first performed by Solomon Spiegelman (1914–). He purified nucleic acid and the replicating enzyme of bacteriophage  $Q\beta$  and implemented their replication reaction in a test tube which contained all essential reagents. Then he replicated the viral nucleic acid in serial transfer experiments by applying constraints that select exclusively for higher replication rates (Fig. 2). Beginning with infectious nucleic acid molecules, sequences were obtained after several rounds that replicated several times faster and had much lower molecular weights than the original molecules but were unable to infect bacteria. On the basis of more detailed knowledge of this replication reaction, Christof Biebricher (1941–) and coworkers, together with Manfred Eigen, repeated Spiegelman's experiments. They applied conditions which were chosen to be optimal with respect to rate of evolution, and they analyzed the stepwise progress of the optimization procedure. Their attempts at facilitating artificial evolution established an "irrational" technique for the design of new drugs known as directed molecular evolution. Eigen and coworkers also performed evolution experiments with various different replicators, either molecules or viruses, and they observed functional organization between replicators and their catalysts as well as some of the molecular mechanisms of evolutionary optimization.

## II. THEORETICAL CONCEPTS

### A. Quantitative Treatment of Darwinian Evolution

The process of complementary reproduction of self-replicating molecular assemblies can be described in detail by applying the mathematical formalism of chemical kinetics. For any population of molecules within a certain environment, consideration from the viewpoint of kinetics accounts for three necessary conditions: metabolism, self-reproduction, and mutability.

Metabolism describes the rates of formation and decomposition of each molecular species taking part in the competition. The term metabolism expresses the continuous influx of energy-rich matter, which main-

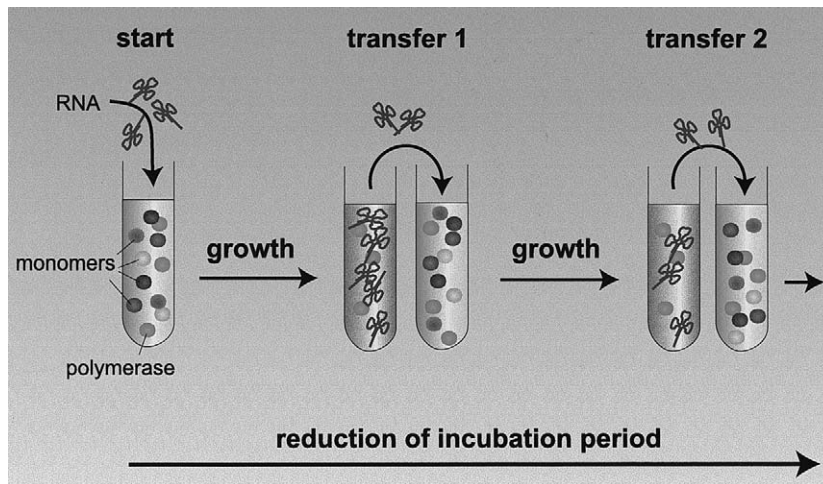


FIGURE 2 Scheme of the serial transfer principle performed by Spiegelman. A series of test tubes contain Q $\beta$  replicase, RNA building blocks (nucleoside triphosphates of A, G, C, and U), and necessary growth factors. The mixture is supplied with RNA template, and the reaction is started by increasing the temperature to 30°C. When a certain product concentration is reached, a small fraction of the mixture is transferred to the next tube and incubated in the same way. The procedure is repeated many times, continually reducing the incubation period. Eventually, a single optimal RNA product is selected.

tains the chemical reaction, and the corresponding efflux of decomposition products. In the case of nucleic acids, the influx is due to their energy-rich building blocks, nucleoside triphosphates. The respective monophosphates result from decomposition by hydrolysis, or by enzymatic cleavage of polymeric strands, and represent the efflux. Metabolism by consuming and processing resources thus effects a dependence of the replicating individuals on their environment.

Self-reproduction is due to the self-accelerating or autocatalytic replication of nucleic acids. The amount of descendant molecules in a defined environment depends on the rate of the replication reaction, which is dependent on the sequence of each replicating molecule. Fluctuations of the reaction rate may also be due to variations of substrate concentrations resulting from influx and outflux within an open system.

Mutability takes into account the quality of replication. The copying fidelity is physically limited because irregularities occasionally occur during the recognition and interaction of complementary strands. Therefore, errors are introduced into the replicating molecules which lead to an increase in the variety of the offspring.

The combined considerations can be expressed in differential equations, which are used to calculate the parallel growth of molecular individuals within an expanding population. Similar to the constancy of general-

ized forces or fluxes in the thermodynamics of irreversible processes, the reaction equations describing autocatalytic growth of nucleic acids can be simplified on the basis of certain constant boundary conditions. In the case of constant fluxes or constant organization, the equations can be solved exactly and result in explicit expressions. The physical interpretation of the results reveals that selection in the Darwinian sense is an inherent property of autocatalytically replicating populations of molecules. Selection can be characterized by an extremal principle which mathematically describes the preferential treatment of those replicators which maximize their fitness with respect to the environmental constraints. These optimal mutants, or more precisely mutant distributions, succeed at the total expense of the others. Analogous to the concept of species that Darwin believed to be the target of "survival of the fittest" within the process of natural evolution, the selected mutant distribution with the most efficient replication pathways behaves like an individual species; therefore, it was called "quasispecies" by Eigen. The substitution of less advantageous by more advantageous mutant distributions results in reaction rate changes in opposite directions—similar to those calculated by Prigogine and Glansdorff for simple autocatalytic processes far from thermodynamic equilibrium. Hence, the instability associated with the selection process is in accordance

with the expectations drawn from the Prigogine–Glansdorff principle: It results in the breakdown of the formerly (meta-)stable quasispecies and its substitution by a new quasispecies.

The quasispecies behaves physically and mathematically like a single species, or more precisely like the selected wild type of a species, but does not define a single species. It includes all mutants of the dominant form and represents a widely dispersed, not necessarily symmetrical, distribution of similar and related, but by no means identical, sequences. A quasispecies is usually characterized by a defined average sequence, called a consensus sequence, and can be dominated by one or more so-called master sequences. Evolutionary progress of the quasispecies results in new mutants that appear on the periphery of the distribution. Selection, or the emergence of mutants which replicate more efficiently than their ancestors, is a manifestation that new genetic information has been generated.

## B. The Informational Aspect

### 1. What Is Information and How Does It Originate?

In the most common sense, information is associated with the content and meaning of a “news” message. From the viewpoint of a recipient, the news can be understood only if the recipient and the transmitter agree on the meaning of the symbols used to code the message. In other words, new information is gained by a recipient when he or she can interpret the received symbols or the elementary units of information. Genetic information is stored in the linear copolymers of RNA and DNA. The number of monomers, representing the symbols, and their sequence within the polymeric chains designate whether or not a molecule encodes a message, and the arrangement of symbols determines the molecule’s information content. The meaning of the symbols originates with the unique ability of RNA and DNA to recognize and “read” themselves. Their reading is based on particular chemical interactions mediated by specific hydrogen bonding that links the four monomers, A, T (U in the case of RNA), G, and C, into the complementary pairs AT (AU) and GC. The encoded messages transmitted by RNA and DNA include instructions that initiate and control all biochemical reaction cascades taking place within each organism.

During the process of selection, mainly the individual abilities to replicate are assessed with respect to the requirements of a given environment. The most efficient

replication pathways, in turn, are “instructed” by genetic information. Therefore, new information is generated by erroneous replication and selection.

### 2. Maximal Information Content of the “Species”

Evolutionary flexibility represents the pivot of the selection process. It is guaranteed by replication error rates that produce a sufficiently large variety of mutants. However, as the mutation rates increase, the accumulation of errors outweighs the selective narrowing down. As a consequence, the genetic information “melts” away into random combinations of symbols that correspond to nonsense messages without any ability to generate and maintain necessary reaction pathways. This is a catastrophe for every living organism because it leads to the evolutionary decay of information. Therefore, natural evolution processes are bounded by an error threshold. Far below this boundary, selection can be reduced to all-or-none decisions. Evolutionary progress is achieved at reasonable velocities only when the error rate is near the threshold value. Indeed, it has been established experimentally by Esteban Domingo (1942–), Charles Weissmann (1931–), Lawrence Loeb (1936–), and several other research groups that viruses (which, due to their short replication periods, are ideally suited for evolutionary studies) behave like quasispecies and replicate close to their error thresholds. In accordance with theory, the reciprocals of the error rates that have been experimentally determined for many viruses define the limiting information capacity of their genomes. Loeb was also the first to prove the hypothesis that an additional increase in the mutation rate would abolish viral replication. By applying deoxynucleoside analogs to HIV replicating *in vitro*, he observed a complete loss of viral replicative potential and thus a lethal mutagenesis.

Generally, the theoretical and experimental results on replication error thresholds are not just true for viruses. They also remain valid for all replicating individuals that can be described by the formalism developed by Eigen (Table I).

The complexity that life has developed by successive evolutionary optimization does not just result from vertically transmitted point mutations. (“Vertically” expresses the hereditary transmission of errors to the direct offspring.) Additional principles for the modification and expansion of genomes exist that enable the horizontal transfer of genetic information, i.e., the transfer between individuals or species. These additional principles are homologous genetic recombina-

TABLE I  
Stepwise Expansion of Information Capacity during the Evolution of Life

Replication process	Subject/organism	Error rate	Maximal genome length	Example
Enzyme-free replication (base pairing)	AU polymer	$10^{-1}$	10	80 (tRNA ancestors)
	GC polymer	$10^{-2}$	100	
Enzyme-catalyzed RNA replication	RNA viruses	$10^{-4}$	$10^4$	4500 (bacteriophage Q $\beta$ )
Enzyme-catalyzed DNA replication	Prokaryotes	$10^{-7}$	$10^7$	$4 \times 10^6$ ( <i>Escherichia coli</i> )
Recombinative, enzyme-catalyzed DNA replication	Eukaryotes	$\leq 10^{-10}$	$10^{10}$	$3 \times 10^9$ (man)

nation, the basic mechanism of sexual heredity, and the transposition and integration of genes via mobile information-carrying elements, such as plasmids and viruses.

### C. Hypotheses on the Hypercycle as an Ordering Principle

Erroneous autocatalytic copying is the reason for selection and for the emergence of cooperation between nucleic acids and proteins. In the beginning of the prebiotic era, nucleic acids probably were the first self-replicating entities. The fidelity of their replication was mainly dependent on the stability of the complementary base pairs AU and GC that formed during the copying procedure. Based on their different hydrogen-bonding energies, very high error rates resulted that reached 1:10 for AU-rich polymers and 1:100 for GC-rich polymers. According to theory, these error frequencies led to a maximal information content comprising 10–100 symbols for nucleic acid replicators. However, even small enzyme molecules are coded by more than several hundred nucleotides, i.e., far beyond the critical value that is defined by enzyme-free replication. Thus, an information crisis resulted from the error threshold relation: Any increase in the coding capacity required a decrease in the replication error rate, and this in turn could be achieved only by application of error-correcting mechanisms. Contemporary replication machineries achieve “proof-reading” with protein enzymes replicating the genetic information which is stored in double-stranded DNA. In contrast to RNA, a newly synthesized daughter strand of DNA remains associated to its parental strand and thereby enables the comparison of both sequences. False incorporations can be identified by enzymes that test the shape of the double strand for true base pairing and that correct unpaired or mispaired regions by substitution. However, how did the first replication machinery with error-correcting

facilities evolve if the coding capacity needed for it could not faithfully be replicated without the very machinery? RNA might represent a major connecting link as was demonstrated recently by David Bartel (1963–) and colleagues. They generated RNA that synthesizes RNA using the same reaction as that employed by protein enzymes that catalyze RNA polymerization. In the presence of appropriate template RNA and nucleoside triphosphates, the ribozyme extends an RNA primer by successive addition of up to six mononucleotides, thereby showing significant template fidelity.

Eventually, the limiting chain length of approximately 100 nucleotides was overcome by the capability for translation of nucleic acid into protein. The appearance of RNA molecules which acted as adaptors or interpreters between nucleic acid and protein enabled the development of feedback loops and therefore the basis for cooperation. Together with these adaptor molecules, known as transfer-RNA (tRNA), the genetic code originated. tRNAs probably arose early and played an essential role in ancient replicating systems. Even today, tRNA and tRNA-like structures are involved in a variety of replicative processes, including replication of RNA viruses of bacteria, plants, and possibly mammals. Alan Weiner (1951–) therefore proposed that tRNA-like motifs emerged as 3' terminal structures that tagged RNA genomes for replication before the advent of protein synthesis. These tags could have served two main roles—providing an initiation site for replication and functioning as a simple telomere, i.e., ensuring that critical terminal regions were not lost during replication.

The first proteins that arose probably required more genetic information than a primitive molecular replicator was able to provide. Because more information needed to be stabilized, cooperation of differentiated nucleic acid molecules emerged that was mediated and regulated by proteins, their translation products. The enforced cooperation of otherwise competing



genes allowed their mutual survival and regulated their growth. It also enabled a more refined competition–selection behavior than that among the individuals of a quasispecies. Instead of single RNA mutants, cooperative ensembles of nucleic acid and protein were evaluated according to their self-replication qualities and according to their stabilities. First, those sequences were selected as fittest that were best able to get themselves replicated as quickly and as accurately as possible by the enzyme responsible for their replication. Second, with the continual introduction of new nucleic acid mutants, new catalytic couplings were constantly been tested.

The cooperation results in a double-feedback loop in which both the enzyme encoded by the nucleic acid template and the sequence information contribute to the replication of the template, a process called second-order autocatalysis. Phenomena of this particular kind were named “hypercycles” by Eigen and coworkers.

Hypercycles exhibit three major characteristics. First, they consist of coexisting and cooperating quasispecies, each of which is replicating autocatalytically and thereby maintaining competitive growth, selection, and its specific genetic information. Second, some of the coexisting quasispecies replicate and evolve independently. Finally, others form cooperative units within the hypercyclic organization and evolve together with mutual advantage. In an extreme case, some species as well as quasispecies may exist only by participating in hypercyclic organization. As a whole, the cooperating units can compete with other hypercyclically replicating systems.

The overcoming of the information crisis of self-replicating nucleic acids probably proceeded via hypercyclic coupling as shown in Fig. 3, and it supported the emergence of “translation” from nucleic acid into protein. It is assumed that hypercycles represent one major step in the organization of complex reaction networks which arose as naturally and continuously as did quasispecies. Experimentally, hypercyclic organization and its efficiency were demonstrated by Eigen and coworkers to mainly determine the infection and replication cycle of a simple bacteriophage such as Q $\beta$ .

#### D. From Hypercycles to Cells

All life on Earth is cellular and reproduces only when this complex replicative unit, the cell, divides. Among the obvious advantages of cellular organization are protection of the cell's content from fluctuations of

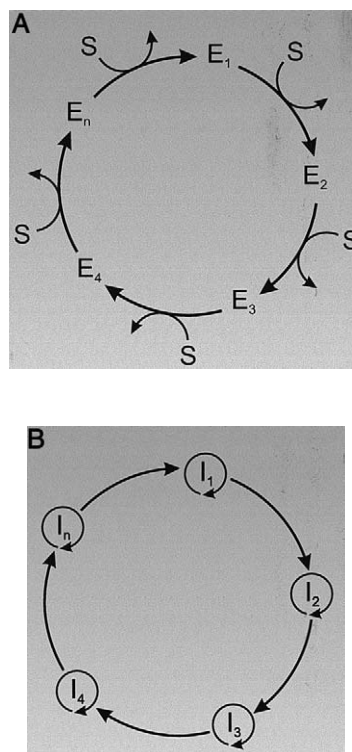


FIGURE 3 (A) Schematic representation of a catalytic cycle which represents a higher level of organization in a hierarchy of catalytic schemes. The constituents of the cycle ( $E_1$ – $E_n$ ) are catalysts which are formed from energy-rich substrates ( $S$ ). Each intermediate  $E_i$  acts as a catalyst for the formation  $E_i + 1$ . The catalytic cycle as a whole is equivalent to an autocatalyst. Replication of single-stranded RNA that involves mutual instruction (self-instruction) by complementary interaction represents an example of biological importance. (B) Schematic representation of a catalytic hypercycle that consists of self-instructive units with two-fold catalytic function. Each of the intermediates  $I_i$  which represent autocatalysts or catalytic cycles of the type shown in Fig. 3A is able to instruct its own replication and to provide catalytic support for the production of the subsequent intermediate. For example, a simple hypercyclic process is found in nature with the infection cycle of bacteriophage Q $\beta$ .

resources in the external environment and the maintenance of internal concentration gradients necessary to drive chemical reactions. Nevertheless, these facts do not explain the necessity for cellular organization: Instead, the origin of spatially limited reaction systems from a homogeneous prebiotic soup appears to be a result of problems in information processing.

Although hypercycles are the essential organizational form for the transition from self-replicating molecules to reproductive, multi-molecular machineries, they do not represent the ultimate optimum of organization. Hypercycles and quasispecies share a substantial evolutionary disadvantage: Both quasispecies competi-

tion and hypercyclic cooperation evaluate only the phenotypic properties of replicating nucleic acids, i.e., their replication rates and their stabilities. Hypercyclic coupling enables the discovery of molecular replicators which are beneficial to one or more others, and it supports the selection of those coexisting replicator combinations which gain mutual advantage by amplifying themselves. However, hypercycles cannot evaluate genotypic properties of nucleic acid sequences or their genetic messages. In other words, they do not selectively

amplify mutant sequences that demonstrate their advantages only before translation.

Primitive translation mechanisms probably gave rise to proteins that were more helpful to self-replication than miscellaneous proteins randomly occurring in a homogeneous distribution. With time, preferences between translation products and certain sequences became pronounced, and these more distinct interactions effected advantages due to more specific catalysis. Finally, the differences among the various template-

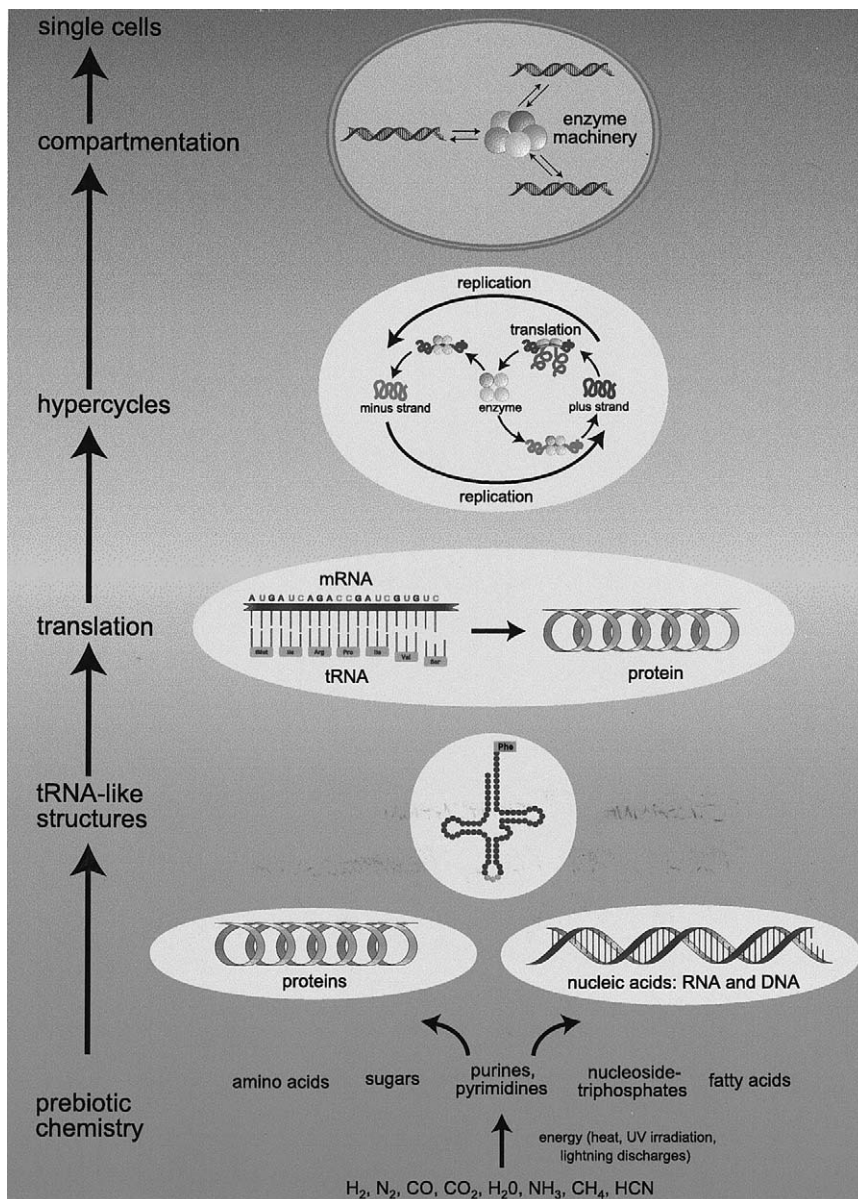


FIGURE 4 Stages of emerging life.

catalyst interactions became so influential that each enzyme had a particular catalytic role. At this stage, selection acted not only on the kinetic characteristics (i.e., the phenotype) of the replicating sequences but also on the information content (i.e., the genotype) of the sequences. Within the enlarged information system, the amount and quality of coding genes were strongly determined by efficiency, fidelity, and rate of the catalyzing enzymes, and these in turn were the translation products of their templates. By spatial separation of individual hypercycles, selective genotypical advantages in competition with other compartments were exploited. Favorably mutated compartments could develop more strongly than others and thus evolve more quickly. The compartmentalization was accomplished by individualization of all replicative units represented by a hypercycle. This event can be viewed as the birth

of the cell, the smallest living entity known to us today. Figure 4 summarizes the complete view that results from these theoretical descriptions.

### See Also the Following Articles

ARCHAEA, ORIGIN OF • BIODIVERSITY, ORIGIN OF • DARWIN, CHARLES • EUKARYOTES, ORIGIN OF • EVOLUTION, THEORY OF • NUCLEIC ACID BIODIVERSITY

### Bibliography

- De Duve, C. (1991). *Blueprint for a Cell: The Nature and Origin of Life*. Carolina Biological Supply, Burlington, NC.
- Eigen, M., and Schuster, P. (1979). *The Hypercycle. A Principle of Natural Self-Organization*. Springer, New York.
- Eigen, M., and Winkler-Oswatitsch, R. (1996). *Steps towards Life: A Perspective on Evolution*. Oxford Univ. Press, Oxford.



# PALEOECOLOGY

Thompson Webb III  
Brown University

---

- I. Introduction
  - II. Scaling Factors
  - III. Zoom Lens Perspective
- 

## GLOSSARY

**AMS dates** Radiocarbon dates obtained by directly measuring the amount of carbon 14 in a sample by using an accelerator mass spectrometer.

**Bryophytes** Mosses and liverworts; nonvascular plants.

**Cyperaceae** Sedge family.

***Fagus grandifolia*** American beech trees.

**Late Quaternary** The past 21,000 years since the last time of maximum glaciation.

**Palynology** The study of pollen and spores.

**Pliocene** The period of geological time from 3.5 to 1.8 million years ago.

**Poaceae** Grass family.

**Quaternary** The past 1.8 million years of geological time.

**radiocarbon dates** Dates of organic matter in geological samples using the radioactive decay of carbon 14, which has a half-life of 5750 years.

**Tracheophytes** Vascular plants.

---

through time on scales of decades to hundreds of millions of years. Paleocologists derive their inferences mainly from fossil and geological data and have assembled data sets with local to global coverage. These data sets provide the possibility of having a zoom lens view of past changes both in space and in time. I use Quaternary pollen and plant macrofossil data to illustrate how these data can be organized spatially, temporally, taxonomically, and numerically to provide a zoom lens view of a variety of ecological phenomena.

## I. INTRODUCTION

Past ecosystems are a continuous creation of those who study them. What we map through time or reduce to time series depends on the data available and the training of those who interpret and display those data. Studies in paleoecology, therefore, are an exercise in perception and interpretation and depend on the sensitivity of the data in time, space, and taxonomy. The ultimate goal is to understand what occurred ecologically in the past, but this understanding can only be obtained if reliable records exist of past taxon distributions (or other components of ecosystems) that can be displayed and interpreted in informative ways. Simple time series are not enough because of the many spatial variations and processes that affect them. Geographic networks of temporally varying data are required to show how past and present taxa have varied in abundance, location, and association within and among their associated ecosystems.

**PALEOECOLOGY** is the study of the composition and distribution of past ecosystems and their changes

The focus here is on the rich data sets of Quaternary pollen and plant macrofossils that provide records of the major changes in vegetation at a variety of space and time scales. These data provide a valuable bridge from modern ecological data to the vast fossil record in geological time that Jabolonski, Sepkoski, and others have compiled from the work of paleontologists. The Quaternary data sets yield a remotely sensed view of past vegetation, landscapes, taxon migrations, and invasions and illustrate various temporal changes in ecological assemblages and communities—all of which show changes and patterns in biodiversity. Direct measurement of changing species through time, however, is not possible with pollen data, even when supplemented by plant macrofossil data, although the extinction and arrival of some species are evident in records dating back to the Pliocene. The focus with these paleoecological data, therefore, is on vegetation and its diverse changing patterns of composition and distribution through time and space and how environmental factors, such as climate and human activities, have influenced these changes. These data therefore show the ecological and environmental setting within which modern patterns of biodiversity have arisen. They also demonstrate, as long claimed by Richard West and Margaret Davis, that the “present plant communities have no long history.”

### A. Data and Sensing System

For map views of the globe and continents, the pollen and plant macrofossil data can be viewed as if they came from a video recorder from high in space with resolution in places of up to 10 m. The images that are retrieved can be of high or low resolution temporally, spatially, taxonomically, and numerically, and they can illustrate local to global changes in plant populations, vegetation, biodiversity, human activity, fire frequency, and plant diseases over decades to millions of years. Because each of these entities or phenomena varies spatially and temporally, records of data covering a breadth of scales in space and time are needed. To obtain the highest quality images of a specific phenomenon requires an understanding of the sensing system that accumulated the data. How does the paleoecological video recording system work and what are the scaling characteristics of the images that it registers? These characteristics include breadth of coverage, sampling resolution, and sampling density in time, space, and taxonomy. Studies of pollen and plant macrofossil data covering a variety of temporal and spatial scales within the Quaternary have helped provide this understanding.

Quaternary paleovegetation data pose several prob-

lems for interpretation. The main one arises because pollen samples in sediments represent death assemblages of immature microgametophytes (i.e., pollen grains) that differ manifestly from the assemblages of sporophytes (i.e., plants) that produced the pollen. Studies of modern data from surface sediments have therefore focused on identifying the features of plant assemblages that appear in pollen. This work parallels that of scientists deciphering the remotely sensed data from satellites because pollen data are remotely sensed data from plant populations and vegetation. Just as the current vegetation emits or reflects radiation that remote sensors on satellites intercept, so too does (and has) the current (and past) vegetation shed pollen that accumulates “remotely” (i.e., well away from the source) in lakes and bogs. Both types of “remote” sensors (satellite instruments and lake sediments) record data with certain sampling characteristics (e.g., spatial and temporal resolution), and their data need calibration and ground-truthing in terms of vegetation attributes such as composition (taxon abundances), structure (height and mixture of growth forms), and pattern (geographic gradients and mosaics and aspect of beta and gamma diversity). Studies of spatial arrays of modern data by Bradshaw, Prentice, and Jackson have provided this calibration and ground-truthing.

Down-core studies of pollen bring time into the picture, and temporal resolution becomes one of the factors controlling what is recorded in the data. Pollen in annually laminated sediments can provide seasonally distinct samples as neatly shown by Peglar, but most samples integrate 10 or more years and can be independently dated back 40,000 years with an average precision of  $\pm 200$  years for data from within the past 12,000 years. However, if vegetational and ecological phenomena are the target for study, then spatially distributed data are needed because vegetation and biodiversity are inherently spatial entities that vary on virtually all time and space scales. Recording their full dynamics requires time series of geographically distributed data that yield a zoom lens space–time perspective. I focus on how the different scaling characteristics of the data sets control what we see and on how data can be organized to yield such a zoom lens perspective. I discuss each of the sampling characteristics in taxonomy, space, and time. I then describe different ways to display and visualize the data and some findings that they have revealed.

## II. SCALING FACTORS

Most Quaternary palynologists study sediments from lakes or bogs that accumulate pollen relatively continu-

ously, for which the data from the surrounding vegetation can be averaged, and whose sediments are radiometrically dated. In these sediments, pollen grains are morphologically distinct and numerous, and plant macrofossils are also found in a subset of these sediments. (Pollen and plant macrofossil data are also available from deposits that are discontinuous in time but can be organized into time series; Betancourt *et al.*, 1990). Once analyzed, the data exist as point estimates of the abundance for each taxon in time series of samples at a site (or at nearby sites and middens for some discontinuous records). These time series can be expanded to transects of time series in latitude-, longitude-, or elevation-time diagrams, in networks of samples from an area for mapping, or to networks of time series, which are also time series of maps, to form a space-time box for displaying the data (Fig. 1). Each of these displays illustrates different views in different dimensions of the space-time variations in the data. A zoom lens perspective can then be achieved in space, time, or both simultaneously by moving from data at a single site for a short time interval to progressively longer time series for either that site or for data on maps from increasingly larger areas. In this way, records of local succession can be seen in the broader context of long-term migrations and global climate changes.

Palynologists control the ability of their data to display selected patterns in vegetation and biodiversity (both in time and in space) by making choices about the numerical aspects of their data and by choosing the temporal, spatial, and taxonomic sampling characteristics of their data. These sampling characteristics have three elements—breadth of coverage, resolution of individual samples, and sampling density—which for a photograph are comparable to its frame of reference, grain size, and number of grains (or pixels) that are exposed, respectively. The temporal and spatial breadth of the data is the total time or area covered by a data set, and the taxonomic breadth is the total set of taxonomic groups (e.g., seed plants) included. The temporal and spatial resolution of each sample is defined in terms of how much time or area is represented in each individual sample within the data set, and taxonomic resolution depends on whether the data are lumped into groups or listed at their finest level of morphological distinction. The uncertainties of radiocarbon dates and age estimates also affect temporal resolution in correlations among sites. The number of samples in time and space defines the temporal and spatial sampling density. The total number of taxa listed in a data set defines the taxonomic sampling density. Choices about these char-

acteristics can influence what changes in vegetation and biodiversity appear in paleoecological times series and map sequences (Figs. 1–3).

### A. Taxonomic and Numerical Characteristics

In most studies of lakes and bog sediments, the potential taxonomic breadth is Tracheophytes and Bryophytes along with certain algal remains (*Pediastrum*) and fungal spores. When non-wetland vegetation is a primary focus, spores, algal remains, and pollen and plant macrofossils from aquatic plants are excluded from detailed study. Careful study and morphological characteristics of pollen grains determine the taxonomic resolution in samples and permit identification of genera and some species, although some grains can only be determined to the family level (e.g., Poaceae and Cyperaceae). Regional restriction of single species, otherwise identifiable pollen morphologically only at the genus level (e.g., *Fagus grandifolia* within the northeastern United States), can allow designation of plant species from pollen, but plant macrofossils, if present, give the best information on species identification.

Palynology is a direct beneficiary of the ubiquity and inefficiency of wind pollination. Millions of grains are produced for every successful pollination, and some of the abundant residual reaches the sediments of lakes and bogs where the pollen grains are well preserved. Entomophilous grains are abundant in honey and on the legs of bees, but too few insects perish in lakes and bogs to leave a record that can be discerned among the overwhelming numbers of anemophilous grains. From the point of view of the remote-sensing metaphor, wind-pollinated plants are bright lights on the landscape and little or no signal comes from the other plants. As a result of this bias, the focus is on using pollen data to record the vegetation rather than species lists. For temperate to boreal forests in which the diversity of wind-pollinated trees is greatest globally, pollen records provide a fairly direct representation of the vegetation; however, in tropical regions and deserts, the pollen records yield a much more indirect representation of the vegetation because most species are not wind pollinated.

In arid lands of southwestern North America, where lakes are rare, packrats create middens with embedded plant macrofossils that can be radiocarbon dated and identified to the species level for those plants collected (Betancourt *et al.*, 1990). These biased records provide valuable information about arid vegetation and environments.

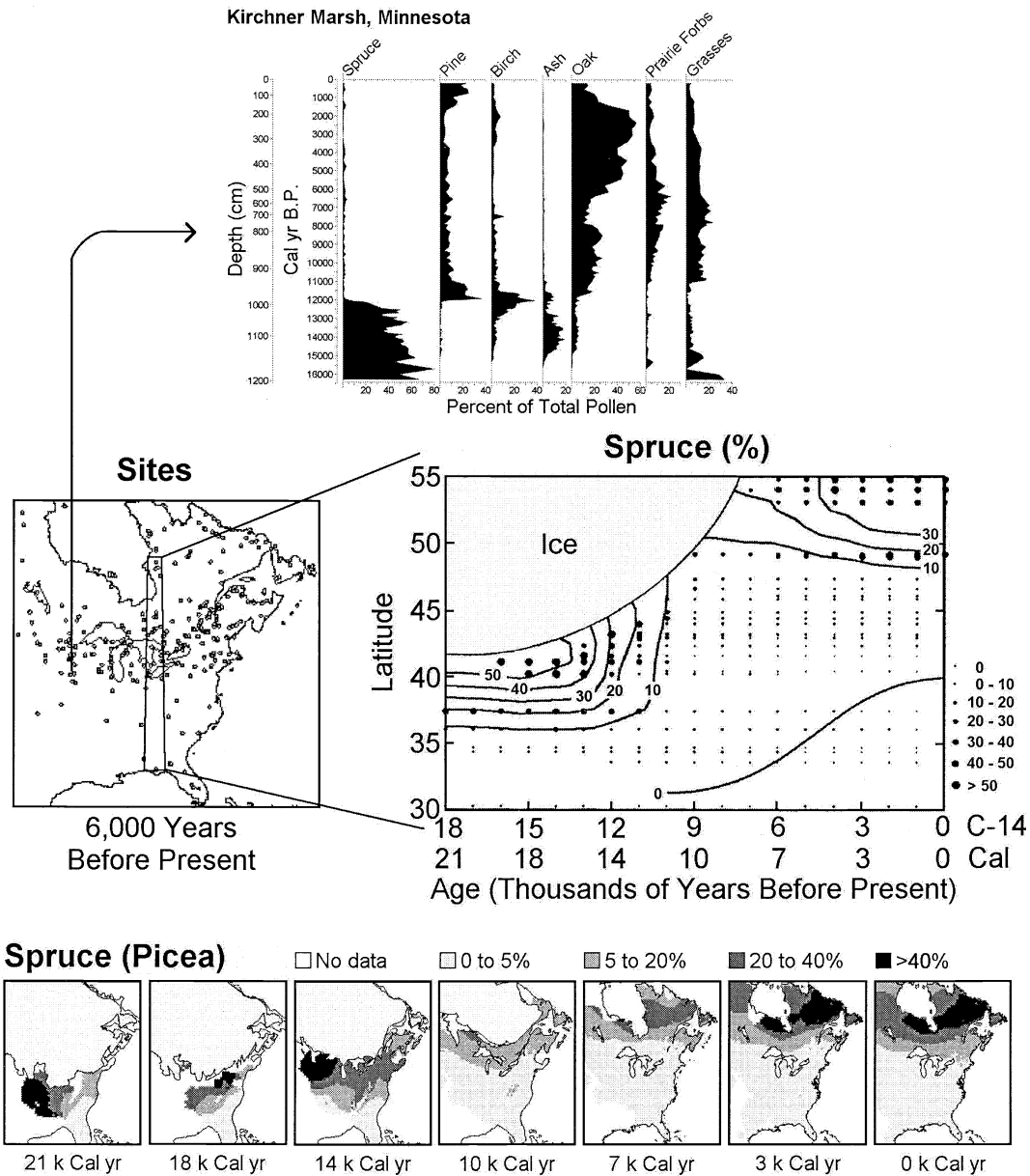


FIGURE 1 Pollen data in time and space. Different views from time series for multiple taxa at a single site to a single taxon (spruce) displayed either for a transect of sites in a latitude–time diagram or for a network of sites in the time series of maps from 21 k Cal yr (thousands of calendar years ago) to present. The shading on the maps indicates pollen abundances of 1–5% (light gray) and >5% (black). The white area at the top of the maps for 21, 18, 14, and 10 ka is the Laurentide ice sheet. When stacked vertically, these maps form a space–time box in which the three-dimensional contours can show the four-dimensional distribution of spruce pollen percentages in space–time.

In most studies of Quaternary pollen, taxon percentages are used. Three hundred to 500 grains are typically counted in a sample, and the counts for each taxon are divided by the total count. Because counts for individual taxa can be thought of as being binomially distributed (i.e., each grain is either taxon x or not taxon

x), the percentages for multiple taxa are multinomially distributed, thus permitting direct calculation of confidence intervals. Pollen concentrations (grains/cc or grains/g) are seldom used because of their dependence on sedimentation rates. When radiocarbon or stratigraphic dates allow estimation of sedimentation rates,

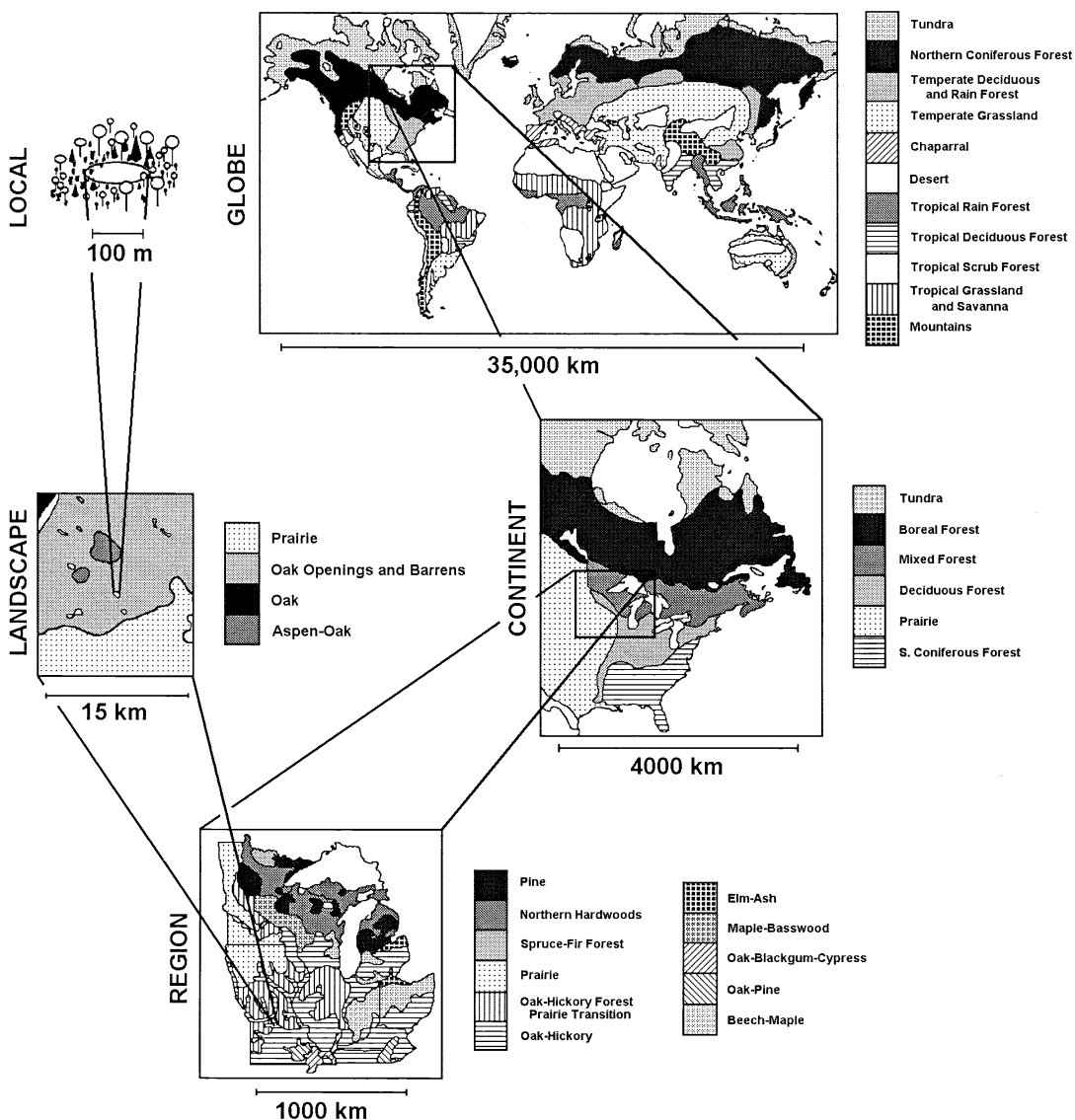


FIGURE 2 Zoom lens view of the vegetation from the scale of single site (where pollen accumulates) up to the global scale. No one view or interpretation of the vegetation is possible because the composition and gradients vary with scale and reflect climate gradients at the global and continental scales, a mixture of soil and climate gradients at the regional scale, and difference in soils and disturbance history at the landscape and local scales (modified from Kutzbach and Webb, 1991).

pollen accumulation rates (grains/cm<sup>2</sup>/year) can be calculated, and procedures exist for calculating and potentially minimizing the confidence intervals for both pollen concentrations and accumulation rates.

Pollen accumulation rates, first introduced by Margaret Davis, have proven valuable for checking the ambiguity of certain changes in pollen percentages. However, pollen accumulation rates are unsuitable when large sets of comparable pollen data are required for mapping. These rates are highly sensitive to sedimenta-

tion differences within and between lakes as shown by W. Pennington and M. Davis, and percentages of pollen taxa are generally used in mapping studies (Figs. 1–3). Many studies show that the percentages for pollen taxa represent well the relative abundance of plants on the landscape. The relationships vary with spatial scale and pollen type, and the uncertainties of these relationships add to the number of uncertainties for the pollen data when estimates of the vegetation are attempted.



Finding reliable quantitative measures for plant macrofossils is much more difficult than for pollen data. Most data are recorded in presence/absence terms or categorically, but sometimes percentages or concentrations are calculated. Local biases can make the data difficult to interpret because, for example, 100 needles may all come from the same tree.

### B. Spatial Characteristics

The breadth of coverage for sets of pollen data can be broad or fine and thus match the different scales for mapping the vegetation (Fig. 2). Continental and global data sets exist (Prentice *et al.*, 2000) along with those at regional (1000–300,000 km<sup>2</sup>) and local (10<sup>-3</sup> km<sup>2</sup>)

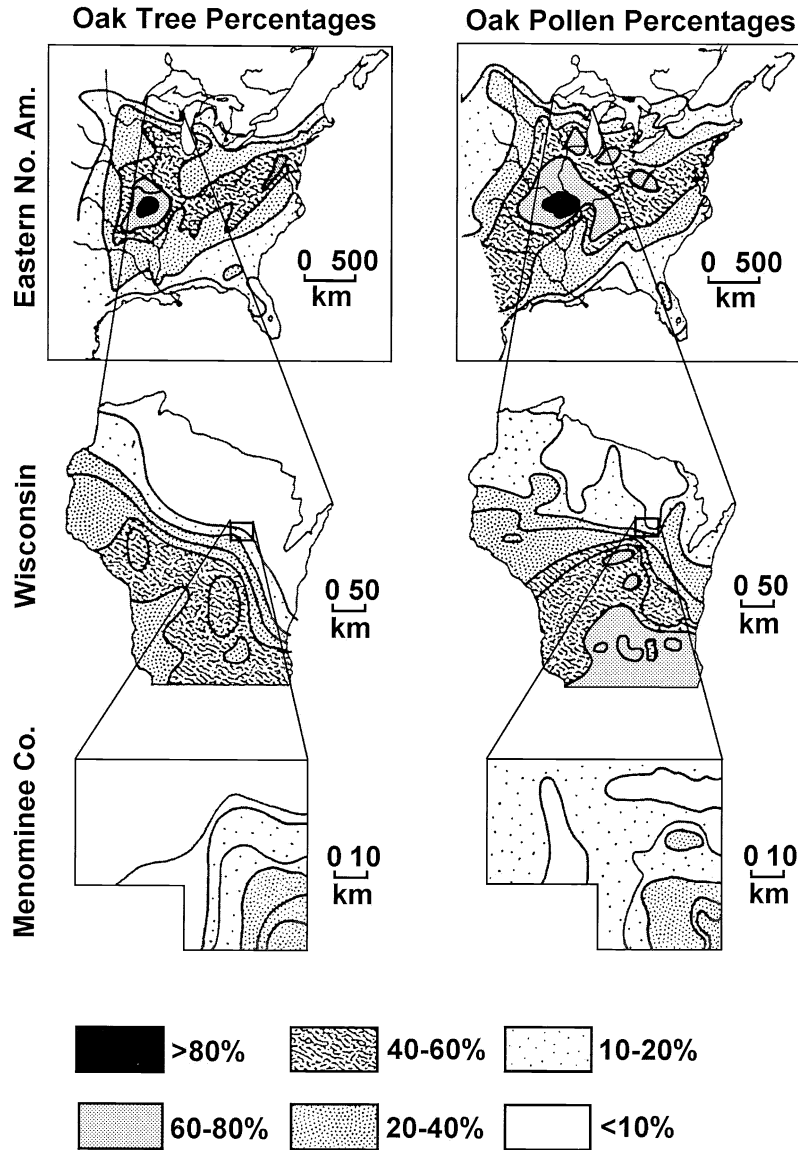


FIGURE 3 Zoom-lens view of the mapped distributions of oak tree and pollen percentages at the scales of a subcontinent (10<sup>7</sup> km<sup>2</sup>), state (10<sup>5</sup> km<sup>2</sup>), and county (10<sup>3</sup> km<sup>2</sup>). These maps show how well pollen percentages can reflect the distribution of tree percentages. Data used in contouring at the subcontinental scale were also used in contouring at the state scale, but a different data set was used at the county scale (modified from Solomon and Webb, 1985).

scales, and an embedded series of these data sets can provide a zoom lens view of past vegetation change if the sample resolution and sampling density are appropriate.

The potential spatial resolution of individual pollen samples varies among pollen taxa depending on the dispersibility of their pollen. Pollen from taxa with large grains is dispersed less far than that with small grains. Individual pollen samples are therefore variable-area samples, with the area of vegetation contributing pollen depending on which pollen type is studied; the percentages for pine pollen, for example, represent a larger area than those for beech or maple. Such differences in sampling area among pollen types can be important for studies of local or even regional vegetation but are less important for networks of data in which the distance between samples is much larger than the average dispersal distances for the different taxa.

The potential spatial resolution also depends on the basin characteristics. Key factors include the presence or absence of a canopy over the basin, the size of the water or wetland surface accumulating airborne pollen, the presence or absence of horizontal mixing of sediments within the basin, and the watershed area that supplies additional waterborne pollen to lakes. Samples from within a forest canopy accumulate most of their pollen in radiuses from 10 to 100 m, and data from samples that preserve pollen are comparable spatially to vegetation data from large permanent plots in forests. In contrast, samples from lakes accumulate their pollen on scales from 10 to 50 km. The former data have enough spatial resolution to illustrate succession within forests, whereas pollen data from lakes integrate across the mosaic of vegetation on landscapes and are sensitive to succession only if it is occurring across much of the pollen-source area. Bogs, mires, and shallow wetlands accumulate pollen at both regional (1–10+ km) and local (10 m) scales because many plants grow on these wetlands. In samples within mires, pollen from local taxa can vary systematically and abruptly, whereas the pollen from regional taxa remains reasonably uniform. (This behavior of pollen percentages for regional taxa led von Post to develop pollen analysis as a method in 1916.) The spatial resolution of pollen samples, therefore, varies with pollen type within certain basin types, but selection of samples by pollen type and basin type can help keep the spatial resolution reasonably uniform. Such choices are key to “seeing” well with the observations afforded by sets of pollen samples.

For plant macrofossils, Dunwiddie and Jackson have shown that needles and seeds in lake sediments have a spatial sampling resolution of 10–20 m, which is much finer than the distance sampled by most pollen in

lakes or open wetlands. Plant macrofossils can therefore help resolve local changes at or near a site and fine-scale elevation differences. Therefore, sampling schemes are possible that mix open and canopy-covered basins, mire and lake sediments, local and regional pollen types, and plant macrofossils in order to represent different scales of spatial pattern in the vegetation. Such mixtures of data sets can show vegetation differences that reflect specific soil and elevation differences within a broader geographic data set.

The geographical sampling density for pollen data varies. To illustrate the variations in recent pollen within a basin or a forest, some studies include samples at 10-m or finer intervals along 500-m or longer transects through a basin or forest. For selected taxa, these studies show high sensitivity in the pollen data to local variations in vegetation cover and biodiversity. The variations are also evident in sets of time series from closely spaced cores in a basin (Simmons, 1993). For studies whose breadth of coverage is regions to continents (Fig. 2), sets of modern data exist with average densities of 1/14 km<sup>2</sup> in 1000 km<sup>2</sup>, 1/500 km<sup>2</sup> in 100,000 km<sup>2</sup>, and 1/5000 km<sup>2</sup> in 10<sup>7</sup> km<sup>2</sup>. These densities are sufficient for the contour patterns of oak pollen percentages to match the corresponding patterns of oak tree percentages at each spatial scale (Fig. 3). Sets of fossil data provide less dense geographic coverage and vary in density from 1/50 km<sup>2</sup> for the Adirondacks to 1/6000 km<sup>2</sup> for the northern Midwest and 1/40,000 km<sup>2</sup> for eastern North America and Europe.

These data sets have been used to illustrate how well pollen data represent or remotely sense spatial vegetational features such as range boundaries, ecotones, and abundance gradients (Figs. 2 and 3). Detecting the range boundary with pollen can be difficult because of pollen transport and relatively high counting uncertainties at low pollen percentages, but Davis *et al.* (1991) used a data set from the Midwest with a sampling density of 1 sample/1000 km<sup>2</sup> to show that the past species limit for beech and hemlock can be identified within an area 20 km wide. For widely dispersed taxa such as oak and pine, such fine-scale resolution may not be possible, but the abundance gradients for these taxa can be used to locate ecotones at scales of 10–100 km (Fig. 3).

### C. Temporal Characteristics

The time range for records from individual cores can vary from 50 years to millions of years ago (Fig. 4), with the bulk of the late Quaternary cores covering 14,000 years and fewer extending back 21,000 years,

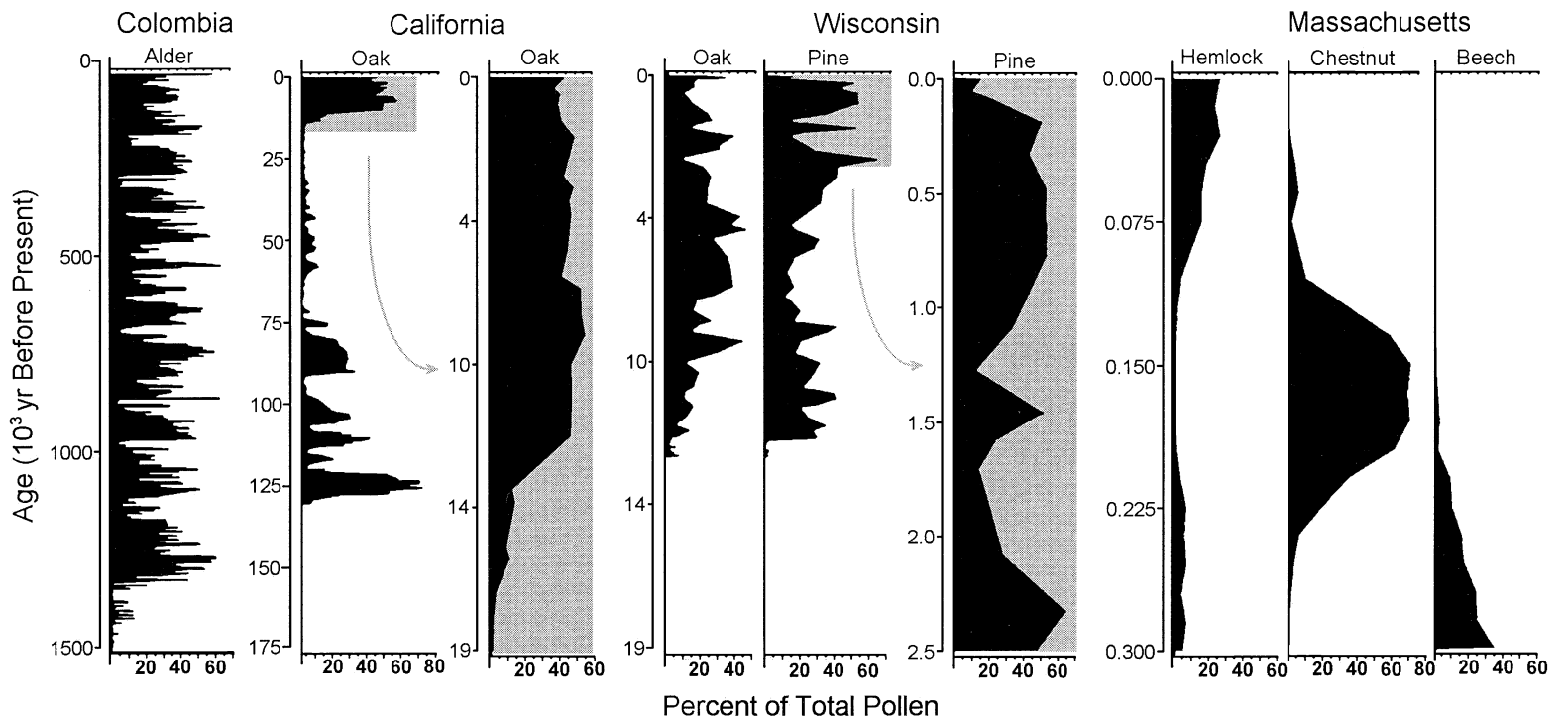


FIGURE 4 Zoom lens view of pollen data in time from a 1.5-million-year record down to a 300-year record. Different ecological processes are evident on different timescales. Temporal coverage, resolution, and density all vary appropriately for the Colombian lake record of alder pollen to show (a) alder's arrival at the site long after North and South America joined and (b) its response to orbitally driven climate changes (Hoogheemstra and Ran, 1994). The California lake record of oak pollen shows its response to orbitally driven climate variations over 130,000 years and the transition from glacial to interglacial climate over 16,000 years. In Wisconsin, pollen data from a lake that is free from bottom-organism mixing are sufficiently high in resolution to record how oak and pine populations responded to climate and regional disturbance events (e.g., fire) over 11,000 and 2500 years, respectively. Local succession is evident in the pollen data from a 300-year record from a small hollow within the forest canopy in Massachusetts (Foster and Zebryk, 1993). Gray indicates a portion of a time series that is expanded in the graph on the right.

the last glacial maximum. Temporal density generally correlates with time range because short cores are often taken to study short-term changes and may contain 10–20 samples for 500 years, which yields 1 sample per 25–50 years, whereas 200 samples in a 2-million-year record can only yield 1 sample per 20,000 years. Twenty to 40 samples for 10,000 years yield 1 sample per 250–500 years, which is typical for most cores in the data sets from eastern North American and European (Huntley and Webb, 1988). At a high level of temporal density are records from millimeter-scale sampling of peats that have yielded 1–5 year sampling for 50–200 years for selected times in the past, as reviewed by Turner and Peglar (1988) and Simmons (1993).

One of the longest continuous records for the Quaternary is from Sabana de Bogota in Colombia and covers 1.5 million years (Fig. 4). This study is remarkable in having relatively high sampling density of 1 sample/1000 years and shows the appearance from North America of *Alnus* and *Quercus* at 1.3 million and 250,000 years ago, respectively. Other long cores with records of 50,000 years or longer exist on all continents but Antarctica; however, only in Europe are these long cores numerous enough to allow some mapping of vegetation from previous glacial and interglacial periods.

At the other extreme in terms of temporal coverage are (i) short cores of the upper sediments of lakes or bogs and (ii) profiles from forest soils that preserve pollen. The former provide records of human impact on the vegetation and of recent changes, whereas the latter show combined fine-scale sampling in both time and space and provide the view most closely approximating scale within which succession into forest gaps occurs. One of the ironies of paleoecological research is how seldom the data reveal succession as the dominant dynamic in the vegetation even though much ecological research has focused on succession.

These short-term records are often analyzed in high temporal density, e.g., 1 sample per 50 years or even 1 per decade. The short cores (50 cm to 2 m) from lakes give coverage in high detail for 100–500 years or more in transects or in networks of relatively high spatial density (1 sample/1000 km<sup>2</sup>) from states or regions. Dating can be fairly accurate and precise because historical dates can be assigned to events in cores, and the date for the core top is usually known. High-resolution mapping in 50-year intervals is possible.

#### D. Temporal Resolution

A key metaphor for thinking about temporal resolution within individual samples and cores is to consider stratigraphic sections of lake sediments as strip charts along

which pollen data accumulate to form “palygraphs.” The sediment accumulation rate (which varies but averages 0.7 mm per year in most Holocene lakes) measures the drum speed (as on a thermograph) that drives the strip chart past the recorder, and the mixing of sediments within 1 to a few cm of the sediment surface before burial measures the degree to which the pen recording the data wiggles up and down vertically in time while registering horizontally the changing abundances of pollen. Within this sediment mixing zone, older sediments are moved both up from below (2–6 cm) and in from other parts of the basin by reworking. Such processes can make for tricky pen work on the palygraph. If the paper in the imagined palygraph were to absorb the ink slowly such that the scatter of tracings in the upper mixed zone is reduced to a single tracing of the abundances below this zone, then our metaphorical instrument would yield what is finally observed down-core. This image of the palygraph helps by indicating the separate components controlling temporal resolution in a core. The mixing depth and sedimentation rate largely control postdepositional time averaging, which regulates the time interval (i.e., temporal resolution) represented by pollen in a sample, whereas the temporal density of samples recorded and the accuracy and precision of assigning dates to each of many cores (strip charts) control analytical time averaging, which determines the temporal resolution among a set of cores used for mapping.

#### E. Dating Uncertainties

The dating uncertainty of an arbitrary pollen sample at some depth in a core depends on the uncertainties of the dating method, what is dated, the uncertainties of the age model, and variations in the sediment accumulation rate. The most accurate and precise dates are from lakes with annually laminated sediments, whose annual layers can be counted. These dates are in calendar years and accurate, with dating errors of only 1 or 2% at most. Unfortunately, lakes with such precision are too sparsely distributed for detailed mapping work. For lakes without varves, the dating near the tops of cores is more accurate and precise than that lower in the core because historical dates can be assigned to events in the core. These dates have uncertainties of 50 years or less and can be pinpointed in the core to 1 or 2 cm.

In studies of late Quaternary data, most cores are dated by conventional radiocarbon dates of bulk sediments. The bulk dates give the dates for the sediment matrix (i.e., strip chart material) and not for the pollen embedded within the sediments (i.e., what the pen records). The bulk dates have counting uncertainties of

20–400 years, with most approximately 100 years. These average “errors” give  $\pm 200$  years as the 95% confidence interval for each date. The dates can also be derived from 2–10 cm of sediment and therefore may average sediments covering 20–1000 years in age depending on the sediment accumulation rate. When dates for “synchronous” pollen events are averaged, such as the elm decline in Europe and the hemlock decline in eastern North America, standard deviations of ca. 300 years result. This uncertainty incorporates both the depth and the counting uncertainties. As Olsson (1986) observed, “There is . . . no point in determining the radiocarbon age with an uncertainty of  $\pm 50$  years if the relevant pollen-analytical level has an uncertainty of  $\pm 200$  years.”

Olsson (1986) also notes the problems with contamination, old carbon, and sampling depths. Old carbon can lead to systematic errors of 2000 or more years and requires correction. Dates on wood and accelerator mass spectrometer (AMS) dates on macrofossils in cores are helping to reduce these errors, and AMS radiocarbon dates are allowing depth intervals of 1 cm or less to be dated. Pilcher (1993) concludes that standard errors for radiocarbon dates from sediments are potentially  $\pm 200$  years but not more than  $\pm 225$  years.

### F. Age Models and Mapping Intervals

Most pollen samples (30–50 per core) are not dated directly by a radiocarbon date (3–6 per core). Rather, their age is estimated by an age model based on linear interpretation, regression, or some other curve-fitting method. For cores with continuous sedimentation, the age model estimates the drum speed for the strip chart on the palygraph.

The uncertainties for all data plotted on the map for 12,000 years ago give an estimate for the amount of time before and after 12,000 years ago that the pollen data may cover on the map. The uncertainties are likely to be between 300 and 500 years. This dating uncertainty for the data mapped matches the average temporal sampling density of one sample per 300–500 years in late Quaternary pollen studies. Given such large uncertainties, maps are best restricted to 1000-year intervals. Attempts to map at finer intervals will only produce maps whose data are not independently observed from the data in the maps for the next earlier and later dates.

For sets of short cores in which historical dates yield a precision of 50 years or less, age models back to 500 years can yield dating sufficiently precise for maps in 100- to 50-year intervals. For regions over which stratigraphic markers such as volcanic ash represent synchro-

nous layers in cores, some fine-scale mapping intervals are possible for the distant past, but even the dating of tephtras has uncertainties of a century or more (Brown *et al.*, 1989). Pollen-assemblage zones are sometimes used chronostratigraphically, especially in Europe, and can yield temporal resolution to 50 years or less among cores, but synchronicity must be assumed for processes (e.g., abundances changes for individual taxa) that are ultimately time transgressive, especially at scales of 200 km or more. The geographic area for such maps is therefore relatively small.

### III. ZOOM LENS PERSPECTIVE

Time series and mapped networks of pollen and plant macrofossils record late Quaternary vegetation patterns at the scale of local vegetation, landscapes, regions, and continents (Fig. 2). A hierarchy of nested data sets can reveal vegetation patterns at a variety of scales and provide information from general to detailed about the spatial vegetation patterns (Figs. 2 and 3). Such hierarchies can link local changes to global events by recording each simultaneously (Kutzbach and Webb, 1991). A paleoecological “zoom lens” can then be used to examine the vegetation patterns that dominate different spatial scales for areas of  $10^0$  to  $10^7$  km<sup>2</sup> with spatial resolution of  $10^{-2}$  to  $10^4$  km over timescales covering  $10^{-1}$  to  $10^4$  years (with time series to  $10^6$  years) with resolution from 1 to 500 years (Fig. 4). From a global perspective over thousands of years with a relatively coarse grid of temporally and spatially averaged samples, investigators can zoom in ultimately to the locally small but statistically significant short-term changes in species abundances within one wetland or forest stand. A zoom lens perspective is possible for eastern North America and Europe and is becoming possible in western North America and other regions.

Two examples using current data illustrate this zoom lens view. The first holds time relatively constant and focuses up spatially from local to continental (Figs. 1–3), and the second zooms in from a 1-million-year timescale down to successional changes during the past 300 years (Fig. 4). The first example covers the past 21,000 years since the last glacial maximum in the northern Midwest and focuses up from a local to a continental perspective. It shows how the local changes in a wetland during the past 12,000 years in central Minnesota are part of regional climate changes that ultimately reflect the global climate changes during deglaciation and the current interglacial period.

At the local scale of the sampling basin, plant macro-

fossils, aquatic pollen, and diatoms at Kirchner Marsh in east-central Minnesota illustrate that 3000 years after the Laurentide ice sheet retreated from the sites ca. 19,000 years ago, a buried ice block melted to form the kettle basin and its associated lake (Kutzbach and Webb, 1991). Water levels were high until approximately 11,000 years ago, when the seeds of damp ground and weedy annuals began appearing in the core. These appeared when the mixture of oak and herb pollen data show the first evidence of oak savanna conditions within the landscape and region near the site (Fig. 1). When the herb pollen values increased sufficiently to indicate prairie vegetation near the site, large fluctuations in aquatic seeds indicate intermittent droughts that continued until a marsh developed at the site 2000 years ago. This history of local wetland development is closely linked to the vegetation and climate changes indicated at the landscape, regional, and continental levels (Figs. 1 and 2).

Here, I shift from the time series view provided by local pollen and macrofossils to the continental view that the maps of pollen data provide. These show how 15,000 years ago spruce trees grew initially in a parkland that became populated with deciduous trees, first ash and then birch and oak, from 14,500 to 12,000 years ago as forests began to develop near the site for the first time (Fig. 1). This early vegetation was unlike any growing today and illustrates the major compositional changes in vegetation that accompanied the major changes in climate. The sudden decrease in spruce abundance near Kirchner Marsh 12,500 years ago reflects the general northward movement and decrease in regional and continental spruce populations as climate warmed and the ice sheet retreated. Pine populations were then invading from the east to replace the birch trees that briefly grew abundantly near the site and in the northern Midwest generally. Oak and elm populations then increased to replace the pines as the climate warmed. The regional addition of herb pollen after 11,000 year ago signals the development of savanna conditions as the climate dried and the ice sheet retreated further north. The savanna grew until prairie vegetation developed approximately 8000 years ago as the grasslands spread from the west across Minnesota and the northern Midwest. By then, the ice sheet had almost disappeared and spruce populations were beginning to grow abundantly in eastern and central Canada where the boreal forest has developed today. With the retreat westward of the prairie-forest ecotone after 6000 years ago, open oak forests began growing near the lake that with further growth in aquatic vegetation became a marsh after 2000 years ago. Oak populations then

decreased regionally as part of a general retreat of oak in the north as conifer populations increased and spruce populations increased south from Canada into the northern Midwest and New England. The broad-scale regional and continental maps show how the local and landscape changes in vegetation fit within the broader context of changes.

The second example is a zoom lens perspective in time (Fig. 4) and shows time series for taxa reflecting the long-term temporal beat in climate at long timescales (1.5 million years to 16,000 years) but then shifting to reflect fires, disease, and other disturbances at shorter timescales, especially when the pollen data are derived from within a forested hollow. Within the latter, the pollen reflects local changes in trees next to the site.

This arrangement and description of the data allow paleoecologists to zoom in or out in space-time to observe different aspects of vegetation dynamics and patterns of biodiversity from long-term changes on several spatial scales (from continental to inside the forest canopy) to competitive interactions after disturbance within communities. With such an ordering of data and images, the interconnected roles and impacts of climate change, disturbance, disease, succession, soil development, competition, and evolution can all be observed and potentially distinguished from one another in studies of past biodiversity.

My own studies in mapping and studying pollen data in space-time have convinced me of the potential for resolving many different views of the past. No single fixed picture is possible because there is no preferred viewing scale or perspective, nor is there any one way to describe the vegetation and biodiversity. Many conventional displays and perspectives are used, but some, such as traditional pollen diagrams, may not be optimal for interpreting key variations in the vegetation. A zoom lens perspective in space and time is required to help explore all the possibilities and to allow for diverse narratives and explanations. Succession and patch dynamics are evident with data scaled to meters and decades, whereas climatically induced migrations and abundance changes appear best over thousands of kilometers and years. To zoom out from a local to a global view, the paleoecological data need to be organized so that highly resolved images of species near a site give way to more generalized averages of data values whose patterns will stand out sharply when viewed across continents or the world. Appropriate study of assembled databases should allow ecologists to resolve and perceive a variety of phenomena, such as ecotones, disturbance horizons, species distributions, and migration patterns, that are important for biodiversity.

## Acknowledgments

An NSF grant from the Earth Systems History Program supported this research. I thank K. Anderson, P. J. Bartlein, S. Jackson, P. Newby, P. Leduc, L. Sheehan, and B. Shuman for technical assistance and constructive comments.

## See Also the Following Articles

BIODIVERSITY, EVOLUTION OF • FOSSIL RECORD • REMOTE SENSING AND IMAGE PROCESSING

## Bibliography

- Bennett, K. D. (1997). *The Pace of Life*. Cambridge Univ. Press, Cambridge, UK.
- Betancourt, J. L., Van Devender, T. R., and Martin, P. S. (Eds.) (1990). *Packrat Middens: The Last 40,000 Years of Biotic Change*. Univ. of Arizona Press, Tucson.
- Bradley, R. S. (1999). *Paleoclimatology*. Academic Press, San Diego.
- Clark, J. S. (1998). Why trees migrate so fast: Confronting theory with dispersal biology and the paleorecord. *Am. Nat.* 152, 204.
- Coope, G. R. (1994). Insect faunas in ice age environments: Why so little extinction? In *Extinction Rates* (J. H. Lawton and R. M. May, Eds.), pp. 55–74. Oxford Univ. Press, Oxford.
- Davis, M. B., Schwartz, M. W., and Woods, K. (1991). Detecting a species limit from pollen in sediments. *J. Biogeogr.* 18, 653.
- Fægri, K., and Iversen, J. (1989). *Textbook of Pollen Analysis*, 4th ed. Wiley, Chichester, UK.
- FAUNMAP Working Group (1996). Spatial responses of mammals to late Quaternary environmental fluctuations. *Science* 272, 1601.
- Foster, D. R., and Zebryk, T. M. (1993). Long-term vegetation dynamic and disturbance history of a *Tsuga*-dominated forest in New England. *Ecology* 74, 982.
- Gaudreau, D. C., Jackson, S. T., and Webb, T., III (1989). The use of pollen data to record vegetational patterns in regions of moderate to high relief. *Acta Bot. Neerlandica* 38, 369.
- Graumlich, L. J., and Davis, M. B. (1993). Holocene variation in spatial scales of vegetation pattern in the upper Great Lakes. *Ecology* 74, 826.
- Grimm, E. C., Jacobson, G. L., Jr., Watts, W. A., Hansen, B. C. S., and Maasch, K. A. (1993). A 50,000-year record of climate oscillations from Florida and its temporal correlation with Heinrich events. *Science* 261, 198.
- Guiot, J., de Beaulieu, J. L., Cheddadi, R., David, F., Ponel, P., and Reille, M. (1993). The climate in western Europe during the last glacial/interglacial cycle derived from pollen and insect remains. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 103, 73.
- Heusser, L. E., and Morley, J. (1997). Monsoon fluctuations over the past 350 kyr: High-resolution evidence from northwest Asia/northwest Pacific climatic proxies (marine pollen and radiolarians). *Quat. Sci. Rev.* 16, 565.
- Hoogheemstra, H., and Ran, E. T. H. (1994). Late Pliocene–Pleistocene high resolution pollen sequence of Colombia: An overview of climatic change. *Quat. Int.* 21, 63.
- Huntley, B. (1996). Quaternary palaeoecology and ecology. *Quat. Sci. Rev.* 15, 591.
- Huntley, B., and Webb, T., III (1988). *Vegetation History*. Kluwer, Dordrecht.
- Jablonski, D., and Sepkoski, J. J. (1996). Paleobiology, community ecology, and scales of ecological pattern. *Ecology* 77, 1367.
- Jackson, S. T. (1994). Pollen and spores in Quaternary lake sediments as sensors of vegetation composition: Theoretical models and empirical evidence. In *Sedimentation of Organic Particles* (A. Traverse, Ed.), pp. 253–286. Cambridge Univ. Press, Cambridge, UK.
- Kutzbach, J. E., and Webb, T., III (1991). Late Quaternary climatic and vegetational change in eastern North America: Concepts, models, and dates. In *Quaternary Landscapes* (L. C. K. Shane and E. J. Cushing, Eds.), pp. 175–217. Univ. of Minnesota Press, Minneapolis.
- Lotter, A. F., Eicher, U., Siegenthaler, U., and Birks, H. J. B. (1992). Late-glacial climatic oscillations as recorded in Swiss lake sediments. *J. Quat. Sci.* 7, 187.
- McDowell, P. F., Bartlein, P. J., and Webb, T., III (1990). Long-term environmental change. In *The Earth as Transformed by Human Action* (B. J. Turner, II, W. C. Clark, R. W. Kates, J. F. Richards, J. T. Mathews, and W. B. Meyer, Eds.), pp. 143–162. Cambridge Univ. Press, New York.
- Olsson, I. U. (1986). Radiometric dating. *Handbook of Holocene Palaeoecology and Palaeohydrology* (B. E. Berglund, Ed.). Wiley, Chichester, UK.
- Overpeck, J. T., Webb, R. S., and Webb, T., III (1992). Mapping eastern North American vegetation change over the past 18,000 years: No analogs and the future. *Geology* 20, 1071.
- Peglar, S. M., Fritz, S. C., Alapieti, A., Saarnisto, M., and Birks, H. J. B. (1984). Composition and formation of laminated sediments in Diss Mere, Norfolk, England. *Boreas* 13, 13.
- Pilcher, J. R. (1993). Radiocarbon dating and the palynologist: A realistic approach to precision and accuracy. In *Climate Change and Human Impact on the Landscape* (F. M. Chambers, Ed.), pp. 23–32. Chapman & Hall, London.
- Prentice, I. C., Jolly, D., and BIOME 6000 participants (2000). Mid-Holocene and glacial-maximum vegetation geography of the northern continents and Africa. *J. Biog.*, in press.
- Roberts, N. (1998). *The Holocene*, 2nd ed. Blackwell, Oxford.
- Russell, E., Davis, R. B., Anderson, R. S., Rhodes, T. E., and Anderson, D. S. (1993). Recent centuries of vegetational change in the glaciated north-eastern United States. *J. Ecol.* 81, 647.
- Simmons, I. G. (1993). Vegetation change during the Mesolithic in the British Isles: Some amplifications. In *Climate Change and Human Impact on the Landscape* (F. M. Chambers, Ed.), pp. 109–118. Chapman & Hall, London.
- Solomon, A. M., and Webb, T., III (1985). Computer-aided reconstruction of late-Quaternary landscape dynamics. *Annu. Rev. Ecol. Syst.* 16, 63.
- Thompson, R. S., Anderson, K. H., and Bartlein, P. J. (1999). Atlas of relations between climatic parameters and distributions of important trees and shrubs in North America. USGS Prof. Paper 1650.
- Tornqvist, T. E., De Jong, A. F. M., Oosterbaan, W. A., and van der Borg, K. (1992). Accurate dating of organic deposits by AMS 14C measurements of macrofossils. *Radiocarbon* 34, 566.
- Tsedakis, P. C. (1994). Vegetation change through glacial–interglacial cycles: A long pollen sequence perspective. *Trans. R. Soc. London B* 345, 403.
- Webb, R. S., and Webb, T., III (1988). Rates of sediment accumulation in pollen cores from small lakes and mires of eastern North America. *Quat. Res.* 30, 284.
- Whitlock, C., and Bartlein, P. J. (1997). Postglacial vegetation and climate in northwest America during the past 125 kyr. *Nature* 388, 57.
- Wright, H. E., Jr., Kutzbach, J. E., Webb, T., III, Ruddiman, W. F., Street-Perrott, F. A., and Bartlein, P. J. (Eds.) (1993). *Global Climates Since the Last Glacial Maximum*. Univ. of Minnesota Press, Minneapolis.



# PARASITISM

Klaus Rohde  
University of New England

---

- I. Related Phenomena
  - II. Types of Parasites
  - III. Adaptations to Parasitism
  - IV. Origins of Parasitism and Complex Life Cycles, the Evolution of Virulence, and Coevolution of Hosts and Parasites
  - V. Host–Parasite Interactions
  - VI. The Ecological Niches of Parasites
  - VII. The Structure of Parasite Communities
  - VIII. Parasite Population Dynamics
  - IX. The Diversity of Parasites (Distribution of Parasites in the Animal and Plant Kingdoms)
  - X. Zoogeography of Parasites
  - XI. Economic and Hygienic Importance of Parasites
- 

## GLOSSARY

**adult parasite** A parasite associated with a host during part or the whole of its mature phase.

**commensalism** An association of animals in which one uses food supplied in the internal or external environment of a host without affecting the host in any way.

**ectoparasite** A parasite living on the surface of a host.

**endoparasite** A parasite living inside a host.

**facultative parasite** A parasite that can also live without a host.

**final (= definitive) host** A host that harbors sexually mature stages of a parasite.

**hyperparasite (of first, second, etc. degree)** A parasite living on or in another parasite.

**intermediate host** A host that harbors sexually immature, developing stages of a parasite.

**intraspecific parasitism** A parasitic association of members of the same species.

**larval parasite** An organism that is parasitic only at a larval stage.

**latent parasitism** Parasitism without obvious symptoms.

**mutualism** An association of organisms in which both partners benefit from the association.

**obligatory parasite** A parasite that cannot survive without a host.

**parasitism** A close association of two organisms in which one, the parasite, depends on the other, the host, deriving some benefit from it without necessarily damaging it.

**periodic parasite** A parasite visiting a host at intervals.

**permanent parasite** A parasite associated with a host for long periods.

**phoresis** An association in which one organism uses another as a means of transport and/or protection.

**symbiosis (sensu lato)** Any association between organisms (parasitism, commensalism, mutualism, phoresis).

**symbiosis (sensu strictu)** An association of organisms in which both partners benefit from the association and cannot live without each other.

**temporary parasite** A parasite found in or on a host only for short periods.

**transport host** A host that harbors sexually immature stages of a parasite that do not develop.

---



**PARASITISM IS DEFINED IN DIFFERENT WAYS** by different authors, usually reflecting their research interest and bias. Parasitism, as used here, is defined as a close association between two organisms in which one, the parasite, depends on the other, the host, deriving some benefit (usually food) from it without necessarily damaging it. Traditionally, fungi, bacteria, and viruses, many of which are parasitic, are studied by microbiologists, whereas parasitologists study protozoan and metazoan parasites. In this contribution, only protozoan and metazoan parasites as well as higher plants (angiosperms) are included. Several types of associations resemble parasitism in various ways and cannot always be clearly distinguished from it, either because of insufficient knowledge or because genuine intermediate forms exist. Such associations are discussed in the following.

## I. RELATED PHENOMENA

### A. Commensalism

A commensal is an organism that uses food supplied in the internal or external environment of the host, without establishing a close association with the host, for instance by feeding on its tissues. Examples are the amoeba *Entamoeba coli*, an endocommensal of humans feeding on bacteria in the lumen of the intestine, and the ciliate protozoan *Ephelota gemmipara*, an ectocommensal on various marine invertebrates.

### B. Phoresis (Phoresy)

In a phoretic association, one organism uses another as a means of transport and/or protection. An example is barnacles living on whales.

### C. Mutualism

A mutualistic association is one in which both organisms derive a benefit, but the association is not compulsory. The cleaner fish *Labroides dimidiatus* feeds on parasites and diseased tissues of various marine fishes, and both partners derive a benefit, the cleaner obtaining food and the host fish getting rid of their parasites and diseased tissues.

### D. Symbiosis

Symbionts live in a compulsory association in which both partners derive a benefit. An example is the symbi-

osis of fungi and algae in lichens. It should be noted, however, that the term symbiosis is sometimes used in a wider sense, including all types of associations between organisms (parasitism, commensalism, phoresis, mutualism).

### E. Predation

A predator is an organism that attacks another, the prey, and usually kills and eats it. Most predators are larger than their prey.

An organism may be a parasite under certain conditions, but a commensal, mutualist, or predator when conditions change. For example, *Entamoeba histolytica* is often a harmless commensal feeding on bacteria in the intestine of man, but may become a dangerous parasite feeding on red blood cells, apparently induced by some changes in the host that are not fully understood. A number of normally pathogenic parasites even improve the health and fitness of their hosts at low infection intensities. Thus, Lincicome (1971, cit. Rohde, 1993) made some controlled experiments using rats and mice infected with two species of trypanosomes and the nematode *Trichinella spiralis* which showed that parasitized animals grew faster, ate more, could compensate better for deficiencies in the diet, had livers richer in certain vitamins, and were more active and more responsive to the human presence. For these reasons, it is best to consider parasitism as a type of association that is not clearly delimited from the others.

## II. TYPES OF PARASITES

There are many types of parasites, distinguished by the site of infection, kinds of hosts, state of development, etc. *Ectoparasites* are parasites that live on the external surface of hosts, for example fleas and lice of various terrestrial vertebrates, and Monogenea and Copepoda of freshwater and marine fishes. *Endoparasites* are parasites that live in the tissues and organs of their hosts, such as tapeworms, flukes, and protozoans of vertebrates. An *obligatory parasite* is a parasite that cannot survive without a host, such as the malaria parasite, and a *facultative parasite* is a parasite that can also live without a host, such as maggots, which normally are saprophagous but can infect living hosts as well. A *permanent parasite* is associated with a host for long periods, whereas a *temporary parasite* is found in or on a host only for short periods: examples of the former are human helminths and blood protozoans, and examples of the latter are mosquitoes and leeches that visit

hosts for blood-sucking only for short periods. *Larval parasites*, like the praniza larva of isopods, are parasitic only at a larval stage, and *adult parasites*, to which most metazoan parasites belong, are associated with a host during part or the whole of their mature phase. *Periodic parasites* (leeches, mosquitoes) visit a host at intervals. *Intraspecific parasites* parasitize individuals of the same species; for example, males of certain deep-sea fish are permanently attached to females of the species and derive food from them. *Hyperparasites* (of the first, second, etc. degrees) are parasites living on or in other parasites. An example of a hyperparasite of the first degree is *Udonella*, a monogenean parasitic on copepods, which themselves parasitize marine fishes. *Microparasites*, i.e., protozoans, bacteria, viruses, and some helminths, are small, with short generation times; they reproduce on or in a host usually at high rates, the duration of infection is usually much shorter than the life span of the host, and they induce immune responses in vertebrates. *Macroparasites*, i.e., most helminths and arthropods, do not reproduce on or in the host, they have longer generation times than microparasites, immune responses are lacking or weak and depend on infection intensities, and infections are often chronic and lead to morbidity rather than mortality. *Parasitoids*, many species of Hymenoptera, lay their eggs into insect hosts which may survive for some time but are invariably killed by the growing larvae of the parasitoids. Hence, parasitoids are predators rather than genuine parasites.

### III. ADAPTATIONS TO PARASITISM

#### A. Size of Parasites

Most parasite species are much smaller than their hosts. Malaria parasites, for example, are microscopic protozoans, and human pinworms are less than 1 cm long. Nevertheless, some species reach a remarkable size. For example, a didymozoid trematode infecting the sunfish, *Mola mola*, reaches a length of 12 m, although its diameter is very small and the volume of the parasite is still much smaller than that of the fish, which reaches a weight of 1 ton. Likewise, the broad fish tapeworm, *Diphyllobothrium latum*, which lives in the intestine of various fish-eating mammals, including man, reaches a length of over 10 m, but its volume is much smaller than that of its hosts. At first glance, perhaps surprisingly, parasites often are considerably larger than their free-living relatives. This phenomenon is clearly shown in flatworms, phylum Platyhelminthes. Most free-living flatworms are very small, from less than 1 mm to a few

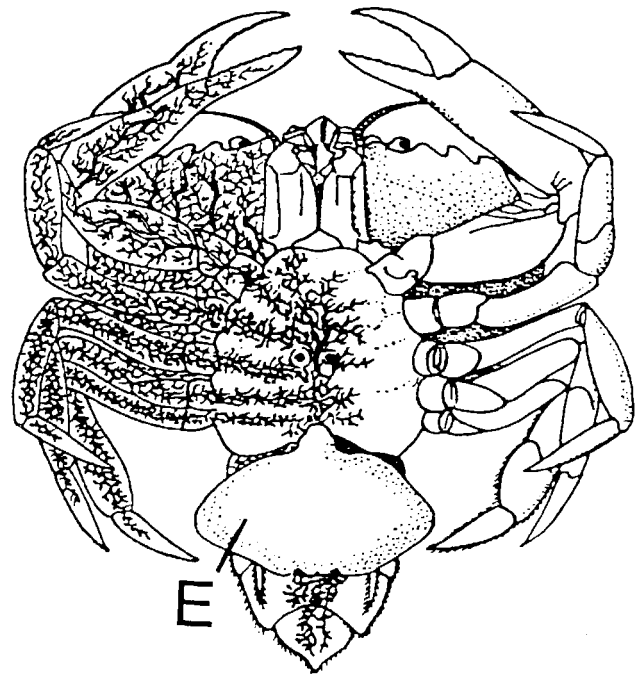


FIGURE 1 The rhizocephalan *Sacculina carcini* parasitic on and in the crab *Carcinus maenas*. E, externa of the parasite containing the gonads. Processes of parasite in crab drawn only on one side of the body. After Boas, from Rohde (1993).

millimeters long, whereas parasitic flatworms such as flukes and tapeworms, as a rule, are much larger, up to several centimeters in length for the flukes and many meters long for the tapeworms. There may be two reasons for this. First, parasitic flatworms in the organs and tissues of their hosts have a much richer and consistent food supply than do free-living species, and food supply therefore does not present a limit to size. Second, selection may have favored multiple and larger gonads and, therefore, a large body size because parasites have to produce many offspring in order to overcome the hazards involved in infecting other hosts. But selection may also favor smaller body size, because parasites depend on living hosts and it is important that hosts survive at least until the parasite has produced offspring that can infect other hosts. A smaller size of parasites relative to that of their hosts is therefore of advantage.

#### B. Reduction and Increase in Complexity

A general misconception is that all parasites have a less complex structure than free-living forms, a phenomenon that has been named *sacculinization* after the parasitic barnacle *Sacculina*, which infects marine crabs and

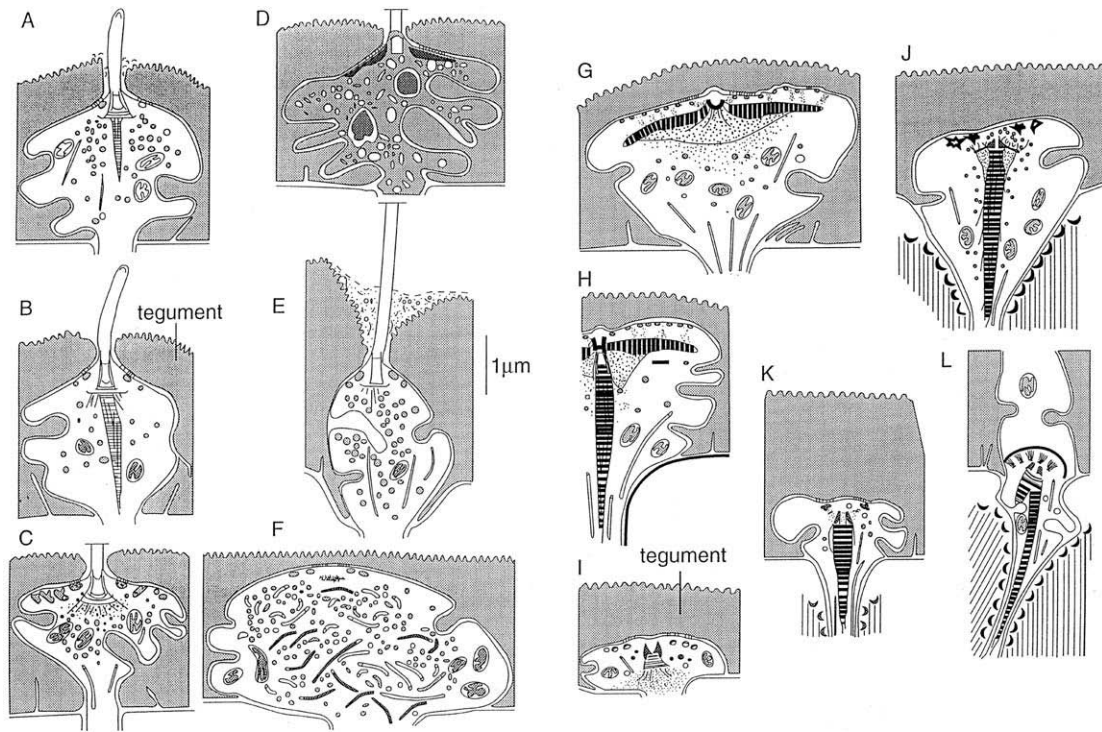


FIGURE 2 Sensory receptors of adult *Lobatostoma manteri* (Trematoda, Aspidogastrea). Redrawn from Rohde (1993).

which indeed shows a remarkable reduction in complexity: The only parts of the parasite visible on the outside of the host are the so-called externa, a saclike structure containing the gonads. Most of the parasite consists of an extensive system of cytoplasmic processes reaching into the various host tissues (Fig. 1). All the crustacean characteristics have been lost, and only the free-living larval stage indicates that the parasite is indeed a crustacean, related to barnacles. However, in most parasites such a reduction in complexity has not occurred, and in many species there is, in fact, an increase in complexity. A well-studied example of the latter case is the trematode *Lobatostoma manteri*, a parasite of marine fishes, about 3–5 mm long, which has not only a remarkably complex nervous system with more longitudinal nerve cords than free-living flatworms but also about 20,000–40,000 sensory receptors, belonging to about a dozen different types distinguishable under the electron microscope (Fig. 2). These numbers are much larger than those of most related free-living flatworms, and this in spite of the fact that there is not a single free-living stage in the life cycle of this parasite. The adult worm lives in the small intestine of marine fish, it produces eggs containing an infective larva, the egg is eaten by a snail in which the larva

hatches and develops to a stage infective to fish, and fish become infected by eating snails.

### C. Increase in Reproductive Capacity

Almost all parasites that have been studied produce a remarkable number of offspring, greater than that of related free-living forms. An example is given in Table I. Free-living turbellarians produce the fewest offspring of any flatworms, ectoparasitic Monogenea produce more, and the highest number of offspring is produced

TABLE I  
Approximate Estimates of Fecundity of Free-Living and Parasitic Flatworms<sup>a</sup>

	Number of eggs	Multiplication of larvae
Free-living Turbellaria	10	None
Ectoparasitic Monogenea	1000	None
Endoparasitic trematodes	10 million	×1000 at least
Endoparasitic tapeworms	10 million	×1–1000

<sup>a</sup> After Jennings, Calow, and Rohde, from Rohde (1993).

by endoparasitic trematodes and cestodes. Most trematodes and some tapeworms not only produce large numbers of eggs, but there is a secondary increase in numbers of offspring in the intermediate hosts by asexual or parthenogenetic reproduction. Thus, a single egg of a trematode can produce thousands of cercariae, i.e., larval stages infective to the final, vertebrate host. A prerequisite for this increased reproductive capacity is the safe and rich food supply available to parasites, as well as the relatively large body size of many parasites compared with their free-living relatives (see above). However, it is likely that selection has favored an increased reproductive output, because the hazards encountered by parasites in their often complex life cycles are enormous. Very few larvae will survive and establish an infection.

#### D. Mechanisms of Dispersal

Dispersal is important for any species, whether free-living or parasitic, because a population restricted to one small area risks becoming extinct if conditions become unfavorable and because dispersal reduces inbreeding and the loss of evolutionary adaptability. For parasites, a third point is important: dispersal may reduce the chances of hosts becoming overinfected. Three aspects of dispersal are important: dispersal over short distances away from an individual host, dispersal in space and range extension over larger distances, and dispersal in time. Trematode larvae illustrate that all three aspects of dispersal can be brought about by the same stage. Larvae (cercariae) are often forcibly ejected in the respiratory currents of the snails in which they have developed, bringing about dispersal away from the host. They actively swim and keep afloat by means of their tails and can thus be dispersed over long distances by water currents. In many species special flotation devices of the tail prolong duration of floating (Fig. 3). Adult flukes produce eggs, and larvae in the snail hosts are produced over long periods, months or even many years, leading to dispersal in time.

#### E. Mechanisms of Infection

It is essential for a parasite to ensure entry into a host, and this is achieved by an amazing variety of mechanisms. Table II lists examples of infection mechanisms used by human parasites. Many parasite species have evolved remarkable behavioral adaptations that facilitate transmission to a host. For example, microfilariae, i.e., larvae of various species of filariae (nematodes), circulate in the peripheral blood of vertebrates, where

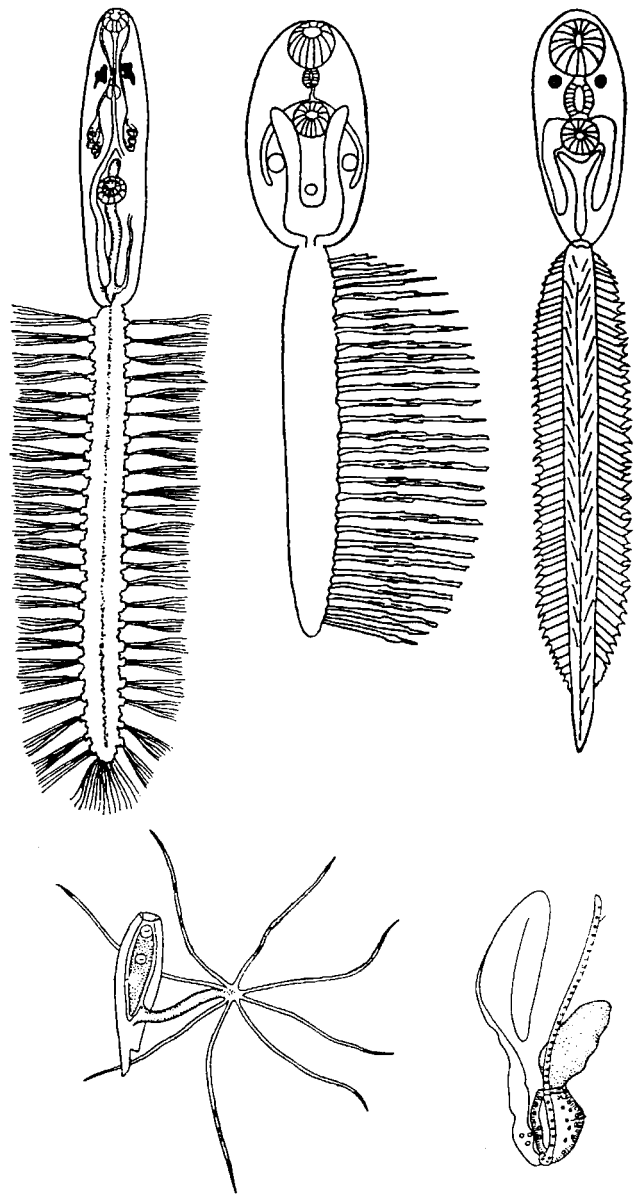


FIGURE 3 Marine cercariae with various flotation mechanisms on the tail. After various authors, from Rohde (1993).

they are ingested by mosquitoes for further development. Depending on the activities of the mosquito species involved, the microfilariae appear in the peripheral blood either during the day or the night, and strains of the same species may be nocturnal in one area and diurnal in another. Larvae of some monogeneans have endogenous hatching rhythms (rhythms not induced by external factors) adapted to the hosts' behavior. Some species (and strains) hatch in the evening, others in the morning, whenever chances to infect a host fish are greatest. Perhaps most remarkable are the adaptations

TABLE II  
Infection Mechanisms of Some Human Parasites

Means of infection	Parasite
Inoculation by arthropod hosts	Malaria parasite ( <i>Plasmodium</i> ); sleeping sickness
Fecal contamination of wounds	Chagas disease ( <i>Trypanosoma cruzi</i> )
Retrofection (through the anus into the large intestine)	Pinworm ( <i>Enterobius vermicularis</i> )
Ingestion of cysts	Amoebic dysentery ( <i>Entamoeba histolytica</i> )
Ingestion of eggs	Roundworm ( <i>Ascaris lumbricoides</i> )
Ingestion of spores	Microsporans
Ingestion of transport hosts	<i>Anisakis</i>
Ingestion of intermediate hosts	Broad fish tapeworm ( <i>Diphyllobothrium latum</i> )
Inhalation	<i>Pneumocystis carinii</i> (likely, but no experimental evidence)
Contact transfer	Mange mite ( <i>Sarcoptes scabiei</i> ), lice
Kissing or joint use of eating or drinking utensils	<i>Trichomonas tenax</i>
Sexual intercourse	<i>Trichomonas vaginalis</i>
Penetration through the skin	Bilharzia ( <i>Schistosoma</i> spp.); hookworm ( <i>Ancylostoma</i> , <i>Necator</i> )
Penetration into the nasal passages	Primary amoebic meningoencephalitis ( <i>Naegleria fowleri</i> )
Intrauterine infection	<i>Toxoplasma</i> , malaria

of some trematodes. For example, the liver fluke of sheep, *Dicrocoelium lanceolatum*, uses terrestrial snails as the first and ants as the second intermediate hosts. Cercariae produced in the snails aggregate in slime balls inside the snails, are expelled by the snails, and are eaten by ants. The first cercariae that enter an ant migrate into the subesophageal ganglion, inducing spastic behavior of the ant, which makes it cling to a plant, where their chances to be eaten by sheep are enhanced. The trematode *Leucochloridium macrostomum* has a larval stage, the sporocyst, which forms colorful outgrowths that extend into the snail's tentacles. They pulsate rhythmically, mimicking worms, the natural food of small birds, which are the final hosts. Birds bite off the tentacles and become infected. Infection also induces snails to move to more exposed sites.

### F. Aggregation

The distribution of organisms in space may be even (more or less equal distances between individuals), random (random distances between individuals), or aggregated (or clustered, or overdispersed; individuals occur in clusters). Applied to populations of parasites, as a rule, parasites show an aggregated distribution in host populations: some individuals are heavily infected, but most are very lightly infected or not at all. Very often, such distributions can be described best by a negative binomial distribution. Reasons for aggregated distributions are manifold: a series of exposures, each with different chances of infection; nonrandom distribution

of infective stages; increase or decrease in chances for further infections by a first infection; variations in susceptibility of host individuals; changes of infection of individual hosts over time. Selection may even have favored aggregation in order to restrict damage to a few heavily infected individuals or to facilitate mating, but experimental evidence does not exist.

### G. Hermaphroditism, Parthenogenesis, and Asexual Reproduction

Contact between parasites and hosts is usually sporadic and, in most cases, only single or a few parasites will manage to infect a host. It is important that populations can be built up from these individuals. A single individual can produce large populations by parthenogenetic or asexual reproduction. In the first case, gametes develop without fertilization; in the second, somatic cells develop. An example of the second case is malaria parasites developing by schizogony in human red blood cells. An example for the first case is (probably) trematode larvae developing in snails. Almost all trematodes, as well as the cestodes and monogeneans, are hermaphroditic, thereby doubling the chance of meeting a mating partner. Furthermore, some species can at least sometimes self-fertilize.

### H. Host Specificity

All parasites are restricted to certain host species, i.e., they cannot infect all species available, although some

parasite species are more restricted than others. For example, the human pinworm, *Enterobius vermicularis*, is found only in humans, whereas the protozoan *Toxoplasma gondii* has been found in a wide range of mammals and birds.

Some parasites, although found in many hosts, nevertheless infect a single or a few host species much more strongly than others. It is therefore useful to distinguish host range (the number of host species infected irrespective of how frequently and how strongly they are infected) and host specificity (taking frequency, or prevalence, and intensity of infection into account). A parasite that infects many host species, but one of them much more strongly than the others, may have a greater host specificity than a parasite species that infects fewer host species, but all of them strongly. Indices to measure host specificity are available (Rohde, 1993).

### I. Site Specificity

There is no “universal parasite” that infects all organs and tissues of a host equally, but the degree of site specificity within or on a host varies greatly. *Entamoeba histolytica* infects not only the intestine of humans but also the liver, lungs, brain, and other organs, whereas adult schistosomes are restricted to the blood vessels. Site specificity is extreme in many Monogenea on the gills of fishes. Different species occupy different parts of the gill on the same host (Fig. 4).

### J. Simple and Complex Life Cycles

Many parasites use a single host; i.e., they have a direct life cycle. Others use a final (definitive) host as well as one or several intermediate hosts; i.e., they have indirect life cycles. Parasites with direct life cycles include fleas and lice on various vertebrates as well as many intestinal nematodes and Monogenea on the skin and gills of fish. Usually, the adult stage is parasitic. It produces eggs or larvae that infect the same or other host individuals. In the case of some isopods, the only parasitic stage is the so-called praniza larva: adults live on the sea floor, larvae attach themselves to the gills of fish, suck blood, and drop off to mature in the benthic environment. Parasites with indirect life cycles include the trematodes, some with a single intermediate host, others with several. An example of the former is the aspidogastreaean trematode *Lobatostoma manteri*: the adult infects the intestine of marine fish; eggs are produced and shed in the feces; they are eaten by marine snails, in which the larvae hatch and grow to (almost) adult body size; and

fish become infected by eating snails (Fig. 5). An example of a trematode with two intermediate hosts is the

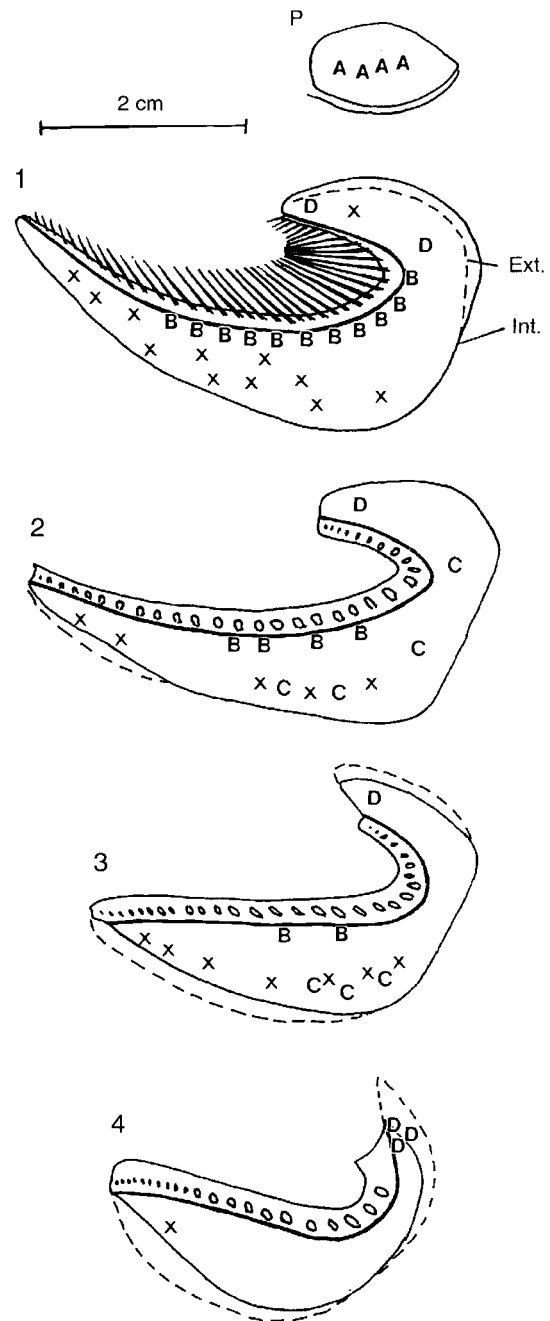


FIGURE 4 Monogenean gill parasites on the gills of mackerel, *Scomber australasicus*, off southeastern Australia. P, pseudobranch; 1-4, gills 1-4; ext, external gill filaments; int, internal gill filaments. A, *Kuhnia sprostoni*; B, *Kuhnia scombri*; C, *Kuhnia scombercolias*; D, *Grubea australis*; x, *Pseudokuhnia minor*. Note that species A-D have identical copulatory organs, and species x has different copulatory organs; A-D are spatially segregated from each other, and species x overlaps with B, C, and D.

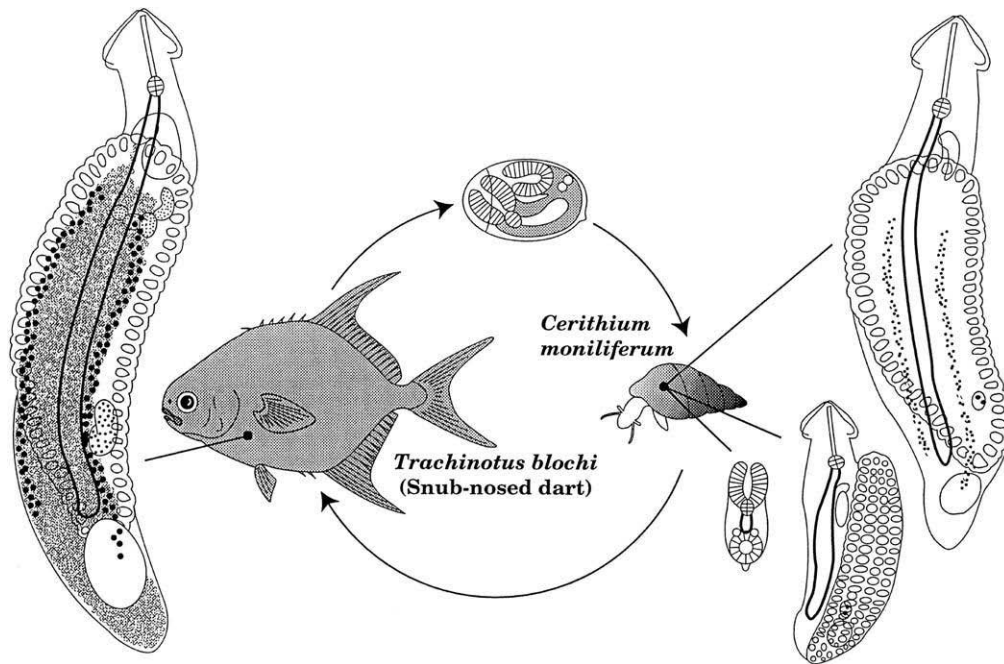


FIGURE 5 Life cycle of *Lobatostoma manteri* (Trematoda, Aspidogastrea). Note one intermediate (snail) and a final host (fish). No multiplication of larvae occurs in the intermediate host.

Chinese liver fluke, *Opisthorchis (Clonorchis) sinensis*. Adults infect the liver of fish-eating mammals, including man; snails are the first and fish the second intermediate hosts. An example of a trematode with three intermediate hosts is the bird fluke *Strigea falconispalumbi*. Adults live in predatory birds, the first intermediate hosts are snails, the second intermediate hosts are tadpoles/frogs, and the third intermediate hosts are amphibians, snakes, mammals, and birds, which are eaten by the final hosts (Fig. 6). The life cycles of some trematodes with four hosts can be extended by incorporating various transport hosts in which larvae accumulate but do not develop further. Nematodes with indirect life cycles include filariae which are transmitted by blood-sucking dipterans. Some nematodes with direct life cycles include species whose larvae undergo a peculiar migration through the body of the host. The human roundworm *Ascaris lumbricoides*, for instance, produces eggs that are swallowed by the host. They hatch in the digestive tract but do not grow there to the adult stage; instead they penetrate the wall of the intestine, invade the blood vessels, are carried into the lungs, penetrate through the alveoli, migrate up the trachea, are swallowed a second time, and now develop to the adult stage in the intestine. Reasons for this curious phenomenon are not known, but it may reflect a different kind of life cycle in the evolutionary past. Hookworms, which infect their hosts by penetrating through the skin, undergo a similar

body migration involving passage through the lung alveoli and final maturation in the intestine.

### K. Some Physiological Adaptations

Parasites belong to a wide range of taxa and infect a wide range of hosts. Hence, their physiological adaptations are of a great variety as well and cannot be discussed in any detail here. Jennings (in Rohde, 1997) has made detailed comparative studies of free-living, ecto- and endoparasitic flatworms. These studies are particularly informative because they show the transition from free-living to commensal to ecto- and endoparasitic forms. Some free-living turbellarians living on the body surface and in the gill chambers of their hosts feed on the same kind of food as their free-living relatives, but in addition, they also feed opportunistically on food scraps of the host. Food reserves and digestive physiology do not differ from those of free-living species. Turbellarians living in the interior of mollusks, arthropods, and echinoderms have become increasingly dependent on their hosts. Some feed on protozoans also found in the hosts as well as on food ingested by the host, intestinal cells, and coelomocytes. Others feed mainly on intestinal cells, using some digestive enzymes from the host, and others entirely lack digestive enzymes and depend on enzymes ingested with host tissues. Most of the endoparasitic turbellarians do not

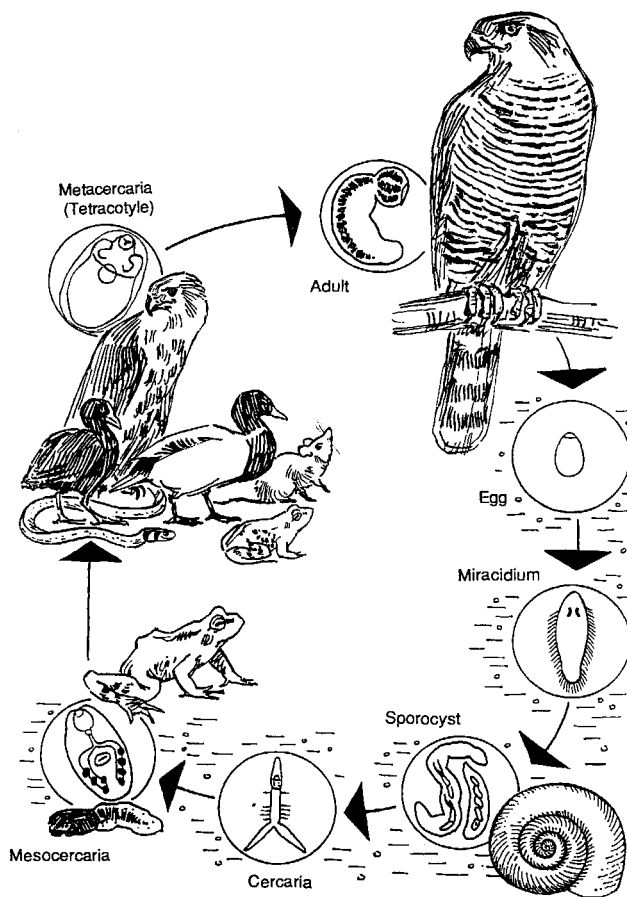


FIGURE 6 Life cycle of the trematode *Strigea falconispalumbi*. Note: Final hosts (predatory birds) containing adult worms which produce eggs in which miracidia develop, first intermediate host (snails) containing the sporocysts that produce cercariae, second intermediate host (tadpoles/frogs) containing mesocercariae, and third intermediate hosts (amphibians, snakes, birds, mammals) containing metacercariae. Modified from Odening (1969 © Spektrum Akademischen Verlag, Heidelberg, Berlin, with permission).

store lipids (as free-living and ectocommusal forms do), but they store glycogen instead. Glycogen storage is also characteristic of the Neodermata (the major groups of parasitic flatworms, including the Monogenea, Trematoda, and Cestoda). Some endoparasitic turbellarians possess physiologically active hemoglobins, which permit preferential abstraction of oxygen from host tissues.

### L. Adaptations of Flowering Plants

Parasitic plants have haustoria, which form a close connection with the vascular system of host plants, either in the roots or in the shoots. They depend entirely or partly on the host for water and inorganic and organic

solutes. "Hemiparasites," i.e., plants only partly dependent on the host, have chlorophyll, whereas "holoparasites," i.e., plants entirely dependent on the host, lack chlorophyll.

## IV. ORIGINS OF PARASITISM AND COMPLEX LIFE CYCLES, THE EVOLUTION OF VIRULENCE, AND COEVOLUTION OF HOSTS AND PARASITES

### A. Origins of Parasitism and Complex Life Cycles

Few fossil parasites are known. They include schistosome eggs from ancient Egyptian mummies a few thousand years ago and galls on the arms of feather stars, probably produced by Myzostomida (parasitic annelids) from the Silurian and Devonian periods, 350–430 million years ago. Conclusions on the origins of parasitism and parasite life cycles must therefore be based on inferences from comparative studies of extant species. The Platyhelminthes have been studied most thoroughly, using DNA studies of several genes, in particular 18-S rDNA and 28-S rDNA, and phylogenetic systematics (cladistics). Phylogenetic systematics seeks to establish branching patterns in phylogeny on the basis of shared acquired characters (synapomorphies). Ultrastructural characters are of particular use because of their complexity. Cladistic and DNA studies by many authors have consistently shown that the Neodermata, the major groups of parasitic flatworms (Trematoda, Monogenea, and Cestoda) all share a common ancestor, i.e., are monophyletic. The Neodermata, or the Neodermata plus some of the parasitic turbellarians, are the sister group of a very large taxon, including most Turbellaria, with which they share a common ancestor. This means that the parasitic groups evolved very early in evolutionary history. Among the Neodermata, the trematodes are the sister group of the other Neodermata and, among the trematodes, the Aspidogastrea are the sister group of the other trematodes, or Digenea. Three of the four families of Aspidogastrea occur in elasmobranchs, whereas almost all digeneans parasitize teleost fishes, amphibians, reptiles, birds, and mammals; very few species of Digenea have been recorded from elasmobranchs, to which they have secondarily adapted. Fossil records indicate that elasmobranchs are 450 million years old and teleosts are 210 million years old. This suggests that the Aspidogastrea are the oldest extant



trematodes and possibly neodermatans. It also suggests that the simple life cycle of *Aspidogastrea* (Fig. 5) is the original life cycle of trematodes, including a final and an intermediate host (which in some species is not obligatory), without multiplication of larval stages in the intermediate host. The complex life cycles of digenean trematodes, including a final and at least one, and up to three, intermediate hosts and several transport hosts, with multiplication of larval stages in the first intermediate host (Fig. 6), may have evolved from this primitive kind of life cycle to make transmission to the final host more effective.

Among the crustaceans, most barnacles are free-living, attached to rocks or other hard substrata. Some barnacles live in a phoretic association, for instance with whales, attached to their skin and feeding on plankton in the environment. Other, closely related species have become parasites. Thus, *Anelasma* parasitizes the skin of sharks, processes of its stalk branching in the host's muscles and extracting food from it. All the approximately 120 species of rhizocephalans, also related to the barnacles, are parasites and strongly modified in adaptation to their way of life (see, for example, *Sacculina*, Fig. 1). This suggests that parasitism in the barnacles may have evolved from free-living to phoretic to parasitic.

## B. Evolution of Virulence

Intuitively, one might suspect that it is not in the parasite's interest to severely damage or even kill its host, because this would also affect the fitness of the parasite. On the other hand, parasite transmission to another host may well be facilitated by such damage to an intermediate or transport host. In other words, evolutionary pressure may have led to an increase in virulence in some cases and to a decrease in others. Anderson and May have developed an epidemiological model that considers virulence, as follows:

$$R_0 = \beta(N)/(\mu + a + \nu).$$

$R_0$  is the fitness of the parasite, its lifetime reproductive success;  $\beta$  is the rate of transmission of the parasite by its host to other hosts, which is dependent on host density  $N$ ;  $\mu$  is the mortality rate of uninfected hosts;  $a$  is the virulence or the mortality rate induced by the parasite; and  $\nu$  is the recovery rate of the host. This model (applied and further developed by various authors) permits some predictions on the evolution of virulence under different conditions. For example, a parasite that can achieve a large increase in transmission

rate by a small increase in virulence should have lower optimal virulence than one that achieves less by such an increase. The transmission rate depends, among other factors, on the likelihood of transmission for instance facilitated by the behavior of the host. Low virulence will evolve when contacts between hosts and thus opportunities for transmission are frequent. High virulence may evolve in vector-transmitted infections, and optimal virulence may differ in parasites that are vertically transmitted (from parents to offspring) from those that are horizontally transmitted. Although the parameters in the model are difficult to measure, some of the predictions have been verified empirically. Thus, parasites transmitted by contact (e.g., lice) are usually less virulent than those transmitted by vectors (e.g., malaria), and sexually transmitted parasites (e.g., *Trichomonas vaginalis*) are usually less virulent and longer lasting than parasites transmitted in other ways. In vertically transmitted nematodes of fig wasps and in vertically transmitted microsporans of mosquitoes, virulence is less than in horizontally transmitted related species.

## C. Coevolution of Hosts and Parasites

The comparison of phylogenetic trees of hosts and parasites has given contradictory results concerning possible coevolution. Reasons may be biological differences between the groups studied, or they may be due to different interpretation of results by different authors. For example, according to several authors, chewing lice infecting rodents of the family Geomyidae, as well as lice of seabirds, seem to have coevolved closely with their hosts. In contrast, lice of rock wallabies in Australia, apparently, have switched hosts frequently and there is little evidence for coevolution. Altogether, knowledge is insufficient to state how common coevolution is.

## V. HOST-PARASITE INTERACTIONS

### A. Cleaning Symbiosis

Hosts use a variety of behavioral methods to rid themselves of parasites. Preening of birds, bathing of birds in dust and water, and passive and active anting (letting ants passively crawl over the body or actively squeezing ants over the plumage, respectively) are thought to help in reducing parasite loads, although evidence is scarce. Some other behavioral patterns (dolphins rubbing against rocks, fish jumping out of the water, etc.) may also play a role. Most widely distributed and well studied,

particularly in the marine environment, is cleaning symbiosis, in which one animal (the cleaner) cleans another (the host), removing its parasites and diseased tissues. Birds remove ticks and other ectoparasites from cattle, hippopotamus, and even large marine fish floating on the surface, and several species of shrimps, as well as over 100 species of fish, are cleaners in freshwater but mainly in the sea. Beside fishes, the Galapagos marine iguana, whales and dolphins, invertebrates, etc. are hosts to cleaners. Cleaner fish often have special morphological adaptations, such as a terminal mouth and fused anterior teeth and conspicuous color patterns, so-called guild signs. In addition, the Indo-Pacific cleaner wrasse, *Labroides dimidiatus*, performs a cleaning dance that attracts host fish. Hosts, on their part, show invitation postures signaling to the cleaner that they are ready to be cleaned, and fish of many species, usually hostile to each other, queue peacefully up at "cleaning stations" (territories where cleaning occurs). They even allow cleaners to enter their mouth cavity. In well-established cleaning symbioses, hosts rarely or never eat cleaners. Some fish were observed to spend as much time at cleaning stations as they spend on feeding, and some cleaner species feed exclusively on parasites and diseased host tissue. Certain fish mimic the color pattern and behavior of cleaners to approach hosts in order to attack them and bite off pieces of fin or skin.

## B. Immune and Tissue Reactions, Resistance

Defense reactions to parasites at the humoral and tissue levels are based on the ability of the host to distinguish self (its own cells) from nonself (foreign cells and material). Vertebrates have three types of such reactions: phagocytosis, inflammation, and adaptive immunity. The first two are nonspecific tissue reactions, and the third is specific to a certain type of nonself material. They are found in all vertebrates, but (particularly immunity reactions) are best developed in birds and mammals. All three defense mechanisms usually interact and occur in most tissues and organs. Typically, the reactions occur in a certain sequence: degeneration or necrosis of cells due to the infection leads to an inflammatory response with edema (swelling of tissue). Phagocytic cells engulf small parasites, but if parasites are not eliminated, a chronic inflammation develops, leading to a connective tissue capsule around the parasite; macrophages in the capsule engulf damaged cells and often the parasite. Special defense mechanisms are active on the surface. Thus, fish continually shed mucoid material from the skin even if uninfected, but in

infected fish the slough increases and leads to the removal of monogenean and other ectoparasites.

Immune reactions are induced by antigens of the parasite that lead to the formation of specific antibodies in the host. In microparasites (protozoans and some helminths, e.g., *Strongyloides stercoralis*, but particularly in bacteria and viruses, which are not discussed here), immune responses are more effective than in macroparasites (most helminths and arthropods), where they are either lacking or only short-lasting. Immune responses have been particularly well studied in trypanosomes. The antigen is a surface glycoprotein that completely covers the parasite. The host's antibodies eliminate most parasites, but a few trypanosomes of different antigenic types survive and build up a new population. This is repeated over and over again, leading to marked fluctuations in infection intensities. The host finally dies, because the immune reactions do not suffice to destroy the whole parasite population. The number of antigenic types is very large and apparently limited only by the host's life span: more than 100 types were shown to develop in a single clone strain. In parasites with complex life cycles, such as the malaria parasite, each stage in the life cycle has different antigenic properties, and there are many variants because of the large numbers of daughter cells produced by schizogony. These are the reasons that attempts to develop effective vaccines have failed so far.

Invertebrates, apparently, cannot acquire specific immune responses; they rely on phagocytosis and capsule formation. Pearl formation in bivalves, for example, is the result of encapsulation of foreign bodies.

Hosts show different degrees of *resistance* to infections that are not due to acquired immunity. For example, some sheep may be less susceptible to nematodes than others because of their genetic makeup, and individuals of the same species of different age may differ in susceptibility. Older individuals are often less infected than young ones, a phenomenon referred to as *age resistance*. For example, the cestode *Austramphilina elongata* uses turtles as final and crayfish as intermediate hosts, but only young crayfish can be infected experimentally: the cuticle of older individuals prevents successful penetration of the larva.

## C. Effects on Host Individuals and Populations

There is a great variety of effects on host individuals, depending on the parasite and host species, site of infection, virulence of parasites, and susceptibility of hosts.

A few examples of human parasites may illustrate this. The mite *Demodex folliculorum*, intestinal nematodes at low infection intensities, *Toxoplasma*, etc. often do not cause symptoms. On the other hand, *Toxoplasma* and the nematode *Onchocerca volvulus* can lead to blindness, filariae can cause elephantiasis, a sometimes enormous swelling of scrotum, legs, and arms, etc. Malaria causes a range of symptoms from influenza-like to death. Hookworms, depending on infection intensity, can cause severe anemia and death. Cysticerci of *Taenia solium*, infecting the brain, may lead to epilepsy-like symptoms and death. Nematodes of sheep often cause death, and monogeneans may severely damage the skin, fins, and gills of fish, particularly in aquaculture. Larval trematodes often cause partial or complete castration of their snail hosts, and they sometimes induce gigantism in the snails, i.e., a markedly larger body size of infected than uninfected snails.

Models predict that microparasites should be highly effective in controlling host populations, and this is indeed often the case. For example, trypanosomes and malaria severely affect human populations and the former also livestock, and oyster beds were decimated by various protozoan parasites in several countries. However, almost all such reports deal with populations in abnormally high densities (humans, oysters) or with populations affected after introduction of a parasite into an area where it was not present originally or after introduction of a host species into a new area (livestock in Africa). The same applies to macroparasites. Host populations under natural conditions may sometimes be severely affected by macroparasites, but evidence is usually circumstantial and mass mortalities may not be due to parasites alone, but due to synergistic effects involving, for instance, environmental degradation, as well as parasites. Best documented are cases of mass mortalities caused by parasites in livestock and aquaculture, as well as after introduction of parasites or hosts into a new habitat. Of historical interest is an epizootic caused by the liver fluke *Fasciola hepatica*, which caused the death of 3 million sheep in Great Britain in 1879/1880 and led to the clarification of the life cycle of this parasite. A well-documented case of an epizootic due to the introduction of a parasite is that of the monogenean *Nitzschia sturionis*, which was introduced into the Aral Sea with sturgeon from the Caspian Sea in the early 1930s. It devastated the local sturgeon population and led to the collapse of the sturgeon and caviar industry in the Aral Sea in 1937, which did not recover for 20 years. Recently, the first experimental proof was given that a gastrointestinal roundworm

can indeed regulate wildlife populations of red grouse in England.

## VI. THE ECOLOGICAL NICHES OF PARASITES

### A. The Niche Concept: Niche Dimensions of Parasites

A niche is defined as the total of an organism's relations to its biotic and abiotic environment. These relations define the organism's place in nature or, in ecological jargon, its place in "multidimensional niche space." Important niche dimensions of parasites are hosts, microhabitats, geographical range, sex of host, age, season, and food. Different parasite species use different host species and different microhabitats within or on the host. Thus, particular nematodes of the cat use different microhabitats (Table III), and even within each microhabitat, there is further subdivision, some species using certain parts of the small intestine, the stomach, etc. Male lambs in the United States have more nematodes of certain species than females, and many parasites prefer hosts of a certain age, occur only at certain seasons, or are restricted to host populations that use certain food.

### B. Saturation of Niches with Parasites

An important question in ecology is whether habitats are saturated with species or whether "empty niches" exist. The most extensive and intensive studies were made of parasite communities on the heads and gills of marine fish. Results on 112 host species showed that maximum component species richness was 27 parasite species but that most species had less than six (Fig. 7). Similar differences were found for abundances (intensities of infection with all parasite species), which ranged from less than five for most host species to over 3000, with no apparent signs of damage in the most heavily infected fish. This strongly suggests that most fish species, at least, could accommodate more parasite species and individuals, and even in fish with the greatest parasite species richnesses and abundances, some parts of the gills that are occupied in other species were not infected.

### C. Proximate and Ultimate Causes of Niche Restriction

Proximate causes are physical and/or chemical factors determining niche selection of a parasite, whereas ultimate causes refer to the biological function of niche

TABLE III  
Nematodes of the Cat: Sites of Infection

Site	Species of nematode
Kidney	<i>Diocotophyma renale</i>
Colon	<i>Strongyloides mystax</i> , <i>S. tumefaciens</i> , <i>Trichuris campanula</i>
Stomach and small intestine	<i>Abbreviata gemina</i> , <i>Ancylostoma braziliense</i> , <i>Gnathosoma spinigerum</i> , <i>Mastophorus muris</i> , <i>Ollulanus tricuspis</i> , <i>Physaloptera brevispiculum</i> , <i>P. felidis</i> , <i>P. pacitae</i> , <i>P. canis</i> , <i>Rictularia cahirensis</i> , <i>Soboliphyme baturini</i> , <i>Spirura rytipleurites</i> , <i>Toxascaris leonina</i> , <i>Toxocara canis</i> , <i>T. mystax</i> , <i>Trichinella spiralis</i> (adults), <i>Uncinaria stenocephala</i> , <i>Ancylostoma tubaeforme</i> , <i>Physaloptera praeputiale</i>
Skin	<i>Dirofilaria repens</i>
Lungs	<i>Aelurostrongylus abstrusus</i> , <i>Anafilaroides rostratus</i> , <i>Capillaria aerophila</i> , <i>Troglostrongylus subcrenatus</i> , <i>Vogeloides massinoi</i> , <i>V. ramanujacharii</i>
Heart	<i>Dirofilaria immitis</i>
Middle ear	<i>Mammomogamus auris</i>
Spinal cord	<i>Guretia paralysans</i>
Lymph vessels	<i>Brugia malayi</i>
Diaphragm and other striated muscles	<i>Trichinella spiralis</i> (juveniles)

selection: why has selection favored one niche over another? There are many proximate causes responsible for niche selection of different parasites, although they have been poorly studied. Parasite larvae, for example, may use chemical stimuli to locate a particular host and, even if they can infect many hosts, they may survive only on the "correct" ones because a factor or factors produced by the "wrong" hosts may kill them. Exsheathment and development of parasite larvae of different species may depend on different stimuli, redox potential, and pH in different parts of the intestine, thus determining microhabitat differences. Males and females of a host species may simply acquire different parasites because of different feeding habits etc. Ultimate causes may be extrinsic, due to interspecific effects, or intrinsic, due to intraspecific effects. Among the interspecific effects, interspecific competition is generally considered to be the most important one, as evidenced by occasional exclusion of one species by another, microhabitat shifts in the presence of other species, etc. However, altogether evidence is scarce. Many cases of microhabitat differences that have been explained by interspecific competition are likely to be the result of reinforcement of reproductive barriers. Thus, only those monogeneans that have identical copulatory organs are spatially segregated in different microhabitats, whereas species with dissimilar copulatory organs coexist in the same microhabitat (Fig. 8, also Fig. 4), suggesting that microhabitat segregation does

not have the function to avoid competition but hybridization between closely related species. Only one intrinsic factor has been suggested: facilitation of mating. Many parasites live at low infection intensities and prevalences (Fig. 7). Restriction to certain hosts and microhabitats may therefore vastly increase the chances of meeting a mating partner. Experimental evidence, however, is scarce.

## VII. THE STRUCTURE OF PARASITE COMMUNITIES

### A. Concepts of Community Ecology

Community ecology deals with the relations between organisms in a certain habitat, in the case of parasites with relations between parasites infecting a certain host. Such relations can be studied at different levels, the level of infracommunity, component community, and compound community. *Infracommunities* of parasites consist of all the infrapopulations within a host individual. *Component communities* consist of all infrapopulations within a host population, and *compound communities* consist of all parasite communities within an ecosystem. An *infrapopulation* is the total of all the individuals of a parasite species within a host individual. Two of the major questions of community ecology are (1) whether communities show predictable patterns of

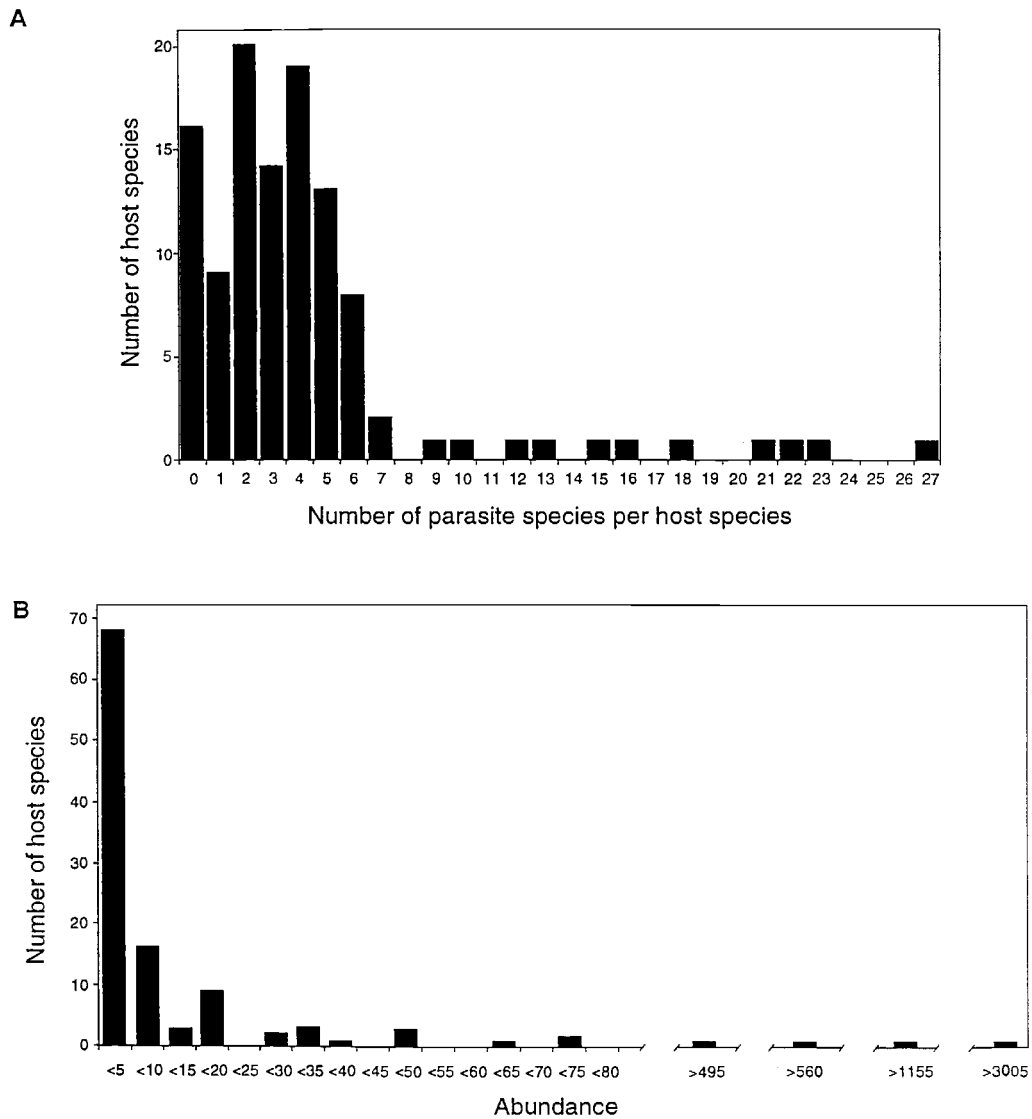


FIGURE 7 Number of ectoparasite species per host species and abundances of infection on 112 teleost species (5666 fish). From Rohde (1998).

species composition, relative abundances, and resource use and (2) what processes are responsible for the patterns. Studies of endo- and ectoparasite communities of various vertebrate hosts have led authors to distinguish *interactive communities*, characterized by species regularly occurring in great densities with much interaction (interspecific competition) between them, and *isolationist communities*, characterized by species occurring at low densities with little interaction between them. Both types, supposedly, are extremes at the ends of a spectrum, with many intermediate kinds of communities between them. Authors also distinguished between *core* and *satellite parasite species*, the former dominant species

with high prevalences and intensities of infection and the latter with low prevalences and intensities of infection.

## B. Empirical Evidence

Even in host species with extremely species-rich parasite communities, core and satellite species usually cannot be distinguished, and evidence indicates that many more parasite species than actually present could be accommodated (i.e., that vacant niches exist; see above). This, and the frequency of positive and the scarcity of negative associations between species, scarcity of nestedness, greater intra- than interspecific aggregation, and no or

only minor effects of the number of parasite species on microhabitat size, strongly suggest that interspecific competition is not of great importance (see above). Concerning the structure of communities as indicated by predictable co-occurrence and relative abundance of species, studies of ectoparasite communities of marine fish showed that the most dominant species usually represented between 60 and 80% of all parasites in an infracommunity but that different species were dominant in different infracommunities. For communities of larval trematodes in snails, Sousa (see Rohde, 1993) has shown that there is a significant difference between communities studied at different scales. Interactions are important at the infracommunity level, but at the level of component community, interactions are not important and communities are largely structured by external processes.

### C. Parasite Communities as General Ecological Models

The vast majority of animal species are probably arthropods parasitizing plants, and they share many character-

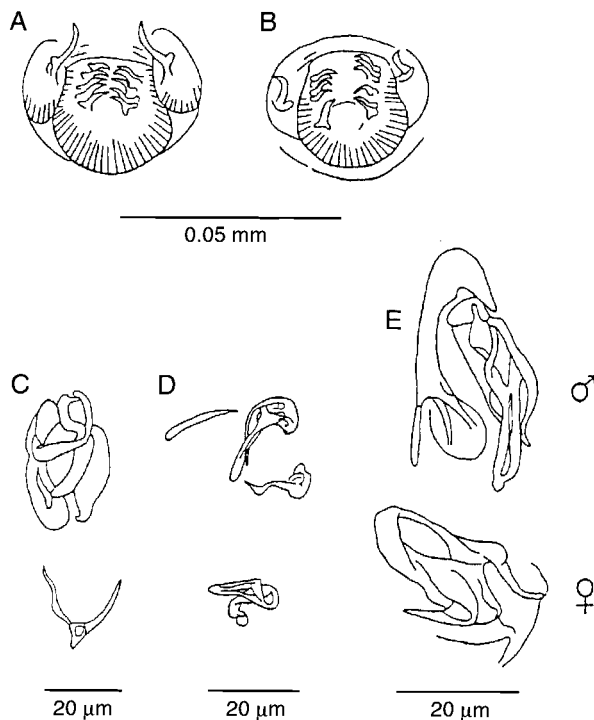


FIGURE 8 (A, B) Copulatory organs of two species of the monogenean *Kuhnia* inhabiting different parts of the gills of mackerel, *Scomber* spp. (C–E) Male and female copulatory organs of three species of the monogenean *Lamellogadus* inhabiting the same parts of the gills of bream, *Acanthopagrus australis*. After Rohde and Hobbs (1986), from Rohde (1993).

istics with parasites of animals: their habitats are resource-rich and seldom exhausted by the parasites. The major problem for such parasites is not to avoid competition with other species but to find the appropriate hosts and sites for feeding and mating. The conclusions based on the study of parasites of animals can therefore be applied to them.

## VIII. PARASITE POPULATION DYNAMICS

### A. Concepts of Population Growth

Population growth is described by the logistic growth equation  $dN/dt = rN[(K - N)/K]$ .  $dN/dt$  is the rate of population growth,  $N$  is the number of individuals at the time  $t$ ,  $r$  is the per capita rate of natural population increase, and  $K$  is the carrying capacity of the habitat (the maximum number of individuals a habitat can support). The equation shows that population growth is exponential when population density is small and that it decreases with increasing  $N$  and approaches 0 when  $N$  approaches  $K$ . In a graphic representation, the curve becomes asymptotic, i.e., it flattens out, when population density reaches the carrying capacity. In habitats in which, for example, populations repeatedly become impoverished as the result of external disturbances, species will be favored that can reproduce and develop rapidly, without great investment in the individual offspring. In stable habitats where population size is usually close to the carrying capacity, it pays off to have few but well-adapted offspring. Selection for large numbers of fast-developing offspring is referred to as  $r$ -selection, whereas selection for few well-adapted offspring is referred to as  $K$ -selection. However, the general applicability of this distinction has been questioned because population density is only one of the factors that determine selection pressure. Nevertheless, the distinction is still useful as an approximation to the real scenario.

### B. Ecological Strategies of Parasites

Most parasites rely on the production of many offspring (see above) and therefore tend to be  $r$ -strategists. In some cases, numbers are controlled by abiotic factors and not by population size; i.e., the factors are density-independent. This has been demonstrated for some helminth species, for example the cestode *Bothriocephalus acheilognathi* in the freshwater fish *Gambusia affinis*, whose prevalence and intensity of infection varied

strongly with water temperature over several years in a lake in North Carolina. However, few such studies have been made and generalizations are therefore premature. Density-dependent factors, i.e., factors dependent on the population size of the parasite, also play a role in some cases. For example, predation by trematode larvae in snails on other trematode larvae, immune responses of hosts that depend on infection intensities (as in many helminth infections), and competition between parasites of the same and of different species are such density-dependent factors. There is experimental evidence for all of these, but studies are too few to permit generalizations on how important such effects are. For example, high infection intensities often lead to stunted growth and reduction in the number of offspring, an example of intraspecific competitive effects. The microhabitat of a parasite species may shift or become smaller if other species are present, or one parasite species may reduce the numbers of another or even completely eliminate it, examples of interspecific competition. However, the importance of interspecific competition, in particular, generally seems to have been overestimated (see above).

## IX. THE DIVERSITY OF PARASITES (DISTRIBUTION OF PARASITES IN THE ANIMAL AND PLANT KINGDOMS)

The first estimate of the total number of parasite species in a fauna was made by Arndt in 1940. He counted 10,000 parasitic species among the total of 40,000 species in Germany but did not include insect "herbivores" intimately associated with plants and therefore, according to our definition, parasitic. Price in 1977, including such species but excluding temporary parasites such as mosquitoes and leeches, estimated that more than half of all British species are parasitic. Parasites, because of their small size, are less well known than free-living forms, and this applies especially to parasites of invertebrates, which have been little studied. Estimates of parasitic species may therefore well be too low. On the other hand, recent studies of the fauna of tropical rain forest canopies have revealed an enormous number of insect species and we do not know how many of these are parasitic. There are no studies of parasites of these insects. Hence, even approximate estimates of the proportion of parasitic species are premature, although we can state with confidence that parasitism is by far the most successful way of life on earth.

Data in Table IV show that 13 large taxa (phyla, subphyla, or classes), some of them very large, consist entirely of parasites, and many other groups include a high proportion of parasitic species. Even some vertebrates are parasitic. Males of a number of deep-sea fishes parasitize females of the same species (Fig. 9). The needle fish, *Carapus acus*, feeds on the viscera of sea cucumbers, which represent an almost inexhaustible source of food because they regenerate, cleaner mimics (fish that imitate genuine cleaners) approach fish and bite off pieces of fin or skin, and gulls were observed feeding on the flesh of whales. Cowbirds and about 50 species of cuckoos are brood parasites; i.e., they lay their eggs into the nests of other birds, which then incubate them. Skuas and frigate birds chase other birds in flight and feed on their regurgitated food ("kleptoparasitism"). Kleptoparasitism in its various varieties is widespread among humans.

Concerning plants, estimates are that 1% of flowering plants, about 3000–4000 species, are parasitic. Parasitism has evolved at least eight times in the angiosperms, and 16 plant families have parasitic species.

## X. ZOOGEOGRAPHY OF PARASITES

### A. Latitudinal Gradients in Species Richness and Abundance

The most dangerous parasites of man are found in warm countries: malaria (originally also found in temperate countries such as those of northern Europe), filariasis, guinea worm, sleeping sickness, Chagas disease, and several other diseases are restricted to tropical–subtropical countries. However, latitudinal gradients in species richness have been quantified best for parasites of marine fishes. Species number of marine fishes is greatest in the tropics. Species richness of metazoan ecto- and endoparasites also rises toward the tropics. However, interestingly, both the number of ectoparasitic species per host fish (their "relative species richness") and their abundance increase at lower latitudes, whereas such a trend does not exist for the endoparasites (Fig. 10). The reasons for these differences are not understood.

### B. Latitudinal Gradients in Reproductive Strategies

Such a gradient has been studied only in marine Monogenea. Most monogeneans produce eggs in which ciliated larvae (oncomiracidia) develop; one family, the

TABLE IV  
Estimated Number of Parasitic Species among Various Groups in the Animal Kingdom<sup>a</sup>

Phylum	Subphylum (Protozoa) or class	Total number of species	Number of parasite species
Subkingdom Protozoa		70,000	
Sarcomastigophora	Mastigophora	?	2,000
	Opalinata	150	150
	Sarcodina	?	250
Labyrinthomorpha		Some	Some
Apicomplexa			
Microspora (Microsporidia)		>7,000	>7,000
Ascetospora (Haplosporidia)			
Ciliophora		?	2,500
Subkingdom Metazoa			
Mesozoa		55	55
	Dicyemida	40	40
	Orthonectida	15	15
Porifera		5,000	1 (few?)
Cnidaria		8,900	20
	Hydrozoa	2,700	few
	Scyphozoa	200	0(?)
	Anthozoa	6,000	few
Ctenophora		80	1 (few?)
Myxozoa (Myxosporidia)		>1,500	>1,500
Platyhelminthes		>ca. 80,000	>ca. 65,000
	Turbellaria	>15,000	>200
	Trematoda	>25,000	>25,000
	Monogenea	>35,000	>35,000
	Cestoda	>2,500	>2,500
Gnathostomulida		80	0
Priapulida		3	0
Entoprocta		60	0
Nemertina		750	10
Nemathelminthes		>12,000	ca. 5,600
	Rotatoria	1,500	20
	Gastrotricha	150	0
	Nematoda	>50,000	>20,000
	Nematoinorpha	<100	<100
	Kinorhyncha	100	0
	Acanthocephala	ca. 500	ca. 500
Annelids		>7,000	ca. 420
	Polychaeta	>4,000	20
	Myzostomida	111	111
	Oligochaeta	2,400	40
	Hirudinea	300	250
Onychophora		70	0
Arthropoda		Many millions	Many millions
	Xiphosura	5	0
	Arachnida	30,000	4,300

*continues*



Continued

Phylum	Subphylum (Protozoa) or class	Total number of species	Number of parasite species
	Pycnogonida	350	<350
	Crustacea	20,000	>2,500
	Myriapoda	10,500	0
	Insecta	Many millions	Many millions <sup>b</sup> (>50%)
<b>Tardigrada</b>		180	1(?)
<b>Pentastomida</b>		75	75
<b>Tentaculata</b>		ca. 5,000	0
	Phoronidea	18	0
	Bryzoa (Ectoprocta)	>4,000	0
	Brachiopoda	280	0
<b>Mollusca</b>		112,000	>100
	Solenogastres	140	few(?)
	Placophora	1,000	0
	Gastropoda	85,000	100
	Scaphopoda	300	0
	Bivalvia	25,000	>10
	Cephalopoda	>600	0
<b>Echiurida</b>		70	Few (intraspecific parasites)
<b>Sipuncula</b>		250	0
<b>Hemichordata</b>		80	0
	Enteropneusta	60	0
	Pterobranchia	20	0
<b>Echinodermata</b>		ca. 6,000	≥2
	Crinoidea	620	0
	Holothuroidea	1,100	1(?)
	Echinoidea	860	0
	Asteroidea	1,500	0
	Ophiuroidea	1,900	2
<b>Pogonophora</b>		47	0
<b>Chaetognatha</b>		50	0
<b>Chordata</b>		62,000	Some
<b>Total</b>		Many millions	Millions

<sup>a</sup> Groups that consist entirely of parasitic species are printed bold. After various authors. Only extant species are included.

<sup>b</sup> Many authors refer to many of the plant-parasitic insects as herbivorous insects.

Gyrodactylidae, consists exclusively of small viviparous species. Whereas gyrodactylids are very rare in warm waters, they represent the vast majority of all species in cold waters (75–89% in the Bering, White, and Barents Seas). This corresponds to a trend in marine benthic invertebrates, which produce large numbers of eggs from which pelagic larvae develop in warm waters and small numbers of larvae by viviparity etc. in cold waters (Thorson's rule).

### C. Latitudinal Gradients in Host Ranges and Host Specificity

Such trends have not been studied and are not obvious in terrestrial and freshwater parasites. Studies of trematodes and monogeneans of marine fish have shown that the former have greater host ranges, i.e., they use more host species, in cold than in warm waters whereas monogeneans do not show such a trend, using very few

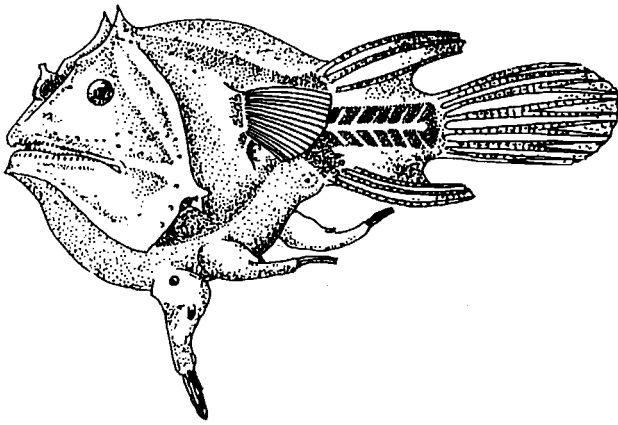


FIGURE 9 Intraspecific parasitism of vertebrates. Three males of the deep-sea fish *Edriolychnus schmidti* fused to a female. After Günther-Deckert, from Rohde (1993).

hosts at all latitudes. Host specificity (i.e., the number of host species infected taking prevalences and intensi-

ties of infection in different host species into account) is the same at all latitudes for both groups.

#### D. Parasites as Biological Markers

Parasites are frequently used as biological markers to study host populations and migrations, particularly in the sea. A prerequisite for such study is that parasites have a long life and are acquired only in the population or area studied. Larval helminths fulfill these conditions and are therefore usually used. Examples for the successful use of parasites as biological markers are studies that demonstrated different populations of herring with different *Anisakis* infections in the Baltic Sea, different populations of Atlantic salmon from different tributaries of the Miramichi River system in Canada, and different populations of sockeye salmon in the North Pacific, one harboring the nematode *Dacnitis truttiae* acquired in Kamchatka

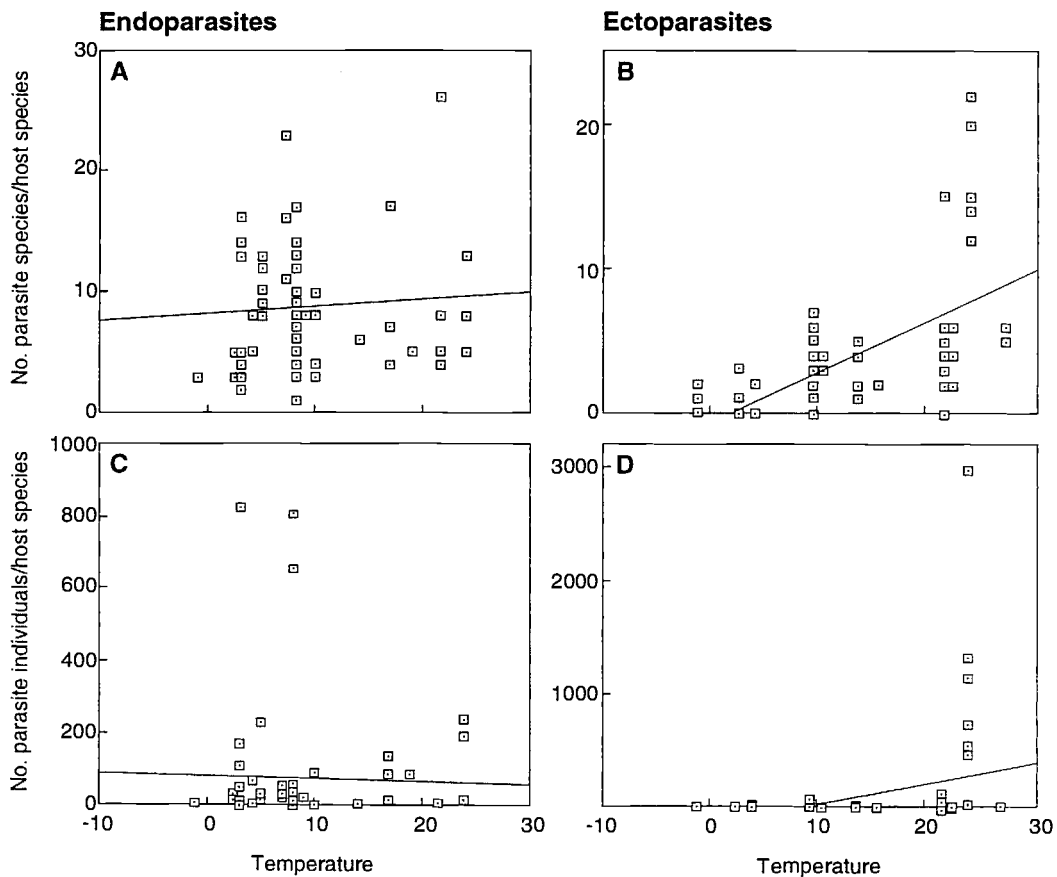


FIGURE 10 Metazoan ecto- and endoparasites on the heads and gills of marine teleost fish at various latitudes. From Rohde and Heap (1998).

TABLE V  
Human Infections with the Most Important  
Parasites Worldwide

Parasitic infection	Number infected
Intestinal roundworms	1400 million
Malaria	300 million
Bilharzia	200 million
Lymphatic filariasis	120 million
Amebiasis	40 million
Food-borne trematodes	40 million
Chagas disease	16–18 million
Leishmaniasis	12 million
Sleeping sickness	0.3 million
Guinea worm	0.1 million

Source: McGill: *The World of Parasites* (<http://martin.parasitology.mcgill.ca>).

and one harboring the larval cestode *Triaenophorus crassus* acquired in western Alaska.

### E. The von Ihering Method

About 100 years ago, von Ihering first used parasites and ectocommensals to clarify places of origin and dispersal of hosts and ancient land connections between land masses. One assumption of the method is that animals have acquired the greatest diversity in the area where they have been longest. Mamaev and Margolis applied the method to Pacific salmon. On the basis of their finding that salmonids harbor many freshwater

parasites, some of them with complex life cycles, but no marine endoparasites, they concluded that salmonids are of freshwater origin. Many similar studies have been made.

## XI. ECONOMIC AND HYGIENIC IMPORTANCE OF PARASITES

### A. Human Parasites Are among the Most Important Disease Agents of Man

The web pages of The World Health Organization, Division of Tropical Diseases (CTD), and of the Centers for Disease Control (CDC) contain information about the current status of the important parasitic diseases, which is continually updated. Information about the most important parasites of man are given in Tables V and VI. Note that some of the most widespread species, such as *Demodex folliculorum*, a mite infecting the skin, and *Toxoplasma gondii*, are not included because of lack of data and usually symptomless infections.

A number of parasite species, most of them protozoans, but also the nematode *Strongyloides stercoralis*, have recently become more important because of AIDS. All these species multiply within the human host and may become fatal in immunodepressed patients. Protozoans involved are, among others, *Cryptosporidium* and *Giardia*, sometimes acquired in polluted drinking water, *Pneumocystis carinii*, Microspora, and *Entamoeba histolytica*.

TABLE VI  
Infections of the Most Common Parasites in Some Countries/Regions

Region	Infection	Number infected
United States	Pinworms ( <i>Enterobius vermicularis</i> )	50 million
	<i>Strongyloides stercoralis</i>	<1 million
South America	Bilharzia	45 million
	Chagas disease	16–18 million
	<i>Strongyloides stercoralis</i>	ca. 7.4 million
Africa	<i>Ascaris lumbricoides</i>	>200 million
	Bilharzia	100 million (mortality, 0.01 million)
	Malaria	23 million (mortality, 2.6 million)
China	<i>Ascaris lumbricoides</i>	ca. 100 million
	<i>Opisthorchis sinensis</i>	5 million
Middle East	<i>Ascaris lumbricoides</i>	100 million
	Hookworms	60 million
	Bilharzia	50 million

Source: McGill: *The World of Parasites* (<http://martin.parasitology.mcgill.ca>).

Note: *Demodex folliculorum* and *Toxoplasma* not included.

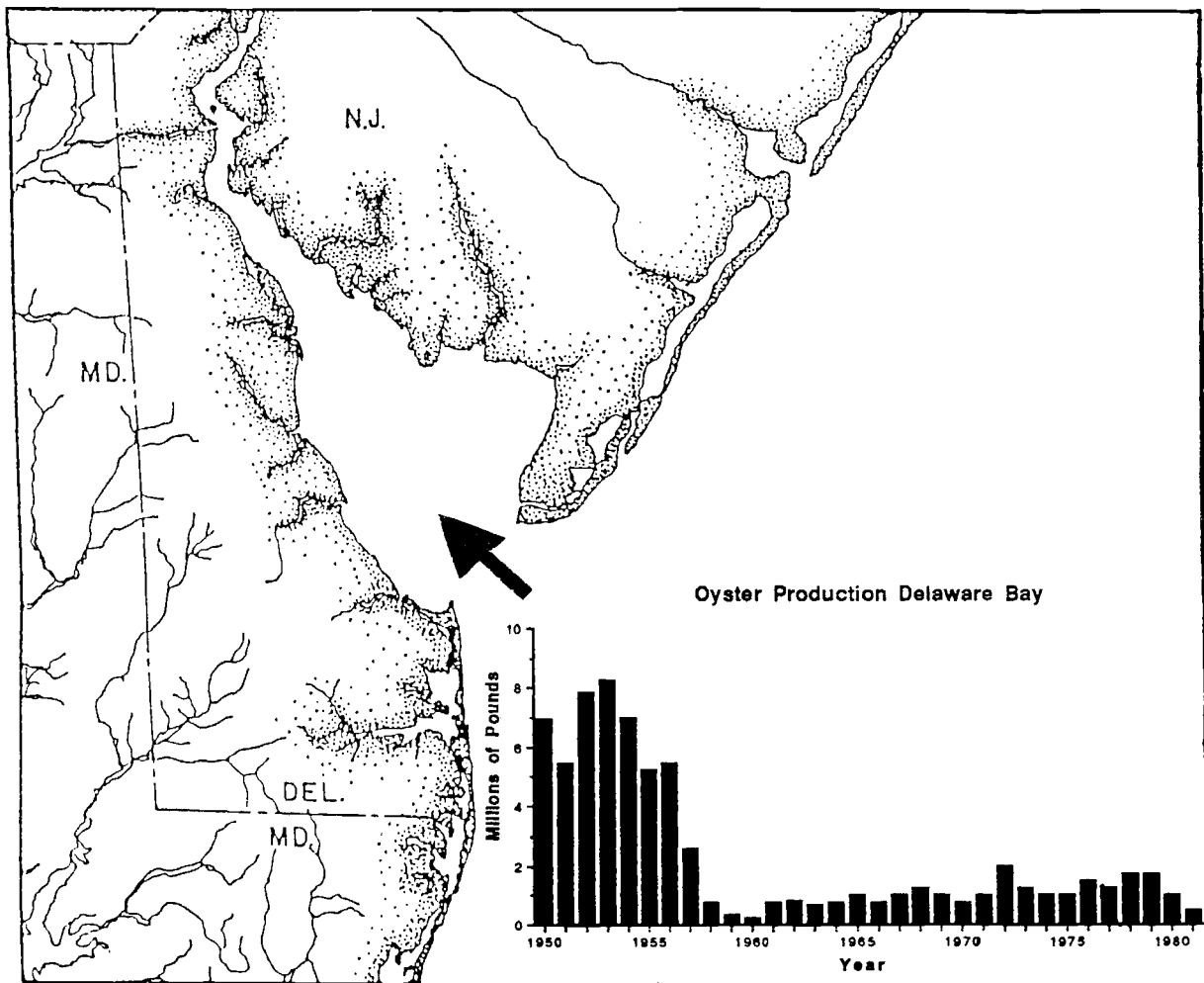


FIGURE 11 Oyster production in Delaware Bay, East Coast of the United States. Note: Collapse of oyster production in the late 1950s due to the haplosporidian *Haplosporidium nelsoni*. Modified after Sindermann, from Rohde (1993).

Resistance to drugs used against various parasite species has developed, and to insecticides used against various vectors. Hence, the epidemiological situation is getting worse. Malaria is becoming more prevalent in many countries, partly due to insufficient financial and human resources for control but also because of man-induced climatic and environmental changes, migration, and war. Likewise, the prevalence of intestinal parasites worldwide is increasing, partly due to increasing urbanization. In contrast, the global prevalence of guinea worm (*Dracunculus medinensis*) has decreased from about 10 million ten years ago to about 150,000 (most of them in the Sudan), as a result of a campaign to eradicate the infection. Almost certainly, malaria and several other important parasite diseases would spread into areas presently not affected

if global warming as the result of the greenhouse effect should occur.

### B. Parasites of Livestock Also Have Very Great Economic Importance

For example, nagana, a disease of many domestic animals caused by *Trypanosoma brucei brucei* and transmitted by tsetse flies in Africa, has made large areas of sub-Saharan Africa unsuitable for livestock production, and nematodes of sheep and cattle cause large economic losses in many countries. *Ostertagia ostertagi*, one of the nematodes infecting cattle, is estimated to cause an annual loss of \$600 million to the cattle industry in the United States alone. The situation is not improving

because of drench resistance of nematodes. Thus, in Australia about 90% of sheep farms have worms resistant to one or more drench classes.

### C. Parasites Have Repeatedly Caused Large Economic Losses in the Fisheries Industries

In particular, aquaculture is affected. On the East Coast of North America, for example, several protozoan parasites have repeatedly decimated oyster culture in several areas. The haplosporidian *Haplosporidium nelsoni* caused a decline in oyster production in the New Jersey waters of the Delaware Bay from about 5–8 million pounds between 1950 and 1955 to 167,000 pounds in 1960, which has never fully recovered (Fig. 11).

### D. Many Parasites, Particularly Nematodes, Are among the Most Important Pests of Plants

For example, wheat is attacked by the nematode *Anguina tritici*, potatoes are attacked by the potato cyst nematode *Globodera rostochiensis*, and rape, rice, etc. all have their specific nematodes that lead to large economic losses. The estimated annual loss to crop production due to nematodes in the United States is \$8 billion (12%), and it is \$78 billion globally ([http://ianrwww.unl.edu/son/nn\\_nema.htm](http://ianrwww.unl.edu/son/nn_nema.htm)).

### E. Parasitic Angiosperms Attack Many Crops

At least 11 species of witchweed, *Striga* spp., attack crops, including all the important tropical cereals; control is difficult because of the high production of seeds and longevity of seeds up to 20 years in the soil. Dwarf mistletoes almost always cause deformities of trees, leading to substantial reduction in yield and quality of timber. The mistletoe *Viscum album* reduced the yield

of a particular apple variety in the United Kingdom by 7–54% (Press and Graves, 1995).

### F. Parasites, Including Nematodes, Are Increasingly Used to Control Insect Pests

*Neoleptana carpocapsae* is bred in the laboratory and sprayed on rape to combat insects attacking it, and *Howardula* sp. effectively kills the dried-fruit beetle *Carpophilus mutulatis*. *Heterorhabditis bacteriophora* kills insects by carrying bacteria into them.

### See Also the Following Articles

COEVOLUTION • HABITAT AND NICHE, CONCEPT OF • PARASITOIDS • SPECIES INTERACTIONS

### Bibliography

- Bogitsh, B. J., and Cheng, T. C. (1990). *Human Parasitology*. Saunders College Publishing, New York.
- Esch, G., Bush, A., and Aho, J. (Eds.) (1990). *Parasite Communities: Patterns and Processes*. Chapman & Hall, London/New York.
- Esch, G. W., and Fernandez, J. C. (1993). *A Functional Biology of Parasitism: Ecological and Evolutionary Implications*. Chapman & Hall, London/New York.
- Kinne, O. (1980–1985). *Diseases of Marine Animals*, Vols. 1–4. Wiley, New York, and Biologische Anstalt Helgoland, Hamburg.
- Miyazaki, I. (1991). *Helminthic Zoonoses*. International Medical Foundation of Japan, Tokyo.
- Odening, K. (1969). *Entwicklungswege der Schmarotzerwürmer*. Akademische Verlagsgesellschaft Geest & Portig K. G., Leipzig.
- Poulin, R. (1998). *Evolutionary Ecology of Parasites*. Chapman & Hall, London/New York.
- Press, M. C., and Graves, J. D. (Eds.) (1995). *Parasitic Plants*. Chapman & Hall, London/New York.
- Rohde, K. (1993). *Ecology of Marine Parasites*, 2nd ed. CAB International, Oxford.
- Rohde, K. (1994). Niche restriction in parasites: Proximate and ultimate causes. *Parasitology* 109, 69–84.
- Rohde, K. (Ed.) (1997). The Origins of Parasitism in the Platyhelminthes, Proceedings of a Special Session of the VIIIth International Symposium on the Biology of the Turbellaria. *Int. J. Parasitol.* 27(6), 677–746.
- Schmidt, G. D., Roberts, L. S., and Janovy, J. (1995). *Foundations of Parasitology*, 5th ed. McGraw-Hill, New York.
- Urquhart, G. M., and Jennings, F. W. (1996). *Veterinary Parasitology*, 2nd ed., Iowa State Univ. Press, Ames, IA.



# PARASITOIDS

H. C. J. Godfray

NERC Centre for Population Biology and Imperial College at Silwood Park

---

- I. Life History Variation
  - II. Host Location
  - III. Host Acceptance and Oviposition Strategy
  - IV. Resistance and Virulence
  - V. Population Dynamics
  - VI. Community Ecology
  - VII. The Importance of Parasitoids
- 

**PARASITOIDS** are a type of animal that have a life history intermediate between that of a predator and prey. As normally defined, parasitoids are invariably insects and their larvae feed at the expense of other insects, with the exception of a few species that attack different types of arthropods or mollusks. The adult female parasitoid lays her egg on, in, or occasionally near the body of the host; the eggs hatch and the developing parasitoid larvae consume the host, eventually killing it. Like a parasite, only a single host is required for full development, but like a predator the host is invariably destroyed. The term parasitoid, originally coined by Reuter at the beginning of the twentieth century, is now used nearly universally to describe species with this life cycle, although in the older literature they are sometimes just called insect or protean parasites. Some solitary wasps have a similar life cycle except that the female parent carries paralyzed hosts to prepared nests or caches, and normally the requirement that hosts are attacked and oviposited on *in situ* is included in the definition of a parasitoid.

More than 50% of parasitoids belong to the insect order Hymenoptera: the sawflies, ants, bees, and wasps. The least derived suborder are the sawflies (Symphyta), which are nearly all phytophagous, although one family, the Orussidae, contains species whose larvae are parasitoids. From a lineage that was probably related to the Orussidae, the remainder of the Hymenoptera (the Apocrita) evolved, and these are in turn divided into a huge group: the Parasitica or parasitoid wasps, which are almost exclusively parasitoids, and the smaller Aculeata in which the ovipositor has become a sting. The Aculeata, derived from the Parasitica, are primitively parasitoids but have radiated into a wide range of feeding habitats, including predators, scavengers, and pollen feeders. The Aculeata contain nearly all the Hymenoptera familiar to the general public, including all the species of ants, bees, and wasps that show advanced sociality (eusociality). The Parasitica are classified into a large number of families and superfamilies, the most important of which are the Ichneumonidae, Braconidae, and Chalcidoidea.

The second major group of parasitoids belong to the two-winged flies or Diptera. The parasitoid habit has evolved on many occasions in this order, with the most important taxon being the family Tachinidae: These are almost exclusively parasitoids, with the adults frequently resembling houseflies (Muscidae, like the Tachinidae in the Calypterata). The parasitoid habit has also evolved on many occasions in the beetles (Coleoptera), although a relative small number of species feed in this way. The most frequently encountered beetle

parasitoids are probably those in the rove beetle family (Staphylinidae). Outside the three insect orders just mentioned, parasitoids are extremely scarce and taxonomically scattered.

As described in more detail later, parasitoids are important members of nearly all terrestrial ecosystems, but there is little agreement about how many species exist. Currently, about 10% of the approximately 1 million described insects are parasitoids, but parasitoids are almost certainly comparatively poorly known compared with other groups—parasitoid taxonomy is legendarily difficult. Most workers today would argue that there are between 4 and 8 million species of insect on Earth, of which 1.5–2 million are parasitoids. To put this into perspective, recall that there are less than 10,000 species of birds and approximately 5000 species of mammals.

## I. LIFE HISTORY VARIATION

A simple way to classify parasitoids is by the host stage that they attack. Holometabolous insects have three juvenile stages (egg, larva, and pupa), whereas hemimetabolous insects have nymphal stages with no pupal metamorphosis. The most common category of parasitoids is composed of those that attack larval or nymphal stages, although egg and pupal parasitoids are also common. The least common type is adult parasitoids, probably because the adult is best able to defend itself against parasitoid attack. Egg parasitoids deposit their own eggs within that of their host and are obviously very small insects. This life history has evolved several times but is most common in two families of chalcidoid wasp, the Trichogrammatidae and the Mymaridae. Members of the latter family, which have very narrow wings with pronounced cilia (fine hairs) and are known as fairy flies, include the smallest of all known insects. Some parasitoids oviposit into the eggs of their host, but their larvae only kill the host later when it is a larva or pupa. These are called egg–larval or egg–pupal parasitoids, whereas another common life history is found in larval–pupal parasitoids.

Most parasitoid eggs are deposited in or on the host, but a substantial minority (especially among the flies and beetles) are deposited in the vicinity of the hosts, and the final stages of host location are carried out by a specialized first-instar larva (a triangulum or planidial larva). There are two main modes of parasitoid larval development: as endoparasitoids or as ectoparasitoids. As their names suggest, the former feed internally within the host and the latter externally. Occasionally,

a parasitoid species may begin life as an ectoparasitoid and switch midway through development to an endoparasitoid or vice versa. Ectoparasitism is found most frequently in species that attack concealed hosts in plant tissue or other refuges, where the external feeding parasitoid has some protection from the environment and predators. Parasitoids can also be classified as idiobionts or koinobionts. Idiobionts kill or permanently paralyze their host at oviposition, whereas koinobionts either do not or only temporarily paralyze the host so that it recovers and continues feeding. The parasitoid suspends its own development, normally as a first instar, and waits for the host to reach a size at which it can support its development, whereupon it resumes growth that results in host death. The koinobiont life history, normally only found in endoparasitoids, allows the adult parasitoid to attack hosts when they are still too small to support the full development of the progeny. However, while in a state of suspended development, the parasitoid larva is at risk from the host insect's immune system.

Part of the definition of a parasitoid is that its larva requires just a single host for development; however, it is possible for several parasitoids to share the same host. Species in which only a single parasitoid develops per host are termed solitary as opposed to gregarious parasitoids. The larvae of solitary parasitoids tend to have large mandibles which they use to attack other parasitoids in the host, whether conspecific or of a different species, and thus the difference between the two types of parasitoid is more than just one of degree. Occasionally, very large numbers of parasitoids emerge from a single host, for example, several hundred braconid wasps from a single hornworm (hawkmoth, Sphingidae) larva, although the largest broods are known from certain specialized wasps with polyembryonic development. In these species (in particular certain genera of chalcidoid Encyrtidae), one or two eggs are laid into the host, but these divide asexually to produce many embryos—occasionally hundreds and even more than 1000.

Unlike predation, in which feeding removes a prey item, a parasitized host may be discovered by a second parasitoid individual prior to its destruction. Sometimes, parasitoids avoid previously parasitized hosts, a practice called host discrimination, and this may be aided by chemical marks deposited by the first female. However, in many circumstances the second female will be selected to add its own eggs to the host. If the second individual is of the same species as that of the first, this is termed superparasitism, whereas if it is of a different species the term multiparasitism is used. Occasionally,

a female may parasitize a host that she had previously attacked (self-superparasitism). Some, but not all, parasitoid species can distinguish between hosts previously parasitized by self and by other conspecifics. In superparasitism and multiparasitism, the parasitoid larvae compete with each other for host resources; when hyperparasitism occurs, one parasitoid larva actually feeds on the other. There are two main categories of hyperparasitism, obligate and facultative. As the name suggests, an obligate species can only develop as a parasitoid of a parasitoid, whereas a facultative species develops as a normal parasitoid when it encounters an unparasitized host and as a hyperparasitoid when it discovers a previously attacked host. Hyperparasitism is sometimes called secondary parasitism (as opposed to primary parasitism) and cases of tertiary and even facultative quaternary parasitism have been recorded. An unusual life history is found in some chalcidoid wasps of the family Aphelinidae: Females develop as primary parasitoids, but males develop as secondary parasitoids of other aphelinids (including females of the same species).

## II. HOST LOCATION

Parasitoids are small to tiny animals that search for hosts that are themselves small or tiny animals in a very complex environment. They have solved the formidable problems of host location by making use of a battery of clues and stimuli that reveal information about the presence or potential presence of hosts. The majority of these cues, especially those acting over long ranges, are chemical, but visual, tactile, thermal, and auditory information is also used. Moreover, the parasitoid is not a hard-wired automaton but constantly updates its estimates of the values of different stimuli to modulate its strategies for host location.

One can envisage a parasitoid as utilizing a hierarchy of stimuli of increasing information content. Perhaps the lowest grade stimuli are to habitats or microhabitats that might contain hosts. Many parasitoids attacking plant-feeding hosts orientate toward green objects, whereas the parasitoids of *Drosophila* feeding in different types of fermenting substrate are attracted to volatile chemicals characteristic of their hosts' microhabitat. Plants damaged by herbivores release a cocktail of volatiles which are often highly attractive to parasitoids. In some cases, parasitoids respond to artificially damaged hosts, whereas in other cases they are only attracted to the "host-host plant" complex. It has been shown that the plant's response to particular herbivores and their salivary and other secretions can be very specific,

prompting the intriguing suggestion that the plant may have been selected to recruit "bodyguards" to protect itself against its herbivores.

Host waste products reveal valuable information about their originator's whereabouts. Many parasitoids are highly attracted to host frass (feces), whereas volatiles associated with salivary or mandibular gland secretions, or with silk, are also used in host location. Some hosts have symbiotic relationships with other organisms that may provide clues to their location: For example, wood-boring sawflies are associated with symbiotic fungi that digest the wood, and their parasitoids are strongly attracted to chemicals released by the fungi.

Clearly, hosts will be under strong selection to reveal as little about their location as possible to searching parasitoids. Many parasitoids use vision in the final stages of host location and camouflage and immobility may help protect hosts against parasitoids and more traditional predators. Vibration is particularly important for species attacking hosts feeding in concealed material, and some hosts "freeze" if they detect the presence of a searching parasitoid. There is evidence that parasitoids may also use thermal gradients to detect the presence of concealed hosts.

Hosts should maintain as much as possible the chemical equivalent of "radio silence," although sometimes hosts have to advertise themselves for different reasons and in doing so may provide clues to searching parasitoids. Some bark beetles (Scolytidae) can only overcome tree defenses by mass attacks of many hundreds or thousands of individuals, and these are coordinated by the beetles releasing an aggregation pheromone. However, not only other beetles but also their larval parasitoids are attracted to the pheromone. Similarly, parasitoids are often sensitive to their host's sex pheromones—both species that attack the adult hosts directly and species that attack other stages and use the presence of adults as evidence of the presence of eggs, larvae, or pupae. A particularly devilish instance of parasitoids subverting their hosts' sexual signaling is that of the dipteran parasitoids of grasshoppers that home in on stridulating males.

Parasitoids are born with an innate hierarchy of responses to different stimuli, but they will change the position of cues in this hierarchy and incorporate new stimuli as they gain experience in host location. For example, if a novel volatile chemical such as vanilla is presented to the parasitoid at oviposition, the parasitoid learns to associate this chemical with hosts and subsequently flies to baits releasing this chemical. In a more natural setting, *Drosophila* parasitoids preferentially fly toward certain potential host microhabitats if they have



previously been successful in finding hosts in that microhabitat. Many parasitoids, as they emerge from their pupa or cocoon, carefully examine with their antennae any remains of the host at the pupation site. They appear to be learning potential host cues because they subsequently fly toward sites containing the same volatile chemicals. An interesting corollary to this behavior is that it will tend to promote the formation of host races or biotypes.

Hosts are often found in patches that parasitoids exploit, but with decreasing returns as the fraction of unparasitized hosts decline. There will thus come a time when the parasitoid will be selected to abandon the current patch and search for a new site. Studying parasitoid patch use behavior has many parallels with predators searching for prey that are distributed in patches. This is one of the classical problems of optimal foraging theory, a field of behavioral ecology. In the simplest case, predators should leave the patch when their instantaneous rate of gain of food equals the maximum long-term rate that can be achieved in that environment. Longer patch residence is thus predicted in poorer environment, for example, when interpatch travel times are long.

The qualitative predictions of patch use theory have been supported by studies of parasitoids, but specific aspects of their searching behavior complicate the theory's application. For example, previously parasitized hosts do not disappear (as do eaten prey) but remain as potential sites for superparasitism. The parasitoid may also have to waste time in identifying them as parasitized, although they may provide useful information about host distributions. Parasitoid behavior may also change critically in the presence of conspecifics: When a parasitoid is searching a patch alone, any superparasitism is normally damaging self-superparasitism, which is not the case if other females are active in the site. Current research in this area using statistical models based on survival analysis seeks to identify the precise factors determining when a parasitoid abandons the patch.

### III. HOST ACCEPTANCE AND OVIPOSITION STRATEGY

After a parasitoid locates a host, it then has to make a series of decisions that are very important for its future reproductive success. First, it must decide whether the host is of sufficient quality to be used as an oviposition site or whether not to waste eggs and search for a better

host, perhaps using the current host as food. Given that the host is of sufficient quality, gregarious parasitoids have to decide how many eggs to lay, while hymenopterous parasitoids have to decide whether to lay male or female eggs. Only Hymenoptera have to make this last decision because they have a genetic system called haplodiploidy in which females develop from fertilized eggs and males from unfertilized eggs. By choosing whether or not to fertilize the egg, the parent can control the sex of its offspring. Unlike groups in which sex is determined by the random segregation of sex chromosomes, natural selection can operate on the behavior of haplodiploid females to produce sex ratios adapted to local conditions.

Before laying an egg, most parasitoids determine whether a potential host is suitable by examining it externally, and in the case of endoparasitoids often internally as well by using sensillae on their ovipositors. At either stage, a host may be rejected if unsuitable, perhaps because it is of the wrong species. As part of biological control campaigns, parasitoids sometimes have to be mass reared, and much effort has gone into investigating whether they can be reared on artificial hosts. Critical to a successful strategy of this type is an understanding of exactly what host stimuli trigger oviposition.

Even if a host can potentially support a parasitoid larva, it may still not be worth laying an egg if the resulting offspring is of poor quality, perhaps very small in size. Understanding how natural selection molds host acceptance behaviors has many parallels with understanding how natural selection influences diet breadth in predators and other consumers, a second major strand of optimal foraging theory. Theory assumes that predators should maximize their rate of gain of energy (or other resources) and hence should not take "time out" from searching for good-quality prey items to attack poorer quality types from which the energy returns do not justify the time it takes to subdue and process them. Applying these ideas to parasitoids, searching females should not waste time ovipositing on poor-quality hosts when that time could be better spent searching for good hosts. Exactly what constitutes a poor host thus depends on the quality of the environment: Attacking a particular host of medium quality might be economically advantageous in a habitat in which very good hosts are rare or absent, but it may not make sense in a environment in which the best hosts are abundant. There is experimental support for this prediction.

In many cases, the simple assumption that parasitoids should seek to maximize their rate of gain of

reproductive success with time is probably overly simplistic. All parasitoids have a limited number of eggs, and maximizing rate of gain of fitness may lead to exhaustion of egg reserves prior to death. More sophisticated models have been developed that maximize reproductive success over the lifetime of the insect and predict that the parasitoid may become more selective as its egg supply diminishes. Again, there is experimental data supporting this prediction, although whether egg limitation is a major feature of parasitoid biology is a highly contentious topic. Another factor that may influence oviposition decisions is host feeding. Many adult parasitoids can feed from hosts, although this normally results in the host's destruction. Host feeding provides the parasitoid with energy and nutrients and hence allows it to produce more eggs. Thus, there is a trade-off: Should the parasitoid use a host for reproduction and possibly risk later starvation or egg exhaustion, or should it forgo an immediate increment in fitness in the hope of deferred payback? In these circumstances, it clearly makes sense to feed from suboptimal hosts.

The question of optimal clutch size in parasitoids can be explored using similar economic arguments, in this case ideas originally developed to understand how natural selection operates on bird clutch size. An obvious prediction is that parasitoids should adjust their clutch size to the amount of resources available, and there is ample evidence that parasitoids do this, with the numbers of eggs laid typically being closely correlated to the size of the host. Careful experiments have shown that parasitoids can often estimate host size from very subtle cues; for example, some egg parasitoids can assess host size by very limited information about the local curvature of the host egg. As eggs are added to the host, the fitness returns from each egg typically decline because the members of the brood compete with each other for a fixed pot of resources. The optimal clutch size is defined as the number of eggs for which the marginal returns from adding additional eggs exactly equal the marginal opportunity costs in terms of wasted time or wasted eggs on the current host. This argument suggests that in good-quality environments in which fresh hosts are easy to discover, smaller clutches should be laid. Similarly, small clutches should be found in circumstances in which egg reserves are low. Experiments on gregarious parasitoids have supported both predictions.

Are solitary and gregarious parasitoids just part of a single continuum in clutch size behavior? There are arguments that suggest this is not so: Solitary parasitoids typically have large mandibles which they use to attack any other parasitoid in the host, including sib-

lings, whereas gregarious parasitoids do not have this aggressive behavior. For evolution to convert a solitary parasitoid to a gregarious parasitoid, natural selection must operate not only on parental clutch size behavior but also on larval morphology and fighting. Although adult parasitoids and their offspring have many evolutionary interests in common, these interests are not identical and there is the potential for evolutionary parent-offspring conflict. Models incorporating this genetic conflict predict that the solitary life history is qualitatively distinct from gregariousness, and that there is hysteresis in the transition between the two states with gregariousness far more difficult to evolve from the solitary state than vice versa.

The study of the sex ratio is one of the most successful areas of evolutionary ecology. The problem is to predict why some species of plants and animals produce sex ratios that differ from 50:50. For a deviation to occur, there must not only be selection for a biased sex ratio but also the flexibility to control son and daughter ratios. Haplodiploid parasitoid wasps have been critical in developing this field because they have both the incentive and the means to produce unusual sex ratios.

As Fisher demonstrated in the 1930s, a 50:50 sex ratio is expected to evolve because were the population to be biased toward females, sons would be a more efficient way of getting genes into future generations (because they would, on average, mate with more than one female), whereas if the population were to be biased toward males, daughters would be a more efficient way of getting genes into future generations (because sons would mate, on average, with less than one female). However, this argument assumes that the population is completely mixed and that all males compete for all females. This assumption breaks down if sons compete for mates, including their sisters. In such situations, the ovipositing female is selected to produce fewer sons to avoid wasteful within-family conflict and also to provide more potential mates for the male offspring. Hamilton, who in the late 1960s developed these ideas, called this process local mate competition and marshaled a long list of examples in which mating among siblings was associated with female-biased sex ratios: 18 of the 26 examples he quoted were parasitoids.

Recent studies have shown that some parasitoids are able to assess the degree of local mate competition that their progeny are likely to experience and to adjust their sex ratio accordingly. For example, *Nasonia vitripennis* is a gregarious parasitoid of fly pupae (e.g., of blowflies that live in birds nests). Wasps emerge in the bird's nest and mate prior to dispersal (the male has limited powers of flight). If a single female colonizes a

bird's nest, all her offspring will mate among themselves and the wasp should produce just enough sons to mate all her daughters; if two females colonize a nest then local mate competition still occurs, although it is less intense, and a sex ratio of approximately 25% males is predicted. With greater numbers of females, the sex ratio quickly asymptotes at 50%. In laboratory experiments in which the number of insects searching together is manipulated, females respond to the presence of conspecifics by producing less female-biased sex ratios.

It had long been noted that parasitoid wasps often lay male eggs in small hosts and female eggs in large hosts, but why this evolved was only understood in the late 1970s by Charnov as part of a wider theory of environmentally correlated sex ratios. In parasitoids, there is normally a strong correlation between host size and the size of the adult wasp that eventually emerges from it. In general, wasps suffer from being small; they tend to have reduced longevity and fecundity and are less able to withstand the insults thrown at them by the environment. However, the negative consequences of being small are probably not experienced equally by the two sexes. In particular, female reproductive success may depend more critically on size than does male reproductive success for the simple reason that eggs are expensive and sperm cheap. This assumption is difficult to test because it requires size-dependent fitness to be measured in the field, which is difficult for such small animals, but the limited amount of evidence available supports this notion. If female fitness does increase more rapidly with size than does male fitness, then Charnov's theory predicts that males should be laid in relatively small hosts and females in relatively large hosts. Moreover, there should be a sharp threshold in host size below which only males, and above which only females, are produced.

Laboratory experiments tend to show a threshold in host size as predicted by theory, although typically it is not as sharp as expected, perhaps because the wasp is using subtly different cues to assess host size than those used by the experimenter. To test whether relative or absolute size is important, many workers have presented medium-sized hosts to wasps in conjunction with either larger or smaller hosts. Theory predicts that the wasp should produce male eggs in medium-sized hosts when they are relatively small and female eggs when relatively large. Some wasps do indeed do this, although others have a much more fixed behavior, responding only to absolute host size. Possibly, these wasps always encounter the same host distribution in the field and hence have never evolved the flexibility to deal with variable patterns of host size.

#### IV. RESISTANCE AND VIRULENCE

Endoparasitoids that delay their development while the host continues to feed and grow in size have to protect themselves from the host's defenses. The main host defense against parasitism is a cellular immune response called encapsulation. Cells (hemocytes) circulating in the blood cavity (hemocoel) recognize a parasitoid egg or larva as foreign and cause other cells to aggregate to the intruder, forming a capsule. The cells in the capsule fuse, and the whole structure becomes hardened and melanized, thus leading to the death of the parasitoid either through suffocation or through the release by the parasitoid of necrotizing substances. Encapsulation is not a specific antiparasitoid measure but rather operates against any foreign body.

There are two broad parasitoid strategies to counter host resistance: passive and active defense. In the first, parasitoids either camouflage themselves so they are not recognized as foreign or insert their eggs in specific organs away from circulating hemocytes. How this camouflage works can be seen in some species in which the egg is coated with a layer of protein prior to injection into the host. If the egg is dissected out of the wasp and the protein layer removed, the egg becomes susceptible to encapsulation if artificially "oviposited" into the host. Many parasitoids carefully position their eggs in the host brain or ganglia where they avoid hemocytes, whereas even eggs placed in the hemocoel may obtain some protection by sticking to fat body and other host organs.

The most common form of active defense is the injection of toxins at oviposition or their later secretion by the developing larva. It is common for the ovipositing female to temporarily paralyze the host so that it is easier to handle, but the same or other substances can also damage the host's immune system. Many parasitoids, though so far only Hymenoptera (particular the Ichneumonidae), inject viruses into the host at oviposition. The best studied type is polydnavirus, so named (poly-DNA-virus) because the genetic material is composed of many separate DNA molecules within a single protein coat. The virus DNA is stably integrated within the wasp's genome, but in certain cells of the female reproductive tract it is copied, greatly amplified, and encapsidated. In the host, the virus does not replicate, but it does invade the hemocytes responsible for encapsulation which it can destroy by triggering apoptosis. How polydnaviruses evolved is not clear: In particular, are they part of the wasp's genome that has acquired virus-like properties or an originally autonomous virus that has been captured and tamed by the wasp? In-

stances are known of wasps injecting viruses that appear unrelated to polydnviruses and also of wasps injecting virus-like particles that resemble the protein coats of viruses but that contain no DNA.

Parasitoid eggs are surrounded by embryonic membranes which normally disintegrate after hatching. However, in some species, the cells of the membranes dissociate and float free in the hemocoel where they grow enormously in size. These cells are called teratocytes, and their primary role is most likely secretory (they are packed full of endoplasmic reticulum), producing substances that limit the host's immune defense. When the parasitoid is fully grown it consumes the teratocytes, which may thus also have a nutritive function.

Hosts and parasitoids almost certainly exert very strong selection pressures on each other, and observed levels of resistance and virulence will be determined by this coevolutionary interaction. Two major factors determining the outcome are the asymmetry in the interaction and the nature of the costs of resistance and virulence. The interaction is asymmetric in that every parasitoid has to overcome its host's defenses in order to survive, whereas not every host is parasitized. Models of the dynamics of host–parasitoid coevolution predict that in certain circumstances hosts may be selected not to defend themselves but to “gamble” on not being attacked. A major determinant of whether this strategy is favored is the nature of the costs to the host of investing in resistance mechanisms and of the parasitoid in investing in countermeasures. Costs are difficult to measure, but artificial selection experiments in the fruit fly, *Drosophila*, have shown that increased resistance to its parasitoids is accompanied by a decrease in competitive ability under conditions of resource stress. It appears that the fly switches limiting resources from optimizing feeding efficiency to immune function.

## V. POPULATION DYNAMICS

Host–parasitoid population dynamics have received considerable study for two main reasons. First, parasitoids are important biological control agents and understanding how they may regulate their host has clear applied relevance. Second, host–parasitoid interactions are a useful model system for answering broader questions in population biology applicable to all resource–consumer interactions. Host–parasitoid interactions are attractive as model systems because of the simplicity of the trophic relationship: In most cases, a parasitoid attack results in one (solitary species) or a certain number (gregarious species) of recruits to the next genera-

tion of parasitoids. This is in contrast with predation, in which it is far more difficult to predict how a single instance of prey capture influences the future number of predators.

Modern consideration of host–parasitoid population dynamics began in the 1930s, particularly through the work of the Australian ecologist Nicholson and his collaboration with the mathematician Bailey. They considered the case of a host–parasitoid interaction with discrete, synchronized generations. Hosts are exposed to parasitoids during a fixed developmental stage, and those that survive parasitism and other sources of mortality produce the next generation of hosts the following years. Parasitized hosts are the source of the subsequent parasitoid generation. In their basic formulation, known as the Nicholson–Bailey model, they assumed parasitoids searched randomly in the environment. The average number of times a host is encountered is then simply a product of parasitoid density ( $P$ ) and a constant ( $a$ ) known as the attack coefficient. Providing parasitoid search is random, the probability of escaping parasitism is simply the zero term of a Poisson distribution with this mean (i.e.,  $e^{-aP}$ ).

The Nicholson–Bailey model is important because it encapsulates the minimum bare bones of a host–parasitoid interaction. It tells us what to expect in the absence of any biological complications, and what it tells us is perhaps surprising. If the Nicholson–Bailey model is iterated over many generations, the densities of hosts and parasitoids oscillate with ever-increasing amplitude until one or both species go extinct. In essence, the parasitoid overexploits the host, which crashes to low densities, and this is followed by a drastic decline in parasitoid number. The host then recovers and in the temporary absence of the parasitoid increases to higher densities than before; this is followed, after a lag, by a huge growth in parasitoid numbers and even more cataclysmic overexploitation. The combination of random search and the time lags inherent in this discrete-generation framework do not allow a host–parasitoid interaction to persist.

The intrinsic instability of the Nicholson–Bailey framework has set the agenda for much of host–parasitoid population dynamics in the past 60 years: What aspects of real host–parasitoid interactions allow the persistence that is so palpably clear in the field? Introducing a realistic functional response (a limit on the number of hosts a parasitoid can attack due to egg or time limitation) does not help, and if anything it makes nonpersistence more likely. In the 1960s and 1970s, many workers thought that interference between parasitoids was the key. Interference is said to occur if female parasitoids disrupt each other's searching, per-

haps because when two parasitoids meet they fight or in other ways interact. Such interference might reduce the efficiency of high-density parasitoid populations and break the cycle of overexploitation and recovery. However, although interference occurs and can be measured in the laboratory, a consensus soon formed that in the field it was unlikely to be strong enough to allow persistence. Another idea was to assume that in addition to parasitism, hosts were also subject to direct density dependence due to competition for food. This additional density dependence can stabilize the Nicholson–Bailey interaction and may very well be important in some host–parasitoid interactions in the field. However, it is unlikely to be the complete answer because at the stable population densities that are predicted by this model, there is still substantial competition for food. This theory can neither explain why so many hosts are regulated at levels well below their food-carrying capacity nor why parasitoids can cause such dramatic reductions in host density after their release in biological control programs.

Most workers today believe that host–parasitoid persistence involves a general phenomenon called heterogeneity of risk (this idea actually dates back to the last papers of Nicholson and Bailey in the 1960s). Overexploitation of the host population occurs because all hosts are equally susceptible to parasitoid attack, and therefore nearly everyone succumbs when parasitoid densities are high. However, suppose that some hosts are in a refuge and protected from parasitoids. These individuals can survive periods of high parasitoid densities and prevent the crash in host density that initiates the diverging oscillations (of course, if too many hosts are in a refuge then the parasitoid will not be able to prevent the host population increasing without restraint).

What might these refuges be? There are many possibilities. First, there may be physical refuges in which hosts are protected from parasitism. For example, the host might be a stem borer and a certain fraction of hosts may burrow so deeply in the plant tissues that they are inaccessible to parasitoids. Second, a certain fraction of the hosts may be physiologically immune to parasitoid attack. Third, there may be a phenological mismatch between host and parasitoid: Some hosts may avoid parasitoids by emerging either sufficiently early or late in the season so that they miss all or most searching parasitoids.

Finally, and probably most important, the refuge may not be physical, physiological, or phenological but statistical: A certain number of hosts, more than would occur if parasitoid search was random, may escape para-

sitism by chance each year. For example, suppose that hosts are distributed across the environment in patches, and that certain patches are more attractive than others to parasitoids (for reasons not associated with host densities, some patches may be in areas that are more likely to be encountered by parasitoids). Provided enough hosts occur in rarely encountered patches—statistical refuges—the interaction will be stable. (A popular way to model this is to replace the Poisson distribution in the Nicholson–Bailey model with a different, more aggregated statistical distribution such as the negative binomial. The resulting interaction is stable provided the variance of the distribution is sufficiently large). All these ways of stabilizing the Nicholson–Bailey model rely on there being differences among individuals in the risk of parasitism; hence their general labeling as models of heterogeneity of risk.

Recent developments in host–parasitoid population dynamics have largely involved the study of interactions with overlapping rather than discrete, synchronized generations and interactions in which the spatial distribution of the population needs to be taken into account. Models for species with overlapping generations typically abandon the discrete time difference equations of the Nicholson–Bailey model and use instead differential equations, although they often incorporate time delays to represent developmental lags. Two new insights have come from this approach. First, an interaction that is otherwise identical to the Nicholson–Bailey model may be persistent if there is a long-lived, invulnerable host stage. The reason for this relates to the arguments regarding heterogeneity of risk discussed previously: The long-lived invulnerable stage provides a refuge that allows the host population to ride out periods of very high parasitoid densities.

The second insight to emerge concerns when hosts and parasitoids differ in generation time, in particular when the parasitoid generation time is approximately half that of the host. The population can then display persistent cycles with a period approximately equal to the host generation time. This is of more than technical interest because such cycles are common among pests of tropical plantation trees such as coconut, oil palm, and cocoa, and these pests tend to be attacked by parasitoids with shorter generation times. The cycles arise because of dynamic interference between the two time lags in the system. Suppose there is a temporary increase in the densities of host: The progeny of this “blip” will form a secondary blip one host generation later. However, the parasitoids will also be able to take advantage of the increase in host population density, and this will result in a secondary

increase in their densities one parasitoid generation later. If the two generation times are equal, the two secondary blips coincide (the Nicholson–Bailey situation), but if the parasitoid generation time is approximately half that of the host the two blips do not coincide, and the parasitoid causes a trough in host density, an “antiblip,” half a host generation later.

There are two main ways to incorporate a spatial component into host–parasitoid models. The first is to assume that the system consists of discrete subsystems linked by migration, whereas the second is explicitly to incorporate space and write equations that describe how host and parasitoid densities change at all spatial coordinates. Species that have isolated populations linked by migration are normally described as having a metapopulation structure. Metapopulations are an area of very active research in population biology, although studies of interacting metapopulations, such as that of a parasitoid and its host, are still rare. The most interesting aspect about host–parasitoid metapopulations is that the interaction they describe can persist, even if every component population would be nonpersistent in isolation. For this to happen, component populations must fluctuate out of synchrony so that one population doomed to extinction can be rescued by immigration from a different population in a different phase of the host–parasitoid cycle. Random, local effects that promote asynchrony favor the persistence of a metapopulation, whereas regional fluctuations in characteristics such as climate will tend to synchronize populations and hence act against persistence.

In fully spatial models, the rescue effect can still operate, but now more elaborate spatial processes can also be observed. For example, a wave of hosts can spread through the environment, chased by a following wave of parasitoids. The hosts escape parasitoids by moving into uncolonized areas ahead of the advancing waves, whereas the parasitoid destroys the host population behind the wavefront and would itself be destroyed were it not able to chase the host population. These waves tend to form spiral patterns, a feature common to many spatial interactions in population dynamics, physiology, and even inorganic chemistry. For certain parameter combinations, the organized spiral waves break down into a seemingly random jumble of interactions and short-lived wavefronts called spatial chaos. Experimental studies of spatially extended parasitoid populations are still somewhat behind theory and one of the greatest challenges of contemporary host–parasitoid population dynamics is to devise ways of studying spatial dynamics in the field.

## VI. COMMUNITY ECOLOGY

Few if any host–parasitoid interactions exist in isolation: Most parasitoids attack more than one species of host, and the majority of hosts are attacked by several species of parasitoid. How does one begin to understand the workings of a large community of interacting hosts and parasitoids? One approach is to build up from the simple one-host, one-parasitoid interactions discussed previously. Alternatively, one can survey the properties of real communities and try to deduce patterns that may reveal how they are structured.

The simplest possible community that is more complex than those discussed previously consists of one host and two parasitoids, or two hosts attacked by a common parasitoid. Both tell us interesting things. In the first case, the simplest models predict that one of the two parasitoids will invariably go extinct. This is a corollary of a basic finding in population ecology that identical species cannot coexist on the same resource—the competitive exclusion principle. For coexistence to occur, other factors must be brought into play. In particular, coexistence can occur if the ecological niche (i.e., the host) is divided in such a way that the parasitoids are no longer identical. Thus, one parasitoid may use the host at one time of year or in one microhabitat, or it may specialize on a different developmental stage of the host than the other. Coexistence can also occur by a more statistical route if hosts differ in their susceptibility to attack by parasitoids so that some are in the statistical refuge discussed previously and if the probability of being protected from attack by one parasitoid is independent of other. Spatial processes can also promote coexistence. Consider two species of parasitoid, one of which is always the winner in a straight competition at a single locality, whereas the other has a superior dispersal capability. Coexistence is now possible in an environment in which new host populations are constantly being generated because the poorer competitor is able to find unexploited hosts: This is known as a competition–colonization trade-off.

I now discuss the two-host, one-parasitoid case. Again, the simple models predict extinction, this time of one of the host populations. This occurs because the equilibrium population of parasitoids maintained on the surviving host species is sufficient to prevent the second species from replacing itself. A recent laboratory experiment cleverly illustrated this using two flour moths attacked by the same species of wasp. This is a similar situation to competitive exclusion, but the trophic structure is reversed: instead of two species com-

peting for the same resource, two species are subject to the same natural enemy. In fact, there are deep biological and mathematical symmetries between these two cases, and this has led to the phenomenon of two species interacting through a common natural enemy being referred to as apparent competition to stress the parallels with direct competition in which two species interact through a common resource. Again, for a two-host, one-parasitoid system to persist, something extra must be added to the simplest models: The two hosts might be present at different times of the year or might be spatially segregated, or the parasitoids might preferentially exploit the most abundant host (behavioral switching).

Much of traditional community ecology has stressed how biological communities may be structured by resource competition, and thus it does not apply to many insect communities in which the majority of species feed on different host plants and thus never come into contact. Apparent competition is significant because it can at least potentially structure a community in exactly the same way as does direct competition. The degree to which this actually occurs is a major theme in current insect ecology.

Progressing from models of three species to those with larger numbers of components becomes increasingly difficult. There are two problems: First, more assumptions have to be made about how different species interact and about the values of large numbers of parameters. Seldom are there field data to reduce this burden of supposition. Second, the dynamic behavior of larger communities becomes increasingly more complicated. For example, a model of a five-species community consisting of two hosts, two specialist parasitoids, and a generalist parasitoid showed the same range of population dynamic behaviors as those exhibited by simpler communities. However, it also showed more complex behaviors in which the full five-species community was unstable with one or more species going extinct, but with the resultant smaller communities, after they had reached equilibria, then being susceptible to invasion by the species that had recently gone extinct. Other multispecies models have shown complex chaotic dynamics. Currently, it is unclear the extent to which the bottom-up approach, modeling explicitly the dynamics of each member of a large community, is a feasible way to approach parasitoid community ecology.

The top-down approach to community ecology consists of searching for patterns in multispecies assemblages that provide evidence of structuring forces. For example, workers have searched for patterns in the number of species of parasitoids attacking different spe-

cies of hosts. The attraction of this approach is that there are numerous studies in the literature providing information on the parasitoid complexes of different insects. The major result to emerge from these studies is that host feeding niche influences parasitoid species numbers. Leaf-mining insects are attacked by the largest number of species, with successively smaller numbers attacking more concealed hosts (gall formers, shoot borers, and root borers) and less concealed hosts (species living in leaf rolls and ties and those living externally like typical caterpillars). There are two explanations for this pattern. One suggests that the number of parasitoid species that can coexist on a single host is influenced by the fraction of that host's population that inhabits a refuge from (all) parasitoid attack. Proponents of this view argue for a correlation between feeding niche and the size of the refuge. Alternatively, many species are likely to occur on hosts that are taxonomically and ecological similar to other hosts because there is less of an evolutionary hurdle to overcome in incorporating a new species in a parasitoid's host range. Leaf-mining insects have a far more uniform ecology than insects feeding in other host niches. They are also relatively taxonomically homogeneous, and this may have facilitated host transfer and broad parasitoid host ranges.

There are less data in the literature on the host range of parasitoids because this requires all potential hosts in a area to be surveyed. However, the data that are available support one major conclusion: Idiobionts have broader host ranges than koinobionts. Recall that idiobionts kill or permanently paralyze their host at oviposition, whereas koinobionts delay development until the host is fully grown. During this suspension of growth, the parasitoid has to cope with attack by the host's immune system, and the need to evolve to be finely attuned to the host to counteract its defenses probably limits koinobiont host range.

A different top-down approach to parasitoid community ecology has recently been taken by a few groups, although it is still too soon to assess its value. There is a long tradition of constructing food webs in ecology, and one of the aims of this research program is to search for patterns that are common across different webs. A major limitation of this research is the heterogeneity in published food webs, which are typically collected in very different ways and which differ greatly in their taxonomic resolution. Host-parasitoid interactions have many advantages for food web studies, the most prominent of which is the relative ease with which trophic links can be identified and quantified. This has led to the construction of several quantitative food webs

in which all hosts, parasitoids, and links are expressed in the same units. The webs published to date illustrate the extent to which different hosts are linked by shared natural enemies and also the extent to which indirect effects such as apparent competition may act as forces structuring the community. Of course, a limitation of this approach is that only one guild of natural enemies is included in the web, but it will be interesting to determine whether common patterns emerge as more communities are studied.

## VII. THE IMPORTANCE OF PARASITOIDS

Why are parasitoids important to students of biodiversity? First, they are abundant members of nearly all terrestrial communities. The populations of few herbivorous insects are unaffected by parasitoid attack, and understanding how they influence and possibly regulate the densities of their host is immensely important in understanding the mechanisms underlying biodiversity. Much of the modern work in parasitoid ecology has been motivated by these considerations.

The second reason why parasitoids are important is also the reason why there is such an enormous and valuable literature on parasitoids dating back nearly a century: Parasitoids are our allies in the war against agricultural and forestry pests. There are two main strategies of parasitoid biological control: classical and inundative. In classical biological control, a parasitoid is released into an area to control a pest. Typically, the pest is a nonnative species that has been accidentally introduced into a new region, and normally the parasitoid is a species that attacks the pest in its region of origin. The parasitoid is released once or on a few occasions until it is established. Although only 10–20% of biological attempts are successful, they can have dramatic effects in controlling pests, and the savings in crop yield are normally many multiples of the investment in the biological control program. There are interesting patterns in the probability of success against different hosts, for example, biological control has an excellent record against scale insects and mealy bugs but is far less successful against aphids.

The second strategy is inundative release in which large numbers of parasitoids are released onto a crop to destroy a pest. Applications are made as often as required, and long-term establishment of the parasitoid (which is normally already there at low density) is not a goal. In effect, the parasitoid is acting here as a biological

insecticide. There are two main situations in which this strategy is used today. The first is when labor is relatively cheap so that the costs of mass rearing the parasitoids are not too great. In many tropical countries, huge numbers of egg parasitoids are mass reared and released to attack cutworms, armyworms, and bollworms. Advances in the mechanization of mass rearing are making inundative release more attractive in developed, high-wage countries. The second situation concerns high-value crops in enclosed growing conditions such as greenhouses. Control of tomato whitefly is typically by parasitoids in northern Europe. Inundative control has been favored in these situations because of the decreased tolerance of insecticide residues by regulatory authorities and because of the general need to reduce chemical input so as not to disrupt insect pollination and the existing successful biological control of other pests.

In the post-*Silent Spring* era biological control has often been viewed as an unmitigated good (particularly by biocontrol workers) because it does not involve pesticides. However, in recent years, increasing attention has been paid to the dangers of biological invasions, particularly by plants, and this has led to a reassessment of biological control and to worries regarding the effect of introduced parasitoids on native insects. These are important issues and need to be carefully scrutinized, although there is a real danger of an overreaction and a return to the automatic reliance on environmentally harmful chemical insecticides. Currently, the major problem is a lack of good data on the consequences of biological control introductions. The challenge to applied ecological entomologists is to develop a sufficient understanding of parasitoid population and community ecology building on the ideas described previously, so that they can provide rules and guidelines that will ensure that biological control has a net positive rather than negative effect on biodiversity.

### See Also the Following Articles

HYMENOPTERA • PARASITISM • SPECIES INTERACTIONS

### Bibliography

- Godfray, H. C. J. (1994). *Parasitoids, Behavioral and Evolutionary Ecology*. Princeton Univ. Press, Princeton, NJ.
- Hassell, M. P. (2000). *Insect Parasitoid Population Dynamics*. Oxford Univ. Press, Oxford.
- Hawkins, B. A. (1994). *Pattern & Process in Host-Parasitoid Interactions*. Cambridge Univ. Press, Cambridge, UK.
- Quicke, D. (1997). *Parasitic Wasps*. Wiley, London.
- Waage, J. K., and Greathead, D. (Eds.) (1986). *Insect Parasitoids*. Academic Press, London.







# PELAGIC ECOSYSTEMS

Andrea Belgrano,\* Sonia D. Batten,<sup>†</sup> and Philip C. Reid<sup>†</sup>

\*The Royal Swedish Academy of Sciences and <sup>†</sup>Sir Alister Hardy Foundation for Ocean Science

- I. Phytoplankton and Primary Production
  - II. Zooplankton
  - III. Pelagic Fish
  - IV. Outlook
- 

## GLOSSARY

**calanoid copepods** These belong to the Crustacea and are very abundant in the zooplankton. They play a major ecological role in the food web of the pelagic ecosystem as grazers of phytoplankton and as a food source for, e.g., larval and adult fish.

**clupeid fish** Pelagic fish, including the anchovy, sardine, and herring, which feed on plankton.

**El Niño Southern Oscillation (ENSO)** Regarded as quasi-periodic fluctuations occurring in the equatorial region of the Pacific Ocean. The ENSO is associated with changes in the sea surface temperature, sea surface levels, and rainfall patterns in the tropics and with a possible influence on the weather at higher latitudes. ENSO events have also been associated with changes in fisheries and ecosystem processes.

**gadoid fish** Includes the cod *Gadus morhua* L.; the life cycle of cod can be summarized as four stages: spawning, larvae, juveniles, and adults. During the larval and early juvenile phases they feed largely on copepods on the nursery grounds; after this phase they become largely stationary and feed on benthic or epibenthic animals.

**North Atlantic Oscillation (NAO)** The difference in atmospheric pressure between Ponta Delgada, Azores, and Stykkisholmur, Iceland. The variations in the NAO are usually associated with a positive or negative phase related to changes in the direction and strength of the westerly wind as well as in sea surface temperature. During a positive phase, winters over Scandinavia are warmer and vice versa. Recently, the fluctuations observed in the NAO have been related to changes in ecosystem processes.

**phytoplankton** Unicellular photoautotrophic plant cells ranging in size from 1  $\mu\text{m}$  to 1 mm. They are divided into 12 taxonomically defined divisions, including diatoms, dinoflagellates, and coccolithophorids.

**primary production** An estimate of the rate of carbon fixation by phytoplankton; this can account for up to 75% of the photosynthetic production on Earth.

**Russell cycle** Named after F. Russell, the Russell cycle describes a major change in the biota of the English Channel from 1925 to 1972.

**zooplankton** From the Greek *zoon* meaning animal, zooplankton comprises those animals that are found passively drifting or weakly swimming in the water column. Zooplankton can be divided into two major categories: holoplankton, which are organisms that spend their entire lives as plankton, and meroplankton, which are organisms that spend part of their life cycle as plankton and part on the seafloor as benthic invertebrate larvae or as nekton (e.g., fish larvae).

*PELAGIC ECOSYSTEMS* cover more than 70% of the surface of the earth. Their species diversity and richness are related to physicochemical and biological processes acting at a range of temporal and spatial scales. They are strongly influenced by atmosphere–ocean (coupling) interactions related to hydrodynamic processes. Climatic oscillations, such as the El Niño Southern Oscillation and the North Atlantic Oscillation, may be associated with changes in ecosystems on a decadal scale. Pelagic biodiversity is part of the ocean's complex, adaptive ecosystem, and the interactions between species diversity and ecological processes such as food web dynamics occur at a variety of temporal and spatial scales. The study of changes in biodiversity patterns in relation to the natural and anthropogenic (e.g., overfishing) variability in pelagic ecosystem structure and function need to be considered for a sustainable global ecosystems policy for the next millennium.

## I. PHYTOPLANKTON AND PRIMARY PRODUCTION

The etymology of the word “plankton” derives from the ancient Greek meaning wandering. Plankton have very limited motility and are dependent on currents and the physical environment for their location. Phytoplankton comprises a very diverse group of single-celled photoautotrophic planktonic plants that are divided into 12 taxonomic divisions (3500–4500 species of oceanic plankton) and live in the surface water of the oceans. The role played by phytoplankton is at the base of the biogeochemical cycles in the ocean. Species range in size from small prokaryotic and eukaryotic cells comparable to bacteria to larger organisms such as diatoms, dinoflagellates, and coccolithophorids (Figs. 1 and 2). A typical phytoplankton community contains a mixture of these, although the species which are dominant vary through time. The majority of the species present in the ocean gain their energy via photosynthesis and, therefore, environmental factors such as light availability, temperature, and the supply of major nutrients in the form of ions are important factors that influence their distribution on both temporal and spatial scales. Several hypotheses have been proposed which explain the coexistence of several species; one, “temporal succession,” supposes that the nutrient regime of the environment changes rapidly so that all species are in different states of approaching to or declining from their maximum abundance. In low-turbulence environments microhabitats may develop which favor the growth of

a distinct patch of plankton. Alternatively, different species may be limited in growth by the availability of different nutrients so that coexistence is possible. In neritic temperate waters of the North Sea and in the open temperate North Atlantic, the mean seasonal pattern of phytoplankton abundance shows a distinct peak in the spring, followed by a summer decline and a second, lesser, peak in the autumn. Work by Colebrook (1986) showed a clear differentiation between those species associated with the spring bloom, which tended to be diatoms, and those associated with the autumn bloom, which tended to be dinoflagellates. There were changes in the North Sea in terms of phytoplankton species composition after 1987, for example, *Ceratium* spp. (Dickson *et al.*, 1992) and *Thalassiothrix longissima* (Reid *et al.*, 1992), which signaled a change in the ecosystem. These changes have been related to an increase in the sea surface temperature (SST) and to changes in the extension of the Baltic/Norwegian waters in the North Sea and an increased input of more oceanic water. Variations in the seasonal cycle are caused by different preferences of the groups for light intensity, water stability, temperature, and nutrient availability. A bimodal pattern of the seasonal cycle is evident across the temperate North Atlantic but other patterns exist in different geographical regions. Polar regions show a single summer peak because light intensity is only sufficient for a short period of phytoplankton growth. The north Pacific Ocean shows no spring peak, and in fact there is little change in the phytoplankton standing stock throughout the year. The principal reason is that the dominant zooplankton (the copepod genus *Neocalanus*) overwinter as adults which produce young in the spring without feeding. The young are then able to immediately take advantage of any increase in the phytoplankton and effectively prevent a spring bloom. The tropics show little seasonality and both phytoplankton and zooplankton fluctuate slightly throughout the year, dependent on local processes. Large blooms of phytoplankton can be monitored via satellite remote sensing (e.g., the coccolithophorid *Emiliana huxleyi*). They have the potential to modify, to some extent, the carbon exchange taking place between the ocean and the atmosphere. Figure 3 shows an example of a cyanobacterium, *Synechococcus* sp., which is an extremely small photoautotrophic cell 1–3  $\mu\text{m}$  in length. Picocyanobacteria and nanoflagellates are responsible for assimilating excretory products of zooplankton consumers in the euphotic zone in nutrient-poor waters characteristic of the central ocean gyres. As noted by Falkowski *et al.* (1998), the fixation of carbon by phytoplankton results in approximately 45 gigatons of organic carbon per year, of

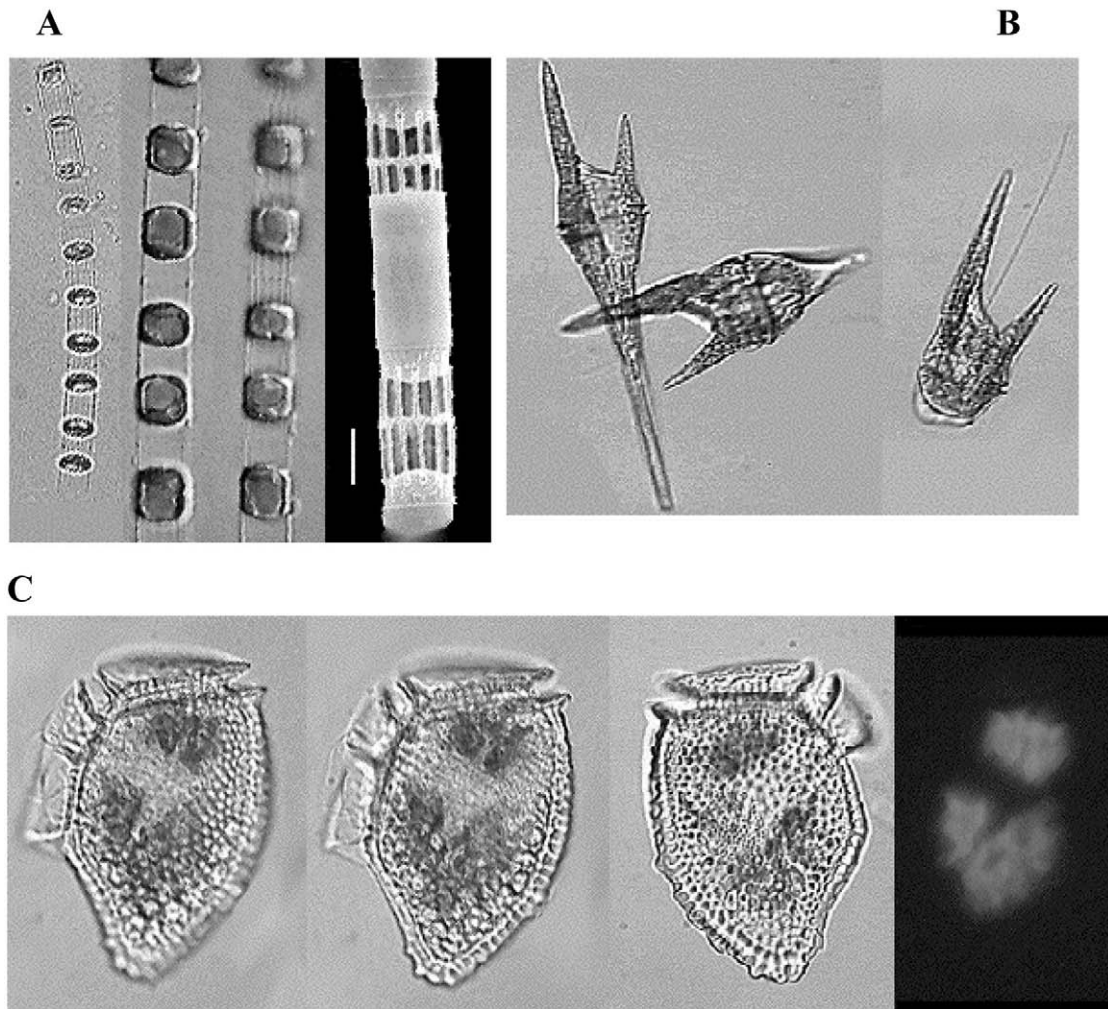


FIGURE 1 Phytoplankton species illustrating the taxonomic diversity as well as the shape and size of these organisms, which are an important component of the pelagic food web. (A) Diatomophyceae, *Skeletonema costatum*, a worldwide diatom species with distinctive features such as long chains of small cells with long external tubes. Diameter varies between 2 and 21  $\mu\text{m}$ . (B) Dinophyceae, *Ceratium furca*, a worldwide dinoflagellate with a solitary or paired life-form, with a straight body, and the epitheca gradually tapers into an anterior horn. The size ranges in length between 70 and 200  $\mu\text{m}$  and in width between 30 and 50  $\mu\text{m}$ . (C) Dinophyceae, *Dinophysis norvegica*, a solitary dinoflagellate, normally found in cold waters, varies in size with length between 48 and 67  $\mu\text{m}$  and width between 39 and 53  $\mu\text{m}$ . *Dinophysis norvegica* can be regarded as a toxic species since it produces a toxin that causes diarrhetic shellfish poisoning and therefore it is a potentially dangerous species when found in coastal waters in large numbers (photos courtesy of Mats Kuylenstierna and Bengt Karlson). See also color insert, Volume 1.

which approximately 16 gigatons is exported out of the surface waters. Changes in the total and export production can affect the marine food web structure, ultimately leading to changes in the fish stocks. The photosynthetic pigments such as chlorophyll *a* can be used as a proxy measurement of phytoplankton biomass, and with the aid of satellite-derived measurements a color index can be derived to illustrate changes in primary production. Behrenfeld and Falkowski

(1997) generated several maps of primary production estimates for the world's ocean (Figs. 4 and 5). Measurements of phytoplankton color on samples taken by the Continuous plankton recorder (CPR) survey from the central northeast Atlantic and North Sea (Fig. 6) shows a similar increasing trend and longer growing season as that of the satellite-derived vegetation index and this trend dates back to 1948. The phytoplankton increase is not ubiquitous because an inverse trend is evident

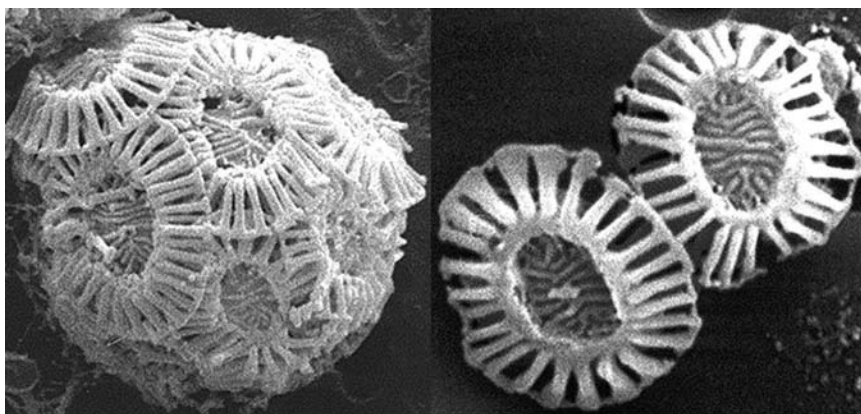


FIGURE 2 A (SEM) photo of a Haptophyta coccolithophorid, *Emiliana huxleyi*, found worldwide the diameter varies between 5 and 10  $\mu\text{m}$ , and the coccoliths are  $3.5 \times 3 \mu\text{m}$  (photos courtesy of Mats Kuylenstierna and Bengt Karlson). See also color insert, Volume 1.

in eastern Atlantic waters between latitudes 59°N and 63°N. This index of phytoplankton change may reflect a possible long-term biological response in the North Atlantic to increasing global temperatures which are causing melting of ice, sea ice, and permafrost in the Arctic (Reid *et al.*, 1998).

In the central North Pacific environment, Venrick (1990) observed 245 species of phytoplankton in the

shallow waters and 231 in the deep zone during the period between 1972 and 1985. In both areas, 21 species accounted for 90% of the individuals and most of the variance in abundance occurred at small spatiotemporal scales and samples within a zone were similar. This study showed that epipelagic populations of the North Pacific central environment were more stable when compared with planktonic populations from other ecosystems. In relation to the pelagic food web structure, phytoplankton is an important food source for the larval stage of pelagic fish and thus plays an important role in the trophic structure of pelagic communities.

## II. ZOOPLANKTON

Zooplankton are the secondary producers in pelagic ecosystems and comprise an extraordinarily wide range of organisms. The zooplankton community of continental shelf waters, for example, may contain larval stages of littoral and benthic invertebrates (meroplankton) in addition to the species that spend all their lives in the plankton (holoplankton). Phytoplankton production in these nutrient-rich waters is generally regulated by the grazing activity of zooplankton (Fig. 7). Most zooplankton belong to the crustacean, with the principal organism group being copepods. Two copepod species, *Calanus finmarchicus* and *C. helgolandicus*, constitute the major components of the northeast Atlantic and North Sea zooplankton in terms of biomass, abundance, and trophic role (Marshall and Orr, 1972). These popula-

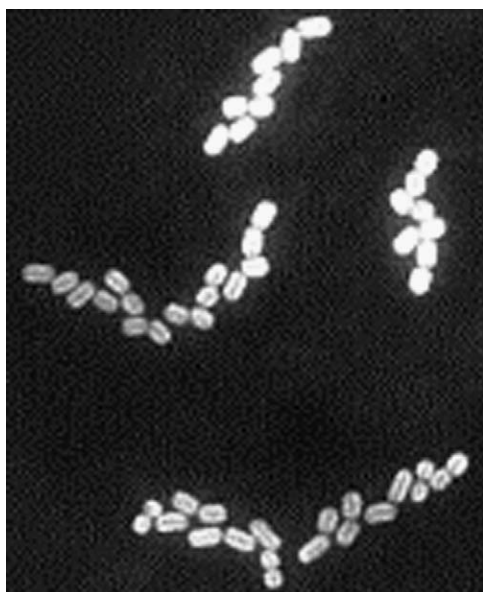


FIGURE 3 An example of a cyanobacteria species, *Synechococcus*, found in the central Skagerrak; the size range is between 1 and 3  $\mu\text{m}$  in length (photos courtesy of Mats Kuylenstierna and Bengt Karlson).

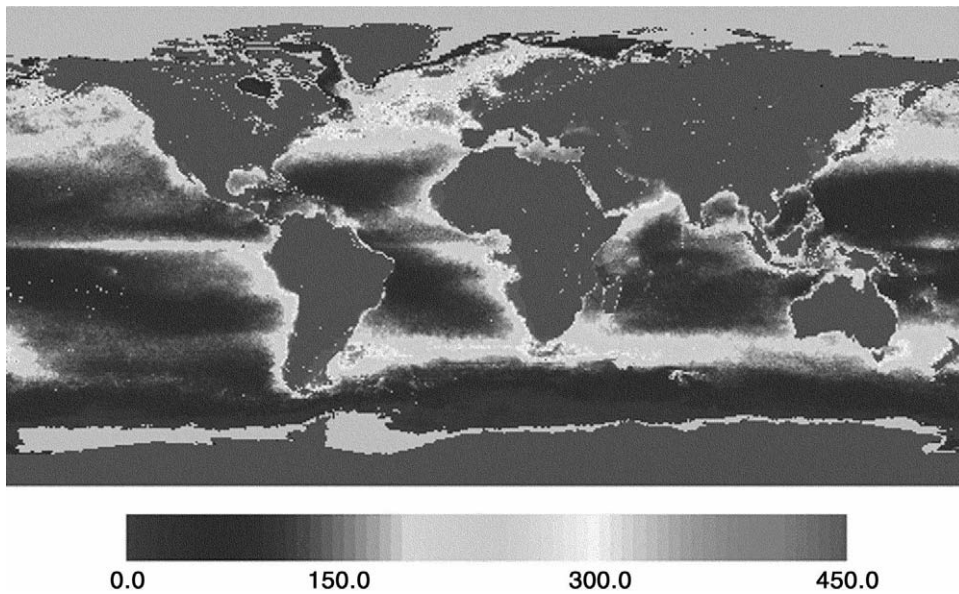


FIGURE 4 A map of annual primary production (PP) generated for the standard parameterization of the VGPM, where the  $P_{opt}^B$  is modeled according to Behrenfeld and Falkowski (1997) and the surface irradiance is corrected for cloudiness (map courtesy of Paul G. Falkowski).

tions follow distinct spatial and seasonal dynamics such that *C. finmarchicus* is a northern spring species, whereas *C. helgolandicus* is located in warm-temperate waters and reaches its maximum abundance in autumn.

The distribution of these species of *Calanus* has been studied during the period 1958–1995 in the North Atlantic (Figs. 8–10) in relation to the year-to-year variability in the North Atlantic Oscillation (NAO). The

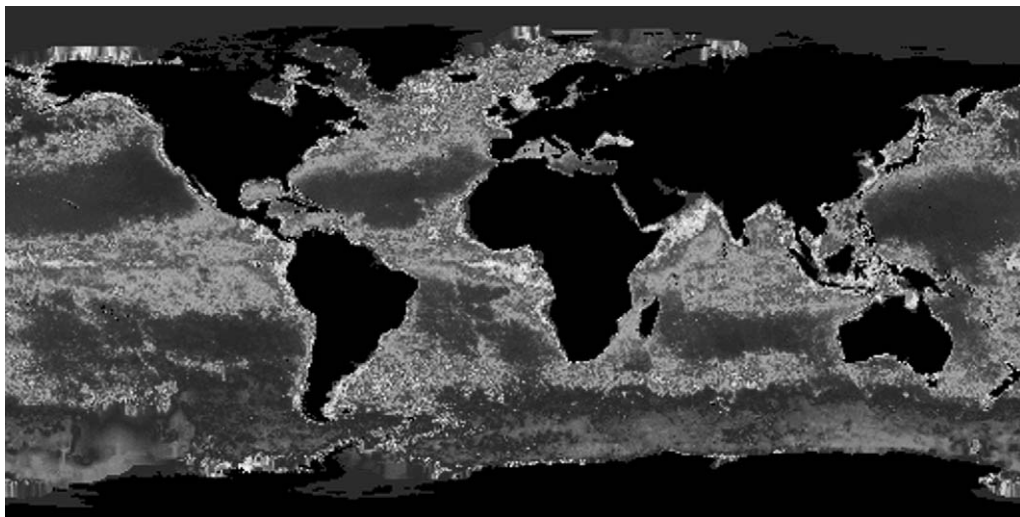


FIGURE 5 A map of primary production generated when  $P_{opt}^B$  is estimated using the relationship between temperature and the maximum phytoplankton specific growth rate and modified for photosynthesis (map courtesy of Paul G. Falkowski).

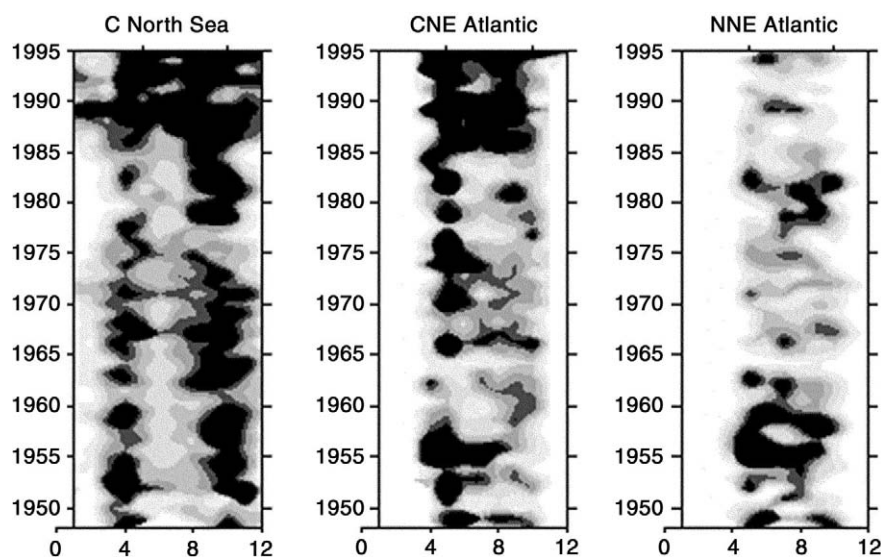


FIGURE 6 Phytoplankton color distribution from the CPR survey from 1948 to 1995 showing an increase in the late 1980s in the central part of the North Sea and in the central northeast Atlantic coinciding with a change in the NAO index from a negative to a positive phase. A reverse trend was observed for the north-northeast Atlantic (redrawn from Reid *et al.*, 1998, courtesy of SAHFOS, with permission from *Nature* 391, 546 (1998).

NAO is a climatic oscillation affecting the hydrology and climate in the North Atlantic Ocean and adjacent region, and it can be compared to El Niño in the Pacific Ocean. After 1987, the zooplankton community in the North Sea changed with, for example, an increase in the abundance of the copepod *Coryceus* spp. and it may have also affected the distribution and abundance of the two *Calanus* species. The increase in temperature,

wind, and alteration of the winter circulation pattern observed during years of a high positive NAO index have resulted in unfavorable conditions for the *C. finmarchicus* population leading to a significant decrease in the abundance of the species. Conversely, these hydroclimatic shifts have proved beneficial to *C. helgolandicus*, the abundance of which has increased during these years (Planque and Taylor, 1998; Stephens *et al.*, 1998).

In the northeast Pacific the plankton community structure (Fig. 11) responded to the warming in 1958 to 1960 in the California Current in relation to an El Niño event that caused a sudden change in the SST distribution (McGowan *et al.*, 1998). The zooplankton abundance has generally decreased in the California Current system while, in a synchronous manner, the zooplankton abundance in the Gulf of Alaska gyre increased. These interdecadal regime shifts in these two systems driven by climatic variations in the atmosphere are reflected first in changes in the physical structure of the ocean resulting in large-scale biological responses in the ocean. These regime shifts have a severe effect on the temporal and spatial distribution of planktonic species, leading to changes in the secondary production as well as in community structure (McGowan *et al.*, 1998). In terms of species richness and diversity in the open area of the Atlantic Ocean, the maximum species richness is found at depths between approximately 1000



FIGURE 7 An example of copepod species found in the zooplankton; two *Acartia tonsa* copepods with ingested food item *Thalassiosira weissflogii* (photo courtesy of Kajsa Tönnesson).

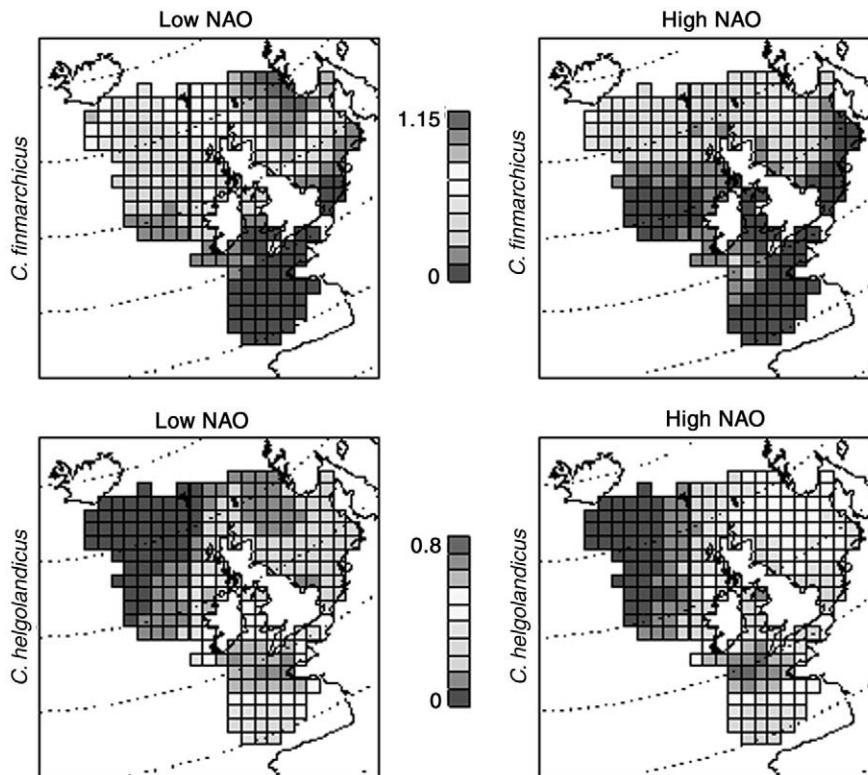


FIGURE 8 The mean spatial distribution of the abundance of two species of calanoid copepod *Calanus finmarchicus* and *Calanus helgolandicus* in the northeast Atlantic and the North Sea, in relation to years of both low and high NAO. The color scale is proportional to the log abundances of the species. High NAO years, 1983 and 1989–1992; low NAO years, 1963, 1964, 1966, 1977, and 1979 (maps from Planque, 1997, courtesy of SAFHOS).

and 1500 m. Angel (1991) showed, for example, that the species richness of planktonic ostracods in the northeastern Atlantic at 20°W occurred between 500 and 1000 m. Moving up the water column, and at higher latitudes, there tend to be fewer species but these occur at higher abundances. The zooplankton, and especially the *Calanus* species, are an important part of the pelagic food web since they are the major source of food for fish species including cod. The nauplii of *Calanus* in particular are a key component of the diet of fish larvae and a recent model of *Calanus* off the Georges Bank (Miller *et al.*, 1998) showed clearly the importance of the availability of *C. finmarchicus* for the cod and had-dock populations along the bank (Lough, 1984).

Recent evidence shows that SST explains almost 90% of the geographic variation in planktonic foraminiferal diversity for the whole Atlantic Ocean (Rutherford *et al.*, 1999). The peak in foraminiferal diversity was found at middle latitudes in the Atlantic, and this pattern may be extended to other taxa such as euphausiids (krill

species), pteropods, and chaetognaths in the North Pacific. This is additional evidence that changes in pelagic diversity in the world's ocean result from changes in the upper-ocean physical structure and especially that water column and ecosystem structure may depend on vertical niche separation.

In the North Sea, particularly the northern areas, there has been an increase in zooplanktonic species richness, as recorded by the CPR survey, since the 1950s. The increase, particularly of calanoid copepods, in the most northerly areas can be attributed to increased inflow from the Atlantic Ocean, but currently the species which are increasing or adding to the numbers are not permanent components of the North Sea plankton. Resident and colder water holoplanktonic species have declined in abundance, and meroplankton (adult and larval) and expatriates from warmer oceanic and mixed waters have increased. This constitutes a significant challenge to analysis of pelagic diversity because the species richness of the meroplanktonic groups



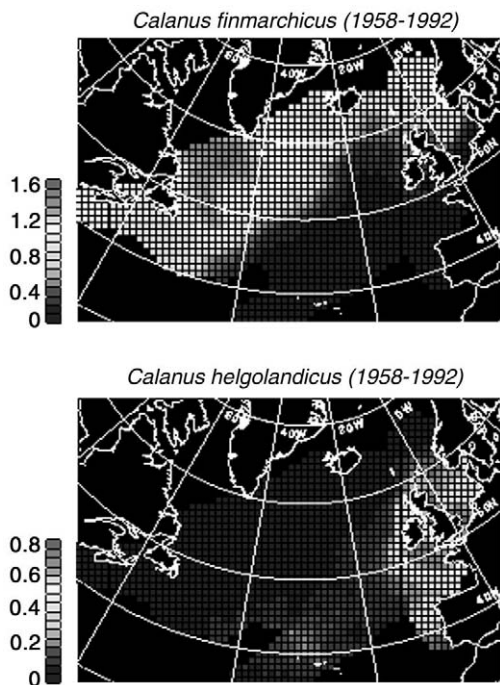


FIGURE 9 Regional distribution of *Calanus finmarchicus* and *Calanus helgolandicus* for the years 1958–1992 (maps from Planque, 1997, courtesy of SAFHOS).

may exceed that of the holoplanktonic groups. For example, in a region of the northwest North Sea, 17 calanoid copepod species were recorded in 10 years of CPR samples, but 24 species of decapod larvae were recorded in 2 years and 34 species of bivalve larvae in just 1 year. Meroplankton species, larval and adult, are linked to the more permanent habitats of the benthic phases of their life cycle, whereas holoplankton live entirely within a dynamic water mass, complicating the description of plankton diversity at a specific geographical location.

### III. PELAGIC FISH

In the marine environment, approximately 80% of the known fish species are found in coastal areas, whereas approximately 2% are epipelagic and the remaining are deep-sea species. For example, at a 1000-m depth in the North Atlantic 44°N 13°W, the fish were the most taxonomically diverse group found in the micronekton (Angel, 1993). In the pelagic food web, pelagic planktivorous fish are regarded as predator species and changes in their prey items in relation to physical changes in the ocean circulation pattern are reflected in fish stock

fluctuations (Fig. 12) and changes in the catches of one species of sardine, *Sardinops sagax* (Kawasaki, 1991; Sharp and McLain, 1993; Strömberg, 1997). The increase in the sardine populations may reflect changes in phytoplankton biomass in the Pacific Ocean during the 1970s and 1980s (Venrick *et al.*, 1987) as well as changes in zooplankton abundances in the northeast Pacific. Zooplankton populations off the Peruvian coast declined from 1972 coincident with the severe decline in the anchoveta fishery. Shifts in the Kuroshio current southeast of Japan, and shifts in the wind stress, may be related to the fluctuations in sardine populations. The effects of El Niño Southern Oscillation events on the fluctuations observed in the fish stocks are unclear as, for example, in the case of the Chilean sardine and the Peruvian anchovetta that do not indicate a clear connection. During the El Niño events in the late 1950s, 1960s, and early 1970s, the Peruvian anchovetta stock was high, whereas the Chilean sardine stock was low. In 1971 and 1972, the El Niño event decreased dramatically and as a result the Chilean sardine stock increased while the Peruvian anchovetta stock crashed. At the same time, both the Japanese sardine and the Californian sardine increased. This trend was confirmed dur-

Regional variations in the seasonal cycle of *C. finmarchicus*

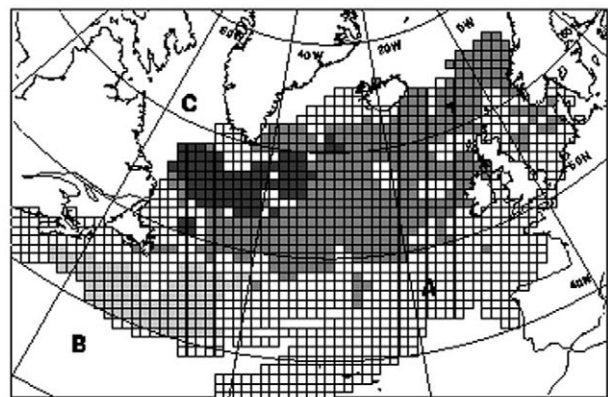


FIGURE 10 Regional variation in the seasonal cycle of *Calanus finmarchicus* showing how the distribution varies in the North Atlantic and how the peak of the log abundance differs between different locations (maps from Planque, 1997, courtesy of SAFHOS).

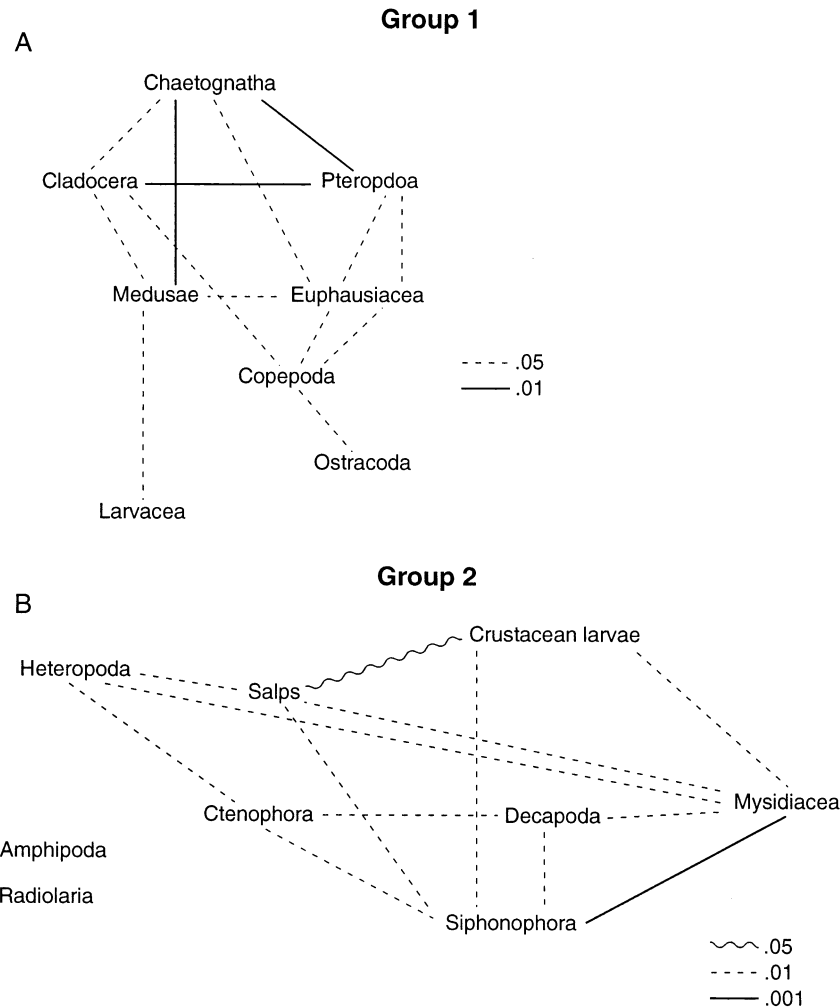


FIGURE 11 An example of changes in a pelagic food web in relation to a large El Niño event in California. (a) Correlations of abundance for zooplankton taxa between the years 1955 and 1959. (b) Similarities found using rank-order abundance of plankton. The years 1955–1957 showed no correlation with years 1958 and 1959. Note that 0.05, 0.01, 0.001 and the respective thickness of the line indicate the significance level of the correlations (redrawn from McGowan *et al.*, 1998).

ing the El Niño events that took place in 1982 and 1983. In the Black Sea the introduction of a new species, the comb jellyfish *Mnemiopsis leidy*, caused a strong decline in the Black Sea anchovy stock. As the abundance of the jellyfish increased, it had a great impact on the recruitment of the anchovy due to changes in the food web and competition for food. These population collapses are very important since pelagic fish such as sardine and anchovy species are among the top 10 species that dominate the world catches.

If we examine the long-term studies of plankton in the North Atlantic observed by the CPR, we find that in

the North Sea after 1987 an increase in phytoplankton biomass coincided with taxonomic changes in the phytoplankton and zooplankton species and a large increase in the catches of the western stock of the horse mackerel *Trachurus trachurus L.* in the northern North Sea. This reflected a northerly expansion of the stocks along the shelf edge from the Bay of Biscay to the North Sea after 1987. The increase in the western horse mackerel may be related to the fact that this species tracked down an increase in food resources available in the North Sea during high NAO years. The presence after 1988 of an unusual plankton species (for the North Sea), *Eucheta*

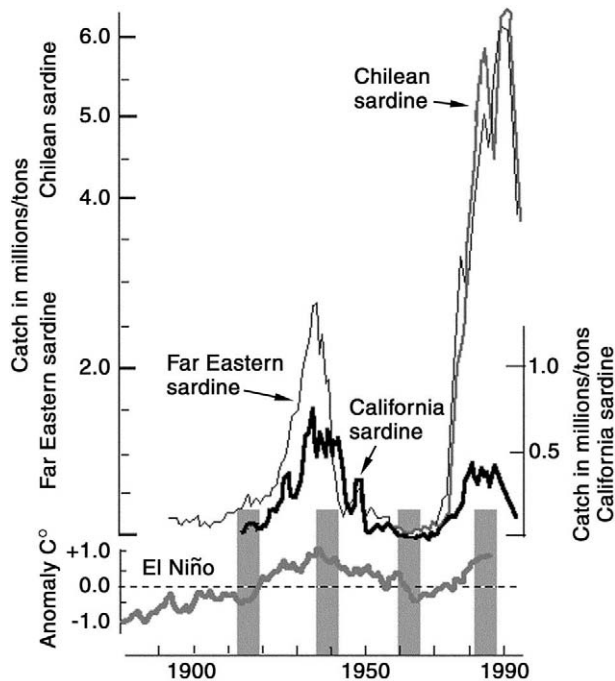


FIGURE 12 The covariation of three different populations of Pacific sardine compared with the temperature anomaly from the North Pacific in relation to ENSO events. The data for the California sardine are part of the California Cooperative Fisheries Investigations (CALCOFI) [data from Kawasaki (1991) and Sharp and McLain (1993); redrawn from Strömberg (1997)].

*hebes*, which normally has a more southerly/oceanic distribution, suggests an increase in the shelf edge current that in turn favored the advection of western horse mackerel from the western margin of Europe, allowing them to reach the northern North Sea. As noted by Strömberg (1997), the causes and events related to the fluctuations in the standing stocks of pelagic fish in relation to changes in the plankton and to production are very complex, and these changes may be related to wind-driven advection processes and to a change in the behavior of predator and prey in response to changes in the small turbulent flow. In the coastal areas, and in particular in the shelf seas, very abrupt changes in the community structure can occur and persist on a decadal scale (Steele, 1998). The observed changes in the dynamics of pelagic fish stocks may be related to low- and high-frequency climatic variability that in turn affects the interannual variation in recruitment, growth, and abundance. For example, in a time series analysis of variation in O-group cod *Gadus morhua* along the Norwegian Skagerrak coast, variation was found to not be directly associated with the NAO and with the abun-

dances of *C. finmarchicus*, but the fluctuations in the cod stock may have been related to changes in the bottom habitat and especially in the seagrass coverage by *Zostera marina* (Fromentin *et al.*, 1998). Further analysis of this data set demonstrated the importance of age-structured interactions such as asymmetric competition and cannibalism between cohorts, causing fluctuations in the abundance of the O-group and 1-group juvenile cohorts of cod (Bjørnstad *et al.*, 1999).

The northern stock of cod on the Labrador Shelf and Grand Banks of the northwest Atlantic declined and, although overfishing occurred, other factors also affected stock recruitment. The match/mismatch hypothesis was revisited by Conover *et al.* (1995). This hypothesis basically states that the spawning of cod occurs at the same time each year and if the spawning of copepods (the major food source) occurs at the same time there is a "match" and a successful recruitment. If the spawning of copepods is delayed through dependence on the occurrence of the phytoplankton spring bloom there is a "mismatch" and the recruitment of cod larvae will be low. Changes in the physical structure in the Labrador Sea and Grand Banks occurred with unusually low temperatures in the Labrador current and this was followed by the appearance in the zooplankton of the copepod *Calanus glacialis* rather than *C. finmarchicus*. *Calanus glacialis* spawns later than *C. finmarchicus* and therefore does not provide a good food source for the cod larvae. Changes in the Labrador current may be the result of climate change and in turn this affects the structure of the pelagic food web by creating a shift in the species composition as just described, thus resulting in a change in cod recruitment. Ultimately, the increase in fishing activities may have a severe effect on fish populations (Strömberg, 1997) since indirect effects considerably alter the biogenic habitat for fish recruitment, as does the removal of top predators. These indirect interactions by the fisheries can cause a trophic change in the food web as a result of top-down control of community structure, whereas a change in the resource availability induced, for example, by the El Niño event represents a bottom-up process that influences the dynamics of pelagic ecosystems. The interplay between these two types of perturbation may be the cause of shifts in species composition which result in changes in the food web structure. In light of these processes, we should also consider the importance of the microbial loop (Azam, 1998) and be aware of bacterial control of the fluxes of organic matter and its role in ecosystem dynamics. For example, fish production in the eastern Mediterranean decreased through a dominant microbial loop, and the uncoupling of bacteria from primary productivity

was found in a fishery off the coastal region of Newfoundland that was causally related to the intensity of the fishery.

#### IV. OUTLOOK

The structure of pelagic ecosystems varies at both temporal and spatial scales. At an ocean basin spatial scale, phenomena such as El Niño can be linked at a temporal scale of approximately 1 year to the stock recruitment of pelagic fish, and species and single species models can be used as predictive tools. At a larger spatial scale and on a decadal timescale, climatic oscillations such as the NAO can be linked to regime shifts in the species composition and community structure of ecosystems, and models should be at a community or ecosystem level. A retrospective analysis of historical data could reveal important biodiversity patterns.

The Russell cycle (Russell, 1973) explained the changes in the biota of the English Channel very well. In 1925, a decline in the herring stock was observed and the Plymouth herring stock (on the southwest coast of the United Kingdom) collapsed in 1936. In 1931, the decline in the macroplankton resulted in a change in the presence of the arrowworm from the species *Sagitta elegans* to *S. setosa*. In the following decade, there was a drastic decrease in the standing stocks of gadoid fish such as cod (*G. morhua*). The English Channel ecosystem essentially shifted from one dominated by herring and that had a good supply of dissolved nutrients to one with less nutrients and that was dominated by pilchards and smaller planktonic organisms. It is interesting that in 1965 there was an increase in the fish larvae concomitant with an increase in nutrient supply. These events were followed in the 1970s by a reverse switch in the *Sagitta* species and an increase in copepod abundance, whereas the number of pilchard eggs decreased. The Russell cycle is a well-documented example that shows how ecosystems can shift or oscillate from one state to another after reaching a bifurcation point (May and Oster, 1976).

Pelagic ecosystems seem to show decadal or longer oscillations and this is reflected in the fossil records and paleoceanographic studies. Pelagic ecosystem diversity and the maintenance of biodiversity appear to be influenced by climatic oscillations and through pressure exerted by overfishing and removal of top predators from the food web. These can be regarded as macroscopic system properties that in turn are related to different flux that may have an effect on the structure and ecosystem functioning (Levin, 1998). An under-

standing of changes in biodiversity patterns in relation to the variability in ecosystem structure and function needs to be incorporated in a sustainable global ecosystem policy for the next millennium.

#### See Also the Following Articles

FISH, BIODIVERSITY OF • FISH STOCKS • FOOD WEBS • OCEAN ECOSYSTEMS

#### Bibliography

- Angel, M. V. (1991). Variations in time and space: Is biogeography relevant to studies of long-time scale change? *J. Mar. Biol. Assoc. UK* 71, 191–206.
- Angel, M. V. (1993). Biodiversity of the pelagic ocean. *Consev. Biol.* 7, 760–772.
- Azam, F. (1998). Microbial control of oceanic carbon flux: The plot thickens. *Science* 280, 694–696.
- Behrenfeld, M. J., and Falkowski, P. G. (1997). Photosynthetic rates derived from satellite-based chlorophyll concentrations. *Limnol. Oceanogr.* 42, 1–20.
- Bjørnstad, O. N., Fromentin, J.-M., Stenseth, N. C., and Gjøsæter, J. (1999). Cycles and trends in cod populations. *Proc. Natl. Acad. Sci. USA* 96, 5066–5071.
- Colebrook, J. M. (1986). Environmental influences on long-term variability in marine plankton. *Hydrobiologia* 142, 309–325.
- Conover, R. J., Wilson, S., Harding, C. H., and Vass, W. P. (1995). Climate, copepods and cod: Some thoughts on the long-range prospects for a sustainable northern cod fishery. *Climate Res.* 5, 69–82.
- Dickson, R. R., Colebrook, J. M., and Svendsen, E. (1992). Recent changes in the summer plankton of the North Sea. *ICES Mar. Sci. Symp.* 195, 232–242.
- Falkowski, P. G., Barber, R. T., and Smetacek, V. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science* 281, 200–206.
- Fromentin, J.-M., Stenseth, N. C., Gjøsæter, J., Johannessen, T., and Planque, B. (1998). Long-term fluctuations in cod and pollock along the Norwegian Skegerrak coast. *Mar. Ecol. Prog. Ser.* 162, 265–278.
- Kawasaki, T. (1991). Long-term variability in the pelagic fish populations. In *Long-Term Variability of Pelagic Fish Population and Their Environment* (T. Kawasaki, S. Tanaka, Y. Toba, and A. Taniguchi, Eds.). Pergamon, New York.
- Levin, S. (1998). Ecosystems and the biosphere as complex adaptive systems. *Ecosystems* 1, 431–436.
- Lough, R. G. (1984). Larval fish trophodynamic studies on Georges Bank: Sampling strategy and initial results. In *The propagation of Cod, Gadus morhua* (E. Dahl, S. Danielssen, E. Moksness, and P. Solemdal, Eds.). Flodevigen Rapp.
- Marshall, S. M., and Orr, A. P. (1972). *The Biology of a Marine Copepod*. Springer-Verlag, New York.
- May, R. M., and Oster, G. F. (1976). Bifurcations and dynamic complexity in simple ecological models. *Am. Nat.* 110, 573–599.
- McGowan, J. A., Cayan, D. R., and Dorman, L. (1998). Climate–ocean variability and ecosystem response in the northeast Pacific. *Science* 281, 210–217.

- Miller, C. B., Lynch, D. R., Carlotti, F., Gentleman, W., and Lewis, C. V. W. (1998). Coupling of an individual-based population dynamic model of *Calanus finmarchicus* to a circulation model for the Georges Bank region. *Fish. Oceanogr.* 7(3/4), 219–234.
- Planque, B. (1997). Spatial and temporal fluctuations in *Calanus* populations sampled by the Continuous Plankton Recorder. PhD thesis, University of Pierre et Marie Curie Paris VI, France.
- Planque, B., and Taylor, A. H. (1998). Long-term changes in plankton and the climate of the North Atlantic. *ICES J. Mar. Sci.* 78, 1015–1018.
- Reid, P. C., Surey-Gent, S. C., Hunt, H. G., and Durrant, A. E. (1992). *Thalassiothrix longissima*, a possible oceanic indicator species in the North Sea. *ICES J. Mar. Sci.* 195, 268–277.
- Reid, P. C., Edwards, M., Hunt, H. G., and Warner, A. J. (1998). Phytoplankton change in the North Atlantic. *Nature* 391, 546.
- Russell, F. S. (1973). A summary of the observations of the occurrence of planktonic stages of fish off Plymouth 1924–72. *J. Mar. Biol. Assoc. UK* 53, 347–355.
- Rutherford, S., D'Hondt, S., and Prell, W. (1999). Environmental controls on the geographic distribution of zooplankton diversity. *Nature* 400, 749–753.
- SAHFOS <http://www.npm.ac.uk/sahfos/sahfos.html>
- Sharp, G. D., and McLain, D. R. (1993). Fisheries, El Niño-Southern Oscillation and upper-ocean temperature records: An eastern Pacific example. *Oceanography* 6, 13–22.
- Steele, J. H. (1998). From carbon flux to regime shift. *Fish. Oceanogr.* 7(3/4), 176–181.
- Stephens, J. A., Jordan, M., Taylor, A., and Proctor, R. (1998). The effects of fluctuations in North Sea flows on zooplankton biomass. *J. Plankton Res.* 20, 943–956.
- Strömberg, J.-O. (1997). Human influence or natural perturbation in oceanic and coastal waters—Can we distinguish between them? *Hydrobiologia* 352, 181–193.
- Venrick, E. L. (1990). Phytoplankton in an oligotrophic ocean: Species structure and interannual variability. *Ecology* 71(4), 1547–1563.
- Venrick, E. L., McGowan, J. A., Cayan, D. R., and Hayward, T. L. (1987). Climate and chlorophyll-a: Long-term trends in the central North Pacific Ocean. *Science* 238, 70–72.



# PESTICIDES, USES AND EFFECTS OF

Paul C. Jepson  
*Oregon State University*

---

- I. Classification and Uses
  - II. Efficiency
  - III. Ecotoxicology and Management
- 

## GLOSSARY

**pesticide** A chemical substance used for controlling, preventing, destroying, or mitigating a pest organism.

---

*PESTICIDES HAVE BEEN IN RECORDED USE* since 1000 B.C. Arsenic was in regular use as a garden insecticide in China by A.D. 900 and chemicals of one form or another have protected humans and their crops and livestock throughout the development of modern civilization. In comparison with this long time scale, we are still in the earliest phases of the use of synthetic organic pesticides, which were first used over large areas in the 1940s. There has, nonetheless, been sufficient time for several generations of pesticide chemistry to evolve, and pesticides have influenced all habitats and have affected the lives of all their inhabitants over this period.

## I. CLASSIFICATION AND USES

Nowhere have the costs and benefits of modern technology been more difficult to reconcile than with pesti-

cides. Benefits that include improved yield, crop quality, and food safety and reductions in vector-borne disease incidence have driven, and will continue to drive, their use. The earlier synthetic organic compounds were, however, flawed in their environmental behavior. They were persistent, they had a very broad spectrum of toxicological activity, and they displayed a tendency to be magnified in concentration through food chains, such that damage was inflicted to animal populations that lived beyond the treated area in habitats that were not intentionally contaminated. The discovery of some of these limiting impacts was only made possible through technical advances in, for example, analytical chemistry and conceptual advances where, for example, the ability to predict environmental behavior from chemical properties only developed after pesticides had been in use for decades. New pesticide discoveries are no longer accompanied by the marvel and optimism that characterized the first synthetic insecticides. Our ability to exploit these materials has, however, advanced considerably in recent years, and the chemicals themselves are increasingly specific in their impacts and appear to pose reduced risks. Scientists still question the value of reliance upon chemical pesticides, however, and modern pest control is characterized in general by a cautious approach to their management and use.

Set in a volume that will be consulted largely by biologists and ecologists with an interest in biodiversity, this article summarizes the chemicals that are in most widespread use and outlines the processes that contribute most to efficient delivery to the biological target. The

TABLE I

## Classification of Pesticides by the Type of Pest Controlled

Acaricide	Mites, ticks, and spiders
Adulticide	Adult insects
Algicide	Algae
Arboricide	Trees, brush, and scrub
Avicide	Birds
Bactericide	Bacteria
Fungicide	Fungi
Herbicide	Plants
Insecticide	Insect and sometimes related arthropod pests, including mites
Ixodicide	Ticks
Larvicide	Insect larvae
Miticide	Mites, ticks, and spiders
Molluscicide	Mollusks, such as slugs and snails
Nematicide	Nematodes
Ovicide	Invertebrate eggs
Piscicide	Fish
Predacide	Vertebrate Predators
Rodenticide	Rodents, including rats and mice
Silvicide	Trees, brush, and scrub
Termiticide	Ants and termites

nature and importance of toxicological and ecological effects are then reviewed, followed by a summary of procedures through which pesticides are regulated and managed. This article does not provide detailed reviews of biochemical mode of action, application, formulation, or environmental fate and behavior in any detail, and readers are recommended to pursue some of the literature in the bibliography to gain insight into these important areas of pesticide science.

## A. Uses of Pesticides

Pesticides may be classified by the types of pest they control (terms often bearing the suffix-cide; Table I) or by the effects that they have upon the pest organism (terms that do not bear the suffix-cide; Table II).

## B. Classification of Pesticides

### 1. Formulations

Pesticides are marketed in complex mixtures, or formulations, containing the pesticide chemical itself, the active ingredient (AI), and additives that enhance mixing and dilution in water or oil and release or deliver the toxic material once it has been applied. Common ingre-

dients in a pesticide formulation include the AI, solvents, carriers, surface-active agents, and specialized additives. Solvent choice is determined by the solubility of the AI, potential toxicity to target plants (phytotoxicity), toxicology, flammability, volatility, and cost. Some solvents are not miscible in water and cause emulsions to be formed (e.g., xylene), whereas others are selected because of their ability to dissolve the AI and their ability to dissolve in water (e.g., isopropyl alcohol). Carriers can include inert clays that disperse the active ingredient through a powder or granular formulation. Surface-active ingredients are used to assist in the process of emulsion formation (the dispersion of the pesticide liquid within the liquid diluent), but they also include wetting agents (materials that reduce surface tension and enhance wetting and coverage of surfaces), dispersing agents (materials that maintain the emulsion as microscopic droplets within the diluted formulation), and spreading agents (materials that enhance coverage of waxy plant foliage and insect cuticles).

The pesticide formulation may be in solid (i.e., powder or granule) form, or it may be a liquid or gas concentrate. Modern pesticides may achieve their intended effects at rates of less than 10 g per hectare (10,000 m<sup>2</sup> = 1 ha), and the formulation blend enables these tiny quantities to be distributed evenly over the intended target surface.

A two-letter code denotes the formulation type on all pesticide labels, and this is a fundamentally important aspect of the selection of a pesticide for a particular use. There are four groups of formulation types:

- Group 1: Concentrates for dilution in water, including DC (dispersable concentrate), EC (emulsifiable concentrate), SC (suspension concentrate), and WP (wetable powder)
- Group 2: Concentrates for dilution with organic solvents, including OL (oil-miscible liquid) and OP (oil-dispersable powder)
- Group 3: Formulations to be applied undiluted, including GR (granules) and UL (ultra-low-volume (ULV) liquids)
- Group 4: Miscellaneous formulations, including RB (bait) and AE (aerosol dispenser)

### 2. Pesticide Types

Pesticides are characterized by their chemical diversity also. The summary below includes many of the most important groups of chemicals used in crop protection, but it is not exhaustive. Some pesticide properties are

TABLE II  
Classification of Pesticides by Effects on Pests

Antifeedant	Inhibits feeding while insects remain on the treated plant
Antitranspirant	Reduces transpiration
Attractant	Lures pest to a specific location
Chemosterilant	Prevents reproduction
Defoliant	Removes foliage without immediately killing plant
Desiccant	Causes plant parts to dry
Disinfectant	Destroys or inactivates harmful organisms
Feeding stimulant	Causes vigorous feeding
Growth regulator	Stops, speeds up, or retards growth in insects or plants
Repellent	Drives away pests without killing them
Semiochemical	Pheromones and other substances emitted by plants or animals that alter animal behavior
Synergist	Substance that enhances the effects of a pesticide

sufficiently uniform within the major classes of chemicals for this broad classification to be used in selection of chemicals for a particular use. Biochemical mode of action, for example, tends to be similar within major classes, and many crop protection programs use materials from several classes to avoid excessive selection pressure for resistance to pesticides. Other properties, are, however highly variable within pesticide classes as well as between them. These include vapor pressure (volatility) and the various partitioning coefficients that determine the distribution and fate of the active ingredient in the environment. These properties determine the effectiveness of the pesticide against a specific target or in a specific climate type or habitat, and detailed knowledge of these properties is required for pesticide selection to be effective. Some aspects of properties will be dealt with in Section II.

### 3. Major Classes of Insecticides

#### a. Inorganic

These are mainly nonvolatile and water-soluble compounds that often have high mammalian toxicity and may also be cumulative poisons. Many are now discontinued or banned. They include materials such as boric acid, copper sulfate, mercurous chloride, and sodium arsenite.

#### b. Organic

*i. Botanical* These pesticides are derived from plant materials, some of which have low mammalian

toxicity. Some have adverse effects on wildlife, including toxicity to fish. Examples include neem tree (*Azadiracta indica*) oil, which is used to protect stored products from insects; pyrethrum, from *Chrysanthemum cinerariaefolium*, a powerful but rapidly degraded insecticide that affects the peripheral nervous system, causing paralysis, known as “knockdown”; nicotine (from *Nicotiana tabacum*), an alkaloid insecticide that is also a neuromuscular poison in vertebrates; and rotenone, from the roots of *Derris* and *Lonchocarpus* spp., also used as a piscicide. Some of these materials are widely used in locally made or commercial formulations, and although they may be highly toxic, many are not persistent in the environment and degrade rapidly in sunlight or when exposed to microbial activity in soil or water. Each material has unique properties, and it is not possible to make generalizations about toxicology or environmental impact.

*ii. Organochlorines* Organochlorines are synthetic pesticides, among the first to be developed. Early materials, including DDT, had dramatic early successes in vector-borne disease control, but most have adverse environmental impacts, resulting from persistence in the environment, accumulation through food chains, global redistribution of residues, and ecotoxicological impacts. Many organochlorines are now banned internationally. Examples include diphenyl aliphatics (DDT, dicofol, methoxychlor), benzene derivatives (HCH, pentachlorophenol), cyclodienes (chlordane, endosulfan, endrin), and polychloroterpenes (camphechlor).



**iii. Organophosphates** Organophosphates, or OPs, are esters of phosphoric acid. Many have high mammalian toxicity and may require frequent application because they are generally not persistent. Organophosphates are nerve poisons, acting through inhibition of cholinesterase. They fall into three groups: (1) Aliphatic organophosphates, which are the oldest group, some with low mammalian toxicity (e.g., malathion, which has been in use for 40 years) but others having high mammalian toxicity, but short persistence. Short persistence makes many OPs of use in short-season crops, where the plant or the consumer has a low tolerance for pests (e.g., vegetable crops). Examples include, acephate, dichlorvos, dimethoate, malathion, and phorate. (2) Phenyl organophosphates, which are more stable and persistent but which include materials with high mammalian toxicity. Examples include fenitrothion, methyl parathion, and temephos. (3) Heterocyclic organophosphates, which may be more persistent and active in soil. Examples include azinphos-methyl, chlorpyrifos, phosmet, and pyrazophos.

**iv. Carbamates** Similar to organophosphates, carbamates are esters of carbamic acid, with short residual life and a wide spectrum of activity. This group also includes some herbicides and fungicides. Some of the pesticides with highest mammalian toxicity are carbamates. Again, they divide into three groups: (1) Methyl carbamates with a phenyl ring structure, including carbaryl and methiocarb; (2) methyl and dimethyl carbamates with heterocyclic structures, including aldicarb and methomyl; and (3) methyl carbamates of oximes, having a chain structure, including bendiocarb.

**v. Formamidines** Formamidines are used as insecticides and acaricides, have the characteristic nitrogen structure  $-N=CHN-$ , and are effective against the eggs and larvae of Lepidoptera (butterflies and moths). They are useful alternatives to organophosphates, where resistance has developed. Examples include amitraz.

**vi. Dinitrophenols** Dinitrophenols are broad-spectrum insecticides that have two nitro groups ( $NO_2$ ) attached. Examples include binapicryl and dinocap.

**vii. Organotins** Organotins are tin-based organic compounds that act as acaricides and fungicides, some with long residual activity. Examples include cyhexatin.

**viii. Pyrethroids** Pyrethroids are synthetic compounds based upon pyrethrins, found in chrysanthemum flowers.

They have similar toxicological properties but tend to be much more photostable and persistent. Natural pyrethrins occur in mixtures of six esters, and resistance is rare in the insects that they are used against. The synthetic pyrethroids are generally marketed as a single ester, and resistance may develop rapidly. Pyrethroids are generally of low mammalian toxicity but are highly toxic to fish and bees and beneficial predatory or parasitic invertebrates. Pest resurgence, or population outbreaks that result from the destruction of natural enemies, may occur in certain crops. Examples include allethrin, bioresmethrin, cyfluthrin, deltamethrin, and permethrin.

**ix. Fumigants** Fumigants are insecticides that kill by vapor or gas action, have low molecular weight, and often contain halogen radicals. They can be highly toxic to vertebrates, particularly in enclosed spaces. Examples include chloropicrin, ethylene dibromide, and methyl bromide.

**x. Petroleum Oils** Petroleum oils, also known as mineral oils, have insecticidal properties. Light applications of refined paraffin oils can be made to trees in leaf. Lower viscosity, semirefined oils can be applied to dormant trees in winter to kill invertebrates and their eggs.

**xi. Antibiotics** One antibiotic, abamectin, derived from *Streptomyces*, is an effective insecticide and anti-parasitic agent also in veterinary use. There is evidence that when used in cattle, it can harm dung-feeding invertebrates. It has some systemic properties (i.e., it can penetrate leaves and be carried within the plant vascular system).

**xii. Semiochemicals** Semiochemicals are behavior-modifying compounds, including pheromones, particularly the sex attractants of female moths. These may attract male moths for monitoring or be dispersed by spraying or formulation into controlled-release devices to establish false trails and disrupt mating of pest species. These compounds have very low environmental hazards, and the development of resistance is rare.

**xiii. Insect Growth Regulators** Insect growth regulators are compounds derived from, or inhibitory to, insect hormones and include methoprene, a juvenoid or juvenile hormone mimic, effective against mosquitoes, and diflubenzuron, which disrupts the synthesis of chitin and kills insects when they molt.

xiv. **Microbials** Included in this group are bacterial agents, particularly *Bacillus thuringiensis*, which is used against Lepidoptera, Diptera, and Coleoptera. A proteinaceous inclusion or endotoxin, which constitutes 40% of cell weight at sporulation, breaks down in insects with alkaline midguts to an active protoxin. This category also includes fungal diseases of insects, including *Beauveria bassiana*, *Metarhizium anisopliae*, and *Verticillium lecanii*, some of which can be formulated and sprayed like conventional insecticides. Some viral agents are also used on a large scale, including nuclear polyhedrosis viruses (NPV) of Lepidoptera. Microbial insecticides tend to have very limited effects on organisms that they do not directly parasitize, and vertebrate effects are almost unknown.

xv. **New Insecticides** A series of new insecticides belonging to several previously unexploited classes of chemistry have been introduced in recent years. Some of these pose greatly reduced environmental risks, although this does not apply to all of the newer materials. These include imidacloprid (in the class chloronicotinylns), a highly systemic pesticide that is now used on a very wide scale. Research on these newer insecticides is limited at present, but there is optimism that ecological impacts and risks to wildlife will be greatly reduced as they become more widely used.

#### 4. Major Classes of Fungicides

##### a. Inorganic

Inorganic fungicides are derived from sulfur or simple metal salts. They are generally stable, persistent, and insoluble in water. They include sulfur, which was originally applied as flowers of sulfur, in dust form, and which is still used, but in a more highly ground colloidal suspension. It has both direct contact and fumigant activity at temperatures above 20°C, but above 32°C the vapor may cause phytotoxicity (toxic harm to the target plant). Environmental damage and toxicological impacts to nontarget organisms are otherwise limited. Copper-based fungicides include Bordeaux mixture, an early fungicide that consists of a solution of copper sulfate and hydrated lime. With 12% copper, this fungicide has low mammalian toxicity. A more stable form of copper (e.g., copper oxychloride) is used in modern formulations, enabling the slow release of copper into leaf surface water film and toxic buildup in fungal tissue. Other inorganic fungicides have included heavy-metal-containing materials that incorporate mercury, zinc, nickel, or chromium. These are normally highly toxic and persistent, and they have been banned internationally.

##### b. Organic

Organic fungicides have the reputation of being safer and less persistent than some of their inorganic counterparts, and they are often used at very low doses.

i. **Dithiocarbamates** Dithiocarbamates are derivatives of sulfur-containing dithiocarbamic acid, in combination with zinc salts, ferric salts, and manganese salts. These fungicides have greater efficacy, better stability, and less phytotoxicity than elemental sulfur. Their toxicity derives from the formation of the isothiocyanate radical ( $-N=C=S-$ ) in breakdown. Compounds in this group include thiram, maneb, methamsodium, and zineb.

ii. **Organometallics** This group of fungicides includes the following: (1) Mercury compounds, formerly popular for disinfective and protective action as well as volatility. They have high mammalian toxicity and are no longer available for any purpose. (2) Organocopper compounds, including copper acetate, which was first synthesized in 1899. These fungicides are not easily washed off leaves, being insoluble in water, and give persistent protection. They act by the nonspecific denaturing of proteins. Examples include cuprobam. (3) Organotin, triphenyltin salts, including fentin, which can be both toxic and phytotoxic.

iii. **Substituted Aromatics** These compounds include derivatives of benzene and phenol, with hydrogen atoms replaced by chlorine, nitrogen, and oxygen, and are suited for seed treatment and control of soil-borne fungi. Examples include chlorthalonil and pentachlorophenol (PCP).

iv. **Dicarboximides** Also called sulfenamides, these compounds are considered to be among the safest pesticides in seed treatment and protectant sprays. Examples include iprodione and vinchlozin.

v. **Phthalamides** Phthalamides are used as nonsystemic, broad-spectrum, foliar fungicides on fruit, vegetables, and ornamentals. Examples include captafol and captan, which is used widely in the tropics, but increasingly subject to restrictions.

vi. **Dinitrophenols** These nonsystemic fungicides are used against powdery mildew and include the compounds binapacryl and dinocap.

vii. *Triazines* This group consists mainly of herbicides, with anilazine as the only fungicide, used as a protectant treatment in vegetables.

### c. Systemic Compounds

These fungicides are absorbed and translocated through plant tissues and provide longer term, protective control, with some curative and therapeutic effects for plants that are already infected. There are many groups, some of which are summarized below:

i. *Oxathiins* Oxathiins control basidiomycete fungi and include the compounds carboxin and meth-furoxam.

ii. *Benzimidazoles and Thiophantes* Benzimidazoles and thiophantes are broad-spectrum fungicides and are widely used in the tropics; intensive use has often led to resistance. Some of these compounds replaced organomercury fungicides as seed dressings. They include benomyl, carbendazim, and thiabendazole.

iii. *Pyrimidines* Included in this class of compounds are bipirimate and ethirimol.

iv. *Acylalanines* This class of compounds includes metalaxyl.

v. *Ergosterol Biosynthesis Inhibitors (EBIs)* Ergosterol biosynthesis inhibitors are a heterogeneous group of compounds with a common mode of action. They can have systemic, protective, and curative properties. They include (1) imidazoles (e.g., imazalil and prochloraz), (2) piperazine, pyridine, and pyrimidine compounds (e.g., pyrifenoxy and triforine), (3) morpholines (e.g., dodemorph and tridemorph), and (4) triazoles (e.g., flutriafol, myclobutanil, and propiconazole).

vi. *Organophosphates* Organophosphates are a group consisting mainly of neurotoxic insecticides but also include fungicides such as pyrazophos.

vii. *Phenylamide and Other Fungicides against Oomycetes* This is another heterogeneous group, sharing the property of toxicity to oomycete fungi. High specificity and systemic activity help to confer resistance when used intensively. Included in this group are (1) phenylamides (e.g., benalaxyl and metalaxyl) and (2) the compounds cymoxanil and prothiocarb.

viii. *2-Aminopyrimidines* This small group of systemic fungicides has activity against powdery mildew and includes bupirimate.

ix. *Quinones* This group includes benodanil and futoril.

x. *Other Organic Fungicides* A number of important fungicides do not belong to the groups above, but comprise a random selection of compounds with unrelated chemical structures. They include chlorothalonil, dodine, guazatine, and thiocyclam. Most are nonsystemic, protective fungicides.

## 5. Major Classes of Herbicides

Herbicides kill or interrupt the growth of plants. Some are selective, but others kill all plants that come into contact with them and are used in industrial settings and in rights-of-way maintenance.

### a. Inorganic Herbicides

This class includes mostly salts that have been in use for a considerable time. Sodium arsenite solutions and arsenic trioxide were popular in the 1960s but are now banned. This class also includes iron and copper sulfates, which are used for foliar application. Sulfuric acid has been widely used in horticulture and cereal crops as a selective herbicide. These materials are water soluble and readily leach from soil. Further examples include sodium salts of boric acid (borates) and sodium chlorate.

### b. Organic Herbicides

i. *Organic Arsenicals* These materials are much less toxic to mammals than inorganic arsenic salts. They inhibit metabolism, competing with phosphate in essential reactions. These compounds include disodium methanearsenate (DSMA).

ii. *Phenoxyaliphatic Acids* Herbicides in this group are a series of compounds in which the phenoxy nucleus is linked with acetic, propionic, and butyric acids. Solubility in water is high relative to that of many herbicides. They have selective, hormone-type effects on broad-leaved weeds, but grasses are tolerant. 2,4-D, 2,4,5-T, and MCPA were immensely popular and once regarded as safe. In the 1970s, dioxin contamination was suspected in 2,4,5-T and its registration was canceled.

**iii. Substituted Amides** This is a large group of organic nitrogenous herbicides, including amides and anilides. Amides primarily act in soil against annual grass weeds, causing stunting. They have low mammalian toxicity and include chlorthiamid and propyzamide. Anilides, which have numerous subgroups, are used in postemergence grass and broad-leaved weed control as well as in preemergence compounds to control germinating weeds and grasses. All have low mammalian toxicity (e.g., alachlor and propanil).

**iv. Diphenyl Ethers** These compounds possess two benzene rings, joined through oxygen or a more complex chain of molecules. They include preemergence and selective postemergence compounds of low mammalian toxicity. They are fairly insoluble in water, do not leach, and may be persistent for several months. Examples include diclofop methyl and oxyfluorfen.

**v. Dinitroanilines** Dinitroanilines are the most widely used group of herbicides in agriculture. They have a dinitroaniline nucleus in common and are effective against annual grasses and broad-leaved weeds when applied preemergence. Degradation occurs through volatilization and photodecomposition. Mammalian toxicity is low, and in soil persistence can be quite long. This class of compounds includes benfluralin and trifluralin.

**vi. Substituted Ureas** Based on the simple nitrogen-containing molecule urea, the first example, monuron, was discovered in the 1950s as a total herbicide. More modern compounds are used as selective, pre-emergence herbicides that are strongly adsorbed to soil. They inhibit photosynthesis, causing chlorosis. These compounds have low mammalian toxicity and their efficacy is influenced by temperature, rainfall, and soil type. Examples include chloroxuron, diuron, isoproturon, and linuron.

**vii. Carbamates** In addition to many insecticides and fungicides, a number of herbicides have been developed from this chemical group. They include pre- and postemergence materials that inhibit germination and cell division. They have low mammalian toxicity and short persistence. Examples include asulam and phenmedipham.

**viii. Thiocarbamates** The thiocarbamates are a group of carbamates containing sulfur, including selective pre- and postemergence compounds. The -allate compounds may persist for many months in soil.

Metham sodium is a soil fumigant, converted in the soil to methyl isothiocyanate. Examples include diallate, metham sodium, and thiobencarb.

**ix. Heterocyclic Nitrogens** These compounds have a ring structure where the carbon atoms are replaced by nitrogen or sulfur. They include triazines, triazinones, triazoles, pyridines, and uracils. *Triazines* are a very selective group of compounds; selectivity depends upon the plant's ability to metabolize the AI. They are soil applied and absorbed through roots, inhibiting photosynthesis once the plant has emerged. Mammalian toxicity is generally low, and intense application in some areas has led to groundwater contamination and restrictions upon use. Examples include atrazine and simazine. *Triazinones* have six-membered rings and include hexazinon. *Triazoles*, with a five-membered-ring structure, are active against broad-leaved weeds (e.g., amitrole). Some *pyridines* are effective against deep-rooted herbaceous weeds, and others are selective brush killers (e.g., picloram and trichlopyr). *Uracils*, another substituted, six-membered-ring family of chemicals with two nitrogen atoms and a double bond, are primarily for preemergence application and uptake by roots. They are effective at controlling annual grasses and broad-leaved weeds over an extended period by inhibition of photosynthesis. Uracils may be very persistent. Examples include bromacil and lenacil.

**x. Bipyridiliums, Also Termed Pyridines** Containing two pyridyl rings, this group includes some extremely well known and widely used compounds (e.g., diquat and paraquat). They have contact action on the above-ground parts of plants and in the plant they are reduced to free radicals that destroy tissue under light. They are widely used as desiccants. These compounds are deactivated by sorption to the soil and slow to degrade, leading to increasing restrictions. Both AIs are hazardous, being the main cause of pesticide-related death in many countries. There is no known antidote.

**xi. Aliphatic Acids** This group includes chlorinated derivatives of acetic acid, trichloroacetic acid (TCA) and dalapon. These compounds are used on non-cropland and in forestry, killing the plant by causing precipitation of proteins within cells.

**xii. Phenol Derivatives** Herbicides in this group are highly toxic, selective, foliar herbicides and include dinitrophenols and one chlorinated phenol, pentachlorophenol. Dinitrophenols, primarily contact herbicides,

TABLE III  
Utilization Efficiency of Pesticides<sup>a</sup>

Pesticide	Method of application	Target organism	Utilization efficiency (%)	Explanation
Demeton-S-methyl	Foliar spray	Aphids on sugar beet	0.000008	Insects located in heart leaves, an effective refuge from direct spraying
Dieldrin	Seed treatment	Wheat bulb fly larvae	0.0015	Pests less likely to encounter seed as plant develops
Dimethoate	Foliar spray	Aphids on field beans	0.03	Relatively low efficiency of spray retention by plant
Lindane	Foliar spray	Capsids on cocoa	0.02	Pesticide losses due to drift when treating large trees
Lindane/dieldrin	Aerial spraying of swarms	Locusts	6.0	Pesticide applied by aircraft within swarm, maximizing chance of contact

<sup>a</sup> From data summarized by Graham-Bryce (1977): Graham-Bryce, I. J. (1977). *Philos. Trans. R. Soc. London*, B 281, 163–179.

inhibit respiration and photosynthesis (e.g., dinoseb (DNBP), DNOC) and are now banned in the United States. Pentachlorophenol is now being banned in Europe and the United States.

*xiii. Benzonitriles or Substituted Nitrites* These compounds consist of a benzene ring with a cyanide (C≡N)- radical and are broad spectrum, acting on various processes of growth and tissue disruption (e.g., dichlobenil and ioxynil).

*xiv. Miscellaneous Herbicides* Herbicides with unrelated chemical structures include some widely used materials such as endothal sodium, an aquatic weed killer, glyphosate, a nonselective, residual, postemergence herbicide, and alloxym sodium, fluzifop-butyl, and other selective, postemergence systemic herbicides, used against perennial and annual grasses.

## II. EFFICIENCY

Ideally, the receiving organism receives a toxic dose of pesticide and nontarget organisms escape injury, achieving high efficiency of use. Efficiency is, however, very rarely measured. When it has been, efficiency is far from ideal when expressed as the proportion of the applied dose taken up by the target organism or the amount needed to directly combat an infestation by treating it directly (Table III).

What should our expectation for efficiency be? High efficiency can-not be expected in all circumstances. For example, many applications are made before serious

infestations ensue, and pests are few in number or widely dispersed. It is generally agreed, however, that there is considerable scope for improvement in the efficiency of pesticide delivery, through attention to application, formulation, and the optimum delivery of the AI through attention to the physicochemical properties of the compound. The interaction between the pesticide and organism that leads to chemical exposure and uptake is termed bioavailability. The term avallance describes the profile of chemical concentration over time, a function of physicochemical properties, transfer, and sorption within the treated environment.

### A. Delivery and Bioavailability

The toxic effects of a pesticide against any intended target or unintended nontarget organism are a function of intrinsic toxicity (the activity when it is directly applied to the organism) and the amount that reaches the organism under the conditions of application. The amount that the organism is exposed to depends upon placement (which is a function of application method), pattern of release (a function of the formulation), redistribution and transfer processes (a function of the physicochemical properties of the pesticide, environmental conditions, and the nature of the substrate), and the location and receiving characteristics of the organism. The latter is a function of the degree of contact or avoidance of the chemical, the activity of the organism in question, and, on a large scale, the spatial dynamics of the organism as it relates to the pattern of treatment and the persistence of the chemical.

## B. Application, Formulation, and Delivery of the Toxic Dose

Through application and formulation, it is possible to manipulate the placement, size, and composition of the discrete units or drops of pesticide in which it is dispensed and the rate of release from those units or particles.

### 1. Soil Treatment

Some of the most difficult challenges relate to the problem of delivering sufficient pesticide to small targets distributed in a bulk medium in which mobility is restricted and the chemical subject to rapid degradation. It is therefore important to localize the pesticide to the vicinity of the target and plant part that needs protection. For example, granular formulations and seed treatments can both be used to increase efficiency of transfer of pesticide to organisms in the soil. Once applied, the chemical must move into the soil matrix and the rate of release determines the concentration in the soil. Diffusion and redistribution can be achieved by bulk flow of water moving down through the soil profile after rain or irrigation. Contact between the water network in soil pores and the coated seed is very important.

Efficiency, or the toxicity of soil-applied AI per unit of material released, is determined by the way in which the chemical partitions between the solid, liquid, and air phases of the soil. The partition coefficients between these different phases can be used to accurately predict the rate of movement of pesticides within a given soil type.

### 2. Spray Application

Spray application is the commonest method of pesticide delivery. It is convenient, flexible, and simple, but it lacks selectivity and there is a risk of contaminating nontarget areas. Hydraulic sprayers give a range of pesticide drop sizes, at the smallest extreme of which there is a tendency to drift away from the target area. Controlled droplet application (CDA) sprayers are much rarer, but they control drop size and as a consequence may reduce drift and increase efficiency.

The main function of spray machinery is to transfer energy to the liquid pesticide formulation to atomize it into droplets by means of a nozzle. The drop size distribution is typical for the nozzle in question and it is characterized by the volume median diameter and the number median diameter of the drops.

- Volume median diameter (VMD): If a sample of spray is divided into two equal volumes, then half

of the volume contains drops with a diameter less than the VMD, and the other half contains drops with a diameter greater than the VMD.

- Number median diameter (NMD): If the drops are divided in half by number, independent of volume, then the diameter of half of the drops is larger than the NMD, and the diameter of the other half is smaller.

The NMD:VMD ratio indicates the spread of drop sizes: the larger the ratio, the broader the spectrum. If the ratio approaches 1, the drops are all of a similar size and behave in a similar way.

Conventional spray application through hydraulic nozzles presents a major challenge for optimization of pesticide use efficiency, and their widespread use accounts for a great deal of pesticide drift and off-target exposure and contamination. Table IV provides details of the drop spectrum from a conventional fan jet hydraulic spray nozzle. Drops of between 1 and 50  $\mu\text{m}$  in diameter account for only 4.2% of the volume, but 90.8% of the total drops in the sample. In addition, 61.9% of the volume of the spray liquid is in drops of greater than 160- $\mu\text{m}$  diameter. In 1 liter of spray liquid, this would yield nearly 90 million drops with a diameter of less than 50  $\mu\text{m}$ , and therefore susceptible to drift.

Drop behavior has been the subject of detailed research. Drops are subjected to various forces as they leave the nozzle and before they impact on a surface.

- Gravitational forces: Small drops have a low terminal velocity. For example, a 5- $\mu\text{m}$ -diameter drop falls at 0.075 cm/sec and often drifts from the target area, whereas a 500- $\mu\text{m}$ -diameter drop falls to the ground at 213.9 cm/sec.
- Movement induced by air movement: Wind carries drops away from the target area. For example, in a wind of 1 m/sec, a 5- $\mu\text{m}$ -diameter drop released from 1 m experiences 350 m of sideways transport, whereas a 500- $\mu\text{m}$  drop moves 0.48 m. To deposit within 5 m of the target area, a drop must have a diameter of greater than 67  $\mu\text{m}$  in a wind speed of 0.7 m/sec and it must have a diameter of greater than 168  $\mu\text{m}$  in a wind traveling at 3 m/sec. The friction of the air rapidly decelerates drops from the velocity at which they left the nozzle. For example, a 5- $\mu\text{m}$  drop is decelerated to its deposition velocity in 0.33 cm and a 200- $\mu\text{m}$  drop is brought to the equivalent velocity in 630 cm.

Drift is worse where relative humidity is low. Drops become smaller as spray liquid evaporates and sedimen-

TABLE IV  
Drop Sizes from a Hydraulic Spray Nozzle Measured with a Laser Particle Size Analyzer<sup>a</sup>

Drop range ( $\mu\text{m}$ )	Volume (%)	Cumulative volume (%)	Number (%)	Cumulative number (%)
563–262	26.17	61.94	0.07	0.80
262–160	35.77		0.75	
160–113	17.52		1.36	
113–84	8.78	33.84	1.82	8.39
84–65	4.81		2.31	
65–50	2.73		2.86	
50–30	2.66		8.45	
30–15	1.12	4.22	21.90	90.81
<15	0.27		60.46	

<sup>a</sup> Data from Micron Sprayers, UK. Nozzle, Fan Jet SS8002 operating at a pressure of 3 bar and a flow rate of 700 ml/min. VMD = 192  $\mu\text{m}$ , NMD = 12  $\mu\text{m}$ , VMD:NMD ratio-16.

tation takes longer. If 1 liter of liquid is atomized to 10- $\mu\text{m}$  drops, then the cumulative surface area is 600  $\text{m}^2$ . If it is atomized to 100- $\mu\text{m}$  drops, then the surface area falls to 60  $\text{m}^2$ . Evaporation is therefore exacerbated by the shrinkage of drops and the increase of relative surface area.

Large drops pose an equivalent problem: not deflected by air currents, they gain momentum and either miss the plant target and hit the soil or strike the plant, shatter, and fall to the ground. The cubed relationship between diameter and volume means that the volume of a 100- $\mu\text{m}$ -diameter drop is 1 million times that of a 1- $\mu\text{m}$ -diameter drop. The variation in pesticide dose between a 4- and 400- $\mu\text{m}$ -diameter drop (common in a hydraulic nozzle drop spectrum) is 1 million fold. Many drops of a diameter greater than 300  $\mu\text{m}$  fall to the ground, and a large proportion of the pesticide applied by a hydraulic nozzle never comes into contact with the pest, disease, or weed it is intended to control.

### 3. Deposition on Obstacles

Objects are surrounded by a cushioning layer of air, and spray drops require considerable momentum to hit a target. Larger drops strike targets, and smaller drops move around them. The probability of striking depends upon the shape of the object; hairs, for example, pick up small drops. Flying insects such as mosquitoes are best hit by drops of 10–20  $\mu\text{m}$  in diameter, resting flies require drops of 30–40  $\mu\text{m}$ , diameters of 90–130  $\mu\text{m}$  are needed where crop penetration is needed, and a diameter of >250  $\mu\text{m}$  is optimal for horizontal surfaces such as weeds.

Continuing reliance upon hydraulic spray technol-

ogy limits the efficiency, effectiveness, and safety of the spray application process.

## III. ECOTOXICOLOGY AND MANAGEMENT

Pesticides pose challenges through being toxic to organisms other than the specific pests, diseases, or weeds they are targeted against. Relative toxicity is expressed in a number of ways, but the most prevalent relates to oral and dermal exposure of mammals, which equates to both human and vertebrate wildlife toxicological risks for organisms that are directly exposed (Table V).

Table VI lists some representative active ingredients by toxicity and demonstrates that in all major pesticide types and range of classes, there is a considerable variety in intrinsic toxicities to mammals.

It is extremely difficult to translate these data, or toxicological data for any other species, into predictions of ecological risk. Very toxic materials may be relatively nonhazardous in the environment, because they are rapidly sorbed or dissipated, whereas less toxic materials may prove hazardous because they readily leach or runoff into water or because they are persistent.

Recent attempts have been made to bridge the gap that exists between basic toxicology and environmental chemistry and the potential ecological impact of pesticides in ecosystems. Van Straalen and van Rijn (1998) computed species sensitivity distributions for soil fauna (soil invertebrates that contribute to biogeochemical cycling), based upon laboratory test data for toxic effects in soil, and calculated the concentration that

TABLE V  
Pesticide Toxicity Classes

Toxicity rating	LD <sub>50</sub> , single oral dose for rats (mg/kg)	LD <sub>50</sub> , single dermal dose for rabbits (mg/kg)	Probable lethal oral dose for humans
6, supertoxic	<5	<20	A taste, a grain
5, extremely toxic	5–50	20–200	A pinch, 1 teaspoon (5 ml)
4, very toxic	50–500	200–1,000	1 teaspoon (5 ml) to 2 tablespoons
3, moderately toxic	500–5,000	1,000–2,000	1 liquid ounce (30 ml) to 1 pint (470 ml)
2, slightly toxic	5,000–15,000	2,000–20,000	1 pint (470 ml) to 1 quart (950 ml)
1, practically nontoxic	>15,000	>20,000	>1 quart (950 ml)

would be below the no-effect concentration (NOEC: the highest dose not to cause a specific effect) for reproductive inhibition for 95% of the species that are theoretically present within the soil arthropod community. Using a simple model for pesticide fate, based upon exponential decay, they then calculated for a series of compounds the time that each product would take to

reach this “95% protection level.” This time, referred to as the “ecotoxicological recovery time,” can be used as a basis for calculating the minimum time for ecological recovery to take place after pesticide application. There is considerable scope for further development of this technique, built upon readily collected toxicological data sets. Such data are, however, surprisingly diffi-

TABLE VI  
Examples of Pesticide Toxicities to Vertebrates from Laboratory Test Data Used in Registration

Pesticide type and class	Pesticide technical name	Oral LD <sub>50</sub> (rat, mg/kg)	Dermal LD <sub>50</sub> (rat (a), rabbit (b))
Insecticides			
Organochlorine	Dieldrin	46	10 (a)
	Endosulfan	80	359 (b)
Organophosphate	Azinphos-methyl	16	222 (a)
	Chlorpyrifos	135	2000 (b)
	Dichlorvos	56	75 (a)
	Dimethoate	320	<130 (b)
	Malathion	2800	4100 (b)
Carbamate	Aldicarb	0.7	5 (b)
	Carbaryl	850	>2000 (b)
Pyrethroid	Deltamethrin	135	>2000 (b)
	Allethrin	930	—
	Bioresmethrin	7070	—
Fungicides			
Dithiocarbamates	Thiram	780	>5000 (a)
	Maneb	7990	>5000 (a)
Dinitrophenol	Dinocap	980	—
Herbicides			
Inorganic	Sodium arsenite	39	—
Bipyridiliums	Paraquat	157	—
Triazines	Simazine	5000	>3100 (b)



TABLE VII  
Rankings of 10 Selected Pesticides According to Three Criteria<sup>a</sup>

Pesticide	Rank based on toxicity	Rank based on persistence	Joint assessment (recovery time)
Lindane (organochlorine insecticide)	6	2	2
Dimethoate (organophosphate insecticide)	8	8	7
Parathion (organophosphate insecticide)	5	6	6
Chlorpyrifos (organophosphate insecticide)	1	7	4
Carbofuran (carbamate insecticide)	4	3	3
Carbaryl (carbamate insecticide)	2	10	5
Methomyl (carbamate fungicide)	7	9	8
Benomyl (benzimidazole fungicide)	3	1	1
Atrazine (triazine herbicide)	10	4	10
Fentin (organotin fungicide)	9	5	9

<sup>a</sup> The score 1 is given for the highest toxicity (the 95% protection level referred to in the text), the greatest persistence (soil half-life), and the longest predicted ecotoxicological recovery time (i.e., the time it takes the pesticide concentration to decline to the 95% protection level).

cult to find in the published literature, and they are not routinely collected for suites of compounds in such a way that they may be exploited within this new approach as a decision-making aid.

Table VII, adapted from van Straalen and van Rijn (1998), summarizes the differences in ranking of toxicity that result if pesticide toxicity to soil organisms and chemical persistence (half-life) in soil are integrated to predict the time when recovery could take place by organisms. A compound with high intrinsic toxicity, such as chlorpyrifos, has a lower ranking once its limited persistence is taken into account, whereas benomyl climbs higher up the rankings because of its persistence. Analytical procedures of this form are beginning to forge a link between the large amounts of laboratory-obtained toxicological data that exist and the field, where functioning communities are exposed to the toxin, rather than individual species.

### A. Farm-Scale Observations of Ecotoxicological Impact

A number of research projects investigating the ecology of farming systems were initiated throughout the 1980s and 1990s in Europe (Holland *et al.*, 1994). Pesticide impacts were investigated in all of these studies, and comparisons between conventional and reduced pesti-

cide inputs were made in at least 13 sites, in Germany, The Netherlands, Switzerland, and the United Kingdom (Table VIII). In all cases, beneficial nontarget invertebrate densities were higher in areas where the total regime of pesticide input had been reduced. In the larger scale studies (e.g., the Boxworth study, UK), with 5.6- to 15.7-ha treatment area sizes, some beneficial species were rendered locally extinct for the full 5-year treatment phase of the project, and recovery was subsequently slow. The level of impact on individual species was shown to be a function of life history attributes that affected pesticide exposure and capacity of that organism to invade the treated area. Predatory capacity was inhibited in the highest pesticide regimes, and there was evidence that this contributed to higher pest densities in some years. Although similar data have been obtained for the invertebrate community of rice systems in the tropics, data sets of this level of complexity are rare in the investigation of pesticide impacts in agroecosystems.

These investigations have revealed the subtler impact of second- and third-generation active ingredients that followed the organochlorines. They reinforce the need to measure ecological impacts on scales that reflect agricultural practice and pesticide use in the real world and that are tuned to the scale of dispersive movement of nontarget taxa. They have also triggered considerable

TABLE VIII

Summary of Data from Farming System Experiment in the United Kingdom (Boxworth, SCARAB, TALISMAN, RISC, LIFE), Germany (INTEX), The Netherlands (Nagele), and Switzerland (Third Way): Results of Integrated Farming, with Reduced Pesticide Inputs, Compared with the Conventional Levels of Use<sup>a</sup>

Project	Beneficial arthropods	Birds and mammals	Earthworms	Soil microorganisms	Soil minerals
Boxworth	+	o			
SCARAB	+		o	o	
TALISMAN	=				
RISC	o				
LIFE	+		+	=	=
Lautenbach	+		+		=
INTEX	+				=
Netherlands	+	+	+		
Third Way	+		+	+	

<sup>a</sup> Key: +, increase; =, no change; o, variable result over the period of the study.

interest in the mechanisms that underlie long-term depletions, even local extinction of certain species. Both scale of treatment and the mode and rate of dispersal of arthropods influence rates of population recovery following pesticide use, and the proximity of local refugia from which recovery can occur, and landscape features conducive to movement and colonization are important factors underlying extinction risk. Modeling may provide an appropriate tool for testing our understanding of invertebrate population processes at the agroecosystem level, but it does not substitute for the need to undertake a far greater number of manipulative experiments and monitoring programs that examine the spatial dynamics of nontarget organisms in sprayed farming systems.

### B. Avian Impacts

The requirement for large-scale, field-based monitoring and experimentation is not restricted to terrestrial nontarget invertebrates. Avian toxicologists have long recognized that field-based studies provide unique and indispensable data for interpretation of pesticide effects on birds (Taub, 1997). Many of the most important effects of pesticides on birds, including eggshell thinning as a result of the bioconcentration of organochlorines through food chains, were originally detected and documented as a result of field-based observations. These effects also include endocrine disruption, food

supply reduction or alteration, variation in sensitivity between species, and the synergistic effects of exposure to multiple compounds.

### C. Aquatic Systems

Similar arguments may be applied to the investigation of pesticide effects on aquatic organisms. The recent decision in the United States to reduce the requirements for ecological data from multispecies test systems in the regulatory process has been criticized because it will reduce the probability of detecting biologically significant effects (Taub, 1997). These effects include indirect trophic-level impacts, compensatory shifts within a trophic level, responses that are associated with seasonal trends in populations, chemical transformations effected by organisms in the exposed system, and impacts that result from long persistence of either the parent product or toxic breakdown products.

In conclusion, there is now abundant evidence that pesticide impacts can evolve at the agroecosystem scale and that this requires the development of appropriately scaled monitoring or experimental systems. Toxicological data are of fundamental value in the initial evaluation of pesticides, but in their real-world applications, pesticides may also elicit ecological effects that reverberate through the system, long after the chemical residues have become undetectable. Laboratory-based, single-species tests are limited in their predictive power for

impacts in the field, and further development of the theory and methodology associated with ecological effects is required.

### D. Tools for Pesticide Management

It is questionable whether we apply our knowledge about pesticides effectively at any stage following the regulatory permission to use a product in a specified way. Given the need to combine knowledge of chemical properties, fate and behavior, environmental attributes such as soil type, toxicology, and ecology, the challenge is considerable. In each of these disciplinary areas, however, there are considerable databases of knowledge and predictive models that can be used to interpret how a pesticide will behave in a given set of field conditions.

Attempts have been made to summarize clusters of pesticide properties in databases that may be used to compare the environmental and ecological risks posed by lists of candidate compounds for specific uses. These databases are not specifically tuned to local conditions; they do, however, combine many factors, including risks to wildlife, in the rankings that they generate, and enhance the capacity of the end user to make more informed decisions about particular uses. The most significant development in this field is the proposal to derive Environmental Impact Quotients (EIQ) for compounds from established databases of pesticide properties (Kovach *et al.*, 1992). The quotient combines farmworker effects (built from acute and chronic toxicity and plant surface half-life) with consumer effects (built from systemicity, soil and leaf surface persistence, and potential for groundwater contamination) and ecological effects (built from aquatic and terrestrial ecotoxicology). There is considerable scope for toxicologists to develop approaches such as this from first principles in order to make data available in a form that can be used to assist decisions in the field. Guided by EIQs, growers and their advisors could then plan pest control tactics that attempt to minimize nontarget impacts and track their progress with a rigorous and quantitative

methodology that could also be explained to consumers.

### See Also the Following Articles

AGRICULTURE INVASIONS • AGRICULTURE,  
INDUSTRIALIZED • ECOTOXICOLOGY • HERBICIDES •  
INSECTICIDE RESISTANCE

### Bibliography

- Bellows, T. S., and Fisher, T. W. (1999). *Handbook of Biological Control*. Academic Press, San Diego.
- Benbrook, C. M. (1996). *Pest Management at the Crossroads*. Consumers Union, New York.
- Calow, P. (1998). *Handbook of Ecotoxicology*, Vols. 1 and 2. Blackwell Sci., Oxford.
- Conway, G. R., and Pretty, J. (1991). *Unwelcome Harvest: Agriculture and Pollution*. Earthscan Publications, London.
- Croft, B. A. (1990). *Arthropod Biological Control Agents and Pesticides*. Wiley, New York.
- Farm Chemicals Handbook* (2000). Meister Publishing Co. Willoughby, Ohio.
- Greig-Smith, P. W., Frampton, G. K., and Hardy, A. R. (Eds.) (1992). *Pesticides, Cereal Farming and the Environment*. H. M. Stationery Office, London.
- Hartley, G. S., and Graham Bryce, I. J. (1980). *Physical Principles of Pesticide Behavior*. Academic Press, San Diego.
- Holland, J. M., Frampton, G. K., Cilgi, T., and Wratten, S. D. (1994). Arable acronyms analysed: A review of integrated arable farming systems research in W. Europe. *Ann. Appl. Biol.* 125, 399–438.
- Kogan, M. (Ed.) (1986). *Ecological Theory and Integrated Pest Management Practice*. Wiley, New York.
- Kovach, J., Petzoldt, C., Degni, J., and Tette, J. (1992). A method to measure the environmental impact of pesticides. *New York's Food Life Sci. Bull.* No. 139.
- Matthews, G. A. (1992). *Pesticide Application Methods*, 2nd ed. Longman, New York.
- Marco, G. J., Hollingworth, R. M., and Durham, W. (1987). *Silent Spring Revisited*. Am. Chem. Soc., Washington, D.C.
- Taub, F. B. (Ed.) (1997). Invited Feature: Are ecotoxicological studies relevant to pesticide registration decisions? *Ecol. Appl.* 7, 1083–1132.
- Van Straalen, N. M., and van Rijn, J. P. (1998). Ecotoxicological risk assessment of soil fauna recovery from pesticide application. *Rev. Environ. Contam. Toxicol.* 154, 83–141.
- Ware, G. W. (1999). *The Pesticide Book*, 5th ed. Thomson Publications, Washington, D.C.



# PHARMACOLOGY, BIODIVERSITY AND

Paul Alan Cox

National Tropical Botanical Garden

---

- I. Evolutionary Perspectives
  - II. History
  - III. Modern Chemical Approaches to Drug Discovery
  - IV. Rational Screens of Biodiversity
  - V. Nature of Biodiversity-Derived Pharmaceuticals
  - VI. Biodiversity and Ethics
- 

## GLOSSARY

**bioassay** A test for pharmaceutical activity in substance conducted either in living organisms (*in vivo* assays) or in the test tube (*in vitro* assays).

**coevolution** A heritable response made by a species through evolutionary time to a different species which is then reciprocated.

**ethnobotany** The study of the uses of plants by different groups of peoples, often indigenous peoples.

**extremophile** A organism that lives and thrives in a habitat characterized by extremes of temperature, acidity, alkalinity, salinity, light, or pressure that would prove lethal to most other organisms.

**fractionation** A method of obtaining a pure compound by extracting components of a mixture with solvents of different solubility.

**herbal** A compilation of medicinal plants and their properties often used in earlier times as a reference for physicians prescribing medical treatments based on plant therapies.

**indigenous intellectual property rights** Rights to intellectual properties belonging to indigenous peoples,

including but not limited to iconographical representations, terminologies, names, phrases, legends, methods, and techniques of traditional cultivation, healing, identification, preparation, and use of biodiversity.

**indigenous peoples** Peoples who have resided in the same geographical area for many generations and who possess legends, proverbs, genealogies, languages, and other unique cultural features linking them to the land.

**phylogeny** A family tree of related organisms tracing their evolutionary history.

**voucher** A representative specimen of a plant or animal that is properly collected, prepared, and preserved in a herbarium or museum to facilitate expert identification of the species.

---

**PHARMACOLOGY**, the science of drugs, studies the ways in which natural products can be employed as medicines. Since ancient times, humans have recognized the pharmaceutical properties of certain compounds derived from plants and animals. Current research now also recognizes that biodiversity must be maintained in order for the environment to continue to be a source of such medicinal substances.

## I. EVOLUTIONARY PERSPECTIVES

Life-forms have evolved a plethora of complex biochemistry to help mediate their interactions with their physi-

cal and biotic environments. Some of the resultant molecules function as materials for membranes and cell walls, photosynthetic pigments, energy storage compounds, neurotransmitters (in animals) and hormonal signals (in plants as well as in animals), defensive compounds, a variety of protective pigments, and as structural materials. The resultant juxtaposition of similarity and diversity of the chemical compounds created by different life-forms is striking. For example, genetic inheritance in all known life-forms is based on nucleic acids; although some enzymes used in replication differ with regard to prokaryotes and eukaryotes, all known living organisms use RNA or DNA to transmit their genetic information. However, differences in other biochemical pathways are profound. These differences result in differential vulnerability to toxic compounds, with ranges in LD50s of specific compounds varying several orders of magnitude between different phylogenetic groups. For example, the concentration of penicillin that completely stops cell wall formation and growth in *Streptococcus* bacteria shows little effect when administered to a centipede or to a geranium, whereas the dose of 2,4-D sufficient to kill a geranium would have negligible impact on either centipedes or bacteria.

Throughout the history of life on Earth, such differences in biochemical vulnerability have generated abundant possibilities for chemical coevolution between different species. Even though life on Earth appears to share a common phylogeny, biochemical differences between species leave open the possibility that a mutation in the biosynthetic pathway of one organism may yield a compound toxic to another. Selection for such protective molecules from predation would likely be greatest in organisms that are unable to flee or lack alternative forms of defense, including sessile organisms such as corals, sponges, tunicates, and other marine invertebrates and nearly all plants. Although poisonous birds have been found in New Guinea, toxic frogs are prominent in the Neotropics, and brightly colored toxic insects are known to most amateur naturalists, such biochemical protection against predation by motile organisms appears to be the exception rather than the rule.

A plant species under pressure from a predator or parasite may perish unless individuals in the plant population possess, by mutation or some other evolutionary accident, a compound toxic or unpalatable to its enemies. If the mutation is heritable, progeny of such plants will increase in time through population, replacing those individuals previously lost to predation. If the species of predator or parasite are highly selective in their choice of prey, they may in turn become imperiled by such chemical innovations, unless individuals in

the predator or parasite population possess a means of detoxifying the defensive chemistry generated by the plants. The resultant coevolutionary race between two species, sometimes termed a “red queen race” after the never-ending race described by Lewis Carroll in *Alice in Wonderland*, has been documented in the interaction between *Passiflora* vines and *Heliconius* butterflies in the Neotropics by Larry Gilbert and coworkers at the University of Texas. Other such tight coevolutionary links are the topic of much scientific interest.

More common than species–species coevolution is what might be termed diffuse chemical coevolution. Instead of evolving defensive chemistry against a single species of predator or parasite, a plant species generates compounds effective against a wide variety of plant enemies. The presence of toxic compounds in a plant cannot, however, be considered *prima facie* evidence of an evolutionary response to predation. Bioactive molecules occur in plants as secondary metabolites as well as evolved chemical defenses against predation, fungal attack, microbial invasion, and viral infection. It is also unlikely that plant compounds known to be toxic to humans represent evolutionary responses to anthropogenic harvesting; given the omnivorous, generalist foraging patterns of higher primates, it is unlikely (with the possible exception of agricultural weeds) that any plants have evolved a specific biochemical response to consumption by humans. However, there are hundreds of thousands—perhaps even millions—of bioactive molecules in nature resulting from plant chemical warfare with viruses, bacteria, fungi, and arthropods. Many of these compounds can be beneficial to humans.

The emerging picture of chemical interactions between sessile prey and motile predators is compelling: what might otherwise appear to be a quiet forest glen or a tranquil coral reef may be the evolutionary equivalent of a battlefield rife with chemical warfare. Sessile prey appear to sequester toxic agents and motile predators develop detoxification strategies at a prodigious rate through evolutionary time. Since many of these bioactive molecules have relevancy to human disease, a coral reef or forest glen could also be said to resemble a large pharmaceutical storehouse, perhaps one that is in great disarray. Tens of thousands of vials of pharmacologically active compounds litter the ground, but the labels identifying the contents and their therapeutic utility have been lost. How can one determine which molecules are useful and for which diseases?

Eloy Rodriguez and coworkers at Cornell University have found evidence that some vertebrates self-dose with pharmacologically active compounds produced by plants. Many mammals, including primates, engage in

self-medication when ill. Ongoing pharmacological self-experimentation by animals may have been observed and mimicked by early humans. Early Polynesian legends indicate that kava, a beverage made from the roots and rhizomes of *Piper methysticum* rich in tranquilizing sesquiterpenes and kava lactones, was discovered by watching the calming effect the roots had on rats. Legends from India claim that in ancient times mongooses were observed to feed on *Rauwolfia serpentina* before engaging in combat with cobras. Copying the reputed feeding behavior of the mongoose, local people found that the shrub could serve as a potent antidote to snakebite. The veracity of such legends is unclear, but it is clear that to trace the use of plant-derived bioactive molecules in medicine we must first examine their use in prehistory by indigenous peoples.

## II. HISTORY

### A. Early Understandings

All known indigenous groups have ethnomedical traditions based, at least in part, on pharmacologically active compounds found in plants and animals. It is therefore safe to assume that early humans also developed significant ethnopharmacopoeias; this assumption is corroborated by archaeological finds of medicine pouches in ancient burial sites. Results of successful experiments with pharmacologically active plants and animals were likely passed from generation to generation. This accumulation of oral information resembles in part the results of a vast, uncontrolled, and unwritten human bioassay experiment. Later knowledge about medicinal plants was transmitted in written form. By the fifth century B.C. Hippocrates codified and disseminated knowledge from earlier times concerning plant medicine in the Mediterranean area. During this period, commerce in medicinal plants was vigorous, especially with traders called *rhizotomii* dealing principally in plant roots. The folk knowledge accumulated by the *rhizotomii* was subsequently studied by scholars, including Aristotle (384–322 B.C.), who laid a foundation for the philosophical consideration of plants. Aristotle bequeathed his library to his student, Theophrastus, who wrote the most important book on plants up to that time, *De Causis Plantarum*, and is said to be the first botanist. In his *Enquiry into Plants*, which in the ninth book catalogs information on medicinal plants, Theophrastus incorporated information from the *rhizotomii*. He was, however, skeptical of folk superstitions concerning the gathering of medicinal plants.

In the first century Padanius Dioscorides, a Greek physician, wrote a seminal work on medicinal plants titled *De Materia Medica*, which describes more than 500 medicinal plants and includes many drawings. Until the beginning of the Renaissance, Dioscorides' work was the final word in medicinal plants for more than 1000 years.

Such early compilations of folk wisdom concerning medicinal plants were not confined to the West. In 2000 B.C., the Chinese emperor Chi'en Nung compiled the *Pen Tsao*, which is perhaps the earliest known herbal, whereas in India, Sanskrit texts on medicinal plants are said to date to 3000 years ago. In the Americas, the Mayans also wrote extensive manuscripts on medicinal plants. Unfortunately, many of these codices were burned by the early Spanish conquistadors.

### B. Herbals and Medicinal Plants

An increase of scholarly interest in herbals and medicinal plants began during the Renaissance, when the first physic gardens (gardens of medicinal plants for therapeutic use and study) were planted in Italy and Germany. Most early Renaissance herbals were largely based on the work of Dioscorides' book, *De Materia Medica*, with additions made from the author's own knowledge. As can be seen in *De Materia Medica*, two pieces of information were disseminated concerning each plant: its purported healing properties and its identification.

Agnes Arber (1953) pointed out that the necessity for precise identification led to an unusual feature of early botanical iconography: for nearly 2000 years, plants in herbals were pictured with the part most likely to be in commerce—their roots. Although these earlier herbals were lovingly copied throughout the ages, the quality of illustrations declined until, in the fourteenth century, it had become increasingly difficult to identify the plants. Innovators such as Fuchs and Mattioli began to inscribe illustrations drawn from living plants on wood blocks for printing. This was a major boon for botanists since images reproduced by wood block printing did not suffer the same image deterioration in successive copies as occurred in hand-in-inked illustrations. The new herbals also contained current information gleaned from European folk medicine.

In the English-speaking world, the first herbal is an eleventh-century Anglo-Saxon codex known as the *Herbarium of Apuleius Patonicus*. The earliest printed English herbal is an anonymous quarto from 1525 imprinted by Richard Banckes (as quoted in Arber, 1953, p. 41): "Here beynneth a newe mater, the whiche shew-

eth and treateth of y vertues and propyrtes of herbes, the whiche is called an Herball.”

A year later, a translation of a French herbal was published by Peter Treversi, and in 1538 William Turner published *Lebellus de re Herbaria Nova*. In 1551, Henry F. Lyte published a translation of Dodoen’s famous herbal, and in 1555 Anthony Askham published a herbal derived from the 1525 English work of Banckes. However, the most popular of all sixteenth-century herbals was that of John Gerard, published in 1597.

Born in 1545, at the age of 32 Gerard was appointed superintendent of the gardens of Lord Burleigh at the Strand in London. Later, he was appointed curator of the Physic Garden of the College of Physicians of London. A catalog of plants in his own garden appeared in 1596, and in 1597 he published *The Herball, or General Natural History of Plants*, one of the most quoted botanical works ever published. Gerard’s *Herball*, with 1392 pages and 2200 woodcut images of medicinal plants, was greeted with tremendous enthusiasm by medical practitioners and the general public.

The importance of Gerard’s *Herball* in the history of discovery of novel bioactive compounds from biodiversity cannot be overstated. His description of the medicinal properties of *Filipendula*—now called *Spirea*, a genus in the rose family—led to the isolation of salicin and the eventual synthesis of “A *Spirea n*” or aspirin. Another example is Gerard’s entry on page 646, “of Foxe gloves”:

Foxe gloue boiled in water or wine, and drunken, doth cut and consume the thicke toughness of grosse and slimie flegme and naughtie humours; it openeth also the stopping of the liver, spleene and milt of other inward parts.

Gerard’s recording of foxglove opening the stopping of “other inward parts” was not systematically examined for 200 years. In 1785, William Withering published “An Account of the Foxglove and Some of Its Medical Uses, &c.,” in which he quoted Gerard’s account of the “vertues” of foxglove and proposed that foxglove could be an important medicine for dropsy, an ailment caused by the retention of fluid due to inadequate pumping by the heart. The connection between dropsy and heart disease was not properly understood in Withering’s day, but Withering observed the action of foxglove on the heart: “It has a power over the motion of the heart, to a degree yet unobserved in any other medicine, and ... this power may be converted to salutary ends.” By any standard, foxglove as administered by Withering was an astonishingly successful treatment for dropsy. J. K.

Aronson at Oxford recently reanalyzed data from Withering’s cases and found a success rate of between 65 and 80%.

The Linnaen name given for foxglove, *Digitalis*, was affixed to the crude drug as well as to cardiac glycosides (steroidal compounds with sugars attached at the 3-position) isolated from foxglove in the early twentieth century. More than 30 cardiac glycosides have been isolated from dried foxglove leaves, including digitoxin and digoxin. These two drugs have never been synthesized and they are still extracted from dried foxglove leaves. Each year, more than 1500 kg of pure digoxin and 200 kg of digitoxin are prescribed to thousands of heart patients.

*Digitalis*, however, is but one drug inspired by Gerard’s *Herball*. Since publication of the *Herball*, 18 different pharmaceutical compounds have been isolated from plants that Gerard described, including 16 drugs prescribed today.

European herbals and folk medicine were not the only source of pharmaceuticals; drug discovery has been greatly enhanced by non-Western traditions such as Aryurvedic medicine in India. For example, *Rauvolfia serpentina* has long been used in India to treat snakebite, insomnia, and as a sedative for hyperactive children. In 1931, Indian chemists isolated a variety of molecules from the plant and later found that *Rauvolfia* powder lowered blood pressure. In 1949, R. J. Vakil published a clinical study of *Rauvolfia* in the *British Heart Journal*. Following up on Vakil’s research, Dr. Emil Schlittler and Hans Schwarz at CIBA extracted from *Rauvolfia* roots an alkaloid, reserpine, which at low doses shows strong activity in lowering blood pressure. CIBA soon introduced reserpine to commerce. Reserpine had a direct effect on the hypothalamus and recently has been prescribed in combination with other antihypertensive drugs such as hydralazine hydrochloride.

### III. MODERN CHEMICAL APPROACHES TO DRUG DISCOVERY

*Digitalis* was discovered through the same sequence of steps used in modern ethnobotanical drug discovery programs: (i) folk knowledge accumulates concerning possible pharmacological activity of a plant, (ii) the plant is used therapeutically by a healer, (iii) the healer communicates this knowledge to a scientist, (iv) the scientist collects and identifies the plant, (v) plant extracts are tested with a bioassay (a preliminary screen for desired pharmacological activity), (vi) a pure com-

pound is isolated by using the bioassay to trace the source of the activity in the plant extract, and (vii) the structure of the pure substance is determined.

This method of discovery and isolation of new drugs from plants persisted until the 1950s because most pharmaceutical research relied heavily on compounds derived from vascular plants, particularly plants used in folk medicine. Indeed, flowering plants and ferns (as opposed to microscopic organisms and fungi) have given rise to approximately 120 commercially sold drugs. Of the top 150 prescription drugs in the United States, 57% contain at least one compound derived from biodiversity, including plants, animals, fungi, and bacteria (Grifo and Rosenthal, 1997). One-fourth of all U.S. prescription drugs contain molecules derived from, or modeled after, naturally occurring molecules in higher plants (Duke, 1993), including reserpine, digitalis, vincristine, and many other compounds derived from plants used in traditional medicine.

In the late 1950s and early 1960s, the search for bioactive molecules expanded to the marine environment. Increased access to underwater breathing apparatus and small submersible craft made studies of novel bioactive chemicals produced by marine organisms far easier. In the underwater realm, most discoveries of novel bioactive molecules had no basis in folk knowledge, even though evolutionary processes similar to those that produced large chemical diversity in higher plants are also believed to be operant in the marine environment, particularly among sessile organisms such as corals, sponges, and tunicates. It is somewhat surprising that few indigenous cultures rely on marine organisms as part of their pharmacopoeias, but perhaps difficulties in harvesting marine resources in the absence of SCUBA technology and the unpredictability of capturing mobile marine organisms resulted in greater indigenous reliance on terrestrial plants.

Natural products derived from biodiversity may contribute to the search for new drugs in three ways: (i) by producing new drugs used in an unmodified state (e.g., vincristine from *Catharathus roseus* [Apocynaceae]), (ii) by providing chemical "building blocks" used to synthesize more complex compounds (e.g., the synthesis of oral contraceptives from diosgenin derived from *Dioscorea floribunda* [Dioscoreaceae]), and (iii) by indicating new modes of pharmacological action that allow complete synthesis of novel analogs (e.g., synthetic analogs of reserpine from *Raulvolfia serpentina* [Apocynaceae]).

Regardless of their origin, recent discoveries of potential new pharmaceuticals have been facilitated by increasingly sophisticated techniques of vouchering

and species identification, bioassay, fractionation, and structural elucidation of molecular entities.

### A. Vouchering

Proper vouchering in pharmaceutical research has sometimes been overlooked with disastrous effects because accurate identification of a species of plant or marine organism has been rendered impossible due to a missing or inadequate voucher specimen. The importance of adequate herbarium voucher specimens cannot be overstated: Any subsequent question or dispute concerning the identity of the species involved can be unequivocally settled by expert examination of a properly collected voucher specimen. Vouchers for fish and marine invertebrates are usually stored as dried or fluid-preserved (FAA or aqueous ethanol) specimens deposited in zoological or natural history museums. Specimens of algae, lichen, and vascular plants are preserved as dried specimens in herbaria. Because of their scientific value for the future (sophisticated bioassays of the future may require only micrograms of material), voucher specimens should be deposited and preserved in well-curated herbaria or museums with duplicate specimens deposited in geographically distant institutions, including institutions in the countries in which the material was collected. Addresses and contact information for herbaria throughout the world are listed in *Index Herbariorum*, published by the International Society of Plant Taxonomists and the New York Botanical Garden.

The collection number of the specimen should be used to label all subsequent pharmacological fractions and residues so that any new discovery or question can be immediately referred to the original specimen. Herbarium specimens should consist of the leaf, fruit, flower, and other plant parts necessary for proper identification by a botanist. Plants are pressed in newspaper and dried in the field using plant presses and a heater. When it arrives at the herbarium, the dried plant material is glued to high-quality rag bond paper and stored in files in herbarium cases. Properly stored, under the correct conditions, herbarium specimens can retain their scientific value for centuries. In the field, it is also important to take careful notes to accompany the voucher specimens, including precise information on the collection location and other information on the plant, such as floral odor or color, that might be lost during drying. The collector's name and date of collection should always be included. For plants or animals used in folk medicine, ethnobiological data on the diseases treated, mode of formulation, and methods of



administration should be sufficient to guide subsequent investigations.

In preparation of samples for pharmaceutical testing, drying is used to stop enzymatic processes of degradation: most cease when water content is reduced to 10% by volume. Rapid drying is preferred, although high temperatures can alter pharmaceutical constituents. Air drying at 20–40°C is preferred, but sometimes in the moist tropics temperatures must be raised far higher to promote drying of bark or roots. With access to the laboratory, freeze-drying of a sample is an attractive alternative, whereas under field conditions in the humid tropics, fluid preservation in aqueous ethanol or formalin is often necessary to prevent spoilage of the sample.

## B. Bioassays

Bioassays vary with regard to type and precision. *In vivo* bioassays involve administering the test substance to a living organism to determine the substance's pharmacological activity. One of the more common *in vivo* bioassays that has been used to great advantage by some investigators is a brine shrimp bioassay, in which the test substance is administered to a culture of brine shrimp, available from any aquarium supplier. The percentage death of the brine shrimp within a short period is recorded, with brine shrimp mortality yielding a crude measure of cytotoxicity. Other *in vivo* bioassays include the Hippocratic screen, a test protocol for oral or peritoneal administration of the test substance in rats or mice. A variety of different behavioral and physiological parameters are measured before and after administration, allowing type and mode of pharmacological activity to be adduced. The most important *in vivo* bioassays for pharmaceutical research are human clinical trials, which are almost always administered after reasonable expectations of safety and efficacy have been determined from animal trials. The power of *in vivo* assays is their ability to reveal unexpected or novel modes of pharmacological activity since the impact of the test substance on the whole organism can be deduced.

More common in modern pharmaceutical use, though, are *in vitro* assays—bioassays that do not depend on a living organism. Although some broad-based *in vitro* assays depend on living tissues or cells, such as the guinea pig ileum, frog sciatic nerve, or human cancer cell lines, far more common is the use of target-specific molecular assays. In such assays, a single point of action in a given biochemical pathway, such as the role of the enzyme phospholipase A<sub>2</sub> in mediating cellular inflammation, has already been identified. Various

substances are then screened to determine if they impact the enzyme, receptor, site, or rate-limiting step of interest. Modern *in vitro* assays are typically costly (given the immense research effort required to identify the best target along a given biochemical pathway), proprietary, and designed for high-throughput rapid testing of numerous samples for the desired pharmacological activity.

In a typical high-throughput bioassay, minuscule portions of a given sample are tested for an enzymatic color change when combined with the test system. These bioassays test by screening, using a computer-controlled camera, hundreds and even thousands of different samples deposited in the microwells of a glass dish that are manipulated by robotic arms. The advantages of modern high-throughput bioassays are their speed, efficiency, repeatability, and specificity. The number of false positives (initial indications of pharmacological activity that do not pan out in subsequent testing) in modern receptor-binding assays is low; however, such high specificity entails a research liability since (unlike tests on living animals) a molecular bioassay is unlikely to reveal any mode of pharmacological action other than the specific activity it is designed to detect.

The use of precise high-throughput bioassays designed by molecular biologists has resulted in some spectacular successes, such as the discovery of the cholesterol-lowering drug Zocor or the new generation of anti-inflammatory drugs based on COX-2 inhibition. However, such screens almost certainly miss many important modes of action. It is unlikely, for example, that the psychotropic action of LSD-25 or the use of Viagra to treat male impotency would have ever been discovered had their effects not been noted through inadvertent administration to living human beings.

## C. Fractionation

Once a "hit" or pharmacological activity is detected in a test substance during a bioassay, the substance, particularly if it is a natural product of unknown chemical composition, must undergo fractionation. During fractionation, solvents of different polarity are used to produce increasingly refined chemical samples of the test substance. Each fraction is then tested against the original bioassay to determine its bioactivity. Fractions exhibiting significant bioactivity are further fractionated until eventually a purified compound is obtained. High-pressure liquid chromatography, gel electrophoresis, and other chemical techniques are also employed in addition to solvent fractionation to purify samples.

This approach to isolating pure bioactive molecules by tracing bioactivity through the purification process is called bioassay-guided fractionation. Typical yields of pure bioactive molecules under research settings are low: Depending on the nature of the active molecule, 1 kg of dried plant material can yield as little as 1–10  $\mu\text{g}$  of pure compound. Thus, on detection of a hit in a bioassay, recollection of a bulk sample (10–40 kg.) is often required to produce sufficient pure substance for structural determination.

#### D. Structural Determination

Once a pure bioactive molecule is obtained through fractionation or chromatographic techniques, the next step is to determine its molecular structure. A definitive technique for determining the structure (if the purified molecule can be crystallized) is X-ray crystallography, but more common techniques involve the use of high-pressure liquid chromatography, gas chromatography (GC), mass spectrometry (MS), or, for novel structures, nuclear magnetic resonance (NMR) spectroscopy. Most modern GC-MS machines have a computer-based library documenting retention time of known molecules: Observed peaks on the GC-MS are quickly and easily compared to those in the computer library. This technique is particularly useful for forensic toxicologists who do not typically expect to find novel molecules in their work. Structural determination in most biodiversity studies, however, relies on NMR since novel rather than known structures are often discovered during the course of pharmaceutical research. The frequencies and resolution capabilities of NMR machines differ, as do the types of experiments possible on NMR equipment, but of particular interest to natural product investigators is the INADEQUATE experiment that, through supercomputer analysis, can resolve structures for molecules in substances that are parts of mixtures rather than pure samples.

### IV. RATIONAL SCREENS OF BIODIVERSITY

Sometimes, such as in the case of penicillin, discovery of new pharmaceuticals from biodiversity is the result of serendipity. More often, though, success in natural product research is the result of carefully designed screening processes. Such planned search strategies to select appropriate organisms for study are termed “rational screens,” of which there are several varieties: random, phylogenetic, ecological, and ethnobotanical.

#### A. Random Screens

Random screens, despite their name, involve a great deal of planning and design. To be successful, they must have a clearly defined disease target and a clear and unambiguous bioassay. One of the best examples of a random screen in recent times is the effort of the Natural Product Branch of the U.S. National Cancer Institute (NCI) to evaluate plants and animals for anti-cancer activity. Using a variety of *in vitro* cancer cell cultures as bioassays, the NCI screened more than 35,000 plant accessions and 6000 marine organisms for activity against human cancers. Although the screen was not truly random in that it was designed to sample a wide range of geographical and taxonomic diversity, it allowed the NCI (with a variety of contracting collecting institutions) to assemble a large archive of materials that could be retested as bioassays continued to evolve. The geographical reach of the NCI was truly broad and led to pioneering efforts to negotiate international protocols for biodiversity prospecting. The NCI “letter of intent” which collaborators used to obtain collecting permits in foreign countries pledges equitable sharing of data and economic benefits with host countries. This letter has emerged as a model for recent international agreements.

Although the low hit rate of the NCI random screens led to criticism of the approach, few can argue with the stellar success of the NCI program in producing what may be one of the most important chemotherapeutic agents discovered in the latter part of the twentieth century, taxol, which was initially isolated from the Pacific yew tree (*Taxus brevifolia*) and which has been licensed by the NCI to Bristol Meyers and approved by the Food and Drug Administration (FDA) for the treatment of ovarian and breast cancer. However, because of the low success rates and high costs of collecting, random screens are currently suitable only for large institutions, even given the advent of high-throughput bioassays.

#### B. Phylogenetic Screens

Phylogenetic screens involve the pharmacological testing of related groups of organisms. Increased precision in elucidating phylogenies, largely due to rapid computer programs for cladistic analysis and the advent of molecular techniques for phylogenetic determination, facilitates the identification of relatives of any species showing pharmacological value. Phylogenetic screens, albeit in a crude sense, have long been utilized. For example, plants in the Apocynaceae, or milkweed family,

have always merited special attention due to the family's abundance of alkaloid-producing species such as *Catharanthus*, which produces the antileukemia drug vincristine. Only recently have modern techniques encouraged investigators to study close relatives of a species for either (a) increased abundance of an important bioactive compound (such as species of *Taxus* for the presence of taxol) or (b) natural homologs of known pharmaceuticals (such as species of *Catharanthus* that may produce variant forms of vincristine or vinblastine).

### C. Ecological Screens

An ecological screen depends on the environmental setting of the organism for clues as to its possible pharmaceutical value. Recently, such screens have focused on extremophiles—organisms that survive under extreme conditions of high temperatures, acidity, alkalinity, or salinity. For example, the polymerase chain reaction, invented by Kary Mullis at Cetus Corporation and widely used for genetic studies, medical diagnosis, or gene therapy, depends on an enzyme, Taq polymerase, derived from a bacterium discovered in the thermal springs of Yellowstone National Park by Thomas Brock of the University of Wisconsin. The bacterium that produces the enzyme, *Thermus aquaticus*, can withstand temperatures up to 95°C, and its enzymes are similarly heat resistant, an important feature for the rapid thermal cycles used in the polymerase chain reaction.

Ecological screens have also been conducted underneath the sea near hydrothermal vents where temperatures can reach in excess of 350°C. The enzyme Pfu polymerase, derived from the Archeum *Pyrococcus furiosus* and discovered by Karl Stetter of the University of Regensburg during exploration of undersea thermal vents, is of increasing interest because of its precision in polymerase chain reactions due to a built-in DNA repair mechanism. *Pyrococcus furiosus* can survive temperatures in excess of 100°C and is resistant to high pH and radioactivity. Enzymes derived from other organisms that survive conditions of extreme cold, alkalinity, acidity, or salinity are of increasing interest to the biotechnology industry (Madigan and Marrs, 1997).

Other ecological screens focus on searches for antifungal compounds from plants that grow in moist tropical environments (successes in this area have been achieved by Alice Clark at the University of Mississippi) as well as searches for inexpensive molluscicidal compounds useful in the fight against schistosomiasis. Kurt Hostettmann at the University of Lausanne discovered saponin-like compounds in several African tree species that show intense toxicity to schistosomiasis-carrying

snails, and he is investigating the possibility of supplying villagers with seeds and the simple techniques to prepare the crude solutions of these compounds which could be used to dose village bathing and washing areas.

### D. Ethnobotanical Screens

An ethnobotanical screen is based on the belief that indigenous peoples through generations have accumulated useful information about pharmaceutically active plants and animals. Since ethnobotanical leads have resulted in the discovery of the active compounds for 25% of all prescription drugs, there is little question as to the efficacy of ethnobotanical approaches to drug discovery. However, although there are strengths to ethnobotanical screens, there are also significant limitations.

In general, cultures with three characteristics seem most likely to have discovered pharmacologically active plants: (i) residency within an area of high biodiversity, (ii) an extended history of residence within the area, and (iii) a cultural mechanism for accurate transmission of ethnomedical information from generation to generation. Under the first criterion, the Kayapo Indians of Brazil merit more attention than the Alyuts in Alaska because of the greater biodiversity of tropical Brazil. Under the second criterion, the Aboriginal peoples of North Arnhem land who have been resident in Australia for thousands of years would be of more interest than the Pitcarin islanders who arrived in their island home only a few centuries ago. Under the third criterion, the Samoan islanders, with a precise matrilineal system of transmitting healing knowledge from generation to generation, would be of more interest than faith healers in the Philippines, who depend on dreams for their knowledge of healing plants.

In an ethnobotanical screen, healing information is recorded by a linguistically adept ethnobotanist trained in both the techniques of ethnographic interviews and field botany. Often, there is little overlap between indigenous and Western disease concepts—for example, there is no Western equivalent of "susto" (a type of soul loss with diverse symptoms) which afflicts indigenous peoples in Central America—and so the ethnobotanist, often in working with a physician, must be adept at mapping signs and symptoms of indigenous disease states into concepts of illness understood by Western peoples. Those plants identified by the indigenous peoples as containing healing properties are collected and identified, with samples prepared for bioassay-guided fractionation. In addition, careful notes about possible adverse reactions, dosages, preparation, and application

techniques are carefully prepared to assist other members of the drug discovery team. Although the ethnobotanical approach seems to work well for diseases (such as skin fungus or acute viral syndrome) that are easily recognizable by indigenous peoples, other diseases, such as brain tumors or lymphoma, which require advanced diagnostic techniques for diagnosis, show less hope of being successively treated by drugs discovered during ethnobotanical research.

There are different approaches to ethnobotanical screens. In the consensus approach, developed by Brent Berlin at the University of Georgia, numerous villagers are interviewed with regard to their knowledge of healing plants. Those medicinal plants that are repeatedly mentioned in interviews are prioritized for testing due to their high saliency in the culture. A variant of this approach is to prioritize for study those plants used in common by different villages or even cultures; an ethnobotanical screen based on the latter approach was successfully used by Shaman Pharmaceuticals in the discovery of Provir, an antiviral drug undergoing phase II clinical trials.

A different approach to consensus techniques, developed by Michael Balick of the New York Botanical Gardens and Paul Cox of the National Tropical Botanical Gardens in Hawaii, relies on specialist knowledge known only to healers rather than on knowledge held in common by villagers. In this approach, the healers who are most highly regarded by their societies, who have the greatest knowledge of diseases, and who use the largest repertoire of plant species in their therapeutic practices are asked to rank the efficacy of their healing plants. In a careful trial in Belize conducted in conjunction with the NCI, Balick found that such "powerful plants" identified by healers had a far greater likelihood of generating hits in bioassays.

## V. NATURE OF BIODIVERSITY-DERIVED PHARMACEUTICALS

### A. Microbial Products

In 1929, Alexander Fleming discovered that a petri dish of *Staphylococcus* had become contaminated with a *Penicillium* fungus: around the fungus in the culture medium was a zone resistant to *Staphylococcus* growth. During World War II, the antibiotic penicillin was isolated. This led to dramatic treatment of infections from wounds and other trauma. Since that time, more than 6000 different antibiotics have been isolated, with more than 1000 having been commercially produced (Sam-

uelsson, 1992). Only a single "natural" penicillin has ever been prescribed as a drug, but more than 100 semisynthetic penicillins have been produced by adding other carboxylic acids to the basic penicillin structure (Sameulsson, 1992). The structure of penicillin, a peptide with a thiazolidine ring, is similar to the structure of the cephalosporins, in which the thiazolidine ring is substituted by a dihydrothiazine ring. *Bacillus brevis* and *Bacillus licheniformis* have been used to produce gramicidin and bacitracin, respectively.

Many new antibiotics and other pharmaceutical products have been obtained from soil samples which are carefully cultivated in sterile petri dishes and fermentation vats. One of the more successful products to be produced from a soil culture is cyclosporin, a cyclic peptide produced from the fungus *Cylindrocarpum lucidum*, originally derived from a Norwegian soil sample. Cyclosporin, a powerful immunosuppressant, was initially seen to have little therapeutic value and remained archived as a refrigerated fungal culture at Sandoz Pharmaceuticals (now Novartis) in Basel, Switzerland until mycologist Michael Dreyfuss convinced management that the product had utility. Today, nearly every heart or kidney transplant patient in the world takes a daily dose of cyclosporin to prevent rejection of the transplanted organ.

Soil samples continue to yield new and interesting bacteria and fungi, whereas exploration of extreme habitats promises to yield important new useful compounds, including those produced by Archaea, distant relatives of bacteria. Using recombinant DNA technology, microorganisms can be employed to produce pharmaceutical compounds which they do not ordinarily produce in nature. Genetic engineers can insert DNA sequences into microorganisms to produce a staggering array of pharmaceutically active compounds. Although most researchers hail such genetically modified organisms as a significant advance, there are public concerns, particularly in Europe, about the safety of food and pharmaceuticals produced by genetically modified organisms.

### B. Marine Organisms

Marine organisms, particularly invertebrates, are known to produce extraordinarily toxic compounds. Currently, three compounds isolated from marine resources are in preclinical development as possible anticancer agents by the NCI. The marine algal species *Portieria hornemaii* from the Philippines has yielded a drug called halomon, a species of the sponge genus *Lissodendoryx* from New Zealand has produced hali-

chondrin B, and a Caribbean tunicate *Ecteinascidia turbinata* has yielded ecteinascidin 743. However, perhaps the most promising pharmaceutical product derived from marine biodiversity is bryostatin, a macrocyclic lactone derived from the Californian bryozoan *Bugula neritina*. Bryostatin has been evaluated in phase II clinical trials in which it has been found to inhibit melanoma, reticulum cell sarcoma, and lung carcinoma. Initial results with human patients have been encouraging but not spectacular. Nevertheless, because of its unique structure and immunomodulatory effects, it is likely that the discovery of bryostatin will lead to new therapeutic concepts in chemotherapy. Currently, 6000 marine samples are archived at the NCI for testing, and there is continuing industrial interest in natural products derived from marine organisms.

### C. Terrestrial Animals

Terrestrial animals have historically played an important role in the production of certain therapeutic hormones and steroids. Progesterone was first isolated from pig ovaries in 1934. Animals were eclipsed by plants, however, as a source of steroidal precursors when in 1940 Russell Marker discovered that Mexican yams of the genus *Dioscorea* could be utilized to produce diosgenin. There has also been a long-term research interest in venoms produced by snakes as sources of pharmaceuticals. Ancrod, a drug used to treat circulatory diseases, is isolated from the venom of the Malayan pit viper *Agkistrodon rhodostoma*. Recent research, such as that conducted by Francis Markland at the University of Southern California, suggests that proteins such as conotoxins found in certain snake venoms may be able to fight cancer tumors. The poisonous secretions of some toads and frogs contain bufotalin, which was traditionally used as a treatment for dropsy before Withering's discovery of digitalis. Ethnozoologist Wade Davis has also found that secretions of species of *Bufo* were used as psychotropic substances by indigenous peoples in Central America. Epibatidine, from the poison dart frog, is being developed as a possible analgesic. It is likely that other venoms and stings will produce new pharmacological insights.

### D. Terrestrial Plants

Plant-derived drugs are used for a broad spectrum of diseases and include quinine (for malarial suppression), digitalis (for treatment of rapid atrial fibrillation), vincristine (for treatment of pediatric leukemia), tubocurarine (a muscle relaxant used in anesthesia), pilocarpine

(used for the treatment of glaucoma), and  $\gamma$ -strophanthin (used for congestive heart failure) (Table I). However, in the latter quarter of the twentieth century, very few plant-derived drugs were released to market, even though there was an increase in consumer interest in North America, Europe, and Japan in "natural" plant medicines. Given that less than 1% of the world's plants have ever been carefully studied for pharmacological activity, it is puzzling that pharmaceutical firms are not rigorously investigating plants as the sources of new pharmaceuticals.

Flowering plants and ferns have traditionally provided the bulk of biodiversity-derived pharmaceuticals, but recent advances in molecular biology and combinatorial chemistry have reduced the pharmaceutical industry's ardor in studying plants. The popular mythology of industry-supported botanists combing the rain forests for new plant-based leads is simply not true: currently, there is not a single major pharmaceutical firm that lists among its employees a PhD botanist or ethnobotanist who works in a drug-discovery program based on plants, although some firms such as Pfizer and Merck have contracted with other institutions to conduct plant surveys. Despite the lack of interest by the pharmaceutical industry, both the promise and the impact of plant-derived pharmaceuticals is profound. The World Health Organization estimates that 85% of the world's population depends directly on plants for medicine, and more than 25% of current prescription drugs have at least one active component derived from a flowering plant. The former statistic, derived principally from developing countries, has contributed significant opportunities to contribute to world health by assisting different nations to evaluate the safety and efficacy of their own pharmacopoeias. Currently, the People's Republic of China, Mexico, Thailand, and Nigeria have decided at the national level to incorporate traditional plant-based medicine directly into primary health care. An example of the possibilities inherent in this approach has been demonstrated in Thailand, where a potent anti-inflammatory compound from the beach plant *Ipomoea pes-caprae*—long used by fisherman to treat stings from the Portugese man-of-war—was isolated and purified using bioassay-guided fractionation and its structure determined using NMR spectroscopy. However, rather than testing a pure compound, the Thai researchers conducted a careful, double-blind controlled study of a crude tincture of the plant. It was found to be both safe and efficacious. The resultant tincture can now be bought in any Thai drugstore for a price far less than that of a synthetic drug.

There are, however, several plant-based pharmaceu-

TABLE I  
Fifty Drugs Discovered from Ethnobotanical Leads

Drug	Medical use	Plant source
Ajmaline	Heart arrhythmia	<i>Rauvolfia</i> spp.
Aspirin	Analgesic, inflammation	<i>Spiraea ulmaria</i>
Atropine	Ophthalmology	<i>Atropa belladonna</i>
Benzoin	Oral disinfectant	<i>Styrax tonkinensis</i>
Caffeine	Stimulant	<i>Camellia sinensis</i>
Camphor	Rheumatic pain	<i>Cinnamomum camphora</i>
Cascara	Purgative	<i>Rhamnus purshiana</i>
Cocaine	Ophthalmic anesthetic	<i>Erythroxylum coca</i>
Codeine	Analgesic, antitussive	<i>Papaver somniferum</i>
Colchicine	Gout	<i>Colchicum autumnale</i>
Demecolcine	Leukemia, lymphomata	<i>Colchicum autumnale</i>
Deserpidine	Hypertension	<i>Rauvolfia canescens</i>
Discoumarol	Thrombosis	<i>Melilotus officinalis</i>
Digoxin	Atrial fibrillation	<i>Digitalis purpurea</i>
Digitoxin	Atrial fibrillation	<i>Digitalis purpurea</i>
Emetine	Amoebic dysentery	<i>Cephaelis ipecacuanha</i>
Ephedrine	Bronchodilator	<i>Ephedra sinica</i>
Eugenol	Toothache	<i>Syzygium aromaticum</i>
Gallotannins	Hemorrhoid suppository	<i>Hamamelis virginiana</i>
Hyoscyamine	Anticholinergic	<i>Hyoscyamus niger</i>
Ipecac	Emetic	<i>Cephaelis ipecacuanha</i>
Ipratropium	Bronchodilator	<i>Hyoscyamus niger</i>
Morphine	Analgesic	<i>Papaver somniferum</i>
Noscapine	Antitussive	<i>Papaver somniferum</i>
Papain	Attenuate mucous	<i>Carica papaya</i>
Papaverine	Antispasmodic	<i>Papaver somniferum</i>
Physostigmine	Glaucoma	<i>Physostigma venenosum</i>
Picrotoxin	Barbiturate antidote	<i>Anamirta cocculus</i>
Pilocarpine	Glaucoma	<i>Pilocarpus jaborandi</i>
Podophyllotoxin	Condylomata acuminata	<i>Podophyllum peltatum</i>
Proscillaridin	Cardiac malfunction	<i>Drimia maritima</i>
Protoveratrine	Hypertension	<i>Veratrum album</i>
Pseudoephedrine	Rhinitis	<i>Ephedra sinica</i>
Psoralen	Vitiligo	<i>Psoralea corylifolia</i>
Quinidine	Cardiac arrhythmia	<i>Cinchona pubescens</i>
Quinine	Malaria prophylaxis	<i>Cinchona pubescens</i>
Rescinnanmine	Hypertension	<i>Rauvolfia serpentina</i>
Reserpine	Hypertension	<i>Rauvolfia serpentina</i>
Senoside A, B	Laxative	<i>Cassia angustifolia</i>
Scopolamine	Motion sickness	<i>Datura stramonium</i>
Stigmasterol	Steroidal precursor	<i>Physostigma venenosum</i>
Strophanthin	Congestive heart failure	<i>Strophanthus gratus</i>

continues

Continued

Drug	Medical use	Plant source
Teniposide	Bladder neoplasms	<i>Podophyllum peltatum</i>
THC	Antiemetic	<i>Cannabis sativa</i>
Theophylline	Diuretic, asthma	<i>Camellia sinensis</i>
Toxiferine	Surgery; relaxant	<i>Strychnos guianensis</i>
Tubocurarine	Muscle relaxant	<i>Chondrodendron tomentosum</i>
Vinblastine	Flodgkin's disease	<i>Catharanthus roseus</i>
Vincristine	Pediatric leukemia	<i>Catharanthus roseus</i>
Xanthotoxin	Vitiligo	<i>Ammi majus</i>

tics that have been recently released or are under current preclinical or clinical evaluation. Taxol, discovered during the NCI random screen, has received FDA approval for treatment of breast and ovarian cancer. For AIDS, currently five plant-derived compounds are in preclinical development. Derived from the resin of *Calophyllum langierum* in Malaysia, calanolide A has shown intense activity against HIV-1, as has costatolide derived from a related tree *Calophyllum teysmanii* from the same region. Michellamine B from the leaves of the Cameroon vine *Ancistrocladus korupensis* has also shown anti-HIV activity, as has conocurvone derived from *Conospermum* shrubs found in Western Australia. Because of its use in indigenous medicine, the fifth member of this anti-AIDS quintet, prostratin, perhaps deserves in-depth discussion as a case study of recent developments in this field.

For many years, Samoan healers have used the stem wood of a small understory rain forest tree, *Homalanthus nutans*, to treat a disease known as "fiva samasama," characterized by yellowing of the eyes, dark urine, jaundice, and fevers (determined to be hepatitis). To prepare the remedy, the wood of *Homalanthus* is macerated into a clean cloth, which, like a tea bag, is immersed into boiling water. After steeping, the contents of the tea bag are discarded, and the resultant tea is drunk by the patient. Samples of the healer preparations and the plants were tested against an *in vitro* HIV-1 bioassay in a collaboration between Paul Cox and NCI researchers Kirk Gustafson, John Cardellina, John Beutler, Peter Blumberg, Gordon Cragg, Michael Boyd, and others at the Natural Products Branch of the NCI (Gustafson *et al.*, 1992). The healer mixture and the crude plant extracts were found to protect cells from death by HIV-1, even though there was no evidence of reverse transcriptase inhibition or other marked inhibition of the mixtures to the virus. Bioassay-guided fractionation

yielded a pure compound, prostratin. Although phorbols similar in structure to prostratin are potent tumor promoters, *in vivo* studies by Peter Blumberg at the National Institutes of Health show prostratin to be an antipromoter, even though it activates protein kinase C. The NCI has advertised prostratin as a potential new component of a combination therapy for AIDS. The NCI will require any pharmaceutical firm which desires to develop the compound to negotiate directly with the Samoan government for a fair and equitable return of a portion of any royalties. In addition, Brigham Young University honored an agreement negotiated between Cox and the Falealupo village in Samoa prior to discovery, promising a minimum of 30% of any royalty income from prostratin to be returned to the village. However, to date there has been little interest by the pharmaceutical community in prostratin. Although this is primarily due to the approval of effective proteases and reverse transcriptase inhibitors, the fact that prostratin is a phorbol and its discovery was linked to traditional medicine contributed to the wariness of pharmaceutical firms. However, as resistance to reverse transcriptase inhibition and other antiviral strategies continues to evolve in the AIDS virus, compounds with a cellular mode of action such as prostratin may gain the interest of the pharmaceutical industry, as may other compounds with broad-spectrum antiviral activity. Thus, although prostratin currently languishes in the limbo of promising drug leads that have not been clinically investigated, in the future it may be developed as a potent antiviral compound.

## VI. BIODIVERSITY AND ETHICS

### A. The Convention on Biodiversity

The Convention on Biodiversity, often referred to as the Rio Treaty, provided a broad basis in international law for the discovery and development of pharmaceutical compounds from biodiversity. In order to promote the conservation and study of biodiversity, the parties to the Convention on Biodiversity agreed that each nation has sovereignty over its own biological resources. Collection and development of these resources can thus proceed only with express permission of the national government. What initially was seen as a mechanism to accelerate investigation and conservation of biological resources by facilitating international cooperation has been criticized by some as an impediment to research. Many countries, fearing uncompensated exploitation of their national biological resources, have either closed

their doors to international scientific exploration or caused the process of obtaining research permits to be extraordinarily slow and difficult. Given the history of colonial exploitation of biological resources, with little thought given to equitable sharing of benefits of discoveries based on biodiversity, such reaction by developing countries is understandable. If the processes of biodiversity were static and extinction quiescent, closing national boundaries to scientific investigation would have little long-term consequence; however, given the rapid rate of extinction (the International Union for the Conservation of Nature estimates that one-eighth of the world's plants are threatened or endangered), nations that close their doors to scientists may unwittingly ensure that their biodiversity treasure will vanish without any significant appraisal of its economic value. An interesting exercise would be to compare the relative difficulties of obtaining permits to collect and pharmacologically analyze 1-kg plant samples from various countries versus the difficulties in obtaining permits to clear-cut and export entire rain forests: too often, those who destroy biodiversity are facilitated while those who study it are hindered.

Given the relative lack of interest of the pharmaceutical industry in investigating biodiversity as a source of new pharmaceuticals, and the occasionally unreasonable expectations placed on pharmaceutical firms for large initial payments and high percentages of royalties, many pharmaceutical firms that once had nascent interest in natural product chemistry have retreated to computer modeling and combinatorial chemistry as the sole engines of pharmaceutical discovery. Although this situation may change if new, important drugs are discovered from biodiversity, it is crucial that investigators adhere to high ethical standards. Training local scientists, obtaining informed consent from village chiefs or elders prior to initiating research, depositing duplicates of all collected specimens with local or regional herbaria, and vigorously protecting indigenous intellectual property help promote understanding and collaboration. Investigators who fail to heed such standards or, worse, ignore local laws and international treaties imperil not only their own research programs but also conservation of the very biodiversity they purport to study.

An interesting, and perhaps unintended, consequence of the Convention on Biodiversity has been its impact on the rights of indigenous peoples. By granting sovereignty over all biological resources to national governments, the convention in a sense can disenfranchise indigenous peoples, particularly in countries with a history of mistreating them. Is it reasonable that a plant

that has been developed and studied through generations of indigenous healers should be considered the sole property of a distant (and perhaps hostile) national government? However, by stating in article 8(j) that indigenous peoples bear some consideration, the architects of the Convention on Biodiversity provided indigenous peoples with one of the first broadly accepted international treaties that explicitly recognizes their existence and aspirations. Article 8(j) requires signatory nations to (i) respect, preserve, and maintain traditional knowledge; (ii) promote wide application of traditional knowledge; and (iii) encourage equitable sharing of benefits from traditional knowledge. Many indigenous groups hope that their voices will increasingly be heard and considered at meetings of the Parties to the Convention.

### B. Bilateral Agreements on Pharmaceutical Research

There have been attempts to directly negotiate with national governments for rights to perform pharmaceutical research on biodiversity. One of the best known of these arrangements was a remarkable agreement entered into by the government of Costa Rica and Merck Pharmaceuticals. Under the IMBIO agreement, Merck received rights to survey the flora and fauna of Costa Rica in return for a share of the proceeds of any new discovery and for a payment of \$1 million used to support the training of local investigators and development of a national inventory of Costa Rican biodiversity. Although no discoveries have yet been announced from Merck, the Costa Ricans have been very innovative in training local people in biological survey techniques. Termed "parataxonomists," these local people, often with profound folk knowledge of organisms in their environment, are trained in proper techniques of collection and vouchering. By devoting the bulk of the Merck funds to biodiversity research, the government of Costa Rica has guaranteed that the people of Costa Rica and indeed the world will benefit for many years.

### C. Indigenous Intellectual Property Rights

In ethnobotanical approaches to drug discovery, it is important that the contributions of indigenous people be appropriately recognized and compensated. Often, indigenous values, particularly views of fairness and equability, differ dramatically from those of Western societies. Many Western societies highly value individualism, rationalism, and candor in expression, whereas

in many indigenous societies values such as respect for village elders, proper regard for ancestors and deities, and recognition of the sacred nature of living organisms play a crucial role. Frequently, indigenous peoples also have very distinct views about appropriate behavior and dress within a village.

It is important that Western investigators dealing with indigenous peoples show appropriate respect for indigenous values. Indigenous mores and folkways should not be violated by investigators, and the position of traditional leaders, chiefs, and village councils should be respected rather than eroded. This may mean, for example, that Western scientists do not publish biodiversity information considered secret by indigenous people or do not venture into areas which are considered to be taboo. Respect can be communicated by completely informing such leaders and councils about the scope, nature, and possible impact of biodiversity research. Informed consent of traditional leaders is crucial when discussions are held concerning fair and equitable returns for indigenous intellectual property: traditional leaders cannot make informed decisions concerning biodiversity research unless they possess accurate information concerning the nature and value of the research. Rather than imposing Western values of fairness and equability on indigenous peoples, investigators should regard indigenous peoples as coequals in any negotiations and seek solutions that meet indigenous rather than Western concepts of fairness and equability. Often, traditional leaders are not familiar with Western culture or conversant in Western languages, but the temptation to use Western-educated elites as surrogates for traditional leaders in negotiations should be avoided.

For many indigenous peoples, conservation is a keenly felt need, a means of both perpetuating their culture and demonstrating respect to both the earth and to their ancestors. Innovative approaches to assisting indigenous peoples in conservation can be successful if indigenous peoples truly control the conservation initiatives. The Terra Nova biomedical reserve in Belize or the Falealupo Rain Forest Reserve in Samoa are examples of indigenous conservation begun by Western scientists that are now completely controlled by indigenous peoples. Culturally appropriate acknowledgment of indigenous collaborators such as including publication of scientific abstracts and reviews in indigenous languages and coauthorship of papers and books with indigenous collaborators are examples of the best practices currently employed by ethnobotanists. Protection of indigenous financial interests often requires vigorous efforts on the part of Western scientists, who need



to assist indigenous leaders in dealing with lawyers, commercial firms, or governments that may not completely appreciate indigenous property rights. Entering into any collaborative project, including ethnobotanical research, with indigenous peoples imposes a responsibility on the part of scientists to protect indigenous rights. It should not be surprising, then, that some of the most persuasive Western advocates of indigenous rights have been scientists who have lived and worked closely with indigenous peoples.

### D. Rights of the Sick and Afflicted

In an era in which life itself increasingly appears to be a patentable commodity, it might appear quaint to raise the possibility that stakeholders in biodiversity research include others than sovereign nations, pharmaceutical firms, and indigenous peoples. However, moral rights are not always harmonious with property rights: Most societies, for example, believe that small children have an inalienable claim on their parents for support, even though children do not possess legal title to any of their parents' possessions, and in former eras the sick and afflicted were believed to have certain claims on healing plants and substances. Although it is unlikely in the modern market economy that the sick and afflicted will ever be granted a significant voice in biodiversity research, ill people remain often unseen but tangible stockholders in biodiversity-based pharmaceutical research. The sick and afflicted, and their families and societies that support them, stand to lose if the world's biodiversity is thoughtlessly destroyed before it can be evaluated for pharmaceutical potential.

### See Also the Following Articles

BIODIVERSITY AS A COMMODITY • BIOPROSPECTING • COEVOLUTION • ECOTOXICOLOGY • ETHNOBIOLOGY AND ETHNOECOLOGY • INDIGENOUS PEOPLES, BIODIVERSITY AND • PLANT SOURCES OF DRUGS AND CHEMICALS

### Bibliography

- Akerele, O., Heywood, A., and Synge, H. (1991). *Conservation of Medicinal Plants*. Cambridge Univ. Press, Cambridge, UK.
- Arber, A. (1953). *Herbals, Their Origin and Evolution*. Cambridge Univ. Press, Cambridge, UK.
- Balick, M. J., and Cox, P. A. (1997). *Plants, People and Culture: The Science of Ethnobotany*. Scientific American Library, New York.
- Battersby, A., and Marsh, J. (Eds.) (1990). *Bioactive Molecules from Plants*, CIBA Symposium No. 154. Wiley, Chichester, UK.
- Cox, P. A., and Balick, M. J. (1994). The ethnobotanical approach to drug discovery. *Sci. Am.* 270(6), 82–87.
- Grifo, F., and Rosenthal, J. (Eds.) (1997). *Biodiversity and Human Health*. Island Press, Washington, D.C.
- Gustafson, K. R., Cardellina, J. H., McMahon, J. B., Gulakowski, R. J., Ishitoya, J., Szallasi, Z., Lewin, N. E., Blumberg, P. M., Weislow, O. S., Beutler, J. A., Buckheit, R. W., Cragg, G. M., Cox, P. A., Bader, J. P., and Boyd, M. R. (1992). A non-promoting phorbol from the Samoan medicinal plant *Homalanthus nutans* inhibits cell killing by HIV-1. *J. Med. Chem.* 35, 1978–1986.
- Janick, J., and Simon, J. E. (Eds.) (1993). *New Crops*. Wiley, New York.
- Madigan, M. T., and Marrs, B. L. (1997). Extremophiles. *Sci. Am.* 276(4), 82–87.
- Nigg, H. H., and Seigler, D. (Eds.) (1992). *Phytochemical Resources for Medicine and Agriculture*. Plenum, New York.
- Prance, G., and Marsh, J. (Eds.) (1994). *Ethnobotany and the Search for New Drugs*, CIBA Foundation Symposium. Academic Press, London.
- Samuelsson, G. (1992). *Drugs of Natural Origin*. Swedish Pharmaceutical Press, Stockholm.



# PHENOTYPE, A HISTORICAL PERSPECTIVE ON

R. J. Berry  
*University College London*

---

- I. Background
  - II. The Making of a Phenotype
  - III. Formal Analysis
  - IV. Genetic and Phenotypic Variability
  - V. Phenotypes and Evolution
  - VI. Taxonomy
- 

## GLOSSARY

**allele** (originally *allelomorph*) A form of variant (sometimes called a mutation) of a gene.

**diploid** Individual who has two sets of chromosomes, usually received from different parents, in contrast to a **haploid** organism such as most microorganisms, the gametophyte stage of higher plants and some parthenogenetic forms (e.g., male ants and honeybees). It may have the same form or allele of a gene on both members of a chromosome pair (in which case it is homozygous) in different forms (alleles) (in which case it is heterozygous).

**gene** Historically, the inherited factor which determines a trait. Tends to be used somewhat loosely; more strictly represents a place or locus on the chromosomes which codes for a particular function.

**gene** (strictly *allele*) **frequency** The frequency of an allele in a population.

**genome** The total genetic composition of an individual.

**genotype** The allelic composition of an individual at a particular locus.

**mutation** Change in an allele, producing a different allele; rate of occurrence affected both physically

(especially by ionizing radiation) and by many chemicals. It may also refer to changes in a chromosome (involving duplication, deletion, or inversion of a segment).

**phenetics** The study of phenotypes, usually describing the grouping of organisms into taxa on the basis of estimates of similarity.

**phenotype** The appearance (function or behavior) of an organism.

---

*THE PHENOTYPE OF AN ORGANISM* is its actual form or appearance to an observer. Usually, the term is used to describe all the features that make up the organism, but it may be used to distinguish a particular characteristic, such as whether a mouse is black or brown. In this case, the adjective refers to a specific characteristic of a mouse, implying that there is variation between individuals or groups of mice. A phenotype commonly refers to the physical appearance of an organism (e.g., a large brown mice with a short tail), but it can also describe nonvisual properties of the organism, such as physiology or behavior.

## I. BACKGROUND

Phenotypes are the appearance and properties of real animals and plants—living and dying, reproducing and bearing fruit, and succeeding or failing—but in themselves they are the products of interactions between the inherited (genetic) material and the environment, both

before and after birth. The distinction between genotype and phenotype is one of the more important ones in biology: A phenotype may be produced by several different genotypes, whereas a genotype may manifest as several different phenotypes, particularly when reacting with different environments.

The distinction between inherited constitution and external appearance (i.e., between genotype and phenotype) was one of the most important demonstrations of Gregor Mendel (1822–1884) in his series of breeding experiments with peas which laid the foundation for genetics, and in which he recognized that a character may be inherited in a dominant or a recessive manner. An organism manifesting a dominantly inherited trait (in Mendel's case, round as opposed to wrinkled pods or colored as opposed to white flowers; more familiar examples are brown versus blue eyes in humans or brown coat color versus albino in rabbits or mice) may carry both alleles for the dominant trait or one for the dominant and one for the recessive trait. That is, it will have the same phenotype but could be genetically homozygous or heterozygous. A distinction between genes and their manifestation was also implicit in August Weismann's (1834–1914) embryological division between germplasm (which gives rise to reproductive cells) and soma (or body). However, the formal nomenclature phenotype and genotype was devised by the Danish botanist Wilhelm Johannsen (1857–1927), who introduced the word gene for the material basis of an inherited character, and thence the terms genotype and phenotype.

Johannsen set out to investigate the relationship between (phenotypic) variation and selection. The rediscovery of Mendel's work in 1900 led to a rift between the biometricians [notably Karl Pearson (1857–1936) and W. F. R. Weldon (1860–1906)], who followed Darwin (1809–1882) in viewing small continuous variation as the raw material of evolution, and the "mendelists" (or geneticists) [led by William Bateson, (1861–1926)], who believed that large discontinuous saltations (or mutations) were the main cause of variation. The problem was the maintenance of continuous variation. Darwin had postulated a continuous replenishment of variation in order for selection to act, but he did not know its source. His proposal of "pangenes" was an effort to solve the problem. Weismann's demonstration of the early separation in development of the reproductive tissue from the rest of the body supported the general assumption at the time that continuous variation was produced by environmental effects. Intraspecific variation (i.e., subspecies or local races) was therefore regarded as environmentally caused; the species

was viewed as monotypic in a way that had much in common with the pre-Darwinian ideal or Linnean type.

Johannsen experimented with a self-fertilized cultivar of the bean *Phaseolus vulgaris*. The implication of this was that all the descendants of a single individual would have the same genes and constitute what Johannsen called a "pure line." Although individual beans might be different (due, for example, to their place in the pod), the mean and variance of all the characters of plants in a pure line were the same and were not affected by attempts at selection, i.e., plants grown from both small beans and large beans produced beans of the same average weight as that of all the plants in the pure line. In contrast, plants grown from crosses between pure lines had different (usually intermediate) characteristics from the parents, although these characteristics remained constant in pure lines derived from each cross. Johannsen argued that selection on continuous variation was inevitably ineffective, and the only variation on which selection could work depended on new mutation. This strengthened the contemporary assumption that evolution was bound up with mutations and their rate of occurrence, and that individuals were largely genetically uniform (i.e., homozygous at most loci for "wild-type" alleles).

This assumption became built into conventional population genetic theory so that when recessive traits manifested in inbred populations (usually in the laboratory or garden, but occasionally appearing under wild conditions) they were assumed to be recent mutations in the process of elimination by natural selection. Most mutations seemed to be deleterious to their carriers (which would be expected if they were random changes in a functioning organism), which meant that high mutation rates would inevitably impose a burden on a population. In 1950, at a time of acute concern about the genetic effects of atomic warfare, H. J. Müller (1890–1967) proposed the concept of "genetic load" and showed that a doubling of the mutation rate in a slow-reproducing species such as humankind could lead to extinction through genetically caused death.

The work of Bateson, Johannsen, and other early geneticists (one of the most influential was the Dutchman Hugo de Vries, 1848–1935) led to the isolation of genetics from evolutionary studies, particularly as represented by paleontology. This was resolved by the theoretical work of Sewall Wright (1889–1988), J. B. S. Haldane (1892–1964), and especially R. A. Fisher (1890–1962). Fisher published significant papers in 1918 and 1922 describing the expected biometrical properties of a Mendelian (i.e., breeding) population

and the effects of an allele substitution on a quantitative (i.e., continually varying) character. (The papers were published by the Royal Society of Edinburgh; the first one was rejected by the Royal Society of London on the advice of Karl Pearson and the geneticist R. C. Punnett.) He went on to argue that dominance (and recessivity) are traits that have evolved and this explains why the majority of new mutations are recessive, detrimental, and have a major effect (i.e., on the phenotype); most of the data available at the time were from laboratory breeding of *Drosophila melanogaster*. Fisher reasoned that there was no intrinsic reason for a mutation occurring for the first time to be either dominant or recessive; the greatest probability is that it will be intermediate, with an effect somewhere between its expression in double dose (i.e., when homozygous) and the unmodified condition. However,

1. Mutations occur repeatedly at virtually every locus. The rare (approximately 1 in  $10^{-5}$ ) mutational events we observe are recurrences of something that has happened thousands of times in the past.

2. When a mutation occurs, it will almost always be present in the heterozygous condition: If an allele has become relatively common in a population so that a fresh mutation to it has a reasonable chance of occurring in an existing heterozygote, then mutation cannot be the only force influencing its frequency.

3. If a newly arisen allele has a beneficial effect on its carrier, combinations of it with other alleles that increase its effect will have a higher fitness than any which decrease it. This will repeatedly occur so that the architecture of the species will become modified to the extent that any new occurrences of that advantageous allele will always produce the maximum effect in its carrier; this will almost always be in the heterozygous condition. In other words, beneficial characters will be selected for dominance and will also spread to replace the previous expression of the trait. Conversely, alleles which are deleterious in the heterozygote are only likely to be transmitted in combinations in which their effect is least, i.e., there will be selection for a small heterozygous effect (in the direction of recessivity).

Fisher put forward these ideas from first principles. They were received skeptically because of the difficulty in believing that selection pressures would be strong enough to allow genes which modify dominance to spread. In pre-1950 days primary selection coefficients were thought to be approximately 0.1 to 1.0% and second-order effects were believed to be much less. Haldane suggested that dominance was more likely to be

the effect of alleles with a biochemical or developmental margin of safety becoming the normal allele. Since they could exercise undiminished action when heterozygous, mutant alleles would be recessive and deleterious. However, Fisher's theory has been proven to be correct on many occasions, and although it may not apply for every allele at every locus it has considerable historical significance in bringing together paleontologists and geneticists, and it has relevance in highlighting an important factor influencing genetical architecture.

Fisher was the first to demonstrate the experimental modification of dominance by crossing domestic poultry with wild jungle fowl for five generations. This changed the inheritance of certain characters so that a degree of heterozygous manifestation occurred where complete dominance had previously prevailed. Fisher suggested that the dominance of the traits he studied had been attained during domestication as a result of selection for the more striking heterozygotes.

A more complete demonstration of the influences of modifying genes on dominance was performed by E. B. Ford (1901–1985) by breeding from the greatest and least expressions of a variable yellow variant (*lutea*) of the Currant Moth (*Abraxas grossulariata*). Although the difference between *lutea* and non-*lutea* can be regarded as caused by a single allelic difference, after only three generations of selection Ford produced heterozygotes virtually indistinguishable from the *lutea* homozygote in selection for the yellowest individuals and ones most like the typical homozygote in the white selection line. In other words, he had changed the heterozygote from a position of no (or intermediate) dominance to complete dominance or recessivity respectively. Ford then crossed his modified heterozygotes with unselected stock, and by the second generation (when the selected modifiers would have a chance of segregating independently) the original variable heterozygotes reappeared: He thus showed that it was the response of the organism rather than the gene that had changed.

Laboratory experiments of this nature have been carried out on a variety of organisms and a range of characters. An experiment particularly informative with regard to genetical architecture in the wild was carried out by H. B. D. Kettlewell (1907–1979) using British and Canadian peppered moths (*Biston betularia* and *Amphidasis cognataria*, respectively; these are fully interfertile). Although there is a melanic form (*swettaria*) in the North American species, it is comparatively restricted in its distribution, and only pale (or typical) moths occur over vast tracts of Canada.

When a melanic heterozygote and a typical homozygote of British origin are mated, the offspring are clearly dark or light: The melanic character is a straightforward dominant. Even when the typical moths come from Cornwall in extreme southwest England where melanics have never been reported, only a slight loss of complete dominance occurs and that only after several generations of crossing melanics back to Cornish stock (i.e., back-crossing a melanic from several generations of melanic  $\times$  Cornish cross with a "pure" Cornish parent). This modification consists of some white dots on the normally jet-black wings of the heterozygote. Perhaps significantly, this slight heterozygous expression of the gene gives specimens similar to those caught in the early days of the spread of peppered moth melanism in the mid-nineteenth century and which are now prized collectors' specimens. This white speckling has long since disappeared in wild-caught British specimens, and modern heterozygotes are indistinguishable from melanic homozygotes. Melanism has become fully dominant during the ensuing decades.

However, there has been no opportunity for dominance to evolve in Canadian peppered moths. When British melanics are crossed to Canadian stock (from areas where *swettaria* does not occur) the first-generation progeny segregate as dark or light in the same way as in a cross between British moths. In the first generation, the dominance modifiers in the British parent will be carried on the chromosomes in the same order as in British stock and produce dominance in the same way. In the next generation, the gametes contain chromosomes which have crossed over between the British and Canadian grandparents. Consequently, in the second generation, the "switch" between pale and black forms does not operate as efficiently. Kettlewell crossed heterozygotes from a British  $\times$  Canadian mating with Canadian moths and repeated this back-crossing for four consecutive generations, after which the heterozygotes ranged from black to pale—there was no sign of dominance. He then reversed the procedure and mated his "broken-down" melanics to British typicals. The dominance of the condition was immediately re-established: The architecture of the British chromosomes shaped a clear segregation between a dark heterozygote and a pale homozygote.

Fisher's theory has proved correct in many similar experiments in both plants and animals, and although it may not be universally operative it provided the genetic basis for the understanding between disciplines that was needed before the neo-Darwinian synthesis could occur.

As late as 1932, T. H. Morgan was asserting that

"natural selection does not play the role of a creative principle in evolution," but 10 years later all but a very few biologists were agreed on an evolutionary theory based firmly on Darwin's ideas knitted with subsequent developments in genetics. This coming together was described by Julian Huxley as the "modern synthesis" in a book with the same name published in 1942. The synthesis can first be seen in three English books: R. A. Fisher's *Genetical Theory of Natural Selection* (1930), E. B. Ford's *Mendelism and Evolution* (1931), and J. B. S. Haldane's *Causes of Evolution* (1932). It was consolidated in three works from America: Theodosius Dobzhansky's *Genetics and the Origin of Species* (1937), Ernst Mayr's *Systematics and the Origin of Species* (1942), and George Gaylord Simpson's *Tempo and Mode in Evolution* (1944). As Mayr noted, it did not occur as a result of one side being proved right and the others wrong but rather from "an exchange of the most viable components of the previously competing research traditions."

## II. THE MAKING OF A PHENOTYPE

In the simplest microbial systems, there is a one to one relationship between gene action and phenotype: Change in a gene is likely to produce a change in its product, which will manifest directly in the organism as an altered character, perhaps the loss or modification of an enzyme. Such simple relationships exist in all organisms, however complex. There are more than 200 "inborn errors of metabolism" in humans, each resulting from changes in a particular gene, producing a change of phenotype in the whole organism. For example, albinism is due to the inability to synthesize melanin, which is made from tyrosine under the influence of tyrosinase, which is under the control of a gene on chromosome 11 in humans<sup>1</sup>; "classical" hemophilia results from the absence of a protein (factor VIII) coded by a gene on the X chromosome. However, most traits are affected by many genes. For example, blood clotting in mammals is effected by a cascade of physiological reactions, each under the control of a different gene(s). Efficient clotting requires all the different stages to be operating, and defects in any stage (especially the genes directing

<sup>1</sup> In fact, human albinism can be due to genes on at least two different chromosomes. There are also mutations at another locus on chromosome 11 which cause albinism, but in this case the tyrosinase enzyme is apparently normal. Both genes produce an albino phenotype, but the phenotypes can be distinguished clinically.

the relevant proteins and enzymes) will lead to a “bleeding” phenotype. In blood clotting it is possible to detect where the error lies because we know the normal determinants of clotting in detail. For most traits we do not have information about the steps in their formation. One of the benefits of genome mapping is that the genetical determinants of complex characters will be analyzable in an orderly way. Currently, most phenotypic analysis is biometric rather than genetic.

For example, the tails of house mice may be shortened by mutations at approximately 40 loci. We know the action of many of these genes: Some affect the notochord, some cell division rates, and some inductive relations between endoderm and mesoderm. We also know that the tail length of mice varies between different inbred strains of mice (which are similar in the genetic sense to Johannsen’s pure lines) and that we can increase or decrease tail length by selecting wild-caught animals or the products of crosses between inbred strains. We do not know which genes are variable (or segregating) in any one population or which genes are being affected by selection. However, it is clear from many selection experiments that the genes affecting any complex trait are distributed throughout the genome (i.e., over many chromosomes) and that selection for any one trait may “unbalance” the genome.

The concept of genomic balance (often called genomic or genetic architecture) is important. When a normal outbreeding species, such as maize, sugar beet, poultry, or *Drosophila*, is made to inbreed by manipulating its breeding system, the individuals characteristically decline in vigor and fertility until they stabilize at a stage before complete homozygosity is approached. The amount of this inbreeding depression varies from one line to another within any species. When two inbred lines are intercrossed, the  $F_1$  (i.e., the first-generation progeny) show a considerable increase in vigor and fertility, known as hybrid vigor or heterosis. In general, the  $F_1$  phenotype is similar to that in the population from which the inbred strains were derived. If  $F_2$ ’s are raised and inbreeding is resumed, inbreeding depression will again occur, although not necessarily to the same extent as in the original lines. Inbreeding depression is often reflected in increased variability not only between individuals but also among repetitive parts such as bilateral characters in animals and floral morphology in plants; fluctuating asymmetry, in which there are differences between the right and left sides of individuals, is commonly used as a rough indicator of inbreeding.  $F_1$ ’s tend to show decreased variability. Michael Lerner argued that this indicates an innate superiority of heterozygotes over homozygotes, perhaps

because of a greater biochemical flexibility in the former. It is a phenomenon which has produced considerable research interest, particularly among marine biologists. Lerner argued that if artificial selection is suspended before much of the variation has been lost (as homozygosity increases) natural selection will tend to restore the character (and hence its genetic determinants) to an equilibrium value, with the mean of the character which was artificially selected tending to revert toward its original value; he called this genetic homeostasis.

Artificial selection tends to accumulate alleles which act on a character in a particular way, e.g., to increase its size or the number of elements making up a repeated trait. If we make the reasonable assumption that any population needs to combine phenotypic uniformity in a stable environment with long-term flexibility should the environment change, the easiest way to do this is for the character in question to be controlled by alleles at different loci, with some alleles acting to increase the expression of the character and others acting to decrease it. In such a case, Fisher showed that selection will favor linkage between the loci responsible, with the evolution of “balanced” chromosomes containing “positive” and “negative” alleles. The simplest situation will involve two segregating loci, A,a and B,b with A,B acting in one direction and a,b in the other (i.e., additive genes). The intermediate type can be either the attraction or the repulsion heterozygote, AB/ab or Ab/aB. However, the latter will be favored since it will be less likely to produce zygotes giving the extreme phenotypes (Fig. 1). For the same reason, any mechanisms bringing about closer linkage between the loci concerned (such as a chromosomal inversion) will be favored.

This theoretical arrangement has been subjected to experimental analysis, mainly through selection experiments in *Drosophila*, and has been generally confirmed; it receives support from the distribution of “quantitative trait loci” identified by molecular techniques. It also accounts for other properties:

1. Selection (natural or artificial) for any character almost invariably produces correlated responses in additional developmentally independent traits of the phenotype.
2. Gene loci affecting viability traits are interspersed along the chromosomes with loci affecting other characters.
3. The highest rate of artificially induced new variation by mutation is many times less than that occurring spontaneously through recombination.

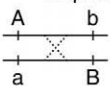
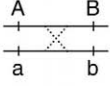
Parents		Offspring					
Chromosomes	Phenotypic score	From non-recombinant gametes		With recombination in one parent		With recombination in both parents	
		Chromosomes	Phenotypic score	Chromosomes	Phenotypic score	Chromosomes	Phenotypic score
<p>Repulsion</p>  <p>2</p>		Ab / aB	2	AB / aB	3	AB / ab	2
		Ab / Ab	2	AB / Ab	3	AB / AB	4
		aB / aB	2	ab / aB	1	ab / ab	0
		aB / Ab	2	ab / Ab	1	AB / ab	2
<p>Attraction</p>  <p>2</p>		AB / ab	2	AB / Ab	3	Ab / Ab	2
		AB / AB	4	AB / aB	3	Ab / aB	2
		ab / ab	0	ab / Ab	1	aB / Ab	2
		ab / AB	2	ab / aB	1	aB / aB	2

FIGURE 1 Effect of linkage on the incidence of the extremes of manifestation of an additively inherited character. If A,B produce a phenotype effect of 1 and a,b an effect of 0, the matings between two repulsion heterozygotes are least likely to produce extreme phenotypes (they will only appear if recombination between the two loci occurs in both parents).

4. There is widespread occurrence of chromosomal inversions in nature.

5. Deleterious gene combinations (even to the point of lethality) may occur solely as a result of recombination.

There are a few cases in which the different genes contributing to complex variation have been identified. This has been done for the determinants of the pin thrum polymorphism in *Primula vulgaris*, in which different members of a linked group of loci control anther height, style length, pollen size, rate of pollen tube growth, and length of the papillae on the stigma; for mimetic patterns in the African swallowtail butterfly *Papilio dardanus*; and for color and banding patterns in the land snail *Cepaea nemoralis*. In house mice, approximately 16% of genes with an identified function are concerned with behavior and have been located on 19 of the 20 chromosomes of the species. However, in most cases it is only possible to conclude that different components of a character complex are inherited as a unit, i.e., that the gene complex is coadapted.

Coadaptation is presumably the reason why the ge-

nomes of interfertile species do not always merge when they meet. For example, races of dark- and light-bellied house mice (referred to as *Mus musculus domesticus* and *M. m. musculus*, respectively, although they should perhaps be given full species status) meet in a narrow zone of intergradation across Jutland (Denmark) and south through Germany. Although the two forms readily interbreed in the laboratory, the hybrid zone has apparently been constant for at least 50 years. Gene frequencies on the two sides of the hybrid zone are very different. Remarkably, the frequencies in the light-bellied form in California, which are descended from the same stock as the light-bellied Danish animals, are more like those in the light-bellied Danish population than are the light-bellied Danish ones from their dark-bellied neighbors. A similar situation exists in deer mice (*Peromyscus polionotus*) in Florida and also in carrion and hooded crows in the same area in which the dark- and light-bellied mice occur in Europe. However, some species lose their identity wholly when brought into breeding contact. This has happened commonly in New Zealand, where much of the terrestrial biota has been

introduced. Hybrid zones are common. Presumably these occur when coadaptation has not evolved.

### III. FORMAL ANALYSIS

In the past, it was suggested that the genes which control major or qualitative traits are different from those which affect quantitative ones (oligogenes and polygenes, respectively). This distinction is now rarely made. In other words, it is assumed that genes affecting quantitative traits (such as weight or size or physiological properties such as metabolic rate) follow Mendelian patterns of inheritance, may have multiple alleles, can mutate, change in gene frequency, show dominance, etc. Quantitative inheritance is merely a general case of the interaction of genes in which the interacting components are little or wholly known.

The number of genes which affect a trait can be estimated by the amount and speed of response in a selection experiment or by the mean and variance of the character as measured in a population. Using these techniques it has been calculated that human skin color may be determined by only 5 or 6 loci, whereas the number of genes affecting oil and protein production in maize may be as high as 54 and 122, respectively. However, such estimates are very dependent on the nature of interactions between the loci concerned and should be regarded as no more than suggestive of a large or small number.

The simplest assumption in multigenic trait determination is that all the loci affect the trait equally and therefore additively. However, detailed analysis has shown that in many (perhaps most) cases, a few genes have a major effect and many genes have a minor effect. For example, in the well-studied case of variation in sternopleural bristle number in *D. melanogaster*, approximately 10 loci account for 75% of the genetic variation in number. The situation is further complicated by pleiotropy: A gene may have a major effect on one character but minor effects on others. For example, phenylketonuria is an inborn error of metabolism producing severe mental retardation in humans and is produced by the nonfunctioning of phenylalanine hydroxylase, which is controlled as a recessive trait by a single gene on chromosome 12. The same enzyme is involved in melanin synthesis, and phenylketonurics have slightly paler hair and complexion than their normal sibs. The gene can therefore be regarded as having a major effect on intelligence but a minor effect on pigmentation.

However, a quantitatively inherited trait will be more

likely than a qualitative one to be affected by environment. A group of individuals having identical genes for growth (i.e., a pure line or a clone) may show considerable variation in size due to differences in available nutrients. Although the same potential for size is present in the initial gene products, the manifestation of the phenotype will be limited by such factors as food availability. Conversely, a population of genetically heterogeneous individuals may grow to the same size if no gene-environment interaction is limiting (or if different interactions compensate for each other). Generalizing, we can express the phenotypic value  $P$  for individual  $i$  in environment  $j$  as

$$P_{ij} = G_i + E_j$$

where  $G_i$  is the genetic contribution of the  $i$ th genotype and  $E_j$  is the environmental deviation resulting from the  $j$ th environment.

A particular genotype may do well in a particular environment, implying a specific interaction between the two. In this case,

$$P_{ij} = G_i + E_j + GE_{ij}$$

In practice, there will be variance of these components so that

$$V_p = V_G + V_E + 2Cov_{GE}$$

where  $V_p$ ,  $V_G$ ,  $V_E$ , and  $2Cov_{GE}$  are the phenotypic, genetic, and environmental variance and the genotype-environment covariance, respectively. The genotype-environment covariance is positive when genotypes with higher values are in better environments and poorer genotypes have poorer environments. This may occur in animals when one member of a litter is large because of its genes and gets more food from its parents or when an animal is socially dominant for genetic reasons and therefore has more resources in food and spaces. A plant genotype which grows faster may have a better environment because it is less likely to be shaded.

In controlled plant and animal breeding, efforts are made to randomize genotypes and environments, so that  $Cov_{GE}$  is minimized. In this situation,  $Cov_{GE}$  can be neglected, and

$$V_p = V_G + V_E$$



Conventionally,

$$V_G/V_P + V_E/V_P = h^2 + e^2$$

where  $h^2$  and  $e^2$  are the proportion of phenotypic variation due to genetic and environmental factors, respectively. The term  $h^2$  is known as heritability in the broad sense,

$$h_B^2 = V_G/V_P$$

In practice the genetic variance is composed of a range of different interactions between loci, which may be additive, dominant, or epistatic. Variability in the narrow sense is defined as

$$h_N^2 = V_A/V_I$$

where  $V_A$  is the variance due to additive genetic factors. It is an important statistic in determining the rate and amount of response to directional selection in breeding programs.

Heritability is a population-specific measurement. It does not measure an invariant property of a particular trait but only the relative contributions of genetic and environmental differences to phenotypic variation in a specific situation. If either genetic or environmental variation changes, heritability estimates will also change; heritability measures the proportion of phenotypic variation in a particular population due to genetic variation.

#### IV. GENETIC AND PHENOTYPIC VARIABILITY

The assumption of the early geneticists that most populations carry little variation and that most individuals are homozygous at all but a few loci at which alleles are either recent mutants in the process of being eliminated by selection or maintained by opposing selection pressures or heterosis is clearly wrong. Evolutionary theory (requiring modifying genes to change phenotype expression) and artificial selection practice (revealing considerable potential for inherited response to selection) imply the existence of considerable genetic variation in and between populations. In the 1950s and 1960s assumptions about genetic load suggested there was a maximum amount of variation which could be tolerated in any population, but the application of protein and later DNA electrophoresis showed that load

theory was too narrow and deterministic and led to reinterpretation so as to incorporate ecological factors (including heterogeneity in time and space and variable stress from biotic and abiotic agents). Empirical data have shown that even inbred organisms (such as obligate self-fertilizers) or ones living in an apparently constant environment (such as the deep sea) may still be heterozygous at a significant proportion of their loci (up to one-fourth for enzyme loci in plants and invertebrates and less for vertebrates).

The implications of load theory led to debate among population geneticists and evolutionists about the significance of the observed variation. Theoreticians impressed by the apparent rigor of load theory and biochemists not used to meaningful variation in their study of chemical pathways tended to dismiss the bulk of genetic variation as neutral and irrelevant to the organism, whereas evolutionary ecologists and practical breeders regarded it as potentially adaptive and as expected by Darwinian understanding. The latter emphasis has been shown to be correct: Far from being constrained by invariant genotypes, phenotypes can (and should) be treated as capable of rapid response to the environment and hence permissive rather than deterministic of survival. A phenotype can be interpreted as the consequence of developmental reaction norms, facilitated in most cases by the width of its underlying genetic base.

This concept is supported by a series of experiments carried out by Bruce Wallace to test if radiation-induced mutations in *D. melanogaster* were inevitably detrimental. He found that mutations induced in flies made homozygous by artificial breeding were often heterotic, although mutations occurring in flies with a heterozygous background were usually deleterious to their carriers, presumably since their gene expression had been adjusted by normal selection. This led Wallace to emphasize the importance of balancing (or stabilizing) selection rather than the directed or cleansing selection which removes unwanted variants.

The concept of reaction norm incorporates the idea of phenotypic plasticity championed by Bradshaw and Levins in the 1960s and provides a much needed synthesis of allometry and ontogeny with ecological realism and fitness components. A particularly good example is the seasonal polyphenism of the African satyrine butterfly, *Bicyclus anyana*, studied by Paul Brakefield and colleagues. The wet and dry season forms of this species are phenotypically distinct in the size of the eyespots and the banding patterns on the wings. Seasonal shifts in rainfall are associated with changes in temperature and with butterfly behavior. In the cooler dry season,

the insects rest on dry grass or leaf litter and rarely fly; they do not breed at this time. In contrast, the adults fly actively in the warmer wet season, searching for mates and oviposition sites in the lush vegetation. Survival differences between the two forms suggest that cryptic matching is more important in the dry season and deception mediated by the false eyes in the wet season.

The wet season form arises through an acceleration of development which can result from an increased or fluctuating temperature, food quality, or hormonal influence. However, eyespot size and plasticity are affected by genetic variation as shown by variation between families in different temperature regimes and by directed selection. The switch between different wing spotting phenotypes in the butterfly *Maniola jurtina* studied over many years by E. B. Ford is probably also of this nature. Ford's extensive field studies showed the interaction between founder population differentiation, natural selection in particular localities, and environmental determinants, although he did not interpret them in this way. Unlike the *Bicyclus* case, the changes in phenotype in *Maniola* are small and unlikely to be important in survival or fitness (although it would be wrong to be dogmatic about this conclusion). However, they serve as a marker of a variable developmental system which may well have a range of functional responses. There are many such examples of trivial (or even wholly cryptic) differences between individuals which have different functional properties (e.g., the possession of B chromosomes or resistance to a newly introduced pesticide or pathogen).

The complicated response system in *Bicyclus* (and other butterflies) is similar in principle to the much simpler situation of eye pigmentation in the amphipod *Gammarus chevreuxi* studied classically by Julian Huxley and Ford. Here, the unpigmented, red-eye form is produced by a single mutant allele which has a slower rate of melanin synthesis than normal. However, pigmented eyes like the wild type arise if genetically red-eyed animals are raised at higher temperatures than normal or if they also carry a gene for small eyes which enables the small amount of pigment produced to cover the whole eye. Another example is the "himalayan" mutation of mammals frequently described in elementary genetics textbooks in which an unpigmented (white) coat becomes pigmented at the cold extremities (feet and ears), as in Siamese cats. Pigmentation may also develop in hair which grows in a shaved area which has a lower temperature than usual. The tyrosinase in this genotype is heat labile and has a maximum activity below normal body temperature. Rate genes of

this nature have been very important in evolution and can be shown to produce major alterations in body form through various forms of allometry.

## V. PHENOTYPES AND EVOLUTION

It is a truism that natural selection can only operate where phenotypic variation exists. Evolution results from changes in gene (strictly allele) frequencies, but these are the consequence rather than the cause of phenotypic differences. It follows that selection on local forms (or ecotypes) which are solely the environmental consequence of a particular environment (e.g., plants grown in sheltered or exposed conditions or subject to a particular dietary lack or a behavioral reaction in an animal) will not lead to adaptive adjustment unless the phenotype in question depends on different genotypes. However, a phenotypic response to environmental conditions may allow a genetically nonadapted population to survive long enough to accumulate variants (through mutation, recombination, or immigration) and then adapt genetically. This idea was put forward independently by Baldwin (1896), Osborn (1897), and Lloyd Morgan (1900): It is commonly known as the Principle of Organic Selection or the Baldwin effect. This has been claimed by some to bridge the directed response which is the basis of Lamarckism and the random source of inherited variation which underlies Darwinism. Certainly, it may mimic Lamarckism, although it is of course wholly Darwinian in its operation.

Most of the early studies of phenotypic plasticity were carried out by botanists: Gaston Bonnier in the Alps and Pyrenees, F. E. Clements in Colorado and California, Turesson in Sweden, and Clausen, Keck, and Hiesey in California. However, the evolutionary implications were not pursued until Gause and Schmalhaussen in Russia and Waddington in Britain began to examine the relationships between genotype, phenotype, and environment in animal experiments. They were able to show not only that natural selection could lead to a character originally induced by the environment becoming an inherited character (Waddington called this "genetic assimilation") but also that there is a causal connection between the environmentally induced change and subsequent genetic changes. They argued that since adaptability—the ability to acquire an adaptive variant during an animal's lifetime—has a genetic basis, the genes underlying flexible adaptive variations may ultimately be responsible for the evolution of fixed adaptations to a new environment, i.e., the environment is much more than a sieve selecting

(or eliminating) chance mutations. These ideas have been criticized (notably by G. C. Williams) on two grounds: That most changes resulting from environmental challenges are not adaptive and that fixation of a genetic response results in a decrease in genetic potential because less (genetic) information is needed to specify a fixed, than a variable response. There is truth in both these objections, although the assumption that plasticity requires more inherited variation than a fixed state is debatable. However, there is no doubt that the link between stress-generated responses and subsequent adaptation requires more study. The problem historically is that ecologists have been primarily concerned with pattern (i.e., spatial relationships within communities), whereas evolutionists have concentrated on processes (i.e., temporal changes within communities), and both have used restrictive definitions of stress. This division between disciplines is not, of course, absolute, but it has led to a rift within population biology and a lack of understanding of coevolutionary possibilities and constraints.

The divergence between ecologists and evolutionists is seen in the models developed by each. In general, ecological models are self-contained because the dependent variables (numbers or density) and parameters (birth and death rates, dispersion differences, rates of predation, etc.) are all measurable by ecological methods. Ecologists have been particularly concerned with identifying criteria for stability or fluctuation when species interact, for invadability, and for extinction when competition occurs. Such models are limited because the parameters are all manifestations of the phenotypic properties of individuals and hence the product of evolutionary processes. Although lip service is paid to the fact that phenotypes may change, in practice they are assumed to be effectively constant in ecological time. Evolutionary models originally concentrated on examining possible rates of evolution; then they became concerned more with the factors involved in the maintenance of genetic polymorphism. Their variables are the frequencies of alleles and genotypes; their parameters are relative fitnesses and rates of mutation, migration, and recombination. For a long time, fitnesses were regarded by the model builders as virtually constant, and it is only comparatively recently that the dependence of selection coefficients on density and frequency has been incorporated. This false assumption of constancy led to the problems associated with genetic load and to controversies over neutralism.

Genetical (evolutionary) models have become increasingly realistic and treat relative fitnesses as capable of varying in time and space as well as with gene fre-

quency and population density. Such models are both genetical and ecological, and therefore they are intrinsically more informative than purely ecological ones. However, they still suffer from a major problem in that they tend to assume a degree of genetical equilibrium or stasis which is unjustified.

The way forward will be to increase ecological reality of evolutionary models, taking into account the characteristics of the niche for particular populations and communities, particularly any genetic constraints. Such models obviously need to include behavioral (sociobiological) input and the notion of evolutionary "strategies."

## VI. TAXONOMY

Classification has become increasingly sophisticated with the increase in traits which can be used to characterize a group (or taxon). The problem is that conventional taxonomic (phenotypic) diversity may bear little relationship to genetic diversity. For example, the seaside sparrow (*Ammodramus nigrescens*) is common on the eastern and southern coasts of the United States. Nine subspecies have been described, including a rare and recently extinct (1987) dusky form, originally regarded as a separate species. Molecular analysis (based on mitochondrial DNA sequences) showed that the dusky sparrow was indistinguishable from other Atlantic forms, but that there was a major and previously unsuspected distinction between Atlantic and Gulf Coast groups.

In contrast, New Zealand tuatara lizards (*Sphenodon punctatus*) are conventionally treated as a single taxon, but molecular (and morphological) criteria indicate they comprise at least three species. For conservation management, it is obviously important that each taxon be considered separately.

There is no clear correlation between genetic and taxonomic diversity. The problems of classification that have always challenged museum workers have been compounded by the possibility of using genetic factors as additional or substitute taxonomic criteria: Closely related species in the genetic sense may have very different niches than those of their near relatives, whereas genetically more distant forms may look alike and interact strongly. These are no clear rules to link phenotype with genotype or phenotypic variety with genetic variety. There is plenty of scope for better multidisciplinary understanding of phenotypes, this will have to involve genetics, development, behavior, biotic and abiotic environments, life history, and phylogeny. Notwithstand-

ing case by case examinations of the determinants and plasticity of the phenotypes in particular species and species groups are still needed.

### See Also the Following Articles

EVOLUTION, THEORY OF • GENES, DESCRIPTION OF • GENETIC DIVERSITY

### Bibliography

- Avise, J. C. (1994). *Molecular Markers, Natural History and Evolution*. Chapman & Hall, New York.
- Berry, R. J., Crawford, T. J., and Hewitt, G. M. (Eds.) (1992). *Genes in Ecology*. Blackwell, Oxford.
- Briggs, D., and Walters, S. M. (1984). *Plant Variation and Evolution*, 2nd ed. Cambridge Univ. Press, Cambridge, UK.
- Falconer, D. S. (1989). *Introduction to Quantitative Genetics*, 3rd ed. Longman, New York.
- Feder, M. E., Bennett, A. F., Burggren, W. W., and Huey, R. B. (Eds.) (1987). *New Directions in Ecological Physiology*. Cambridge Univ. Press, Cambridge, UK.
- Gould, S. J. (1977). *Ontogeny and Phylogeny*. Harvard Univ. Press, Cambridge, MA.
- Hamilton, W. D. (1996). *Narrow Roads of Gene Land*. Freeman, New York.
- Hoffman, A. A., and Parsons, P. A. (1997). *Extreme Environmental Change and Evolution*. Cambridge Univ. Press, Cambridge, UK.
- Jablonka, E., and Lamb, M. J. (1995). *Epigenetic Inheritance and Evolution*. Oxford Univ. Press, Oxford.
- Mather, K. (1973). *Genetical Structure of Populations*. Chapman & Hall, London.
- Mayr, E. (1982). *The Growth of Biological Thought*. Harvard Univ. Press, Cambridge, MA.
- Raff, R. A. (1996). *The Shape of Life*. Univ. of Chicago Press, Chicago.
- Schlichting, C. D., and Pigliucci, M. (1998). *Phenotypic Evolution*. Sinauer, Sunderland, MA.
- Waddington, C. H. (1957). *The Strategy of the Genes*. Allen & Unwin, London.
- Yablokov, A. V. (1986). *Phenetics*. Columbia Univ. Press, New York.





# PHOTOSYNTHESIS, MECHANISMS OF

John A. Raven  
*University of Dundee*

---

- I. Light Harvesting and Excitation Energy Transfer to Reaction Centers
  - II. Reaction Centers and Primary Photochemistry
  - III. Membrane-Associated Reactions Leading from Primary Photochemistry to ATP and NAD(P)H
  - IV. Use of NAD(P)H and ATP in CO<sub>2</sub> Fixation and Other Reactions
  - V. Conclusions
- 

tions of photosynthesis, the earliest products being the oxidized reaction center (bacterio)chlorophyll at the outside/thylakoid lumen side and a reduced (bacterio)chlorophyll (RC1) or reduced (bacterio)phaeophytin (RC2) on the cytosol/stroma side of the photosynthetic membrane.

---

## GLOSSARY

**ATP synthetase** Membrane-associated protein complex that couples exergonic fluxes of H<sup>+</sup> across the membrane to ADP phosphorylation, or vice versa.

**carboxylase** Enzyme that catalyzes the formation of a C—C bond between inorganic C (CO<sub>2</sub> or HCO<sub>3</sub><sup>-</sup>) and some organic molecule.

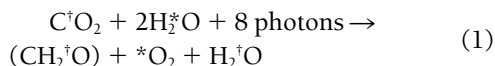
**chemiosmotic energy coupling** Exergonic (photo)-chemical redox reaction, or an exergonic hydrolysis of a phosphate anhydride or thioester, coupled to endergonic ion (normally H<sup>+</sup>) transfer across a membrane, or vice versa.

**chromophore** Organic molecule that absorbs photosynthetically active radiation and either carries out photochemistry or transfers excitation energy to another chromophore that carries out photochemistry. Always bound to a polypeptide in the photosynthetic apparatus.

**reaction center** (Bacterio)chlorophyll–protein complex that performs the primary photochemical reac-

**PHOTOSYNTHESIS, OR BUILDING WITH LIGHT**, involves the use of electromagnetic energy to bring about endergonic reactions. The concept is generally extended to the use of the more immediate photochemical products to reduce CO<sub>2</sub> and (in most photosynthetic organisms) oxidize H<sub>2</sub>O and evolve O<sub>2</sub>, this extended definition is used here. This article deals with the partial processes of photosynthesis in the order in which they occur subsequent to initial photon absorption. Starting with the characteristics of the photon-harvesting apparatus, we then consider the photochemical reaction centers and their immediate products, which are an oxidant and a reductant oxidation–reduction and/or an ion (H<sup>+</sup> or, rarely, Cl<sup>-</sup>) gradient. These light-harvesting and photochemical reactions are invariably associated with membranes. Next in temporal sequences are the reactions that lead to the phosphorylation of ADP and generation of strong [NAD(P)H] reductant; these reactions are also membrane related. Finally, we deal with the use of ATP and NADPH in CO<sub>2</sub> fixation and other endergonic reactions; these reactions are either in aqueous solution or associated with membranes other than

those in which primary photochemistry occurs. In  $O_2$ -evolving organisms, the basic photosynthetic equation is



## I. LIGHT HARVESTING AND EXCITATION ENERGY TRANSFER TO REACTION CENTERS

For photochemical reactions to occur, photons must be absorbed by a pigment. In the majority of photosynthetic organisms, only one in 100 to one in 1000 pigment molecules can actually bring about useful (to the organisms) photochemical reactions. The photochemically inactive antenna pigments serve to absorb radiation and transfer the energy as excitation energy (i.e., not in stable chemical forms) to photochemically active pigments in reaction centers. This division of labor presumably relates to the generally rather "dilute" nature of solar radiation, such that each pigment molecule absorbs a photon on the order of a few times each second, yet the photochemical machinery and the immediate downstream reactions can react hundreds or thousands of times per second. Accordingly, the resource-expensive (energy, nitrogen, iron) reaction centers and the related catalytic machinery that is energetically and temporally downstream of the reaction centers are present as a small number of copies of the catalysts relative to the number of molecules of light-harvesting pigment; this arrangement is presumably related to optimal allocation of resources in natural selection.

A very small number of species, all of them members of the Archaea, do not accord with this generalization about antenna pigments and reaction centers. These Halobacteria use carotene pigments associated with membrane-spanning proteins to catalyze their primary photochemical reaction, with each protein molecule associated with its pigment molecule acting as a light-energized  $H^+$  (bacteriorhodopsin) or  $Cl^-$  (halorhodopsin) pump (see Section II).

By contrast, all other photosynthetic organisms are (eu)Bacteria and their symbiotic eukaryotic descendants that use the antenna plus reaction center system. In all of these organisms the pigment involved in the primary photochemical reactions is a Mg- (very rarely Zn-) binding cyclic tetrapyrrol, chlorophyll (in  $O_2$ -evolvers) and bacteriochlorophyll (in non- $O_2$ -evolvers). Chlorophylls always occur in a functional photosyn-

thetic apparatus in association with proteins in, or more rarely on, bilayer membranes. In all chlorophyll-containing organisms there are some chlorophyll molecules that act as antenna pigments. In many such organisms (e.g., all green algae and higher plants) the great majority of light harvesting is carried out by chlorophyll-protein complexes, in this case with chlorophylls *a* and *b* as the major light-absorbers with minor contributions from carotenoids. In the case of algae with chlorophyll *a* and, usually, chlorophyll *c*, carotenoids generally have a more significant role in light harvesting by pigment-protein complexes. In a final group of organisms (mainly cyanobacteria and red algae), the light-harvesting pigments in addition to chlorophylls are open-chain tetrapyrrol pigments covalently bound to proteins, attached to the outer (stromal in red algae, cytosolic in cyanobacteria) side of the photosynthetic membranes, or in the thylakoid lumen of cryptophyte algae. The variations in the chemistry and location of light-harvesting pigment-protein complexes are indicated in Table I.

A number of energetic and structural constraints determine the spatial relationship of the different pigments in light-harvesting complexes. The energetic constraint is that energetically efficient excitation energy transfer among antenna pigments, and from antenna pigments to reaction centers, involves an unavoidable energy loss during the transfer. Thus, the peripheral pigments in the antenna have a higher energy content of their excited states than do the inner antenna pigments, and those in turn have a more energetic excited state than the reaction center pigment. The situation is complicated by the occurrence of two readily accessible excited states in some photosynthetic pigments, and especially in the chlorophylls. Here the second excited state, produced by absorption of blue radiation, has a higher energy content than the first excited state, corresponding to the *in vivo* absorption in the red (chlorophylls) or infrared (bacteriochlorophylls) region of the electromagnetic spectrum. It is the first excited state that performs the primary photochemistry in (bacterio) chlorophyll-based photosynthesis. Thus, photons that excite (bacterio) chlorophylls to the second excited state are not directly used in photochemistry via this state, but only after they become converted to the first excited state; this also applies to excitation energy transfer among pigments.

These energetic requirements are reflected in the structure of the light-harvesting apparatus. Thus, the pigments are arranged spatially such that those absorbing shorter-wavelength (thus higher energy) photons are generally at a greater distance from the reaction center, whereas those absorbing at longer wavelengths

TABLE I  
Chemistry, Taxonomy, and Location of Light-Harvesting Pigment-Protein Complexes

Chromophores	Taxonomy	Location
O <sub>2</sub> -Evolvers		
Chlorophylls <i>a</i> and <i>b</i>	Chlorophyta, Euglenophyta, Chlorarachniophyta, Embryophyta; "chloroxybacterial" cyanobacteria	Integral membrane protein of thylakoid
Chlorophylls <i>a</i> + (usually) <i>c</i>	Cryptophyta, Heterokontophyta (Bacillariophyceae, Chrysophyceae, Phaeophyceae, Tribophyceae), Haptophyta, Dinophyta	Integral membrane protein of thylakoid
Chlorophyll <i>a</i> , peridinin Phycobilinogen	Dinophyta Cyanobacteria <i>sensu stricto</i> , Rhodophyta Cryptophyta	Peripheral protein on stromal surface of thylakoid Phycobilisomes on stromal/cytosol surface of thylakoid Thylakoid lumen
Non-O <sub>2</sub> -Evolvers		
Bacteriochlorophylls <i>a</i> and <i>b</i>	Proteobacteria: purple sulfur bacteria (Chromatiaceae), purple nonsulfur bacteria (Rhodospirillaceae)	Integral membrane protein of "chromatophores"
Bacteriochlorophyll <i>g</i> Bacteriochlorophylls <i>c</i> , <i>d</i> , and <i>e</i>	Heliobacteriaceae Chlorobiaceae, Chloroflexaceae	Integral membrane protein of plasmalemma Chlorosomes on the cytosol side of plasmalemma

(more similar to that of the reaction center pigments) are close to the reaction center. Furthermore, the smaller-scale arrangement of individual chromophores in the pigment-protein complexes is also a function of the distance between chromophores over which effective excitation energy transfer can occur; these arrangements have been shown for a number of light-harvesting pigment-protein complexes using X-ray crystallographic techniques.

A related consideration is that of how many chromophore molecules can be involved in light harvesting for a given reaction center. A *minimum* estimate of the upper limit of this number comes from determining the ratio of pigment molecules serving a given kind of reaction center and dividing it by the number of the type of reaction center that they serve. This is clearly most readily determined when there is only one kind of reaction center in the organisms concerned, as is the case in organisms that do not evolve O<sub>2</sub>, but it is also applicable to O<sub>2</sub>-evolvers, which invariably have two kinds of reaction center (see Section II). Such computations suggest that up to 1000–2000 chromophores occur per reaction center of a given type. Even more

pigment molecules can be involved in excitation energy transfer to a given reaction center if pigment molecules are not all dedicated to a given reaction center but can be involved in excitation energy transfer to several nearby centers.

Another size-related aspect of light harvesting occurs at all scales, but becomes most obvious at a larger scale: this concerns the optical thickness of the cell or tissue in which the photosynthetic apparatus occurs. A small optical thickness involves a smaller physical thickness for a given pigment concentration per unit volume, or a lower concentration of pigment per unit volume if there is a fixed physical thickness, or a combination of the two. A large optical thickness means more self-shading of pigment molecules, so that the average specific absorption coefficient of each pigment molecule is lower. Furthermore, in a given radiation environment, it takes longer for a pigment molecule to absorb enough photons to cover the energy cost of its synthesis in a cell or organ with more self-shading. Finally, the cells or tissues with a small optical thickness, and hence smaller package effect, have a greater opportunity for the spectral diversity of pigments that they contain to



be manifest in the specific wavelengths of radiation that are absorbed. The ultimate in large optical thickness is found in the thalli of some seaweeds that, with more than one millimole of pigment per square meter, look black to the human observer (because of low reflectance or transmittance, with minimal wavelength dependence of these two processes due to high absorptance at all visible wavelengths) regardless of whether their pigments are of the kind found in green, brown, or red algae.

A final aspect of light harvesting concerns the occurrence of pigment molecules in photosynthetic cells that (of necessity) absorb radiation but that show little or no transfer of excitation energy to reaction centers. Examples of such pigments are  $\beta$ -carotene, and at least some portion of the other carotenoids. The pigments serve as photoprotectants, quenching the triplet excitation states of chlorophyll and singlet oxygen in the case of  $\beta$ -carotene, and as nonphotochemical quenchers of excess (singlet) excitation energy of antenna pigments in the case of phototransformable carotenoids such as violoxanthin–antheroxanthin–zeaxanthin and diadinoxanthin–diatoxanthin in certain  $O_2$ -evolvers.

## II. REACTION CENTERS AND PRIMARY PHOTOCHEMISTRY

We have seen in Section I that the halobacteria, with their halorhodopsin or bacteriorhodopsin pigments, do not have separate antenna and reaction center pigments. These organisms have active transport of ions ( $H^+$  out of the cell or  $Cl^-$  into it) as the sole product of photochemistry external to the pigment–protein complex. The other photosynthetic organisms, with (bacterio)

chlorophyll as their photochemically active pigment, do have reaction centers; these organisms are predominant in terms of number of species, habitats occupied, and quantity of energy and materials transformed.

In all of the (bacterio)chlorophyll-based reaction centers the primary photochemical event is the production of a reduced compound on the side of the membrane abutting the cytosol (prokaryotes) or chloroplast stroma (eukaryotes) and an oxidized compound on the side of the membrane adjacent to the external medium (many prokaryotes) or thylakoid lumen (almost all  $O_2$ -evolvers). The photochemical redox process thus involves the transfer of an electron from near one side of the membrane to near the other side of the membrane. Furthermore, the photochemically active pigment is bacteriochlorophyll *a*, *b*, or *g*, or chlorophyll *a* (or apparently, in one case, chlorophyll *d*), bound to protein (Table II).

While all (bacterio)chlorophyll-based reaction centers share this basic mechanism, such centers can be divided into two types termed RC1 and RC2 (Table II). In the RC1 type (e.g., the reaction centers of the Chlorobiaceae and Heliobacteriaceae, and PSI of  $O_2$ -evolving organisms), the (bacterio)chlorophyll catalyst of primary photochemistry that is oxidized generates an oxidant [oxidized (bacterio)chlorophyll] at a redox potential of +0.25–+0.45 V, while the primary reductant is another form of (bacterio)chlorophyll with a redox potential of –0.6––0.7 V. In the RC2 type (e.g., the reaction centers of the Chloroflexaceae and Proteobacteria, and PSII of  $O_2$ -evolvers), the oxidized (bacterio)chlorophyll product of primary photochemistry has a redox potential of +0.45 (bacteria that cannot evolve  $O_2$ ) to +0.1 V ( $O_2$ -evolving organisms), while the reductant [a (bacterio)pheophytin, i.e., a (bacterio)chlorophyll without Mg] has a redox potential of –0.8 V

TABLE II  
Reaction Centers of Various Photosynthetic Organisms That Contain (Bacterio)chlorophyll

Type of reaction center	Chromophore involved in primary photochemistry	Taxonomy
<i>O<sub>2</sub></i> -Evolvers		
RC1 (as PSI)	Chlorophyll <i>a</i> (P700)	All <i>O<sub>2</sub></i> -evolvers
RC2 (as PSII)	Chlorophyll <i>a</i> (P680)	All <i>O<sub>2</sub></i> -evolvers
Non- <i>O<sub>2</sub></i> -Evolvers		
RC1	Bacteriochlorophyll <i>g</i> (P798)	Heliobacteriaceae
RC1	Bacteriochlorophyll <i>a</i> (P840)	Chlorobiaceae
RC2	Bacteriochlorophyll <i>a</i> (P840)	Chloroflexaceae
RC2	Bacteriochlorophyll <i>a</i> or <i>b</i> (P870 or P890)	Chromatiaceae and Rhodospirillaceae

(bacteria that cannot evolve  $O_2$ ) to  $-0.5$  V ( $O_2$ -evolvers).

The primary charge separation is relatively unstable, and reduction of a more stable compound with a higher redox potential, and which is less prone to back-react yielding chemiluminescence, occurs within  $\sim 100$  ps. For RC1 this secondary acceptor is a phylloquinone (vitamin E) followed, within  $\sim 100$  ns, by electron transfer to iron-sulfur centers. For RC2 the secondary acceptor is a ubi- (non- $O_2$ -evolvers) or plasto- ( $O_2$ -evolvers) quinone bound to the reaction center; the resulting semiquinone free radical reduces a second, chemically identical quinone within  $\sim 300$   $\mu$ s.

The reduction of the primary oxidized product of photosynthesis takes longer than does oxidation of the primary reduced product. For RC1, and for the RC2 of organisms that do not evolve  $O_2$ , the compound that reduces the primary oxidant is a *c*-type cytochrome or (in many  $O_2$ -evolvers) the cuproprotein plastocyanin; this reduction takes  $\sim 200$   $\mu$ s. For RC2 in  $O_2$ -evolvers (i.e., PSII) the reductant of the primary oxidant is a tyrosine residue termed  $Y_z$  in a polypeptide ( $D_1$ ), which, with polypeptide  $D_2$ , binds the reaction center chlorophyll *a*. The action takes  $\sim 200$  ns, and the oxidized tyrosine is re-reduced by a protein-bound cluster of four Mn in  $\sim 200$   $\mu$ s. All of the RC1 reaction centers have, as known from molecular genetic data, a common ancestry, whereas RC2 reaction centers also have a common ancestry that differs from that of RC1. Further discussion of evolutionary points are beyond the scope of this article.

### III. MEMBRANE-ASSOCIATED REACTIONS LEADING FROM PRIMARY PHOTOCHEMISTRY TO ATP AND NAD(P)H

Dealing first with RC1, a role that these reaction centers can perform in most, if not all, of the organisms in which they occur is cyclic electron transport. This process involves ubi- (non- $O_2$ -evolvers) or plasto- ( $O_2$ -evolvers) quinone and a cytochrome  $b-c_1$  (non- $O_2$ -evolvers) or cytochrome  $b_6-f$  ( $O_2$ -evolvers) complex. The link between the iron-sulfur reductant of RC1 and the quinone involves the soluble FeS protein ferredoxin (less commonly the Fe-free flavodoxin), and other redox catalysts that are incompletely characterized and that are probably variable among or even within taxa. One of these could be NAD(P)H dehydrogenase. The oxidizing

end is much better understood, with the cytochrome  $b-c_1$  ( $b_6-f$ ) complex reducing a cytochrome *c* (or plastocyanin) and hence the primary oxidant of the photochemical reaction.

The bioenergetic role of cyclic electron transport is the pumping of  $H^+$  from the cytosol (or stroma) to the external environment (or thylakoid lumen) compartments. Two  $H^+$  per electron are pumped via the quinone and the cytochrome complex. If the NAD(P)H dehydrogenase is involved, then a further  $2H^+$  per electron could be transferred. A major role for these pumped  $H^+$  is to move back across the membrane through the ATP synthetase;  $4H^+$  energize the phosphorylation of one ADP. If all the absorbed photons are used to energize RC1, each photon could produce 0.5–1.0 ATP, depending on the involvement of NAD(P)H dehydrogenase. An alternative use of the  $H^+$  gradient is to energize secondary active transport coupled to exergonic  $H^+$  fluxes.

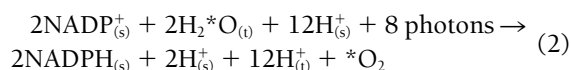
Another role for RC1, and apparently the major one in almost every situation in which RC1 occurs, is to generate a reductant at a low enough redox potential to reduce  $CO_2$  to carbohydrate. As will be seen in Section IV, in many cases the reductive step in  $CO_2$  fixation involves an energetic input from ATP, and here the reductant is NAD(P)H. The path from the reduced protein-bound iron-sulfur involves ferredoxin and a membrane-bound ferredoxin-NAD(P)H reductase. The reduction of ferredoxin and NAD(P)H, and hence the reduction of  $CO_2$  (or  $NO_2^-$ , or other electron acceptors), requires an input of (weaker) reductant at the oxidizing end of the photosystem. For members of the Chlorobiaceae and Heliobacteriaceae, the reductant is ultimately some exogenous organic or inorganic electron donor (other than  $H_2O$ ) such as  $S^{2-}$ , which supplies electrons to the oxidizing end of the photosystem via cytochrome *c*, with or without the involvement of the cytochrome  $b-c_1$  complex. For the  $O_2$ -evolvers, PSI (photoreaction I) is supplied with reductant by PSII (photoreaction II) via plastoquinone, the cytochrome  $b_6-f$  complex, and cytochrome  $c_6$  or plastocyanin.

RC2 does not produce a stable reductant with a redox potential that is sufficiently negative to reduce ferredoxin or NAD(P)H $^+$ . For the photosynthetic Proteobacteria and the Chloroflexaceae, a major function of the RC2 reaction center is in a cyclic electron transport pathway. This uses the ubiquinol produced at the reducing end of the reaction center to reduce the cytochrome  $b-c_1$  complex, and then cytochrome *c*, and finally the oxidant produced by the reaction center is reduced. This mechanism pumps  $2H^+$  per electron, and hence per photon, whose excitation energy is transferred to

the reaction center. One use of the  $H^+$  gradient is, as for cyclic electron flow and  $H^+$  pumping by RC1, to generate ATP. Another is to use a weak reductant (e.g.,  $S^{2-}$ ) to generate a strong reductant (NADH). Here electrons are transported against an energy difference using the energy from the flux of  $H^+$  from the medium back to the cytosol. This process is the reverse of the active  $H^+$  transport caused by the respiratory, exergonic, electron transport from NADH to  $S^0$  (generating  $S^{2-}$ ). Similar uses of the  $H^+$  gradient can be envisaged in the Halobacteriaceae.

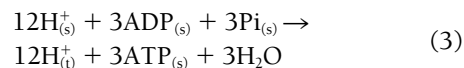
The other, and quantitatively predominant, occurrence of RC2 in the biosphere is as PSII of  $O_2$ -evolvers. Here the usual immediate sink for the electrons from reduced plastoquinone (=plastoquinol) is the cytochrome  $b_6-f$  complex, with subsequent sequential reduction of cytochrome  $c_6$  or plastocyanin, PSI, ferredoxin, and  $NADP^+$ , or, finally to a smaller extent, some other oxidant such as  $NO_2^-$  or  $O_2$  (see the following). The corresponding electron donor to the oxidizing side of PSII, after the polypeptide-bound tyrosine  $Y_Z$  and polypeptide-bound Mn, is the very weak reductant  $H_2O$ . It appears that PSII does not engage in cyclic electron transport involving the cytochrome  $b_6-f$  complex in the manner of RC2 in the Proteobacteria and Chloroflexaceae, although it may perform cyclic electron transport via cytochrome  $b_{559}$ , a redox catalyst with no analog in the RC2 of non- $O_2$ -evolvers. This cyclic electron transport probably occurs as a photochemical means of dissipating excess, and potentially damaging, excitation of PSII.

Putting PSII and PSI together, in terms of photon requirement, the transport of one electron from  $H_2O$  to  $NADP^+$ , and hence  $CO_2$ , and the associated  $H^+$  transport, requires the energy of two photons, one used by PSII and the other by PSI. The stoichiometry of  $H^+$  transport is that one  $H^+$  is left in the thylakoid lumen, and one  $H^+$  is taken up in the cytosol or stroma, per electron moved from  $H_2O$  to  $NADP^+$  and then to  $CO_2$ . In addition, for each electron moving through plastoquinone and the cytochrome  $b_6-f$  complex, two  $H^+$  are transferred from the cytosol or stroma to the thylakoid lumen. Thus, per electron moved from  $H_2O$  to  $NADP^+$  (and thence  $CO_2$ ),  $3H^+$  are moved from the cytosol or stroma to the thylakoid lumen:

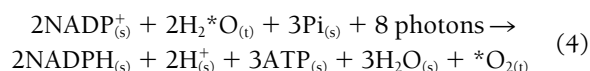


where the subscript (s) means stroma (or cytosol of cyanobacteria) and the subscript (t) means thylakoid lumen.

With an  $H^+/ATP$  of 4 in ATP synthetase, the ADP phosphorylation corresponding to the recycling of  $12H^+$  can be described by



Thus, each electron moved from  $H_2O$  to  $NADP^+$  generates 0.75 ATP, so that 3ADP are phosphorylated per  $2NADP^+$  (and hence  $1CO_2$ ) reduced. This can be seen when Eqs. (2) and (3) are added together to yield



This has important implications for the minimum photon cost of  $CO_2$  fixation in  $O_2$ -evolvers, which is 8 photons absorbed per  $CO_2$  reduced to carbohydrate.

In addition to the reduction of  $CO_2$  and  $NO_2^-$  by PSI, there is also the reduction of  $O_2$  to form, ultimately,  $H_2O$ . The Mehler Peroxidase reaction involves two electrons from the reducing end of PSI passing to two  $O_2$  to form two superoxide anions, which are then dismutated by the enzyme superoxide dismutase to form one  $H_2O_2$  and one  $O_2$ . The  $H_2O_2$ , plus ascorbate, are then acted on by the enzyme ascorbate peroxidase, which converts one  $H_2O_2$  and one ascorbate to two  $H_2O$  and one dehydroascorbate, involving (via other enzymes) two more electrons from PSI to regenerate the ascorbate co-substrate from dehydroascorbate. Overall, the reaction involves the evolution of one  $O_2$  at the oxidizing end of PSII and the uptake of one  $O_2$  at the reducing end of PSI, so that there is no net evolution or uptake of  $O_2$ . The reaction can be quantified by the use of  $^{18}O_2$  tracer. The Mehler Peroxidase reaction is a means of disposing of active, and potentially damaging, oxygen species produced in thylakoid redox reactions and, to the extent that superoxide production by PSI can be increased above the unavoidable basal rate, a means of photochemical energy dissipation and of ATP generation not paralleled by  $NADP^+$  reduction as an alternative to cyclic photophosphorylation.

The photon cost of  $CO_2$  fixation in terms of incident irradiance depends on the photon cost of  $CO_2$  fixation in terms of absorbed photons (8 in the case mentioned here) and the fraction of the incident photons that are absorbed (see Section I). The catalytic capacity of the reactions generating NAD(P)H and ATP can, under many circumstances, constrain the rate of  $CO_2$  fixation when photon supply is not limiting. However, in some

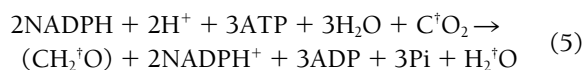
cases it is the supply of CO<sub>2</sub>, or the capacity of CO<sub>2</sub>-assimilation reactions (see Section IV), that constrains the CO<sub>2</sub> fixation rate at saturating light supply. The capacity for light harvesting relative to that for primary photochemical reactions and downstream thylakoid reactions and CO<sub>2</sub> fixation reactions (see Section IV) depends on the irradiance that the plant has encountered in its evolutionary history (genetic adaptation) and during the life of an individual (phenotypic acclimation). In both adaptation and acclimation, the response to low light is an increased capacity for light harvesting relative to downstream reactions, and vice versa for the response to high light.

In O<sub>2</sub>-evolvers the rate and (incident) photon cost of CO<sub>2</sub> fixation at limiting incident irradiances, and the rate of CO<sub>2</sub> fixation at saturating irradiances, can be reduced by photoinhibition. This phenomenon has its basis in damage to the PSII reaction center by one in 10<sup>6</sup> to 10<sup>7</sup> of the photons whose excitation energy reaches the reaction center. In the absence of any protective mechanisms, this photodamage would occur more rapidly at higher incident photon flux densities, but would result in a greater fractional inhibition of photosynthesis when measurements on organisms exposed to high incident photon flux densities are subsequently made at lower (rate-limiting) irradiances. The observed phenomenon of photoinhibition, that is, a reduced rate of photosynthetic CO<sub>2</sub> fixation during, and following, exposure to (especially high) irradiances, is not solely a function of photodamage occurring faster than it can be repaired. Thus, the decreased photosynthetic rate can be a result of diversion of the products of PSII photochemistry to processes other than CO<sub>2</sub> fixation at high incident irradiances, for example, to cyclic electron flow round PSII or to the Mehler Peroxidase reaction, decreasing the photochemical unusable "excess" excitation of PSII. More commonly, excitation energy is prevented from reaching the PSII reaction center by dissociation of some of the light-harvesting pigment-protein complexes (Table II) from the reaction center, or by nonphotochemical quenching of excitation in the light-harvesting complexes by processes triggered by very low pH in the thylakoid lumen. This increased acidification of the lumen is a result of photochemically driven H<sup>+</sup> pumping in excess of the rate at which H<sup>+</sup> is recycled through the ATP synthetase due to restricted ATP consumption and hence limited ADP availability, and can dissipate excitation energy via a xanthophyll cycle mechanism in all but cyanobacteria and red algae and, in all O<sub>2</sub>-evolvers, via mechanisms not related to a xanthophyll cycle.

#### IV. USE OF NAD(P)H AND ATP IN CO<sub>2</sub> FIXATION AND OTHER REACTIONS

Most of the ATP and NAD(P)H generated in the thylakoid reactions in O<sub>2</sub>-evolvers is used in CO<sub>2</sub> reduction to (CH<sub>2</sub>O). Other uses for the ATP resulting from photochemical reactions include photoassimilation of exogenous organic compounds, as well as other transport and biochemical reactions. Photochemically produced reductant can be used for reductive biosyntheses, such as reduction of NO<sub>3</sub><sup>-</sup>, NO<sub>2</sub><sup>-</sup>, and SO<sub>4</sub><sup>2-</sup>. These uses of ATP and reductant notwithstanding, the major global biogeochemical role of NAD(P)H and ATP generated in photochemical reactions is the conversion of CO<sub>2</sub> to carbohydrates; a net primary production of ~100 Pg carbon per year worldwide is carried out by O<sub>2</sub>-evolvers. However, some CO<sub>2</sub> is fixed by non-O<sub>2</sub>-evolving autotrophs. In the Chlorobiaceae, the reverse tricarboxylic acid cycle is used, whereas the Chloroflexaceae use the hydroxypropionate cycle (Table III). These two processes each involve more than one carboxylation reaction in converting CO<sub>2</sub> to carbohydrates. By contrast, the photosynthetic carbon reduction cycle used in the Proteobacteriaceae and in all O<sub>2</sub>-evolvers involves but a single carboxylase, that is, ribulose-1,5-bisphosphate carboxylase-oxygenase (RUBISCO), in converting CO<sub>2</sub> to carbohydrates, although some essential biosyntheses involve parallel, quantitatively minor, carboxylation reactions catalyzed by other carboxylases.

RUBISCO, as its name suggests, not only catalyzes a carboxylase reaction but also has an oxygenase activity. The oxygenase activity is competitive with the carboxylase activity. Thus, the oxygenase is not a significant reaction in proteobacterial photosynthesis in their typical high-CO<sub>2</sub>, low-O<sub>2</sub> habitats where heterotrophic organisms dominate and organic carbon sedimentation into the habitat exceeds O<sub>2</sub> diffusion into it. Furthermore, the oxygenase activity would not have been significant in the high-CO<sub>2</sub>, low-O<sub>2</sub> world prior to 2 billion years ago, that is, for the first 1.5 billion years or so of the occurrence of RUBISCO. When oxygenase activity is negligible, the conversion of 1CO<sub>2</sub> to the carbohydrate redox level by the photosynthetic carbon reduction cycle needs 2NAD(P)H and 3ATP:



Equation (5), showing the consumption of NADPH and ATP in CO<sub>2</sub> fixation, can be combined with Eq. (4), describing the production of NADPH and ATP by light-

TABLE III  
Taxonomic Distribution and Energy Costs of CO<sub>2</sub> Assimilation Reactions in Photosynthetic Organisms

CO <sub>2</sub> assimilation reaction	Taxonomy	NAD(P)H cost per net CO <sub>2</sub> fixed	ATP cost per net CO <sub>2</sub> fixed
Reverse tricarboxylic acid cycle	Chlorobiaceae	2	?
Hydroxypropionate pathway	Chloroflexaceae	2	?
RUBISCO and photosynthetic carbon reduction pathway			
(i) In high CO <sub>2</sub> -low O <sub>2</sub>			
(ii) In low CO <sub>2</sub> -high O <sub>2</sub> (present atmosphere or water in equilibrium with it)	Proteobacteriaceae (Chromatiaceae Rhodospirillaceae)	2	3
(a) Diffusive CO <sub>2</sub> entry to RUBISCO, with photorespiratory carbon oxidation cycle	C <sub>3</sub> land plants; some algae	2.8	4.2
(b) Active transport of inorganic C to RUBISCO by a process that does not involve C—C bond formation	Cyanobacteria; many algae	2	4–5
(c) Active transport of inorganic C to RUBISCO by a process that does involve C—C bond formation	C <sub>4</sub> higher plants CAM higher plants	2 2	5 6.5

driven processes in the thylakoid, to give the overall equation for photosynthesis [Eq. (1)].

While the reductant requirement per CO<sub>2</sub> assimilated is the same as for the reverse tricarboxylic acid cycle and the hydroxypropionate cycle, the ATP requirement is greater for the photosynthetic carbon reduction cycle (see Table III).

The greater energy requirement for the photosynthetic carbon reduction cycle than for the other carbon fixation cycles is exacerbated in the present atmosphere, or in natural waters in equilibrium with it. If CO<sub>2</sub> enters by diffusion (C<sub>3</sub> physiology) then there is significant oxygenase activity of RUBISCO, although there is significant variability among RUBISCOs of different organisms in selectivity for CO<sub>2</sub> relative to O<sub>2</sub>. The oxygenase activity produces phosphoglycolate, and thence glycolate. Recycling this glycolate to carbohydrate has a significant energy cost; the excretion of glycolate in aquatic organisms imposes an even greater energy cost for net CO<sub>2</sub> fixation. The energy cost of the net fixation of 1CO<sub>2</sub> from the present atmosphere by a vascular land plant with C<sub>3</sub> physiology (CO<sub>2</sub> reaching RUBISCO by diffusion) is some 2.8 NADPH and 4.2 ATP. This contrasts with the 2NADPH and 3ATP required when CO<sub>2</sub> is saturating and, regardless of O<sub>2</sub> concentration, there is no oxygenase activity (see Table III).

C<sub>3</sub> physiology occurs in the great majority of species of terrestrial embryophytes, a few lichens and asymbio-

tic terrestrial algae, and some aquatic algae and embryophytes. Some red algal RUBISCOs have the highest reported selectivity for CO<sub>2</sub> over O<sub>2</sub>, so that glycolate production may be minimal in red algae with C<sub>3</sub> physiology. The algae, lichens, and bryophyte gametophytes with C<sub>3</sub> physiology have limitations on photosynthesis by CO<sub>2</sub> diffusion that vary with the structure of the photosynthetic apparatus, whether the medium supplying inorganic carbon (CO<sub>2</sub> + HCO<sub>3</sub><sup>-</sup>) is air or water, the CO<sub>2</sub> concentration in the medium, and the thickness of the diffusion boundary around the photosynthetic structure. In sporophytes of embryophytes in water the situation is as for gametophytes, whereas for sporophytes in air there is the additional complication of stomata. These variable conductance apertures connecting intercellular gas spaces to the atmosphere are crucial in maintaining land plants in a hydrated state when soil water availability is low and/or the atmosphere is very dry, and permit CO<sub>2</sub> fixation in the light when the plants can remain hydrated with open stomata. In most cases the C<sub>3</sub> vascular plant sporophytes on land fix CO<sub>2</sub> under conditions in which the rate of fixation is about one-third constrained by CO<sub>2</sub> diffusion and two-thirds constrained by biochemical reactions. In other words, were CO<sub>2</sub> diffusion constraints to be reduced by 10%, the rate of photosynthesis would increase by 3%; if the constraints from biochemistry were reduced by 10%, the rate of CO<sub>2</sub> fixation would increase by 7%. The fractional limitation of photosynthesis by CO<sub>2</sub> diffusion

in aquatic plants with  $C_3$  physiology is more variable than in the terrestrial vascular plant sporophytes with this photosynthetic physiology.

In a number of species of  $O_2$ -evolvers the supply of  $CO_2$  from the environment to RUBISCO involves some energized process that gives a higher  $CO_2$  concentration, and  $CO_2/O_2$  ratio, near RUBISCO than that in the medium. Such mechanisms suppress oxygenase activity, in some cases almost completely, and frequently increase the affinity of *in vitro* photosynthesis for  $CO_2$  relative to the situation with  $C_3$  physiology (diffusive supply of  $CO_2$  to RUBISCO).

One group of these  $CO_2$ -concentrating mechanisms does *not* involve covalent bond formation between  $CO_2$  and an organic compound before RUBISCO activity. These mechanisms have considerable diversity, and can involve active transport of  $CO_2$ ,  $HCO_3^-$ , or  $H^+$  as means of concentrating  $CO_2$ . Such processes occur in many aquatic  $O_2$ -evolvers, including all of the cyanobacteria, many of the algae, and some of the submerged vascular plants found in aquatic environments, and in a few terrestrial  $O_2$ -evolvers, including all of the cyanobacteria, many of the terrestrial algae and lichens, and, among embryophytes, certain hornworts. There is an energetic cost of operating the  $CO_2$ -concentrating mechanism; this has not been well defined for any mechanistic variant. In the case of cyanobacteria and dinoflagellates, the kinetic properties of RUBISCO mean that  $C_3$  physiology is not an option for growth in an air-equilibrium solution. In some of the other organisms, the selective significance of a  $CO_2$ -concentrating mechanism is quantitatively less clear (see Table III).

A second group of  $CO_2$ -concentrating mechanisms *does* involve covalent bond formation between  $CO_2$  and an organic compound in the delivery of  $CO_2$  from the environment to RUBISCO. These mechanisms are essentially restricted to vascular plants, with  $C_4$  metabolism occurring only in certain flowering plants, whereas CAM (Crassulacean Acid Metabolism) is found in a wider range of vascular plants. Here the initial carboxylation reaction preceding RUBISCO is almost invariably catalyzed by phosphoenolpyruvate carboxylase (PEPC) to yield oxaloacetate, which is subsequently converted to malate or aspartate. In the case of  $C_4$  metabolism, the formation of malate or aspartate occurs in mesophyll cells, with transfer of the  $C_4$  acids to bundle sheath cells that contain RUBISCO but have little capacity for  $CO_2$  exchange with the atmosphere. In the bundle sheath cells the  $C_4$  dicarboxylate releases  $CO_2$ , which is then fixed by RUBISCO, and the  $C_3$  monocarboxylate residue is returned to the mesophyll, where it can be used to form more  $C_4$  dicarboxylate. In  $C_4$  plants the spatial

separation of carboxylation and decarboxylation between two cell types involves only a temporal separation of seconds or tens of seconds. For CAM carboxylation to produce the  $C_4$  dicarboxylate, malate and its decarboxylation to regenerate  $CO_2$  occurs in a single cell type, but with an  $\sim 12$ -h diel separation of carboxylation (in the dark) and decarboxylation (in the light). The energy cost of these biochemical  $CO_2$ -concentrating mechanisms in addition to the 3ATP and 2NADPH required for the photosynthetic carbon reduction cycle in the absence of oxygenase activity is  $\sim 2$ ATP for  $C_4$  and  $\sim 3.5$ ATP for CAM, giving a total per  $CO_2$  fixed of 5ATP and 2NADPH for  $C_4$  and 6.5 ATP and 2NADPH for CAM (see Table III).  $C_4$  plants are almost all terrestrial; while there are very large numbers of terrestrial CAM plant species (more than 20,000), there is also a significant aquatic representation ( $\sim 100$  species). In the case of terrestrial plants,  $C_4$  plants lose less water in transpiration per unit carbon fixed than do  $C_3$  plants, but with the similar or greater productivities; CAM plants are even more economical in terms of water loss, but generally have lower productivities when similar life-forms are compared.

## V. CONCLUSIONS

The core reactions of primary photochemistry (RC1, RC2 in Bacteria and eukaryotes; bacteriorhodopsin and halorhodopsin in Archaea) show relatively little variation in structure or overall function, although there are significant differences (e.g., RC2 in  $O_2$ -evolvers can dehydrogenate water). Similarly, the quinones, the cytochromes  $b-c_1$  and  $b_6-f$ , cytochrome  $c$ , and ferredoxin also show little variation, as do plastocyanin (alternative to cytochrome  $c_6$  in some  $O_2$ -evolvers) and flavodoxin (alternative to ferredoxin in some phototrophs). ATP synthetase also shows great similarity among bacterial and eukaryotic phototrophs. There is more variation among the reactions upstream of the photochemical reactions that are involved in photon harvesting and excitation energy transfer.

The core  $CO_2$  fixation reaction using RUBISCO in  $O_2$ -evolvers and Proteobacteria is supplemented in some  $O_2$ -evolvers by a diversity of  $CO_2$ -concentrating mechanisms. In many aquatic plants there are  $CO_2$  pumps not based on preliminary carboxylation and decarboxylation, whereas these reactions are found in  $C_4$  and CAM. Bacteria other than Proteobacteria fix  $CO_2$  via reactions not involving RUBISCO.

### See Also the Following Articles

ARCHAEA, ORIGIN OF • BACTERIAL BIODIVERSITY • CARBON CYCLE • EUKARYOTES, ORIGIN OF

### Bibliography

Badger, M. R., Andrews, T. J., Whitney, S. M., Ludwig, M., Yellowlees, D. C., Leggat, W., and Price, G. D. (1998). The diversity and coevolution of Rubisco, plastids, pyrenoids, and chloroplast-based CO<sub>2</sub> concentrating mechanisms in algae. *Can. J. Bot.* **76**, 1052.

Blankenship, R. E., Madigan, M. T., and Bauer, C. E. (eds.). (1993). *Anoxygenic Photosynthetic Bacteria*. Kluwer, Dordrecht, Netherlands.

Falkowski, P. G., and Raven, J. A. (1997). *Aquatic Photosynthesis*. Blackwell Science, Malden, Massachusetts.

Hall, D. O., and Rao, K. K. (1994). *Photosynthesis*, 5th ed. Cambridge University Press, Cambridge, United Kingdom.

Lawlor, D. W. (1993). *Photosynthesis: Molecular, Physiological and Environmental Processes*, 2nd ed. Longman Scientific, London.

Nicholls, D. G., and Ferguson, S. J. (1992). *Bioenergetics 2*. Academic Press, London.



# PHYLOGENY

Kevin C. Nixon  
*Cornell University*

---

- I. Historical Overview
  - II. Analytical Methods
  - III. Measures of Support
  - IV. Data Sources
  - V. Phylogenetic Interpretation
  - VI. Phylogeny and Classification
  - VII. Applications of Phylogeny
  - VIII. Conclusions
- 

## GLOSSARY

**homology** Any similarity that is explained by common ancestry.

**monophyletic group, monophyly** A group of taxa that includes all of the descendants of the most recent common ancestor for the group.

**paraphyletic group, paraphyly** A group of taxa that includes some, but not all, of the descendants of the most recent common ancestor for the group.

**synapomorphy** A character state that is derived relative to the ancestral state (plesiomorphy). Synapomorphies are special cases of homology that are evidence of monophyly among the taxa that bear them relative to taxa that bear the plesiomorphic state(s).

---

**A PHYLOGENY IS A PATTERN** of historical evolutionary relationship among species and higher level taxa that is often presented as a tree diagram, or phylogenetic

tree. The term phylogenetics is often applied to the study of such relationships. Historically, phylogenetic trees were often produced by indirect methods and were not reproducible. Classifications often had little if any direct relationship to phylogeny. Modern phylogenetics utilizes cladistic methods to construct phylogenetic trees based directly on morphological and molecular data. For those who distinguish cladistics from phylogenetics, cladistics refers only to the methods by which the branching patterns are generated (e.g., parsimony or maximum likelihood) while phylogenetics refers to the interpretation of such diagrams as historical patterns. This is a useful distinction, since cladistic methods are neutral to the type of data and the resulting interpretations, and could be applied to nonphylogenetic problems (e.g., recovering or imposing hierarchic structure within any system of objects with shared variation). Phylogenies are analogous to genealogies on the scale of species and higher level taxa (e.g., genera and families). Phylogenies are usually presented as treelike branching diagrams, in which taxa that are on the same branch are thought to be more closely related to each other than taxa that occur on different branches. Interpreting such diagrams as historical patterns requires a basic understanding of hierarchy, and phylogenetic trees are often incorrectly assumed to support particular historical suppositions (e.g., one modern taxon is “primitive”) that are not indicated by the results. Phylogenetic trees are increasingly useful in a broad array of biological studies as a basis for experimental design as well as the framework on which to generalize results. Additional uses of phylogenetic information include



measures of phylogenetic diversity, which can be used in making conservation and habitat preservation decisions.

## I. HISTORICAL OVERVIEW

In the late eighteenth century, taxonomy became stabilized with the work of Linnaeus and others, and the modern hierarchic system of classification (often called the Linnean system) became standard. Hierarchical classification systems in the late eighteenth and early nineteenth centuries were often illustrated with treelike diagrams, but in general these were not intended to show evolutionary relationship and should not be interpreted as such. Following Darwin and the ultimate acceptance of evolution as the explanation for the diversity and variation of life, efforts began to produce phylogenies (diagrams of higher level evolutionary divergence among taxa), which also were often presented as treelike diagrams. The methods by which these early phylogenies were produced varied, most often based on a combination of raw similarities among taxa and either implicit (or explicit) ideas about the way in which characters had evolved. For example, Bessey in the early twentieth century published a list of dicta indicating trends in morphological characters of flowering plants (e.g., flower parts primitively with many parts, becoming fewer) and used these dicta as an implicit means of deriving phylogenies. Anecdotal theories about character evolution have been utilized by many systematists as an explicit or implicit basis for major phylogenies even to the present day (e.g., Cronquist, 1981).

Early phylogenies were usually considered independently from systems of classification, or classifications were only partially based on the results of phylogenetic analysis. One of the most prominent features of classifications from this period that persists in many current classifications is the recognition of paraphyletic groups (see below).

In a modern discussion of phylogeny, there are three main aspects that need to be addressed. First is the issue of how the phylogeny is derived: the source of evidence and how that evidence is analyzed. Second, there is the nature of the resulting phylogenetic diagram, or pattern, and what information such a diagram conveys. Third, there is the relationship between a phylogenetic system (a phylogenetic tree) and a classification that may or may not be derived from such a phylogenetic pattern. All of these issues remain controversial at some level, with different workers taking different combinations of views on each topic. While some may

view these controversies as overwrought, one healthy consequence for science is the competitive improvement of methods and the need for explicit theoretical justifications by all participants in these arguments.

## A. Modern Phylogenetic Analysis

Beginning with the popularization of the works of Hennig in the late 1960s and early 1970s, the idea that raw similarity should be the basis for phylogenetic reconstruction was quickly overturned. One might refer to this as the “cladistic revolution.” By the mid-1980s Hennig’s notion that only special similarities, or synapomorphies, should be used to reconstruct relationships had become the dominant force in phylogenetic analysis. Although one still finds phenetic analyses published, these are often focused on infraspecific patterns where cladistic methods have not yet been fully developed. The vast majority of phylogenetic analyses in the late 1980s to the present were undertaken with cladistic methods, in particular parsimony (see below). Other related methods such as neighbor-joining and maximum likelihood have been embraced by a subset of workers, but for various reasons are still not as widely used as parsimony. Thus, the discussion here will focus primarily on parsimony.

## II. ANALYTICAL METHODS

### A. Phenetic Methods

Phenetics had its basis in the statistical literature that developed out of the “New Synthesis” and came to fruition with the advent of large and accessible computers in the late 1950s and early 1960s. The various methods of phenetic analysis are too many to enumerate here but generally can be classified into methods that produce treelike diagrams (clustering) and methods that produce scatter diagrams on two or more axes (ordination). Phenetic methods do share the premise that all similarities, whether derived or primitive, should be used equally in an analysis. It is debatable as to whether the original proponents of phenetic methods actually intended the results of such analyses to be interpreted as phylogenies, but cluster methods produce treelike diagrams that were often directly interpreted as phylogenetic trees. Various justifications for phenetic methods have been proposed, and indeed, the methods as well as the theoretical underpinnings are eclectic. It became apparent in the 1970s and early 1980s that phenetic methods were not defensible as methods for

recovering phylogenetic history, and parsimony analysis became the method of choice for most systematists interested in reconstructing patterns of divergence. Those interested in these debates can find entry into the literature by examining almost any issue of the journal *Systematic Zoology* from ca. 1970 to 1985. By the 1990s, phenetic methods had become rarely used except in particular subspecific and populational studies, and need not be considered further here.

## B. Phylogenetic Methods

From the standpoint of methods of analysis, there is little to be gained by separating phylogenetics from cladistics although it should be noted that cladistics is the broader of the two. Cladistics is derived from the Greek term *clado* for “branch” and was coined by opponents of such methods to refer to those who were more interested in the pattern of divergence than in the process and degree of divergence. Originally the term cladistics was mostly applied to parsimony analysis (and some related methods that have fallen out of favor) but at the present time cladistics might be more broadly interpreted to include at least two other methods, neighbor-joining and maximum likelihood. However, as with all such terminology, there are differences of opinion, and some workers would rather restrict the term cladistics to parsimony-based methods only. For the purposes of this discussion, we will use the term in the broader sense. Any method that results in a (nonreticulate) tree diagram that is interpreted as representing the evolutionary history and relationships of taxa can be considered phylogenetic. Because of problems in interpreting most phenetic diagrams as phylogenies, those methods cannot be considered phylogenetic.

### 1. Parsimony

Parsimony analysis remains the method of choice for analysis of character data by most morphological systematists, and perhaps still even for molecular systematists. Modern parsimony analyses utilize matrices of either molecular or morphological data, or a combination of the two, to produce cladograms. Although Hennig did not propose parsimony analysis as we now implement it, his works provided the basis for the mathematical and theoretical underpinnings of parsimony analysis that were largely developed by Farris and others from 1969 to 1981 in a series of papers published in *Systematic Zoology* and elsewhere. Parsimony is derived from Occam’s razor—the idea that the simplest explanation is always preferable. Parsimony analyses (when successful) result in cladograms (branching diagrams)

that have the fewest number of steps (character transformations) when characters are optimally mapped onto the diagrams. Such trees are usually more or less directly interpreted as phylogenies, and character steps on these trees represent the minimum possible number of evolutionary changes to explain the distribution of the character. Such an interpretation does not require the assumption that evolution itself is parsimonious; indeed, most such trees based on real data have multiple origins for many, if not most, characters in the matrix. Parsimony simply eliminates the need for ad hoc explanations for multiple character origins by reducing them to the minimum possible number.

### 2. Parsimony Analyses

One major difficulty in parsimony analyses (and to an even greater degree in maximum likelihood) is the computational difficulty in finding most parsimonious solutions. There currently exist no methods for directly calculating most parsimonious trees. Because of this, the problem is directly related to the number of possible trees to be examined and is considered to be NP-complete. Because there is no direct way to calculate shortest trees, brute force methods that examine many possible trees must be used. These methods are generally implemented by “branch swapping” on a starting tree and examining the length of each new tree generated by moving branches to different positions on the tree. For very small data sets (<15 taxa), it is possible to examine all possible trees and be certain that one has collected all of the most parsimonious trees. For slightly larger data sets (possibly up to 30 taxa, depending on the quality of data), it is possible to use a “branch and bound” algorithm that reduces the number of trees examined but still guarantees that all shortest trees are found. However, for the vast majority of analyses, more than 30 taxa, and in recent analyses, as many as several hundred taxa (or “terminals”), are included. It is important to note here that the number of characters (e.g., the length of a DNA sequence) only affects computation in a roughly linear manner, while computation time is affected logarithmically as the number of taxa increases. With larger analyses, branch swapping is used under various “quick” strategies to find locally optimal trees and these trees are then used as starting points for additional searches. Recent advances in search strategies have resulted in programs that can easily analyze data sets of several hundred taxa with a very high confidence that shortest trees have been found (see Goloboff, 1999). Currently, for matrices larger than 1000 taxa, quick estimation methods such as the jackknife are used instead of directly searching for shortest trees.

### a. Homology

The term homology, like the term monophyly, is used in various ways in the systematic literature. However, most recent workers accept the term to mean a shared similarity that is due to (explainable by) common descent. Thus, any similarity that is not explained by common ancestry is not homology. Characters that are initially thought to be homologous but are found by analysis to be nonhomologous are termed homoplasious. Homoplasy is best considered as error in homology assessment, which can be explained by parallelism, convergence, or other more complex processes such as hybridization and retained polymorphism. Homologous characters by definition are evidence of relationship. One theoretical basis for cladistic analysis using parsimony is that parsimony maximizes the homology statements presented in the data matrix. Thus, when we score two terminals (taxa, species, etc.) as having the same state for a character, we are asserting a hypothesis of homology (the "primary" homology of some workers). Parsimony then resolves conflict among such statements by picking the tree (cladogram) that maximizes the number of correct homology assessments (or minimizes the number of incorrect assessments). In so doing, it minimizes the number of ad hoc explanations that are necessary to explain similarities that appear to be homologous at first, but are not based upon the most parsimonious tree. Farris and others have demonstrated that the shortest trees maximize explanatory power as well as information content of a classification and are therefore preferable even outside of an evolutionary context. It is important to note that parsimony does not assume that evolution is parsimonious, only that parallelism and convergence are not correlated in such a manner as to be misleading. Given randomly distributed parallel and convergent false homologies, with enough "true" homologies parsimony will resolve the correct tree (the tree that most accurately reflects phylogenetic history). This assertion has been challenged recently by proponents of maximum likelihood, but this controversy remains unresolved at this time and it appears that all methods (parsimony and maximum likelihood included) fail under certain circumstances.

### 3. Neighbor-Joining (NJ)

Neighbor-joining is a method that is arguably cladistic and is favored by many molecular biologists because it is rapid and generally (at least as commonly implemented) results in a single tree. The basic algorithm of NJ is similar to the Distance Wagner method of Farris and constructs a tree by calculating distances among terminals and adding terminals in such a way as to minimize

overall distance. While NJ is an expedient method for quickly obtaining an estimate of the most parsimonious tree, there seems to be no justification for preferring an NJ tree over a shorter tree obtained by parsimony analysis.

### 4. Maximum Likelihood (ML)

Maximum likelihood as a method of analysis is popular among many molecular biologists, particularly those who are more interested in models of evolution than in the actual phylogenetic pattern of taxon relationships. ML requires an explicit model of character transformation, with associated probabilities for each possible transformation from one state to another. Trees are searched in a manner similar to methods used in parsimony (e.g., branch swapping) but each tree is evaluated not by overall length, but instead by a measure of compound probability. Those trees with the highest compound probabilities (maximum likelihood) of character distribution are selected as best. Within certain theoretical frameworks, parsimony can be viewed as a particular form of maximum likelihood with reduced assumptions and infinite parameters. Under such circumstances, parsimony analyses and maximum likelihood analyses will give the same results. Under most circumstances, maximum likelihood and parsimony analyses of the same data sets have provided very similar results. However, at the present time maximum likelihood is not feasible for larger data sets due to massive computation times (at least with today's hardware and software). The computational problems are unlikely to be resolved by the mere improvement of hardware and will require advances in software that are probably not possible with current engineering capabilities in this field. However, it is possible that recent advances in algorithms for parsimony searches can be incorporated into maximum likelihood programs with similar relative levels of improvement. Because of the additional complexities of ML, even with such improvements, the ability to perform maximum likelihood analyses on the ever-larger data sets being produced with molecular techniques will lag behind parsimony for the foreseeable future.

## III. MEASURES OF SUPPORT

### A. The Bootstrap

Many workers wish to place some measure of statistical confidence on the phylogenetic trees that they construct. Because of theoretical issues, it is preferable to

refer to such statistical measures as support values as opposed to confidence intervals. The most widely used support measure in recent literature is the bootstrap. The bootstrap is calculated by permuting the original data set numerous times by sampling characters with replacement and recalculating the most parsimonious trees for each permuted data set. Usually, from 100 to 1000 separate permuted data sets are analyzed, and support for a particular clade can be evaluated as the number of separate replicates in which that clade occurs in the most parsimonious solutions. Thus, if a group is found in parsimony analysis of 90 of 100 permuted data sets, its bootstrap value would be 90. Theoretical issues surrounding the bootstrap include the assumptions about character distribution and sampling and the exact interpretation of the bootstrap values that are obtained for each branch in the tree. Because the relationship between the bootstrap values and the underlying "true tree" is not apparent, many workers prefer to view bootstrap values as an assessment of the strength of signal within a particular data set, and not as a statistical measure of confidence except in a very restricted sense. This has resulted in development of an alternative terminology, under which bootstrap (and jackknife) values are referred to as "measures of support."

### B. The Jackknife

The jackknife is closely related to the bootstrap and is performed also by permuting the data set and counting the occurrence of groups (clades) among replicates. The jackknife differs from the bootstrap in that instead of resampling characters, characters are removed at a set level of probability—the probability of removal usually set to be the same as the probability of not being sampled in a bootstrap. The jackknife has some theoretical advantages over the bootstrap and is generally faster to calculate because of very fast software that is available. The jackknife has been used recently as the sole analytical method for some very large molecular data sets (e.g., Kallersjö's analysis of an rbcL data set for green plants with more than 2000 taxa). Indeed, some proponents of the jackknife have suggested its exclusive use in lieu of more intensive parsimony analyses, under the view that only well-supported clades need be recovered. This is not an abandonment of parsimony, but rather a rejection of poorly supported conclusions under a parsimony optimality criterion.

### C. Bremer Support

Another measure of support commonly used in conjunction with parsimony analyses is Bremer support,

sometimes incorrectly termed the "decay index." Bremer support is measured by collecting suboptimal trees (trees of longer length than the shortest) and determining which groups are present. If, for example, a clade is present in shortest trees and all trees up to 2 steps longer, but is not present in all trees 3 steps longer, the Bremer support value for that group is 3 steps. While Bremer support is appealing because it utilizes parsimony directly, it is difficult to compare Bremer support values among different data sets. In general, the bootstrap and jackknife have been preferred in recent years as measures of support, and especially for large data sets, the former two are much easier to calculate than Bremer support.

## IV. DATA SOURCES

Phylogenetic analyses have evolved from being entirely or largely based on morphological data to being more and more based on molecular data, in particular DNA sequence data. There are both theoretical and practical reasons for this shift. First, many believe that molecular data are more objective and that character definition and scoring of morphological characters is both more subjective and subject to error. Because the expense of sequencing DNA has dropped dramatically, it is arguable now that it is actually more cost-effective to extract and sequence DNA than to spend large amounts of time tediously collecting morphological data. One problematic aspect of DNA sequence data is the lack of such data for fossil taxa. Fossil taxa, when available and when well preserved, can often be included in morphological cladistic analyses and may provide additional insight into patterns of relationship. Occasionally, such fossils even may have dramatic effects on the position of particular taxa. A major caveat of the inclusion of fossil taxa in cladistic analyses is the large amount of missing or incomplete data for fossils, especially in "soft" features that do not preserve well. Terminals with large amounts of missing data have the potential to "move" on the most parsimonious trees, resulting in many equally parsimonious solutions and in essence "deresolving" the tree. Such taxa that move because of ambiguity in data are often referred to as "wildcards" because of the propensity for them to match multiple terminals in the analysis. In general, it is best not to include such fossils when there is evidence that they are not contributing relevant information to the analysis but instead are reducing the precision of the results.

In the early 1990s a debate developed about the best way to treat disparate data sets, such as different gene

sequences, or gene sequences and morphology. One side suggested that data sets should be analyzed separately unless it could be shown that they were congruent, while the other proposed that data sets should be combined as the best method to resolve such incongruence. Currently, it appears as though the latter position is more popular, since most multigene studies combine data sets in a simultaneous analysis regardless of whether tests indicate congruence.

The availability of vast quantities of molecular data across broad groups of taxa holds the promise of providing more or less complete large-scale phylogenies in the near future. Theoretically, larger analyses with more taxa should provide better estimates of relationship that are less sensitive to the addition of new data. Using parsimony jackknifing, Lipscomb and co-workers (1998) analyzed all of the eukaryotes in the Ribosomal Database Project (480 eukaryotes, 15 prokaryotes as outgroups). This analysis confirmed previous morphological and molecular studies that indicated that the protista and fungi were not monophyletic. However, this analysis showed that several of the startling and “new” findings of molecular systematics were probably the result of analysis of small data sets. In particular, there was no evidence for placing *Giardia* and its relatives as the most primitive eukaryotes nor any clear indication that the Dinoflagellates were sister to the Apicomplexa.

## V. PHYLOGENETIC INTERPRETATION

One aspect of phylogeny and phylogenetic reconstruction that often presents difficulty is the interpretation of phylogenetic trees as evolutionary diagrams. Imprecision in both terminology and concepts has resulted in interpretations that are often inaccurate and unsupported. One of the most common mistakes in interpreting cladograms (or phylogenies) is the tendency to view the trees as providing a linear sequence of events, with a “first branch” that bears a more primitive group, followed by a “second branch,” etc. Such terminology generally confounds group size with time of origin, usually placing the smaller extant group of two concordant groups at the “base” of the tree. Thus, recent molecular analyses of the angiosperms have often been characterized as placing *Amborella* at the base or as the first branch of the angiosperms (see Fig. 1). However, the correct way to view such diagrams is in a *dichotomous* linear time scale, with each successive dichotomy younger than the preceding. In Fig. 1, time can be viewed as proceeding from left to right. *Amborella* is not itself the first branch anymore than the angiosperms

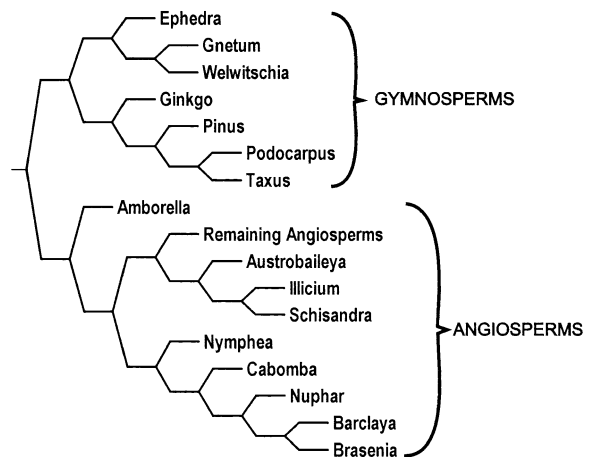


FIGURE 1 Reduced phylogenetic tree based on a recent large-scale molecular parsimony analysis of seed plant relationships. The results suggest that *Amborella* is the remaining extant sister taxon of the remainder of angiosperms; note that this tree does not indicate that *Amborella* is “the most primitive” flowering plant. See text for discussion.

minus *Amborella* is the first branch—the first two branches of the angiosperm tree bear a single species, *Amborella*, on one side and a few hundred thousand species (the remainder of angiosperms) on the other side. In terms of primitive or derived features, *Amborella* might be more derived than numerous persisting species on the other, larger branch of angiosperms and should not be viewed simply as the “most primitive” angiosperm. For instance, any one of several taxa near the base of the diagram (e.g., *Nymphaea*) might be more similar to the primitive angiosperm than is *Amborella*. However, *Amborella* might be considered to be the “most unique” angiosperm since it is the sole representative of one of the two basal clades (based on extant species alone) of angiosperms, and in that sense is the equivalent of the remainder of angiosperms. This simple and direct view of branching patterns in a phylogeny is not only the correct way to interpret phylogenetic diagrams, but it also provides the basic concept of phylogenetic diversity, or the measure of phylogenetic uniqueness, that is outlined below.

Another common mistake in dealing with phylogenies is the interpretation of polytomies, or multifurcations in cladograms, as indicating rapid radiation from a single ancestral plexus (see Fig. 2). Often, such patterns are referred to as “star phylogenies.” Although evolution may not always produce dichotomous divergences, it is not possible to positively identify multifurcations as rapid divergences. Instead, such patterns may simply reflect a lack of sufficient data to resolve more completely bifurcating trees. In general, polytomies re-

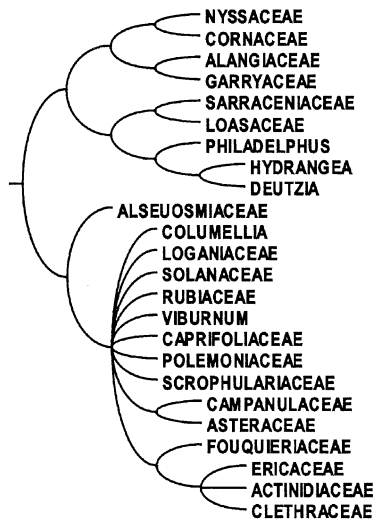


FIGURE 2 A portion of a phylogenetic tree based on a consensus tree from a parsimony analysis of a selected group of families of flowering plants. Note the large polytomy, or “star phylogeny” as such patterns are sometimes called. Such structures do not necessarily indicate rapid divergence from a common ancestor, but instead reflect lack of sufficient evidence, or conflicting evidence, regarding relationships. The graphic tree style was selected to more clearly indicate polytomies.

reflect a lack of conclusive evidence that is due to either character conflict or lack of data. Polytomies are best viewed simply as areas of the tree that need additional data to resolve, unless other independent data support a hypothesis of rapid divergence at the point of the polytomy.

Another frequently misunderstood aspect of phylogenetic analysis is the role of *a priori* character polarization. In the 1970s, many of the computer programs used to calculate phylogenetic trees required *a priori* character polarity (the assignment of “primitive” and derived states). However, strict parsimony analysis will produce the same topology regardless of the ultimate character polarity of the characters; in fact, parsimony computer programs do not consider character polarity during tree searches. Beginning in the late 1970s, it became apparent that the correct method for rooting trees was to select one or more outgroups that were included in the analysis, and the resulting trees rooted between the outgroup(s) and the ingroup. Although the best choices for outgroup taxa are those that are closely related to the ingroup, virtually any taxa not part of the ingroup may be used. However, the more distant the outgroups, the less accurate the rooting is likely to be, and particularly with molecular data, if the outgroups are too far removed, virtually random patterns of rooting may emerge. Character polarity be-

comes a matter of reading character distributions from the resulting trees when the outgroup method is employed, not an *a priori* exercise. Not only is this far easier than attempting to assign polarities to characters directly, but it reduces a major source of subjectivity in the analysis of data. Maximum likelihood models that allow equal probability for changes in any direction are also calculated independent of *a priori* rooting, but any models that have asymmetric models of character change are sensitive to a choice of a root, or starting point.

## VI. PHYLOGENY AND CLASSIFICATION

The issue of how to classify taxa into formal named systems may seem trivial to some biologists, but it has a profound impact on the way that we think about living things, and therefore does have an impact on the design and implementation of many biological studies. Statements such as “reptiles are cold blooded” or “reptiles have scales” or “the dinosaurs went extinct in the late Cretaceous” all have different interpretations depending on the classification in use. Characterization of paraphyletic groups on the basis of widespread and notable characters (e.g., the scales of “reptiles” or the naked seeds of “gymnosperms”) may provide useful means of communicating, but in formal classifications paraphyletic groups often provide misleading implications about relationship. Thus, if we include crocodiles in the group reptiles, we might guess that crocodiles are more closely related to lizards than to birds; this in fact is not the case, based on several lines of evidence when analyzed cladistically that reveal that crocodiles are more closely related to birds than to other “reptiles”—in other words, crocodiles and birds form a monophyletic group that excludes other reptiles and mammals (see Fig. 3). The advantage of a phylogenetic classification lies in the information content; if we know that a classification adheres strictly to the rule that all named groups are monophyletic, then all members of the group will be more closely related to each other than to any taxa *not* included in the group. Traditional groups such as “reptiles” (or formally, Reptilia) do not meet such criteria, and unfortunately it is impossible to know from such a nonphylogenetic classification that crocodiles (when placed within the “reptiles”) are more closely related to birds than to other “reptiles.” This observation at once supports both the need to abandon paraphyletic groups in formal classifications, and in doing so, to adopt a system of classification that is based consistently on phylogeny—with named groups always

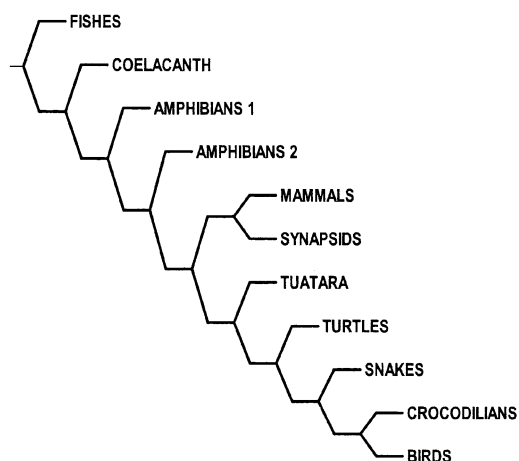


FIGURE 3 A phylogenetic representation of relationships among major groups of vertebrates, based on several published studies of morphological and molecular data. Note that crocodilians are more closely related to birds than to other “reptiles.”

corresponding to groups based on available evidence that are believed to be monophyletic.

Much of the difficulty that some systematists and many outside of systematics have with the abandonment of paraphyletic groups comes from the intuitive nature of such groups and their often easy characterization (e.g., reptiles are scaly, air-breathing, cold-blooded vertebrates that lay eggs). A classification that recognizes paraphyletic groups must do so based on primitive (plesiomorphic) features, as in the example just provided. The difficulty, and the arbitrary factor in such classifications, comes not in what to include in a paraphyletic group but instead in what to exclude. Because there is no objective way to determine what groups are significant enough to be excluded, the final decision is left to the authority, and in the strictest sense of the term, such classifications are authoritarian—and each authority may recognize different groups. Thus, removal of the birds from the group “reptiles” is based on a particular viewpoint about the importance of the characters (e.g., feathers) that the groups possess. However, another authority might just as easily argue to remove snakes (they lost those encumbering limbs) or turtles (the ultimate innovation in defensive armor). Decisions to exclude (and thereby elevate) groups such as birds are ultimately based on a view on the importance of particular characters, and such classifications do not accurately convey information about relatedness. Unfortunately, because paraphyletic classifications are based on arbitrary assignments of significance to characters, they are not even useful in the prediction of innovation or levels of anagenetic change (there is no measure

of the amount of change involved in the removal of concordant groups, only its significance). The realization that paraphyletic classifications are at their base arbitrary is the first step toward the acceptance of the need for classifications based solely on monophyly.

## A. Ranks and Priority

The need for a useful classification to be based on phylogeny has been confounded in recent years with other aspects of classification, such as ranking and the priority of names. Recently, a movement has emerged that has been self-proclaimed as “Phylogenetic Systematics,” which putatively embodies the ideals of Hennig and requires strict translation of phylogenies into named classifications. However, the need for maintaining only monophyletic named groups has been intertwined with other proposals, such as the abandonment of ranks, and proclamations that the existing Linnean system is incapable of accommodating a truly hierarchic system. Of course, the Linnean system is a strictly hierarchic system, and the claims that it must be abandoned are both premature and poorly reasoned. Unfortunately, many outside of this controversy have been misled to believe that the new proposals will provide greater stability in names, which is untrue. Stability in names in any phylogenetic system will come with stability in the underlying phylogenies on which the system of classification is based, and we are on the verge of achieving such stability with the vast quantities of molecular data and improved analytical algorithms that are becoming available.

## VII. APPLICATIONS OF PHYLOGENY

### A. General Evolutionary Context

In recent years, phylogeny has become an important aspect of many evolutionary and ecological studies. Phylogenetic studies focus on the pattern of history, but provide the foundation for both generalizations and specific conclusions about processes of evolution and diversification. Generalizations about evolutionary process based on phylogenetic trees are more predictive because phylogenetic trees provide information about character distributions as well as relationship. Phylogenies also provide the ultimate test of particular ideas or theories about the course of evolution. By mapping character distributions in the most parsimonious possible manner on a well-supported phylogeny, it is possible to understand whether characters of interest have evolved once or more than once and whether particular

characters are correlated in a historical sense. Attempts in the past to accomplish these goals were often based on paraphyletic classifications that were misleading or based merely on ranked classifications with unknown or unstated bases. Without a clear understanding of phylogenetic pattern, these studies often miscounted the number of times evolutionary steps occurred within and among taxa and came to incorrect conclusions about character evolution. Precise phylogenies now provide a means to improve such studies.

### B. Adaptation

Phylogenetic trees have become a standard tool in the study of adaptation, and such uses are often referred to as the “comparative method.” First, it is necessary to establish that a particular “adaptation” is distributed as an apomorphy within the group in question and then, if there are multiple origins, to determine if these origins are correlated with other characters and/or environmental variables. While numerous statistical approaches have been suggested for such studies, they all assume that multiple independent origins of characters correlated with environmental or historical factors are evidence of adaptation. Indeed, some workers maintain that it is only possible to discuss adaptation in a historical context, i.e., based on explicit phylogenetic trees. Undoubtedly continued work in these areas will result in improved statistical tests for adaptation based on character distributions on phylogenetic trees.

### C. Vicariance Biogeography

Phylogenetics has also become the main tool for biogeographic studies, and indeed the field of biogeography is now almost synonymous with vicariance biogeography. Vicariance biogeography is rich in theory and a thorough explication is beyond the scope of this article. The basic premise is that historical patterns of land connection and movement are recoverable through finding repeated patterns among the phylogenies of various plant and animal groups. Repeated patterns of clade distribution suggest a common history that is interpretable as supporting wider continuous distributions in the past. A well-known example of a broad vicariant pattern of distribution is found in the plant genus *Nothofagus* (Nothofagaceae) in the Southern Hemisphere, with species distributed in South America, New Guinea, Australia, New Caledonia, and New Zealand and fossils known from Antarctica. Similar patterns in other plants (e.g., Proteaceae) and animals (e.g., ratite birds) can be evaluated in a phylogenetic context to provide evidence of interchange among different “areas of endemism”

and develop an “area cladogram” that presents a pattern of relationship among the land areas. These patterns can then be interpreted in the context of plate tectonics as indicators of historical fission of land masses and subsequent plate movements.

### D. Phylogeny and Biodiversity: Phylogenetic Diversity

Phylogenies also provide a framework for alternative ways of looking at biodiversity. Most measures of biodiversity use species richness, either in a geographic (either broad or local) or ecologic sense. Other views of diversity may focus on the nature and breadth of adaptation. Such measures unfortunately require a subjective view of the importance of particular adaptations—e.g., the birds might be considered “diverse” because they include so many adaptations for use of the bill. However, phylogeny provides another perspective on biodiversity that allows an objective way to compare uniqueness and diversity of taxa. Although various specific measures of phylogenetic diversity have been proposed, most share a basic approach by which phylogenetic trees are used to evaluate species richness in concordant groups. It means little to say that “orchids are highly speciose” or “monotremes are species depauperate” unless we have some idea as to the relationships of the two groups being compared. Phylogenetic pattern provides the basis for such comparisons.

Currently, the use of phylogenetic diversity measures is largely limited to theoretical discussions and there have been few efforts to actually apply such measures to conservation. This is partly due to the relative paucity of high-quality phylogenies that are available across broad groups of taxa and partly because of a distinctly ecological bias in most studies of biodiversity. As molecular data sources provide better and more complete phylogenies for use by other workers, this is likely to change. It is probable that in the near future measures of phylogenetic diversity will become standard components, in combination with more traditional measures of ecologic uniqueness, species richness, and sensitivity, in the formulae that are used to evaluate conservation priorities for areas and endangered species.

## VIII. CONCLUSIONS

An understanding of phylogeny is increasingly important as a basis for experimental design and for interpretation of results in a broad array of biological studies. New methods of analysis, improved computer hardware



and software, and new and improved sources of data, most notably DNA sequences, have revolutionized our ability to construct phylogenetic trees. It is likely that our understanding of the origins and diversification of most major lineages of plants, animals, and microorganisms will improve dramatically over the next century, and we will arrive at a stable and robust reference system for all of life on earth.

### See Also the Following Articles

BIODIVERSITY, EVOLUTION AND • CLADOGENESIS • DARWIN, CHARLES • EVOLUTION, THEORY OF • TAXONOMY, METHODS OF • VICARIANCE BIOGEOGRAPHY

### Bibliography

- Cronquist, A. (1981). *An Integrated System of Classification of Flowering Plants*. Columbia Univ. Press, New York.
- Goloboff, P. A. (1999). Analyzing large data sets in reasonable times: Solutions for composite optima. *Cladistics* 15(4), 415–428.
- Hennig, W. (1966). *Phylogenetic Systematics*. Univ. of Illinois Press, Urbana, IL.
- Lipscomb, D., Farris, J. S., Kallersjo, M., and Tehler, A. (1998). Support, ribosomal sequences, and the phylogeny of the eukaryotes. *Cladistics* 14, 303–338.
- Nixon, K. C. (1996). Paleobotany in cladistics and cladistics in paleobotany: Enlightenment and uncertainty. *Rev. Paleobot. Palynol.* 90, 361–373.
- Schuh, R. T. (2000). *Biological Systematics: Principles and Applications*. Cornell Univ. Press, Ithaca, NY.



# PLANKTON, STATUS AND ROLE OF

C. S. Reynolds

*Freshwater Biological Association and NERC Institute of Freshwater Ecology*

---

- I. The Structure of Planktonic Communities
  - II. Habitat Constraints in the Plankton
  - III. Form, Function, and Selection in the Phytoplankton
  - IV. Form, Function, and Selection in the Zooplankton
  - V. Function in the Bacterioplankton
  - VI. Temporal Patterns in the Organization and Diversity of Planktonic Communities
  - VII. Mechanisms Promoting and Maintaining Diversity in the Plankton
  - VIII. Conclusions and Implications
- 

## GLOSSARY

**autotrophy** The ability of organisms to grow and reproduce independently of external sources of organic carbon compounds.

**eukaryote** An organizational state of cellular organisms in which the genome of the cell is stored in chromosomes enclosed in a membrane-bound nucleus; all protists (algae and protozoa), fungi, plants, and animals are eukaryotes.

**euphotic** The top layer of a water body through which sufficient light penetrates to support net photosynthetic gain and the growth of photosynthesizing organisms. Rarely more than 100 m in depth, the euphotic layer can be as little as 1 m in turbid waters.

**heterotrophy** The ability of organisms to grow and

reproduce on organic carbon sources, taken in dissolved or particle form.

**metazoan** Literally, a multicelled animal.

**mixotrophy** The ability of a normally autotrophic organism to switch, circumstantially, to phagotrophy, or to support an otherwise meager food supply by resorting to the ingestion and assimilation of bacteria or their products.

**pelagic** The (open-water) part of the aquatic environment that is far from the shore and the bottom bed.

**phagotrophy** A type of heterotrophy that involves the consumption of protists, plants, or animals as food.

**photoautotrophy** A type of autotrophy in which organisms gather light energy in order to reduce carbon dioxide to organic carbon; characteristic of green plants, most algae, and some prokaryotes.

**prokaryote** Organizational state of cells lacking a membrane-bound nucleus and certain other organelles. Bacteria, including the Cyanobacteria, are typically prokaryotic.

**picophytoplankton** The smallest (<2  $\mu\text{m}$ ) size class of photoautotrophic plankton.

---

**“PLANKTON” IS A COLLECTIVE TERM** for organisms adapted specifically for a life passed mainly in suspension in the open waters (the pelagic zone) of the sea and of such inland waters as lakes, reservoirs, and rivers. Planktonic organisms include protists (allegedly simple, unicellular, or colony-forming algal primary pro-

ducers and their protozoan consumers), microorganisms, and certain types of small metazoan animals, all sharing a common liability to passive entrainment in water currents, generated by tide, wind, convection, gravity, and the rotation of the earth. The inherent physical variability of open-water habitats typically favors absolutely short life histories; rapid changes in dominant species composition, in response to fluctuating environmental conditions, contribute to the maintenance of high biological diversity in individual habitats and to the survival of a high species richness among planktonic assemblages in general.

## I. THE STRUCTURE OF PLANKTONIC COMMUNITIES

The functional definition of plankton, ventured at the introduction to this chapter, has superseded the original, nineteenth-century allusions to plankton “floating” in water. Nevertheless, even this is still unsatisfactory, for its implication that the suspension is either complete or continuous is strictly erroneous. However, genuinely planktonic organisms—which include the plantlike, chlorophyll-containing primary producers of the phytoplankton, the heterotrophic, decomposer microorganisms of the bacterioplankton, and the more animal-like consumers of the zooplankton—are too small (often <20 mm) for their own intrinsic movements to be able to overcome, often or at all, the dispersive effects of water movement. Thus “embedded” within the tireless and unconstrained motion of open water, planktonic organisms broadly go wherever the flow takes them. In this way, the ecology of plankton is inextricably related to the physical properties of the medium, the extent and limits of its motion, and the environmental conditions set within these bounds (Reynolds, 1997b). This situation contrasts with that of most fish and other larger (>20 mm) animals of open water—the “nekton”—whose swimming strength is usually able to overcome normal movement of the water.

The older literature also promulgated a view that suspension in water was necessarily beneficial, supposing water to be something of an ideal habitat. Living in water does confer some notable positive advantages over terrestrial or aerial habitats. These include the mechanical (“Archimedean”) support water provides, as a consequence of its much greater density in comparison with air; its slow temperature fluctuations, as a consequence of its much higher specific heat than air;

and its solvent properties, which maintain nutrients and metabolic gases in readily assimilable state.

In truth, however, the planktonic ways of life have evolved to accommodate several problems and drawbacks associated with living in open water. Dominant among these is the issue of turbulence. Water molecules experience strong mutual attraction, which makes the liquid relatively viscous when compared to other fluids. Seeing waves break on the shore, or watching “white” water plunging through riverine rapids, we may be casually impressed by the fluidity of water flow but, without the driving energy, calm is rapidly reestablished as viscosity overcomes the residual motion at the molecular level. What happens is that the introduced mechanical energy is dissipated through a cascade of propagating eddies, of diminishing size and velocity, until molecular attraction imparts order over chaos and the molecular movement is overwhelmed. This behavior is now measurable and it has been mathematically described (see, for instance, Mann and Lazier, 1991). What is of particular interest here is that, depending upon the intensity of persistent wind- or gravitational forcing, viscosity overcomes inertia within the range 0.2 to 3 mm (see Reynolds, 1997b, for examples). This means that the immediate environment experienced by organisms smaller than this (i.e., most of the phytoplankton, bacterioplankton, and the smaller components of the zooplankton) is wholly viscous: far from being fluid, the forces acting on the microorganism are comparable to those experienced by a human immersed in treacle or unset cement. The organisms do not experience turbulence, neither are their delicate morphologies threatened with physical damage, but they remain entrained in the turbulent field and continue to be randomized throughout its spatial extent. Larger zooplankton (say >0.2 mm), though still too feeble to resist entrainment consistently, are sufficiently tough and flexible to tolerate the millimeter range of turbulence and to exploit it effectively in food gathering (Rothschild and Osborne, 1988).

Beyond the selective constraints imposed by the physical properties of pelagic, open-water environments, it is also necessary to recognize that, with respect to the obligate material components of living cells, the aqueous concentrations of some of these (especially carbon, nitrogen, phosphorus, iron, and fifteen or so micronutrient elements) are often so dilute that their availabilities place a severe constraint on the assembly of planktonic biomass. Moreover, despite its alleged transparency, the absorbance of solar energy by pure water (see, for instance, Kirk, 1994) is such that, at

depths of  $>100$  m, it is always as dark as night. Biological productivity in lakes is often severely constrained by rarefied resources or by deficiencies in processing energy, or by both. Far from being an ideal environment, the pelagic is a rather unpromising medium for successful exploitation by living communities.

Yet, within this general constraint, there is a remarkable richness of individual species inhabiting the plankton of the world's lakes and seas. Not all have even been adequately described and separated. The extraordinary diversity and phyletic representation of planktonic organisms may only be hinted at in the following subsections. As a preface to any such review of the planktonic biota, however, it is necessary to emphasize that the familiar separation, first of plants and animals and then their subdivision among phyletic divisions, cannot be applied too rigidly. This is not simply a reluctance to take sides with rival claimants about whether photosynthetic protozoa or bacterivorous algae are essentially plant or animal: most can be regarded as members of an ill-defined kingdom of Protista, which comprises mainly unicellular eukaryotic organisms, including the more plantlike photoautotrophs and the more animal-like phagotrophic consumers of other organisms or their products. The convenience of distinguishing planktonic "plants" and "animals" by function (Tables I and II) does not necessarily correspond to any fundamental evolutionary or phyletic separation. Even the names applied to the subdivisions follow past convention rather than convey any emerging understanding of the molecular affinities of the various groups of protists. The perplexity is yet greater when referring to the bacteria: whereas the lack of a membrane-bound nucleus and of certain other intracellular organelles provides a robust separation of their prokaryotic organization from the cells of eukaryotes, it is still difficult to distinguish among most bacteria other than by their biochemical activities and their affinities at the molecular level. Relatively "safe ground" is reached only among the more distinctive metazoan phyla, though even there, affinities among the main groups are often still obscure.

### A. Phytoplankton

Taking the capacity for photoautotrophy, the ability to manufacture organic carbon compounds through the photosynthetic reduction of carbon dioxide, to be the sole distinguishing criterion for separating them from other planktonic organisms, the phytoplankton is still extremely diverse. More than 4000 species of marine phytoplankton have been named and described (Sour-

nia *et al.*, 1991). The total number recorded from inland waters is not certainly known, but it is estimated that there are quite 4000 of these as well (Reynolds, 1996). Few genera and still fewer species are common to both fresh and salt waters. Even if a fairly conservative view of their classification is adopted, the species are drawn from at least six distinct protist phyla and at least two major prokaryote subdivisions (see Table I). The Purple (Chromatiaceae) and Green Sulfur Bacteria (Chlorobiaceae) are represented in specialized, anoxic habitats. Of the planktonic genera of Cyanoprokaryotes (formerly classed as Cyanophyceae, or "blue green algae," and now most commonly referred to as "Cyanobacteria") most occur in lakes, though several are also common in the low-salinity ( $<11$  parts per thousand) areas of the Baltic. "Sea sawdust" (*Trichodesmium* spp.) is found in low-latitude seas. The Cyanoprokaryotes are also well represented among the smallest marine and freshwater primary producers (the picophytoplankton: cells  $<2 \mu\text{m}$  in diameter; Waterbury *et al.*, 1979).

The more conspicuous components of the phytoplankton of the open sea belong to the Pyrrophyta (including a wide variety of dinoflagellate species) or to the Chrysophytes. This large division is taken to include the large number of diatom species, drawn from one or other of the two main orders (the centric Biddulphiales and the pennate Bacillariales), as well as a diversity of elaborate, scale-bearing Coccolithophorids. Besides diatoms and Cyanoprokaryotes, the more conspicuous components of the freshwater phytoplankton may be contributed from among the many chlorophyte and chrysophyte orders. However, cryptomonads, peridiniids, chloromonads, and euglenoids can all occur in large numbers at certain locations.

### B. Marine Zooplankton

The number of species of represented in the zooplankton of the sea is considerably enriched by the distinctive dispersal stages of many marine animals that spend their adult lives in the littoral or the benthos. To differing extents, these larvae (the amphiblastulae of sponges, the medusae and ephyrae of the cnidarian coelenterates, the pilidia of nemerteans, the trochospheres of polychaetes, the cypris larvae of cirripedes, the phyllosomae and zoeae of the eucarid malacostraca, the veligers of the lamellibranch mollusks, the various auriculariae, bipinnariae, and plutei of the echinoderms and the appendicularian larvae of the ascideans; see Table II) share the diminutive size ranges, membranous translucence, and feeble swimming movements charac-

TABLE I  
Phytoplankton in Freshwater and Marine Systems

Freshwater phytoplankton	Marine phytoplankton	Freshwater phytoplankton	Marine phytoplankton
Prokaryota		Class: Synurophyceae	
Division: Anoxyphotobacteria		Order: Synurales	
Family: Chromatiaceae		Synura, Mallomonas	
Thiopedia, Thiodictyon		Class: Bacillariophyceae	
Family: Chlorobiaceae		Order: Biddulphiales	
Chlorobium, Pelodictyon		Urosolenia, Aulacoseira,	Rhizosolenia, Cyclotella,
Division: Cyanoprokaryota		Cyclotella	
("blue-green algae")		Stephanodiscus	Chaetoceros, Thalassiosira,
Order: Chroococcales			Skeletonema, Ethmodiscus
Synechococcus, Microcystis	Synechococcus	Order: Bacillariales	Order: Bacillariales
Order: Nostocales		Asterionella, Synedra,	Asterionella, Nitzschia
Anabaena, Aphanizomenon	Trichodesmium	Fragilaria	
Cylindrospermopsis		Class: Haptophyceae	
Gloeotrichia		Order: Prymniales	
Order: Oscillatoriales		Chrysochromulina,	Chrysochromulina, Isochrysis,
Planktothrix, Limnothrix,		Prymnesium	Phaeocystis
Pseudanabaena, Lyngbya,		Order: Coccolithophorales	
Phormidium			Emiliana, Florisphaera,
Eukaryota			Gephyrocapsa, Umbellosphaera
Phylum: Cryptophyta		Class: Xanthophyceae	
Class: Cryptophyceae		Order: Mischococcales	
Order: Cryptomonadales		Monodus, Ophiocytium	
Cryptomonas, Chilomonas	Cryptomonas	Order: Tribonematales	
Rhodomonas		Tribonema	
Phylum: Pyrrophyta		Phylum: Euglenophyta	
Class: Dinophyceae		Class: Euglenophyceae	
Order: Peridinales		Order: Eugleninales	Order: Eugleninales
Peridinium, Ceratium,	Peridinium, Ceratium,	Euglena, Phacus,	Eutreptia
	Ornithocerus,	Trachelomonas	
Glenodinium	Dinophysis, Scrippsiella,	Phylum: Chlorophyta	
	Gymnodinium, alexandrium,	Class: Prasinophyceae	
	Gonyaulax, Gyrodinium	Order: Pedinomonadales	
Class: Adinophyceae		Pedinomonas	
Order: Prorocentrales		Order: Halosphaerales	
	Prorocentrum, Pyrocystis		Halosphaera
Phylum: Raphidophyta		Class: Euchlorophyceae	
Class: Raphidophyceae		Order: Volvocales	
Order: Chloromonadales		Chlamydomonas,	Dunaliella, Nannochloris
Gonyostomum		Volvox, Eudorina	
Phylum: Chrysophyta		Order: Tetrasporales	
Class: Chrysophyceae		Gemmellicystis	
Order: Bicosoecales		Order: Chlorococcales	
Bicosoeca		Chlorella, Ankyra,	
Order: Chromulinales		Pediastrum,	
Chromulina, Ochromonas,		Coelastrum, Dictyos-	
Dinobryon,		phaerium,	
Chryso-sphaerella, Uroglena		Scenedesmus,	
Order: Hibberdiales		Sphaerocystis	
Class: Dictyochophyceae		Order: Ulotrichales	
Order: Pedinellales		Geminelle	
	Pedinella	Order: Zygnematales	
Order: Dictyochaes		Closterium, Staustrum	
	Distephanus, Dictyochoa		

TABLE II  
Zooplankton in Marine and Freshwater Systems

Marine zooplankton	Freshwater zooplankton	Marine zooplankton	Freshwater zooplankton
Phylum: Mastigophora	<i>Bodo</i> , <i>Peranema</i>	Class: Branchiopoda Order: Anostraca	<i>Chirocephalus</i>
Phylum: Craspedomonadina (Choanoflagellates)	<i>Monosiga</i>	Order: Diplostraca (Cladocera) <i>Evadne</i> , <i>Podon</i>	<i>Diaphanosoma</i> , <i>Holopedium</i> , <i>Bosmina</i> , <i>Daphnia</i> , <i>Ceriodaphnia</i> , <i>Monia</i> , <i>Simocephalus</i> , <i>Bythotrephes</i>
Phylum: Rhizopoda Order: Amoebina	<i>Pelomyxa</i>	Class: Ostracoda <i>Gigantocypris</i>	<i>Cypria</i>
Order: Foraminifera <i>Globigerina</i>	<i>Arcella</i> , <i>Difflugia</i>	Class: Copepoda Order: Cyclopoidea <i>Oithona</i> <i>Chondracanthus</i>	<i>Mesocyclops</i> <i>Ergasilus</i>
Order: Radiolaria <i>Acanthometra</i>	<i>Actinophrys</i>	Order: Calanoidea <i>Calanus</i> , <i>Temora</i> , <i>Centropages</i>	<i>Eudiaptomus</i> , <i>Eurytemora</i> , <i>Boeckella</i>
Order: Heliozoa	<i>Nassula</i> , <i>Colpoda</i>	Class: Branchiura	<i>Argulus</i>
Phylum: Ciliophora Order: Holotricha <i>Colpoda</i>	<i>Euplotes</i> , <i>Caenomorpha</i> , <i>Halteria</i> , <i>Metopus</i> , <i>Strombidium</i> , <i>Tintinnidium</i>	Class: Cirripedia [ <i>Cypris</i> larvae]	
Order: Spirotricha	<i>Carchesium</i>	Class: Malacostraca Order: Peracarida (Suborder: Mysidacea) <i>Leptomysis</i> (Suborder: Cumacea) <i>Diastylis</i> (Suborder: Isopoda) <i>Eurydice</i> (Suborder: Amphipoda) <i>Apherusa</i>	<i>Mysis</i>
Order: Peritricha <i>Vorticella</i> , <i>Epistylis</i>		Order: Eucarida (Suborder: Euphausiacea) <i>Nyctiphanes</i> , <i>Euphausia</i>	<i>Macrohectopus</i>
Phylum: Porifera [amphiblastula larvae]		Phylum: Arthropoda Class: Insecta [all larvae] Order: Megaloptera	<i>Sialis</i>
Phylum: Coelenterata Subphylum: Cnidaria Order: Leptomedusae <i>Obelia</i>	<i>Limnocoidea</i> , <i>Craspedacusta</i>	Order: Diptera <i>Pontomyia</i>	<i>Chaoborus</i>
Order: Anthomedusae <i>Hydractinia</i>		Phylum: Mollusca Class: Gastropoda Order: Opisthobranchiata <i>Limacina</i> , <i>Clione</i>	
Order: Trachylina		Class: Lamellibranchiata [Veliger larvae]	<i>Dreissena</i> [veliger]
Order: Siphonophora <i>Veella</i> , <i>Physalia</i>		Phylum: Chaetognatha <i>Sagitta</i>	
Subclass: Scyphozoa <i>Aurelia</i> , <i>Cyanea</i> , <i>Pelagia</i>		Phylum: Echinodermata [larvae: plutei, auricularia, bipinnaria, etc.]	
Subphylum: Ctenophora Class: Tentaculata <i>Pleurobranchia</i>		Phylum: Chordata Subphylum: Tunicata Class: Ascidiacea [appencularia larvae]	
Class: Nuda <i>Beroë</i>		Class: Larvacea <i>Oikopleura</i>	
Phylum: Platyhelminthes Class: Turbellaria <i>Convoluta</i> , <i>Microstomum</i>		Class: Thaliacea <i>Doliolum</i>	
Phylum: Nemertea [pilidium larvae of certain Hoplonemertines]		Subphylum: Vertebrata Class: Actinopterygii [larval fish]	
Phylum: Nematoda [A few shelf-water species are described]			
Phylum: Rotatoria Order: Monogonota (Suborder: Flosculariacea) (Suborder: Ploima)	<i>Flinia</i> , <i>Conochilus</i> <i>Brachionus</i> , <i>Keratella</i> , <i>Kellicottia</i> , <i>Synchaeta</i> , <i>Asplanchna</i>		
Phylum: Gastrotricha [Encountered in plankton of small water bodies]			
Phylum: Annelida Class: Polychaeta <i>Tomopteris</i> [trochosphere larvae]			
Phylum: Crustacea			

teristic of the species which are planktonic throughout their lives. Among the smallest ( $<20\ \mu\text{m}$ ) examples of these heterotrophic protists are the nanoflagellates (some being closely allied to the photoautotrophic phytoflagellates, classified here as phytoplankton) and choanoflagellates. The microzooplankton fraction (in the size range,  $20\text{--}200\ \mu\text{m}$ ) includes the rhizopod foraminiferans and radiolarians, and a range of ciliate and suctorian ciliophorans. More conspicuous ( $0.2\text{--}20\ \text{mm}$ ) in the marine zooplankton are the ctenophoran "comb jellies" and sea-gooseberries, the chaetognath "arrow worms" (e.g., *Sagitta*), some specialized turbellarians (e.g., *Convoluta*, *Microstomum*) and polychaetes (e.g., *Tomopteris*), certain opisthobranch gastropods (such as *Clione*, *Limacina*), and the larvaceans (e.g., *Oikopleura*) and salps (e.g., *Doliolum*).

The most prominent of the animals of the marine plankton, however, are crustaceans. Most of the species are copepods, including calanoids, like *Calanus*, *Temora*, and *Centropages*, and cyclopoids, such as *Oithona*. There are also some representative cladoceran (e.g., *Evadne*, *Podon*) and ostracod (such as *Gigantocypris*) genera. The malacostracan orders are distinctively represented by organisms that remain planktonic in their adult stages. Mysidaceans (such as *Leptomysis*) are locally abundant and a few cumaceans (e.g., *Diastylis*) are remarkable for being nocturnally planktonic and diurnally benthic in coastal waters. Amphipods (e.g., *Apherusa*) and isopods (e.g., *Eurydice*) have planktonic representatives. Perhaps the most renowned component of the plankton of the high-latitude oceans, mainly for being one of the major food sources of the filter-feeding whales, is krill (*Euphausia* spp.). Elsewhere, the smaller euphausiids form an important food for pelagic fish: *Nyctiphanes* is one of the common genera in European coastal waters.

Some of the cnidarian jellyfish may be regarded as being the largest animals in the plankton ( $>20\ \text{mm}$ , perhaps to  $2\ \text{m}$ ). Though some of these are rather larger than some of the swimming organisms ("nekton": fish, cephalopods) that are excluded from the understanding of "plankton" (discussed earlier), the large jellyfish qualify for their poor ability to control their own movements in the sea. The siphonophores, like *Verella* and the Portuguese man o'war, *Physalis physalis*, are little more than drifting "polyp colonies." The true jelly-fish, which move themselves by slow, rhythmic pulsation of the umbrella-like manubrium, include the distinctive *Aurelia*, *Cyanea*, and *Pelagia*.

Finally, the young stages of several species of pelagic fish are of such diminutive size and swim so feebly and with weakness of motility that, for the first part of their lives, they are reasonably included among the plankton:

Clupeids (herrings, sardines) and Scombrids (Mackerel) fulfill this description; ultimately demersal Gadids (cods and allies) and benthic flat fish (Pleuronectids, Soleids) also pass planktonic dispersal stages.

### C. Freshwater Zooplankton

The phyletic representation in the zooplankton of lakes is well known to be relatively much poorer than in the sea, but mostly this reflects the poorer representation of the animal phyla among fresh waters anyway. The summary in Table II shows considerable coincidence of representation at the higher phylogenetic levels but, as with phytoplankton, there is almost no commonality of species.

Protists usually figure prominently in the plankton of inland waters, supposedly as a function of the organic matter available, including detrital particles and their associated bacteria. The smaller heterotrophic flagellates generally consume free-living bacteria, but many of the common planktonic microciliates, like *Coleps* and *Tintinnidium*, feed also on small algae and flagellates. In doing so, they fulfill an important linkage in the so-called microbial loop (Azam *et al.*, 1983): in many systems, the excretion of excess organic carbon fixed in photosynthesis, its assimilation by bacteria and its successive transfer through closely coupled flagellate-ciliate consortial linkages to copepod consumers and, eventually, planktivorous fish, demonstrably exceeds the carbon transfers along the conventional phytoplankton  $\rightarrow$  zooplankton  $\rightarrow$  fish trophic pathway.

Occasionally, the larger Amoebae and ciliates, such as *Nassula*, may dominate the zooplankton; this may owe to the prevalence of a particular food source, or to the fact that a crustacean plankton has not yet developed, or that some other factor (low oxygen concentration, for example) restrains potential competitors for the same food resource.

At least two freshwater coelenterate genera have planktonic medusae. The Gastrotrich phylum of minute, wormlike but unsegmented animals is occasionally represented by planktonic specimens. The phylogenetically close phylum of the Rotatoria is prominently represented in the freshwater plankton by some two dozen genera, drawn mainly from the Order Monogononta. The most ecologically important genera include *Asplanchna*, *Brachionus*, *Filinia*, *Keratella*, *Kellicottia*, and *Synchaeta*. Some colonial rotifers also feature in the plankton (e.g., *Conochilus*). Some are specialist feeders; most of those mentioned browse or filter-feed on bacteria, detritus, and planktonic algae, generally within species-specific size ranges (Pourriot, 1977).

As in the sea, the most prominent planktonic animals

are crustaceans. The Branchiopods are represented by the anostracan “brine shrimps” (e.g., *Chirocephalus*), especially in temporary waters, and by the Cladocera, the familiar “water fleas.” This grouping (its monophyletic origin is now doubted) includes the mainly herbivorous, filter-feeding species of *Daphnia*, *Ceriodaphnia*, *Moina*, *Simocephalus*, *Bosmina*, and *Holopedium*, and the predatory *Bythotrephes* and *Leptodora*. Planktonic Ostracods are noted from lakes in Southeast Asia (*Cypria javensis*) and from Laguna de Petén, Guatemala (*Cypria petenensis*). The Copepods are particularly well represented by calanoids (such as *Eudiaptomus*, *Eurytemora*, *Boeckella*) and by Cyclopoids (e.g., *Mesocyclops*); some parasitic Cyclopoids (e.g., *Ergasilus*) are dispersed through the plankton. The Branchiurans (e.g., *Argulus*), which are ectoparasitic on open-water fish, are certainly to be considered essentially planktonic. Though principally marine, the mysids are represented in the plankton of several high-latitude lake systems, where *Mysis* is regarded as a relict from the last glaciation. The Amphipod flag is carried by *Macrohectopus*, which is endemic to the plankton of Baykal Lake.

Among the arthropod insects, several genera of Diptera have larvae, which pass most of their time in the plankton. The most specialized of these are the juvenile chaoborines (e.g., *Chaoborus*) or phantom midges, whose transparent bodies reveal the internal provision of buoyancy-providing air sacs. The dispersal stages of aquatic larvae of other orders of insects sometimes show adaptations that are unmistakably planktonic: a striking instance is provided by the first instars of *Sialis* (Megaloptera; see Elliott, 1996).

#### D. Explaining Species Diversity

The foregoing passages confirm the richness of phyletic representation and the very large number of individual species that collectively contribute to the overall diversity of planktonic organisms. The challenge now is to account for the richness of the planktonic species and to explain the mechanisms by which it is maintained. At one level, we might be surprised that this problem arises at all. It should not be at all unreasonable to anticipate that, in a supposedly fluid and isotropic medium, fully accessible to suitably adapted species, Darwinian selection should move the structure of the organismic assemblage toward just a small number of specialists, each being the best-fit survivor in its key community role. Every less-fit competitor might be supposed to suffer progressive exclusion by the stronger species; overall, diversity should be suppressed. Thus, to confront the remarkable richness of planktonic plant-

and animal-species surveyed previously is counter-intuitive and paradoxical (Hutchinson, 1961).

Neither is it clear for how long such richness has distinguished the biota of open waters. Owing to the facts that most microorganisms do not form robust fossils, that given freshwater habitats are, in geological terms, very transient features, and that, at the relevant evolutionary scales, even the present ocean floors are relatively young, it is difficult to be categorical about the origins of planktonic communities. However, they are likely to be old. By the beginning of the Cambrian period, some 600 million years before present, when some of the oldest fossiliferous sedimentary rocks were formed, most of the invertebrate phyla represented in modern plankton had already appeared. These were wholly aquatic, though not necessarily planktonic. The Cyanobacteria had been established long before this, with primitive oxygenic photosynthesis coccoid species already converting the reducing conditions of the early planetary environment to an oxic one, between 2 and 2.5 billion years ago. Many of the Protistan groups, including several of the eukaryotic algae, had also appeared by the Cambrian (Ragan and Chapman, 1978). Presumably, some of these were free living, in suspension in the water. The step to a truly planktonic existence is supposed to be short and, with such a diverse phylogeny of modern plankton, it is reasonable deduction that it was taken within each evolving group, perhaps several times. It is probable that the first planktonic communities began to come into existence about a billion years ago, when many new opportunities for functional specialisms and the adaptive radiation of species were available.

Interestingly, the various Chrysophyte groups are considerably younger, there being no undoubted records from before the Mesozoic period (Tappan, 1980). Both the diatoms and the chromulines expanded and diverged during the Cretaceous era (135–65 million years ago); the coccolithophorid Haptophyceae also appear to have originated during the Mesozoic; it is their biomineralized remains which predominate the chalk deposits that have lent their name to the Cretaceous. Some very significant changes in the chemical composition of the sea must have occurred during that period.

While the same species may not have held sway throughout, or even for the past 100 million years, it is clear that the factors favoring a high planktonic biodiversity are recurrent, if not ongoing. Many theories have been advanced to explain Hutchinson's (1961) paradox of the plankton, but the conundrum for long remained unsolved. Either the number of occupiable niches must be far more numerous than had been sup-



posed or the competition was somehow incomplete in its effects.

Hutchinson himself suspected that the assumption of a homogeneous environment with steady-state properties was flawed. Collectively, the environments that planktonic populations inhabit are subject to huge variability in their chemical makeup and in their physical characters. They may be physically separated from each other (as are freshwater catchments), enjoying quite different climates or, even if contiguous (like the seas), their systems may be close to mutual isolation by currents and circulations. On the continents, lake basins are created and destroyed at, relative to evolutionary rates, high frequency. The skeletal understanding of the history of the world's oceans is that they have undergone large oscillations in metabolism and productivity associated with changes in the biospheric carbon cycles (Thierstein, 1989). Within the limits of habitat suitability that evolutionary specialization allows, the potential ranges of individual species should be distinct from those that have evolved separate specialisms. Superimposed on the longer-term changes are relatively higher frequencies of periodic forcing. Within these smaller habitat units, the higher frequency environmental oscillations might lead to alternations in the favored specialisms and consequent interspecific transfers of competitive advantage. Physical limitation of range (endemism) or otherwise (cosmopolitanism) would then represent the two extremes of dispersal efficiency across physical barriers and habitat preferences.

Current understanding of plankton ecology conforms to this general view in that it is certainly possible for the experienced observer to determine the broad habitat provenance of given planktonic assemblages: their species composition has an indicative value because similar assemblages characterize similar pieces of water. Students of the plankton have usually accepted a prevailing view that planktonic species are generally cosmopolitan and disperse freely, so that, on balance, most are able to establish quickly wherever suitable conditions arise. Dispersal mechanisms among the protists and prokaryotes are certainly effective (review of Kristiansen, 1996), and it is true that many conventionally identified morphotypes enjoy worldwide distributions. As more molecular information becomes available, the cosmopolitan nature of plankton is called increasingly into question. Besides, the notion that "everything is everywhere" is surely a matter of degree: ease and frequency of dispersal is not equal among planktonic species. Isolation and regional endemism certainly occur (Tyler, 1996), especially where long physical distances or significant, hostile barriers sepa-

rate sites of suitable habitat. It often takes anthropogenic intervention to bridge these gaps, as the recent "breakout" of the lamellibranch, *Dreissena polymorpha*, and hitherto endemic mysids from their original Caspian-basin locations into the waterways of Europe and, now, North America graphically reminds us. Equally remarkable is the arrival of the first freshwater Cladocera on Easter Island in 1780, which had failed to bridge the distance from the next nearest lake (over 3000 km distant) prior to Captain James Cook's requirement to replenish his ship's supply of drinking water (Dumont, 1999).

The same student of plankton will also be well attuned to the seasonality of the species composition of samples, as dominance moves frequently among the species of the assemblage. The differential responses of individual species of plankton to temperature or day length or nutrient resources are generally recognized to underpin what is perhaps the most familiar feature of planktonic communities—the so-called seasonal succession. As with other aspects of plankton ecology, the identity of its driving variables has been actively pursued and described by conceptual word models; the most compelling of these has been the PEG-model of Sommer *et al.* (1986). Like its less successful contemporaries, it is nevertheless founded on an implicit acceptance of the differences among planktonic species and the suitability of species-specific adaptations to particular habitat constraints.

The successful pursuit of viable explanations for the origins of biodiversity and the mechanisms of its maintenance must acknowledge a distinction between "local diversity," a measurable property of the temporal compositional fluctuations in response to local, within-patch variability, and the total species richness, or overall biological diversity, which is supported by the aggregate of patches and its continued ability to offer an adequate number of accessible habitats to satisfy the dynamics of survival of each species. The way to explain planktonic biodiversity is through a simultaneous recognition of the variables constraining habitat suitability and the adaptive specialisms and limitations of species for which the habitat constraints will select and those against which they will discriminate.

## II. HABITAT CONSTRAINTS IN THE PLANKTON

The assumption of a uniform, hospitable, steady state is erroneous so far as most open-water habitats are

concerned. On the contrary, many are as hostile as their desert-like barrenness conveys. Relative infertility may be locally or regionally attributable to an inadequacy of light energy to sustain planktonic primary production and the import of organic carbon is too modest to support the heterotrophs. Light income may be scarce for reasons of latitude (short day length, low solar declination), or it may be severely "diluted" over a deep mixed layer, or its penetration may be abruptly curtailed by a high concentration of inert particles (turbidity). Commonly, the concentrations of assimilable sources of certain primary nutrients place a very low ceiling on the biomass that can be assembled (nitrogen, phosphorus, and iron are cited most frequently as capacity-limiting factors). Chronic nutrient deficiency also interferes with the production of microbial biomass, even supposing there to be an adequate flux of organic carbon. For the planktonic animals, the problem is to be able to forage sufficient food; the concentrations and the size distributions of potential food particles influence profoundly the nature and abundance of the consumer populations. A dearth of primary-producer plant biomass is no basis for the intense production of zooplankton, neither will it underpin large harvestable crops of prime pelagic fish. What we find is that the locations where significant net primary production is possible and that the occasions when it is exploitable by heterotrophic consumption are, in fact, strongly circumscribed.

Besides the spatial differentiation of planktonic habitats, there is usually a considerable temporal variability. Conspicuous are the seasonal changes consequent on latitude: lengthening spring days and higher flux densities are generic to all temperate ecosystems but for small organisms with short life spans, the changes are perceived by successive generations and not as some perennial amplitude of fluctuation registered by a forest tree, for example. Ambient temperature may increase too, incidentally raising the threshold of wind energy required to keep the water fully mixed. The onset of thermal stratification, as a consequence of a shifting balance between the buoyancy forces brought through surface warming and the kinetic dissipation of the work of the wind energy, precipitates a train of environmental effects, leading to enhancement of the segregation of the warm, insulated, aerated, and increasingly resource-depleted epilimnion from a colder, darker, and potentially less oxidizing hypolimnion. Neither is such seasonality confined to high latitudes: even quite small differences in wind prevalence, cloud cover, humidity, and hydraulic exchange result in seasonal variations in ther-

mal stability and down mixing in the tropics (for examples, see Talling and Lemoalle, 1998).

So it is that the different kinds of water body (lake or pond, river, estuary, coastal shelf, upwelling, or the open oceans) each offer, to the most appropriately adapted species, or to those simply furnishing the largest inocula, varied and sometimes very transient opportunities to build their local populations. Furthermore, because the medium is fluid and the movement induced by wind or gravity is subject to fluctuations in strength, the persistence and vertical extent of an upper, differentiated layer is a highly variable character of the environment, which alters not just from season to season but from day to day and from hour to hour. Interest is also growing in the effect of year-to-year variations. At all these scales, the intensity and frequency of the environmental variability determine and, potentially, modify the critical habitat constraint and, hence, the attributes of organisms most likely to benefit dynamic performance.

It follows that selective advantage is likely to move among species. Environmental variability is a powerful influence on the assembly of planktonic communities and thus on their biological diversity. The causal linkages between observable patterns and processes in the maintenance of species diversity in the plankton may be usefully explored through a diagrammatic representation of the two principal variables characterizing particular open-water environments, namely, the fluctuating fluidity or vertical extent of the uppermost water layer and the availability of the resources to support the assembly of biomass. The layout of Fig. 1 is descended from Margalef's (1978) original scheme for marine phytoplankton, in which a "nutrient" axis is set against one of "turbulence"; the arrangement is amenable to tracking habitat variability and the changing species composition through time. The scheme has been developed for fresh waters to accommodate habitat conditions in which nutrient resources and the light availability in the surface-mixed layer are sufficient to saturate the fastest *in situ* algal growth rates (in the top left corner of Fig. 1) and to distinguish them from conditions that become either increasingly resource-constrained (moving downward in the matrix), or increasingly energy-deprived (moving rightward), or fall deficient in both (bottom right corner). The selection of given species or groups of species was found broadly to be correlated with the matrix space thus represented; high-performance, fast-growing algae would be favored in the relative paradise of the upper left-hand corner; algae favored by the second and third contingencies show some specialist adaptations, respectively, for op-

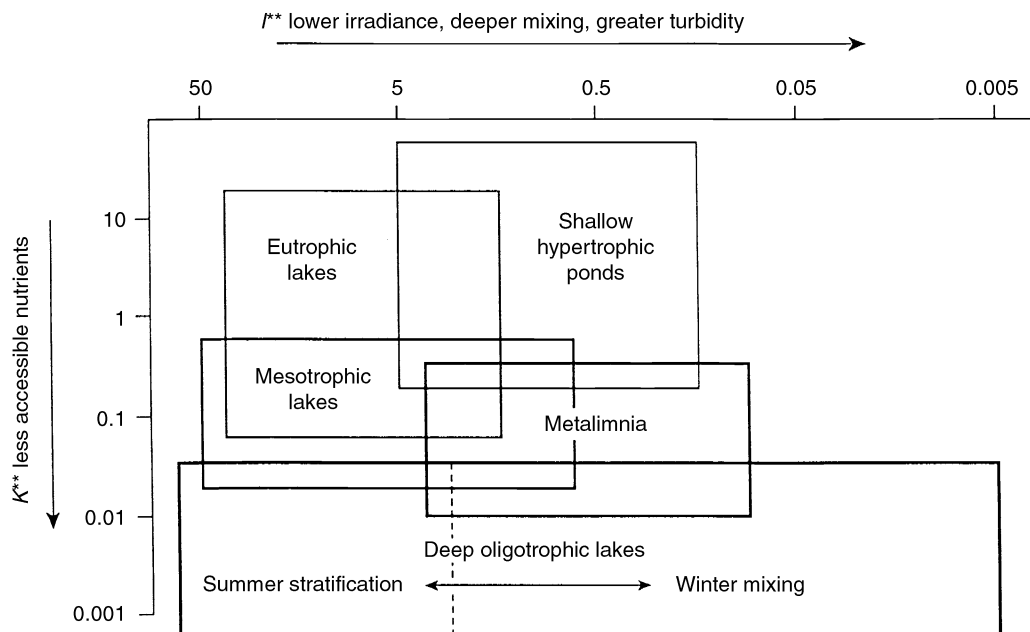


FIGURE 1 Two-dimensional representation of limnetic habitats, which, starting in the top left, distinguishes well-insolated, resource-replete water masses from (working down) long-stratified, resource-segregated water columns and from (working across) increasingly light-deficient environments.

erating under conditions of low concentrations and remote sources of nutrients, or for harvesting light from low irradiances or from infrequent short periods in the light field. The fourth, low-nutrient, low-light contingency is unfilled and is something of a phycological desert, evidently untenable as a habitat for a photoautotroph (Reynolds, 1988b).

It was never imagined that such a two-dimensional matrix could describe the entire spectrum of freshwater variability, but it has proved to be a consistent template for predicting the structural responses of planktonic assemblages to the main dimensions of environmental variability. Latterly, effort has been invested to improve both the utility and the quantification of the axes. The horizontal axis in Fig. 1 describes the integral,  $I^{**}$ , which, as the product of the physical depth of mixing and the vertical attenuation of a finite input of light energy, is sensitive to the constraints both of deep mixing and of high turbidity; quantities diminish rightward on a logarithmic scale. The vertical axis in Fig. 1 accommodates another composite scale,  $K^{**}$ , being the quotient of the concentration of the critical nutrient in the medium (usually the surface mixed layer) and the gradient of concentration between the top and bottom of the entire trophogenic layer. This scheme differentiates

sites that are chronically deficient in nutrients (points plotted well down the  $K^{**}$  axis) from, on the one hand, those in which steep gradients develop as a consequence of near-surface uptake (points track from high to low on the  $K^{**}$  axis) and, on the other, those in which the nutrient is scarcely exhausted (points located consistently toward the top of the  $K^{**}$  axis).

Against these axes, it is possible to characterize the "signatures" of seasonal changes in various kinds of water body, including of small, "eutrophic" lakes, in which nutrient availability is seasonally reduced, of shallow, fertile systems, in which the turbidity is the predominant variable, and of deeper, oligotrophic systems, where the depth of mixing is the strongest seasonal variable. An analogous (and in many ways, a more self-evident) approach has been developed recently for the sea (Fig. 2). This distinguishes energy-limited, well-mixed, high-latitude oceans from the highly stratified, nutrient-segregated waters of the tropics and will represent processes and interactions in coastal and shelf waters, including frontal zones and major upwellings of deep circulation currents.

The premise to be developed is that the biodiversity of plankton may be fitted to such templates and, moreover, that environmental change and variability select for alternative species. Reverberations in the selection

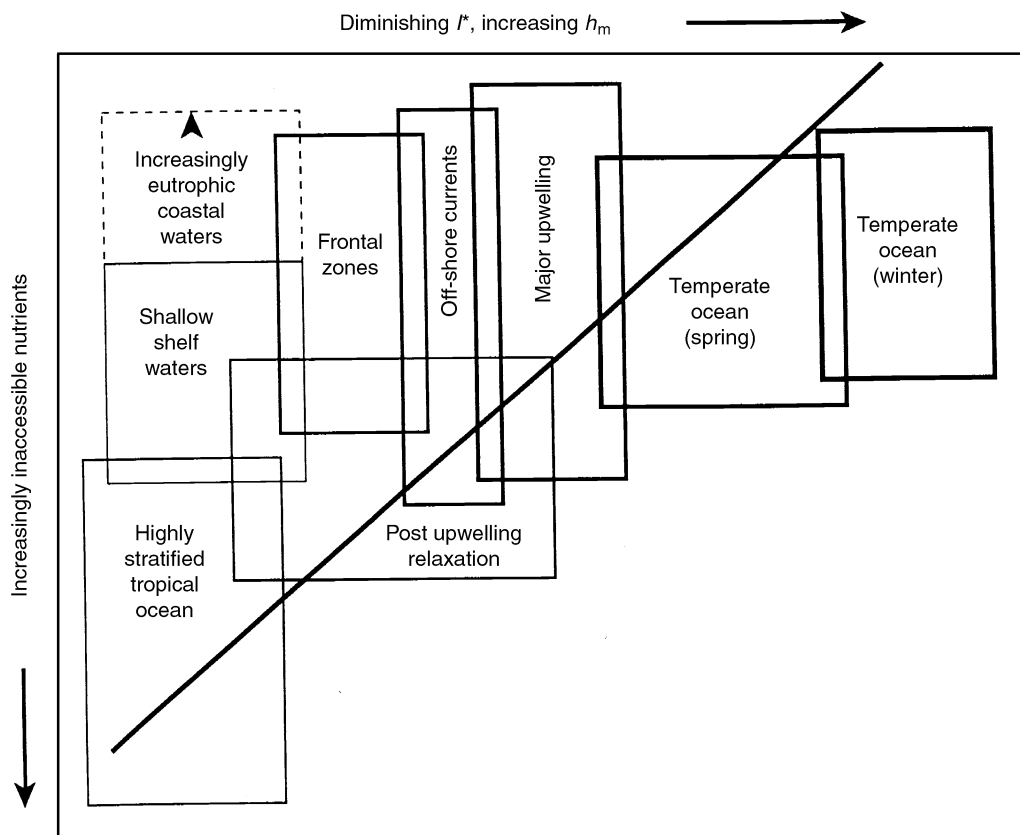


FIGURE 2 Two-dimensional representation of some marine habitats, which separates the mixed (light deficient) and well-stratified (nutrient-segregated) from shallow, nutrient-rich shelf waters, as well as upwellings and frontal zones. As in Figure 1, moving away from the top left corner, habitats become increasingly resource-stressed (working down) or processing-limited (working across). Toward the bottom right (beyond the diagonal), habitats are ultimately untenable to photosynthetic autotrophs.

of the consumers and structural variance spread through the whole planktonic community.

### III. FORM, FUNCTION, AND SELECTION IN THE PHYTOPLANKTON

#### A. The Basic Adaptations of Planktonic Photoautotrophs

The essential environmental requirements of planktonic photoautotrophs, to be able to grow in numbers and increase in total biomass and to be able to perpetuate and disperse their genes, do not differ fundamentally from those of plants in other ecosystems: access to water, exposure to adequate levels of photosynthetically active wavelengths of light, a source of assimilable car-

bon and adequate supplies of each of a score of other elements. Problems over the adequacy of the water supply to phytoplankton may be safely discounted (water relations can be adversely affected by salinity changes, where these result in the loss of cell water to the medium by osmosis). The main macronutrients (C, H, N, O, P, S, Na, K, Mg) involved in the synthesis of proteins, of cell protoplasm, and of the various organelles (all of which have to be replicated at each cell division), the key micronutrients (Fe, Mn, Mo, Cu, Zn, B, Va) mediating the assembly processes, and the elements (Ca, Si) condensed in the elaboration of calcareous and siliceous skeletal biominerals must be drawn into the cell from the bathing medium, mostly against significant concentration gradients. Elemental carbon, which constitutes nearly 50% of the dry mass of protoplasm, is usually distinguished as a separate resource to terrestrial plants

because of the atmospheric source of carbon dioxide; for phytoplankton, the proximal source of carbon dioxide is that which is dissolved in the water and which, once again, has to be drawn into the cell against a steep concentration gradient, itself sometimes exacerbated by high organismic demand. So it is, in the dilute world of the plankton, that diffusion alone is rarely able to supply the resource requirements of plankton. Adaptive mechanisms for gathering chronically deficient resources include the enhancement of uptake affinity and of the maximization of storage. Mechanisms enhancing access to remote reserves or for exploiting alternative sources of nutrients are valuable against a depleting resource base.

The biochemistry of the assembly processes is not the primary concern of the student of biodiversity, but it is important to understand the essence of the metabolic machinery. Every activity, from resource uptake, through internal transport, protein synthesis, organelle assembly, and operation, to the replication of new cells, requires the controlled expenditure of energy, which, as in just about all living things, is supplied by the respirational oxidation of carbohydrates. Fulfilling this basal metabolic requirement constitutes an energetic maintenance cost to the organism. In the case of the heterotroph, additional energy is consumed in the foraging of foods—the essential requirement is that the investment of energy yields a greater return of potential energy in the organic carbon thus derived. The distinguishing feature of the photoautotroph is the ability to first generate the carbohydrate through the photosynthetic reduction of carbon dioxide. The key reaction is the stripping of reductant (electrons) from water molecules. The energy to do this comes in sunlight (pretty well, the visible wavelengths) harvested in the chlorophyll-protein complexes of the photosynthetic apparatus. Again, the functional architecture of the apparatus is not of primary concern here, but we should recognize that what constitutes an adequacy of exposure to photosynthetically active light depends partly on the efficiency of its interception and photochemical conversion to carbohydrate and partly on the fraction of photosynthate that is required to satisfy the basal metabolic costs of its operation.

These criteria assume considerable ecological relevance to phytoplankton where, because of the sharp underwater attenuation of downwelling sunlight, only the near-surface layer is readily supportive of net primary production. Even then, seasonal, diurnal, and weather-determined fluctuations in the intensity and duration of incident sunshine means that the spatial and temporal extent of this euphotic layer is highly

variable. To operate habitually at more rarefied light levels clearly requires the provision of more, or more efficient, light-harvesting capacity. Moreover, because it is frequently the case that the depth of the wind-mixed layer extends beyond the euphotic layer, convective entrainment of phytoplankton often diminishes its aggregate exposure to the light field, at times leaving cells unable even to compensate their basic metabolic energy requirement. Here, too, great adaptive importance attaches to the sufficiency of the biomass-specific light-harvesting provision and its presentation to the light field. At the same time, the planktonic photoautotroph still needs a suite of biochemical and physiological defenses to deal with the risk of overexposure of enhanced light-harvesting that the extreme variability that pelagic open water habitats present.

## B. Specializations among Planktonic Photoautotrophs

Because the fundamental requirements of (allegedly) simple plants are more readily measurable than are those of more complex terrestrial ones, it is possible to quantify some of the generalized assertions about the selective value of species-specific adaptations to meet the stress of deficiencies in the supply of resources and the processing constraints set by operating in a fluctuating environment. Starting with the quantum yield of photosynthesis, the theoretical output of 1 mol of reduced carbon for the investment of 8 mol photons of photosynthetically active radiation harvested—or  $0.125 \text{ mol C (mol photon PAR)}^{-1}$ —is demonstrably approached in experiments with phytoplankton; typical measurements are  $0.07\text{--}0.09 \text{ mol C (mol photon PAR)}^{-1}$ ; (Bannister and Weidemann, 1984). Over the temperature range,  $10\text{--}30^\circ\text{C}$ , and at the order of light intensities required to saturate chlorophyll-specific photosynthesis in phytoplankton ( $50\text{--}200 \text{ mol photon PAR m}^{-2} \text{ s}^{-1}$ ), photosynthetic yields are found to fall in the range,  $2\text{--}15 \text{ g C (g chlorophyll)}^{-1} \text{ h}^{-1}$ . The carbon-fixation rate provides the capacity for assembling new producer biomass. Taking an average 50:1 carbon-to-chlorophyll ratio, exponential cell-specific growth capacities of  $0.04$  to  $0.26 \text{ h}^{-1}$  may be deduced, which potentially might support biomass doublings from one every 17 h to one every 3 h. The maximum cell replication rates in laboratory measurements on the unicellular chlorophyte, *Chlorella*, one of the fastest-growing freshwater eukaryotes, are equivalent to biomass doublings every 19–4 h over the same temperature range (see Reynolds, 1997b), comfortably within the theoretical

carbon assimilation capacity. The shortfall is mainly attributable to consumption of fixed carbon during growth (respiration).

For such rates of cell replication to be realized and sustained, not only must the light levels and the carbon dioxide supply be upheld, but all the other chemical components resources have to be available in concentrations that will also saturate the uptake demand. For instance, the widely accepted stoichiometry of Redfield (1958) leads us to suppose that for every 42 mg C incorporated, the new generation would, ideally, also require roughly 1 g P and 7 g N if the normal cell stoichiometry was to be preserved. Taking the amount of new carbon dioxide that could migrate across the water surface to balance the consumption by primary producers (about  $100 \text{ g C m}^{-2} \text{ y}^{-1}$ ; higher areal productivities in lakes and coastal waters are reliant on imports dissolved in inflows and the reuse of respiratory carbon dioxide), the potential new production also requires the supply of some 2.4 g P and  $14 \text{ g N m}^{-2} \text{ y}^{-1}$ . Moreover, the minimum direct PAR investment in biomass assembly would be not less than  $100 \text{ mol photons m}^{-2} \text{ y}^{-1}$  (equivalent to a mean underwater photon flux of  $10\text{--}20 \mu\text{mol photon m}^{-2} \text{ s}^{-1}$ ; Reynolds, 1997b).

Such theoretical calculations of the stoichiometric capacities provided by the fluxes of each main resource are helpful to the identification of which of them is most likely to set the constraint on growth generally and, hence, that to variations in which will evoke the most sensitive changes in growth rate and supportable biomass. In this way, the factor group most likely to limit maximal attainment is revealed. In many kinds of lake and seas, the total annual area-specific loads (including recycled nutrient) of either nitrogen or phosphorus or both fall far short of the respective hypothesized saturation requirements of  $14 \text{ g N}$  and  $2.4 \text{ g P m}^{-2}$ . So indeed the typically much lower standing crops of planktonic biomass are constrained by the rates of supply and reuse of critical nutrients. On the other hand, even in very clear water and under blue skies for 12 h per day, the depth of the surface mixed layer and its impact on the insolation of fully entrained algae continue to impose a severe capacity limitation on planktonic primary production (Reynolds, 1997b: Fig. 45). Cold, oligotrophic seas and lakes are not exclusively nutrient limited. However, it is where nutrients are abundant, or are rapidly cycled in a shallow mixed layer, that we expect to see a level of phytoplankton sufficient to challenge the carbon invasion rates, with an attendant rise in pH. In small lakes, internal recycling (carbon dioxide from community respiration, plus the carbon dioxide supplied in solution in inflowing

streams) helps to offset the effects of capacity limitation by carbon.

The distribution of metabolic limitations can be fitted to the  $I^{**}$ -vs- $K^{**}$  ordination of planktonic habitats (Figs. 1 and 2). In the energy- and resource-replete environments represented toward their upper left corners, the selective advantage resides with species adapted to the maximization of the opportunities for rapid resource exploitation and conversion to new biomass. Away from these areas, specialisms permitting continued operation under markedly subideal conditions are selected. In the downward direction, the most useful of these provide an advantageous measure of tolerance of resource stress—including the abilities to conserve nutrient resources, to search for them more efficiently, and to tap into other sources of resource. In the rightward direction, the derivation of harvestable energy is progressively more interrupted and truncated: adaptations for dealing with the disturbance to the processing ability—including the enhanced ability to capture photons from poor or fluctuating light fields—are increasingly demanded.

It is possible to show consistent morphological and adaptive traits among phytoplankton species that are variously exploitative, disturbance tolerant, and stress tolerant. It is also possible to quantify the impacts of adaptation and to demonstrate the satisfying coherence among the form, function, and ecology of planktonic algae (Reynolds, 1988b, 1995). For example, the species that appear in “new water bodies” (from rain puddles to tidal pools), in hydraulically (flood plain lakes after the river drops back) or seasonally refreshed systems (lakes at the onset of thermal stratification) are typically exploitative: they are characterized by ready and effective dispersal mechanisms and by a facility for rapid growth ( $>10 \times 10^{-6} \text{ s}^{-1}$  at  $20^\circ\text{C}$ ). Typical freshwater representatives include such algae as *Chlorella*, *Ankyra*, *Koliella*, *Chlamydomonas*, *Rhodomonas*, *Chrysochromulina*, *Monochrysis*, *Monodus*, and *Synechococcus*. They are unicellular or form small coenobia (generally  $<10^3 \mu\text{m}^3$ ), offering a surface/volume ratio ( $>0.5 \mu\text{m}^{-1}$ ) favorable to rapid solute exchange and nutrient assimilation. The area (e.g., of light field) projected by the algal cell mass, usually a mark of its potential efficiency as a light-gathering antenna, is equivalent to at least  $6.5 \text{ m}^2 (\text{mol cell C})^{-1}$ . Reynolds (1995) proposed that these traits were indicative of an “invasive” or, in the terminology of Grime (1979), C-type life-history strategy.

The adaptations that help species to survive developing nutrient stress include the physiological flexibility to overcome deficiencies in the supply of carbon

(such as the ingestion of bacteria by the facultatively phagotrophic dinoflagellates and chromulines), nitrogen (“fixing” the gas dissolved in the water, as occurs in the facultatively produced heterocysts of Nostoclean Cyanobacteria), and phosphorus (producing phosphatases to break the chemical bonds immobilizing orthophosphate in various particulate complexes). The ability to conserve assembled biomass through reduction in sedimentary and grazing losses is contributed by combining motility (usually swimming) with large size: ease of disentrainment and self-regulated migratory ranges give access to resources in the remoter parts of the water column that may be denied to smaller species. Freshwater examples (which include larger species of *Peridinium*, *Ceratium*, *Microcystis*, and other colonial, bloom-forming Cyanobacteria and, arguably, such colony-forming algae as *Uroglena*) are characteristically “large,” having cells or coenobia that are  $>10^4 \mu\text{m}^3$  in volume and often much more than  $30 \mu\text{m}$  in diameter). Their consequent rather low surface/volume ratios ( $<0.2 \mu\text{m}^{-1}$ ) leave the algae with slow rates of growth ( $<8 \times 10^{-6} \text{ s}^{-1}$  at  $20^\circ\text{C}$ ), sensitivity to low temperatures, and poor antennal projection ( $\leq 2.5 \text{ m}^2 (\text{mol cell C})^{-1}$ ). The occurrence of growing populations of these stress-tolerant—S-type, or (“acquisitive”; Reynolds, 1995)—strategists tends to be restricted mainly to warm, well-insolated shallow lakes and epilimnia.

The phytoplankton species tolerant of, if not dependent on, near-continuous entrainment within deep or turbid mixed layers have well-developed capabilities for maintaining growth despite intermittent brief exposure to light. “Attuning” (Reynolds, 1995) strategists tolerate this scale of pelagic disturbance through the projection of a large mass-specific surface area ( $>8 \text{ m}^2 (\text{mol cell C})^{-1}$  in the case of 1-mm threads of *Planktothrix agardhii*;  $\sim 30 \text{ m}^2 (\text{mol cell C})^{-1}$  in the case of an 8-celled colony of *Asterionella formosa*). They have high surface-to-volume ratios (generally  $>0.5 \mu\text{m}^{-1}$ ) though these are not necessarily attained through small unit-size but by morphological complexity (filaments, attenuated or fenestrated coenobia, protuberances, etc.) and so continue to benefit from reasonably rapid rates of cell-specific nutrient uptake and growth ( $>10 \times 10^{-6} \text{ s}^{-1}$  at  $20^\circ\text{C}$ ). The potential for energy conversion may be enhanced by an increased cell-specific chlorophyll content and accessory pigmentation to widen the spectrum of harvestable wavelengths. Other typical R species include larger diatom (e.g., *Aulacoseira*, *Synedra*), desmid (e.g., *Closterium*, *Staurastrum*), and Chlorococcalean genera (e.g., *Pediastrum*) and solitary, filamentous members of the Oscillatoriales (like *Limnothrix* and *Pseudanabaena*).

Not all phytoplankton species fit perfectly within one or other of the three categories but show properties intermediate between them. What is interesting, however, is that the intermediacy in morphological and physiological adaptation matches well the intermediacy in their ecologies. The “space” between the invasive and the stress-tolerant acquisitive species is occupied by genera such as *Dinobryon*, *Dictyosphaerium*, *Sphaerocystis*, *Gemelliscystis*, arguably *Volvox*, *Eudorina*, *Aphanizomenon*, and *Gloetrichia*, which diminish in surface-area/volume ratio and maximum growth rate but increase in their abilities to exploit and conserve nutrient resources. Freshwater algae between the invasive and the attuning poles include species of *Cyclotella*, *Scenedesmus*, and *Coelastrum*, which could be said to be increasingly large, more convoluted, and increasingly disturbance turbidity-tolerant algae. The axis between stress- and disturbance-tolerance is occupied by relatively large, acquisitive, but self-regulating species, which can persist for months to years on stable density gradients, notably *Planktothrix rubescens*, *P. mougeotii*, and species of *Lyngbya* and *Phormidium*. *Cryptomonas* shows traits almost intermediate between all three extremes.

### C. Seasonality of Planktonic Autotrophs

The utility of the functional classification is best demonstrated in relation to the complex issue of seasonal change. Despite its celebrated species richness, the phytoplankton is usually dominated by very few genera at a time (it has been suggested that 95% of the extant standing biomass will be incorporated in no more than eight species at any one time; often it will be in rather fewer, as few as one or two; Reynolds, 1997b). Yet it is well recognized that the dominance moves among different species through time and that, in a given system, the sequence will be similar from one calendar year to the next. Moreover, similar patterns may be observed in similar but often geographically remote lakes. Numerous such cycles have been described in the literature (Reynolds, 1984b; Sommer *et al.*, 1986); it is sufficient to mention a single archetypal example. In a small, calcareous, temperate lake in Britain (Croze Mere), the diatom *Asterionella* dominates the early spring growth, increasing from a few tens to several thousand cells per milliliter, over a period of six to eight weeks of lengthening days and intensifying insolation. Then, in mid-April, when the work of the wind is no longer sufficient to discharge the increasing buoyancy of the heat flux to the surface, the lake will become thermally stratified. *Asterionella* settles out of a mixed

depth, which is too truncated to keep it in suspension, leaving the clear water open to the establishment of such algae as *Rhodomonas* and *Monodus*, *Cryptomonas*, and of the motile colonies of *Eudorina*. As the summer solstice approaches, these too are replaced, first by nitrogen-fixing *Aphanizomenon* or *Anabaena* and by the dinoflagellate, *Ceratium*, which will dominate the annual biomass maximum. In the autumn, the shortening days and declining temperature lead to a weakening of the stratification, deeper wind mixing, and the restoration of the depleted nutrients: by the late autumn, diatoms, including *Aulacoseira* and *Asterionella*, are generally the most abundant algae, dominating the shrinking residual biomass.

No two years will be exactly the same, and the relative proportions of simultaneously dominant and co-dominant species will fluctuate. Yet the pattern is robust and is amenable to diagrammatic summary (Fig. 3A): against axes empiricized in terms of  $I^{**}$  and  $K^{**}$ , or even analogized to mixed depth versus the biologically available concentration of the critical nutrient, the time trajectory of the changing coordinates traces the extent of the seasonally changing environment, from the well-mixed, nutrient-replete starting condition, through the slow resource depletion of the spring period, and the rapid depletion after the onset of stratification. Finally, the effect of enhanced mixing weakens the light income but gradually restores the resource base. The partial independence of the two axes is conveniently emphasized by the inclusion of the winter "loop" when the light income falls below a level that will support new growth, while the system may be accumulating new external resources, only after  $I^{**}$  has increased to the point where net growth can be supported at the cost of net resource reduction.

The development continues by the imposition of the trajectory on the plot showing the distribution of the morphological traits of the algae (Fig. 3B); as these prove such useful predictions of physiological performance, which, in turn, anticipate their ecologies, the fit should not be surprising. Nevertheless, the correlation of function and the dynamic response to environmental change is an extremely satisfying one. It will

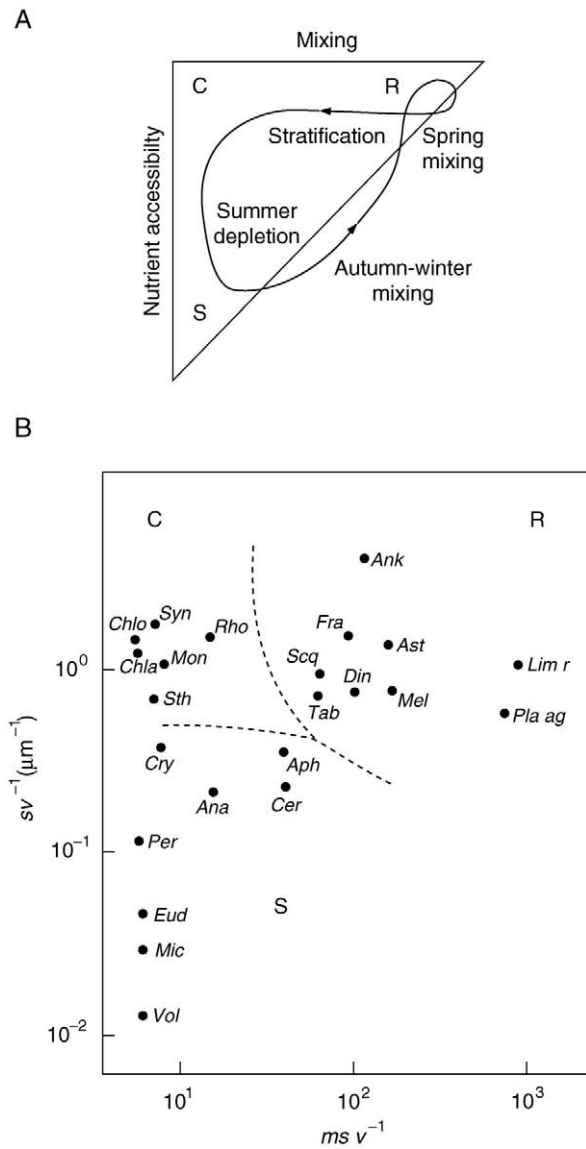


FIGURE 3 (A) How growth conditions in a monomictic deep lake (mixed in winter) track through a calendar year, selecting sequentially for a low-light tolerant spring bloom (of R species), then opportunist C species after the onset of stratification, leading to the biologically enhanced segregation of nutrient resources tolerated by S species, before autumnal downmixing again forces conditions back toward the top right: nutrients may be further increased (by inflow) in winter, when primary production is still increasingly light limited. (B) Some of the algal species that might be selected are shown in an analogous matrix that is based purely on shape- and size- characters of the algae concerned ( $sv^{-1}$  versus  $msv^{-1}$ , where  $s$  is the surface area of the alga in  $\mu m^2$ ,  $v$  is its volume in  $\mu m^3$ , and  $m$  is the maximum length dimension, in  $\mu m$ ). A strong correlation exists between the morphologies of these algae and their functional ecologies. The algae shown are represented as follows: Ana = *Anabaena flos-aquae*, Ank = *Ankistrodesmus falcatus*, Aph = *Aphanizomenon flos-aquae*, Ast = *Asterionella formosa*, Cer = *Ceratium hirundinella*, Chla = *Chlamdomonas*, Chlo = *Chlorella*, Cry = *Cryptomonas ovata*, Din = *Dinobryon divergens*, Eud = *Eudorina unicocca*, Fra = *Fragilaria crotonensis*, Lim r = *Limnothrix redekei*, Mel = *Aulacoeira subarctica*, Mic = *Microcystis aeruginosa*, Mon = *Monodus*, Pla ag = *Planktothrix agardhii*, Per = *Peridinium cinctum*, Rho = *Rhodomonas pusilla*, Scq = *Scenedesmus quadricauda*, Sth = *Stephanodiscus hantzschii*, Syn = *Synechococcus*, Tab = *Tabellaria flocculosa*, Vol = *Volvox aureus*.



provide us with one of the most important clues to understanding how local diversity is maintained in nature.

### D. Traits in the Marine Phytoplankton

Before that, however, it is necessary to establish that the rest of the planktonic biota respond to environmental cues in broadly analogous ways. Marine phytoplankton would seem appropriate for first consideration. Evidence of a similar coherence among morphological form, physiological performance, and population responses to environmental forcing was presented in Margalef's (1978) seminal discussion of plankton life-forms, while the insights underpinning the "mandala" model (Margalef *et al.*, 1979) continue to intrigue plankton ecologists. Nevertheless, the apparent functional and morphological correlations have scarcely been quantified. Preliminary considerations of the types of phytoplankton characterizing the various subdivisions of the sea (Fig. 2), however, do reveal entirely parallel trends between morphology and performance. For instance, as days lengthen over the open seas of the temperate latitudes, an initial typical spring bloom of such chain-forming diatoms as *Thalassiosira nordenskioldi*, *Skeletonema costatum*, and *Chaetoceros* (all manifestly having R-type, attuning characters) gives place in summer to relatively large, motile, S-like peridiniids (e.g., *Scrippsiella trochoidea*) and ceratians (e.g., *Ceratium tripos*), which are often followed, in autumn, by chain-forming (*Chaetoceros* spp.) or needle-like (*Rhizosolenia*) diatoms with high surface-area-to-volume ratios and antennal enhancement. In the high-nutrient, high-energy waters of coastal margins, potential dominants are generally small, usually unicellular, centric diatoms (e.g., *Cyclotella caspia*, *Thalassiosira weissflogii*), green flagellates (e.g., *Dunaliella*, *Nannochloris*), euglenoids (*Eutreptia*), and gymnodinioids (*Gymnodinium*); all qualify as fast-growing opportunist (C-) species. The estuarine prorocentroids are also conspicuous. Blooms of *Phaeocystis* may follow, before giving place to ceratians. In the highly stratified, tropical oceans, dominance passes quickly from diatoms to *Emiliana* and to the large *Ceratium* spp. and *Ornithocercus*. The big, buoyancy-regulating, stress-tolerant, S-species of diatom (*Ethmodiscus*) and dinoflagellate (*Pyrocystis*) are at the culmination of this sequence.

Spatial differences—among latitudes or between neritic shelf waters and the open oceanic systems—probably also attest to the functional differentiation of the phytoplankton. Areas of frontal activity and of surface upwellings of the deep oceanic circulation currents

tend to combine intensive mixing with renewed resources and to select species in the C-R intermedium through much of the year: *Gyrodinium* spp., *Alexandrium tamarense*, *Thalassiosira leptoporos* can be typical assemblage members. Away from the epicenters of the upwelling areas and where, indeed, the water is less mixed but still charged with nutrients, such dinoflagellate taxa as *Dinophysis*, *Gonyaulax* and *Gymnodinium catenatum* provide prominent markers of the (R-S) transition between stressed and disturbed conditions.

## IV. FORM, FUNCTION, AND SELECTION IN THE ZOOPLANKTON

### A. The Basic Adaptations of Planktonic Animals

In the terms of species success, the survival strategies, and selective biases, the ecological constraints governing the lives of phagotrophic zooplankton are wholly analogous to those acting on the phytoplankton. Specific requirements differ substantially, as do the time-scales over which the organisms respond to environmental fluctuations. The essential requirements of the phagotroph are to be able to encounter and capture sufficient appropriate food to supply the elemental assembly of its biomass, to maintain and defend that biomass and perpetuate its genetic instructions, and to secure the energetic demands of the relevant processes. Resource adequacy for nonphototrophs may be judged principally in terms of the availability of ingestible particulate biomass (living or detrital), being the only practical source of reduced carbon, although the judgment of organic-resource quality must acknowledge the capacity of its blend of proteins to fulfill the biochemical requirements for nitrogen, phosphorus, sulphur, iron, and so on. Resource gathering, whether essentially herbivorous, carnivorous or detritivorous, or whether actively hunted or sedentarily filtered, carries a high energy demand: "maintenance costs" of phagotrophs are proportionately high, not untypically accounting for some 90% of the resources consumed. The energy comes from the original photosynthetic investment in carbon bonds, whose controlled oxidation sets the principal processing constraint in animals, which is to supply the tissues with sufficient oxidant.

In terms of the lives of planktonic animals, already contending with the problems of operating in a fluid, viscous medium, the key issues are still about staying alive, about finding adequate amounts of suitable foods, with a sufficiency of individuals reaching reproductive

age, and about having the collective residual fecundity to be able to recruit the next generation. The size range of planktonic animals embraces the flagellated protists (whose phylogeny merges with the dinophycean, synurophycean and euglenophycean mixotrophs, and photoautotrophs), whose dimensions leave them substantially embedded within the viscous range of aquatic environments, through to crustaceans and young fish whose lives reach well into the turbulent range. Whereas the former rely on encounters of suitable food particles (organic detritus, not necessarily autochthonous, bacteria, small algae) within the same viscous neighborhood, animals living mainly within the boundaries of the turbulent world have a modicum of control over their own movements. They also exploit turbulence in their foraging by generating local circulations entraining small food particles (Margalef, 1997; Rothschild and Osborn, 1988).

Within this general scheme of classification, there remain numerous opportunities for differentiation of foods and foraging mechanisms, as well as substantial feeding specialism, especially among the rhizopods and the rotifers. There is also considerable differentiation among the strategies for survival and resource exploitation. In addition, predation by planktivorous fish and invertebrates has a central role in structuring zooplankton assemblages: whereas larger animals ostensibly have more foraging opportunities for a wider choice of potential foods, they are also more vulnerable to visual predators (Brooks and Dodson, 1965; Hall *et al.*, 1976). The latter is readily demonstrable in experiments, but the part played by food preferences and availabilities has proved less tractable. However, Romanovsky's (1985) analysis of life-history strategies of the planktonic Cladocera distinguished among those species that invest in high survivorship rather than in rapid juvenile growth (exemplified by *Diaphanosoma brachyurum*), those ultimately large-bodied species that sustain rapid rates of juvenile recruitment and development (typified by *Daphnia hyalina*), and those smaller-bodied, fast-reproducing inhabitants exploiting the fluctuating opportunities of temporary ponds and water bodies (including species of *Moina*). Taking particular account of the interspecific trait variability in adult cladoceran size and the negative correlations ("consistent tradeoffs") between egg number and yolk investment and between larval development time and resistance to starvation, Romanovsky recognized three basic life-history strategies. Following terminology attributed to L. G. Ramezsky, he referred to these as being "patient," "violent," or "explerent." It needs little insight, however, to recognize that this imaginative perception is but a quite indepen-

dent assessment of the expression of "resource-stress tolerance," "exploitative invasiveness," and "disturbance tolerance" among animals. In this way, it is possible to fit planktonic animals to a habitat template, analogous to that for the phytoplankton, on the basis of a wholly functional classification.

## B. Life-History Strategies

The resident zooplankton of the open seas and of larger oligotrophic lakes live in environments where the producer biomass is chronically resource limited: the potential food resources for zooplankton (planktonic algae, bacteria, and particulate organic detritus) are generally  $< 0.1 \text{ g C m}^{-3}$ . To be successful under persistent conditions of dilute resource, animals must blend energy-efficient foraging with controlled resource exploitation—they must adopt the "patient" strategy of resource-stress tolerance, especially with respect to the recruitment of juveniles. Investment in a small number of relatively large eggs with a relatively long development time is typical. Physical and behavioral defenses reduce the vulnerability to predation. Such traits distinguish the life histories not only of *Diaphanosoma* but of many of the calanoid copepods also. The sophisticated browsing mode of feeding they adopt has been shown to be sufficient to sustain natural populations and support their reproduction at food concentrations in a range equivalent to  $0.01$  to  $0.08 \text{ g C m}^{-3}$  in both the sea (Huntley and Lopez, 1992) and lakes (Hart *et al.*, 1996). For reference, a producer biomass of  $0.08 \text{ g C m}^{-3}$  may be supported by photosynthesis in a mixed layer of  $\leq 60 \text{ m}$  (i.e.,  $\leq 5 \text{ g C m}^{-2}$ ), so long as the microbial loop is able to maintain a bioavailable nutrient base exceeding  $0.1 \text{ g P}$  and  $1 \text{ g N m}^{-2}$ . Arguably, the ciliates (e.g., *Halteria*, *Strombidium*), which constitute a significant part of the diet of calanoids in oligotrophic lakes, must be, similarly, stress-tolerant patient strategists.

In the more productive waters of fertile lakes and rivers, estuaries, nutrient-enriched coastal-shelf waters, and oceanic upwellings, the capacity to support autotrophic biomass may, for some of the time at least, significantly exceed  $0.1 \text{ g C m}^{-3}$ . This is an approximate threshold, above which the energetic return from filter-feeding becomes steadily more favorable. Among lakes, for example, Daphniids tend to become much more abundant, absolutely, and relative to calanoids. Most species of *Daphnia*, when adequately nourished, are able to grow and mature rapidly, to increase egg production and to recruit the next filter-feeding generation over a few days (George and Reynolds, 1997). The 12-fold

increase in the *Daphnia*-dominated community-filtration rate in 13 days, measured by Reynolds *et al.* (1982; note, aggregate filtration capacity doubles every 3 to 4 days), conforms to the understanding of an exploitative, or "violent," life-history strategy. Among the noncrustaceans, the ciliates of microaerophilous environments (e.g., *Metopus*, *Caenomorpha*) react in an analogous way to a food abundance in media from which other sorts of consumer are mainly excluded.

Many noncrustacean members of the zooplankton take advantage of short generation times to track bursts of algal or bacterial abundance. Protists including *Difflugia*, *Coleps*, *Tintinnidium*, and *Nassula* and rotifers such as *Keratella*, *Kellicottia*, and *Brachionus* will often respond to an expansion in suitable algal foods, but this occurs before the Daphniids can ascend to dominance. Their disturbance-accommodating lifestyles are well suited to the direct exploitation of small phytoplankton and organic detritus in the middle reaches of large rivers, provided travel times permit (Viroux, 1997).

## V. FUNCTION IN THE BACTERIOPLANKTON

Microbial populations constitute the third integral component of the planktonic communities of lakes and seas, where they are often numerous (typically within the range  $10^5$  to  $10^7$  cells  $\text{ml}^{-1}$ ). Microbial diversity, however, is not yet well understood, principally because the bacterial taxa are still insufficiently distinguishable to be separated routinely to species level. For many years, bacteria have been identified as much by what they do as by their morphological affinities. Given that, for most of them, the utilization of dissolved reduced-carbon substrates provides the main source of energy, even the functional approach appears to have its limitations for bacterioplankton. "Organic carbon compounds" are derived from a very wide range of the breakdown products of biogenic materials, many originating from the land, as well as a host of very unnatural anthropogenic organic substances. The ability to break down particular classes of compounds or kinds of bonds is unlikely to be shared by all bacteria. Every such process presumably requires the action of one or more discrete enzymes, and the ability to produce each enzyme requires one or more genes dedicated to its production. As the genetic complement of microbes is relatively modest, it follows that the number of separate, substrate-specific bacterial strains is likely to be high.

Modern molecular approaches are slowly sorting out "who does what" but, for the present, the diversity of the bacterioplankton is anticipated to be high.

Among oxic freshwaters, the free-living bacteria are known to include members of the coccoid and rodlike Bacilli, the Flavobacteria, Pseudomonads, and Vibrios (Atlas and Bartha, 1993). All are heterotrophs, which exploit sources of biogenic organic carbon. These sources were always recognized to constitute the waste products and cadavers of the biotic components of the aquatic system generally. As understanding of pelagic carbon metabolism has developed, it has become possible to separate the active role of microbial heterotrophs in transferring dissolved organic compounds derived from low-molecular weight, mainly algal photosynthetic intermediates (especially dissolved glycollate), as the food supply to the flagellate—ciliate—copepod trophic links, from that of degradation of biogenic particulates (organic detritus, fecal pellets, etc.), usually by sessile or stalked bacteria, such as *Caulobacter*, actinomycetes, and fungi, that attach to the particles and whose activities liberate carbon dioxide and mineral nutrients into the water. Both contribute substantially to the cycling of carbon and nutrients, but in contrasting ways. Oxic breakdown of abundant biogenic material releases resources to the autogenic synthesis of new mass. The phagotrophs of the microbial loop depend on the close coupling that the viscous scale facilitates, but most of the energy is consumed internally and the material transferred to higher trophic levels is absolutely small.

In smaller lakes and ponds, the nature and diversity of organic carbon sources are enriched by direct transport of biogenic materials from the catchment area. However, current understanding is also having to accommodate the recognition that often the largest fraction of organic carbon present in stream and lake waters comprises plant-derived humic and fulvic acids washed in from catchment soils (Wetzel, 1995). They are particularly abundant, of course, in brown and "black" waters draining forests and peatlands. The relative size of this fraction indicates its high resistance to microbial breakdown; it persists in measurable amounts in the largest oligotrophic lakes and even in the open oceans. Refractory dissolved organic carbon from terrestrial sources does break down slowly; some is rendered labile by exposure to ultraviolet radiation.

Away from aerobic mixed layers, freshwaters furnish other microhabitats exploitable by bacteria. In the microaerophilous environments of the bottom sediments and in the deep water of stratified, productive lakes,

anaerobic microbes are common and typically numerous; they include *Clostridium*, the sulphate-reducing *Desulfovibrio*, and the Archaean Methanogens. Interacting gradients of light and of redox superimposed on relatively stable density gradients may also provide vertical sequences of microhabitats, each niche potentially differentiated according to the microorganisms it supports and whose physiological adaptations it most suits. To be able to carry out photosynthesis in a reducing environment favors *Rhodospirillum*; but to be able to use sulfide or sulfur, as photosynthetic electron donor, can bias the selection in favor of Chromatiaceae and Chlorobiaceae). The involvement of specialist chemolithotrophs in nitrification (*Nitrosomonas*, *Nitrobacter*) and in the oxidation of sulfur (e.g., *Thiobacillus*, *Beggiatoa*) and iron compounds (*Ochrobium*) further adds to the microbial diversity of freshwaters.

In the open waters of the sea, most bacteria are aerobic—relatively few anaerobes are ever found in the surface waters. Many of the same genera represented in freshwaters include marine species: *Pseudomonas* and *Vibrio* are often found to dominate and species of *Flavobacterium*, *Alcaligenes*, and *Cytophaga* are found in high numbers (Atlas and Bartha, 1993). Sediments receiving organic inputs are bacteria-rich; moreover, where these fall anaerobic, marine sulfate-reducers and methanogens are present in substantial numbers.

In both marine and freshwaters, the vital contribution of the bacterioplankton to ecosystem function arises not only from the mineralization of organic carbon products but from the potential remobilization and renewed bioavailability of nitrogen, phosphorus, and the metals required in the assembly of new biomass.

Most bacteria are supposed to be capable of rapid self-replication, with the potential to undergo several doublings per day. The relative ease of dispersal contributes to their apparent ubiquity and to their fidelity to appropriate habitat opportunities. Microbial biomass is, nevertheless, clearly subject to environmentally imposed rate and capacity limitations. In many instances, their requirements are not dissimilar from those governing the growth of photoautotrophs. However, small cell size and superior uptake kinetics favor the performances of bacteria. A greater collective versatility in the energetics and potential sources of reductant and oxidant opens up more of the aquatic environment to chemotrophs than to phototrophs. The distinguishing dependence is the one for appropriate organic carbon skeletons. Their availability remains the regulation on the numbers and dynamics of planktonic bacteria. The nature of the carbon compounds, together with

the metabolic constraints set by the redox environment, are presumed to be decisive in species distributions.

## VI. TEMPORAL PATTERNS IN THE ORGANIZATION AND DIVERSITY OF PLANKTONIC COMMUNITIES

### A. Quantifying Structure in Planktonic Communities

This survey of the wealth of species to be found in the plankton serves to emphasize the collective breadth of the adaptive specialisms themselves but which, individually, provide advantage only to the certain species that possess them and only at the certain times when the specialism affords some operational benefit over others that do not have them. A reasonable deduction that may be made is that the collective diversity of planktonic species reflects the number of distinctive habitats, or niches, that it is possible to define. This is, roughly, the niche differentiation theory of biodiversity: every species has an optimum performance, which is facilitated when each of the component processes is at the species-specific optimum. When the conditions simultaneously satisfy all the species-specific optima of a given species, then that species is uniquely favored to outperform all the others for which the niche conditions are suboptimal. There can then be as many successful species as there are tangible niches, for each stands to be the fittest competitor under its favored blend of environmental conditions. This means that in any given niche location, the local diversity is likely to be actually quite low.

An alternative view might be moved to accommodate the observation that local diversity is sometimes actually quite high, even when the available niches seem to be few in number (this, it will be recalled, was Hutchinson's paradox). It should also accommodate another of the observations emphasized in the foregoing sections—that is, there is a great deal of temporal environmental variability, sufficient, indeed, to move the selective advantage from one species to another, often before any has been able to assert its competitive dominance. Sometimes, moreover, the variability is plainly driven externally, beyond organisms' power to regulate—the changes are brought about by physical forcing with sufficient magnitude to “disturb” the existing internal organization. At the appropriate intensity and fre-

quency, such disturbances might provide a mechanism for rotating the competitive advantage among species and for maintaining local diversity. This, indeed, is the essential basis to the main counterview to niche-differentiation theory.

Two further features of the present appreciation of planktonic communities that are attractive to the student of biodiversity have also been noted in the survey. One is the planktonic timescale: the frequency of generations allows the rates of internal compositional changes (competition, succession, predation) to be measured and the resistance to external forcing, as well as the resilience to recover from forcing, to be quantified (Reynolds, 1997a). The other is the quantitative information that is now available on the performance capacities of representative planktonic organisms—from their resource thresholds to their resource-regulated responses and the onset of their resource-saturation. These help us to distinguish internal structural changes from externally forced restructuring and to express them in energetic terms. In this section, we may deduce why the two main theories of biodiversity are not at all mutually exclusive and when we may presume either is valid. To do this, however, it is necessary formally to distinguish between the diversity of community structure at a given location and the richness of species present among a series of such locations at the regional or even global level.

## B. Quantifying Diversity in Planktonic Communities

Diversity is a concept drawn from communication theory (Shannon, 1948), referring to the amount of discrete information in a particular location. In communities, it could refer to the different fragments of genetic information. A diverse community is one with many different kinds of genes (species) present simultaneously. This biotic diversity ( $H''$ ) can be quantified with great precision, using the specialized Shannon-Weaver function:

$$H'' = -\sum b_i/B \log_2(b_i/B)$$

Diversity increases with the number of species ( $s$ ) each contributing biomass ( $b_i$ ) to the total,  $B$ , in the sample of lake or seawater. The more species present, the greater is the diversity, until  $H''_{\max} = \log_2 s$ . In practice, it is necessary to set a bar on the efficiency of search for the rarer species (more diligent searchers and more expert taxonomist-hunters would always score higher diversities). Nevertheless, the evenness of the

interspecific distribution,  $E = H''/H''_{\max}$ , is itself a useful measure.

Local species richness (the sample, the pond, the lake) may run to a few tens of phytoplankton and, if the phagotrophic protists are included, perhaps 100 or so species of zooplankton. Regular sampling may raise this total considerably, as seasonal changes are encountered and the probability of encountering rare species is increased (Padisák, 1992); “rare” in this context may mean one individual per liter, but there could still be a billion such individuals in a small, 10-ha lake. As observed earlier, most ( $\geq 95\%$ ) of the total biomass of phytoplankton and that of zooplankton will each be invested in eight or fewer species. Most of the consideration of evenness concerns how the total biomass is shared among those eight. Thus, whether local species composition is species rich or otherwise, the perception of its diversity is greatly affected by the relative evenness of the most common species.

Often without clear confirmation, the issue of “protecting biodiversity” generally extends to the totality of species, or the totality of species that could live in a particular habitat (Wilson and Peter, 1986). Unlike diversity, it has no formal mathematical expression. However, we may quickly recognize that if a single patch of the aquatic environment—say, the open water of a 10-ha lake—can furnish habitat sustaining viable populations of a few tens of species of planktonic species, then a great many such ponds must exist if all the species are to be supported, many of which will be common to a wide selection of the water bodies.

In this way, two questions need to be answered. The first is about the mechanisms by which within-patch diversity is upheld, against the tendency for the optimally adapted functional specialists to exclude its competitors. The second questions the mechanisms by which between-patch diversity is maintained, so that individual species have more opportunities to maintain stocks at viable and sustainable levels.

## C. Assembly and Autopoiesis of Communities

The essential step in logic that separates the niche-diversification and the persistent-disturbance theories of species richness rests on an acceptance that the supportive capacity is not always filled; conversely, although the capacity of the resources is always finite, the underexploited resource base is, by definition, *not* “limiting.” Many students of the phytoplankton persist in comparing the quantities of nutrients in the water relative to the ideal composition of algal biomass, then

judging that the nutrient in least supply is “the limiting factor.” That phytoplankton concentration is ever able to increase at all is because, in reality, there are times when nutrient resources are available to support the additional biomass and unharvested light energy to fuel its assembly. Moreover, these opportunities are attributable to natural fluctuations in the resource base: the spring increase in temperate waters, the most conspicuous feature of lakes and seas at higher latitudes (Fogg, 1965; Sverdrup *et al.*, 1942) is the consequence of lengthening days and strengthening insolation of waters, which have themselves demonstrably been recharged with nutrients during the winter months of hydraulic circulation and augmentation from inflow, at the same time as meeting low biological demands. The spring bloom, at least initially, is the response of the producer community to an expanding capacity of the resource and energy bases. Later in the year, the resource supply may well fail to meet the biological demand and severe constraints are then imposed on its further growth and maintenance. Alternatively, it is left to shortening days and convective mixing in the autumn months to erode the supportive capacity of the insolation flux. We will return later to the subject of these constraints but, for the moment, it is important to consider in more detail how the phytoplankton behaves when its production is *not* limited by nutrients and when the production of zooplankton leaves large amounts of phytoplankton relatively still unharvested. The organismic responses to the opportunities provided by a luxury of supply in all resources are instrumental in molding the structure of ascendent communities.

The key feature of an expanding (or merely replete) resource base is that its exploitation is relatively straightforward. Species do not compete for the resources in the sense that their garnering by the individuals of one species denies an adequacy of the nutrients simultaneously supplied to the individuals of a second or an *i*th species. The only question of ecological importance is “Can it obtain enough?” This is not to say that opportunistic “luxury uptake” and resource storage might not become important at some later juncture, only that species 2 to *i* can also perform to their capability at the same as species 1 is also performing. Nevertheless, rapidity of growth and reproductive efficiency are crucial to the exploitative outcome: having a large starting inoculum and, especially, a high yield/resource conversion potential is vital to the fitness of colonist and invasive species. If we judge it to be “outcompeting” the slow-growing species present, we are giving a second nuance to the word “competition.” Really, we need another word, perhaps “fitness”: who was the better com-

petitor in Aesop’s fable, the tortoise or the hare? When a lake, just like Crose Mere (discussed earlier), stratifies in April or May, it opens the stage to many more players. Is it not those species, such as *Chlorella*, *Rhodomonas*, and *Cryptomonas* that can grow relatively rapidly, by virtue of their C-type morphological and functional biases, which actually do so (perhaps doubling mass every 24 hr) to become initially prominent? Many field data (considered by Reynolds, 1984a; Sommer, 1981) point to the simultaneous growth of other species, of *Ceratium*, *Peridinium*, and *Microcystis*, for example, but at maximal rates that are far slower than those which *Chlorella* can attain. The tradeoff for invasiveness is acquisitiveness—though the relatively large, motile life-forms of acquisitive, stress-tolerant algae stifle rapid conversion, such that some four to five days may be needed to achieve each biomass doubling, yet they do not sink out and they are too big for most planktonic phagotrophs to ingest. The slowly assembled biomass is carefully conserved and, in some cases, protected by elaborate chemical defenses. By exploiting the structure of the water column to the capacity of its resources and by continuing to grow for longer, often in the face of intensifying resource shortage, the S-type adaptive strategy allows the organisms to achieve a larger, climatic biomass than that achieved by any of the earlier invasive species. The fact that the dominance can become total, other species are effectively excluded by the ultimately superior competitor and the Shannon diversity falls to a minimum, is fully in keeping with the anticipated outcome of events within the pelagic niche.

Another relevant observation, that relatively few species ever attain such outright dominance of the phytoplankton, leads to the deduction that there are many routes of internal change but few ultimate outcomes. This last point is taken to be a powerful indicator of the system’s ability to self-organize, a measure of its *autopoiesis*. This is a property of all ecosystems (Jørgensen, 1992) but its mechanical basis is not well understood. However, the further examination of the processes of capacity-filling behavior of the phytoplankton yields a conceptual and preliminarily empirical view of internal organization. If the developmental progress of a producer-dominated pelagic system is considered against axes representing the standing crop (in units of organic carbon) and the harvest of photosynthetically active radiative flux (in  $\text{W m}^{-2}$  or  $\text{J m}^{-2} \text{d}^{-1}$ ), then the coordinates corresponding to the “open stage” at the start of the summer stratification may be set close to the origin (Fig. 4A). The standing crop is far below the potential of the resource-limited carrying capacity (horizontal axis) and thus is collectively capable of

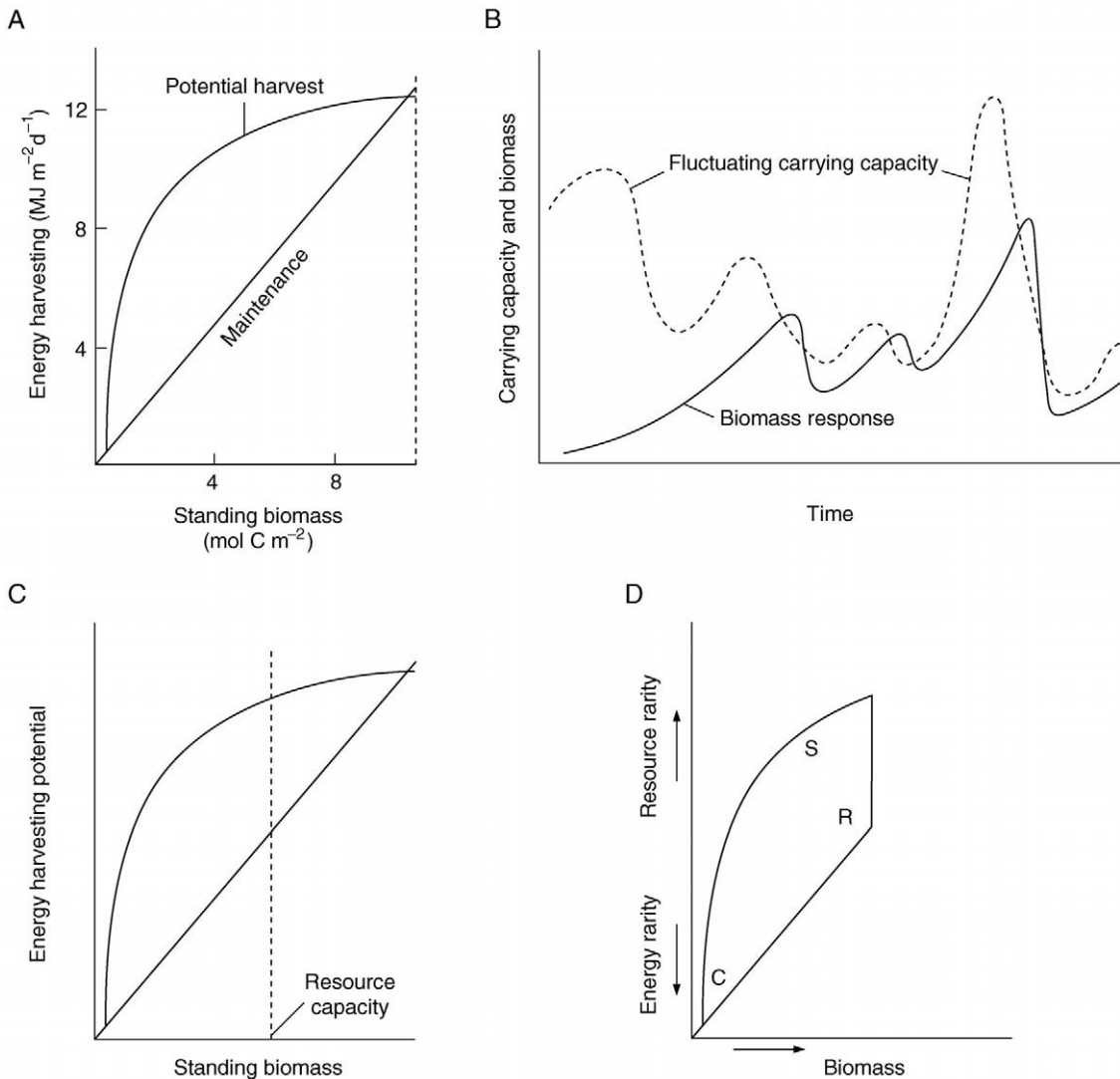


FIGURE 4 (A) A simple plot to show the maximum light that can be harvested by a given biomass of planktonic algae relative to the energetic costs of its maintenance and loss rates; the difference is the exergy of the system. For a time, developing systems increase biomass and exergy, until the reduction in the exergy cushion leaves them increasingly sensitive (B) to fluctuations in the harvest-determined carrying capacity; (C) resource limitation truncates the cushion but the "sail shape" (D) can be used to subsume the triangle from Figure 3.

intercepting very few of the photons that penetrate the water surface before their energy is absorbed by the water. The notion of capacity filling is that the structure will move away from the origin and toward the upper right corner, whose coordinates are determined either where all of the least available of the resources has been incorporated into biological standing crops (the resource capacity) or where every available photon is intercepted by planktonic light-harvesting centers and is driving photosynthetic carbon reduction (the processing capacity). If we now imbue our producer com-

munity with the physiological performance of a *Chlorella* culture, the unimpeded biomass-specific light harvesting and carbon fixation rates are able to accumulate new cell carbon sufficient to double or possibly quadruple the plant biomass within 24 hr. It is possible to suppose, in these early developmental stages, that the doubling of the biomass will roughly double the area of light harvesting surface, so the coordinates representing the supportable biomass move upward and rightward. The new biomass incurs an increment in respiratory losses, so the coordinates representing the

maintenance costs—assumed, for simplicity, to be directly proportional to the biomass—diverge from the slope of the harvesting capacity. Given strong light and a steady water temperature, producer biomass begins to increase exponentially. It is predictable that, as accumulation proceeds, the light-harvesting centers are increasingly probable to be shaded (meaning that even with no diminution in the surface flux, individual light-harvesting centers are activated less frequently). The provision of additional biomass is no longer rewarded with proportionate energy harvest.

The limiting condition is approached when the maintenance of the expanded biomass consumes all the energy it is able to capture. This biomass, shown by the broken line in Fig. 4A, is the maximum that can be supported. The coordinates are proposed on the basis of photosynthetic and respirational properties solved for *Chlorella* and the highest daily aggregates of light income daily income of photosynthetically active radiation that can be realistically proposed ( $12.6 \text{ MJ m}^{-2} \text{ d}^{-1}$ ;  $10.5 \text{ mol C m}^{-2}$ ) through relationships developed in Reynolds (1997a).

#### D. System Exergy and Disturbance

The precise quantification of Fig. 4A may be misleading and not universally applicable. This is less important, however, than the form of the traces of the maximum light-harvesting curve and the maintenance diagonal and of the geometrical shape that they bound. It represents the only part of the plot in which population growth can be sustained. Above it is a void of underpopulated opportunity (exploitation of which requires the expansion in the light-harvesting capacity by increasing producer biomass in the rightward direction); below is an area representing unsustainable biomass in excess of the current energy income (far from being able to increase its total biomass, the system here must shed biomass, so that the maintenance of what remains is brought back within means of the current energy flux to sustain). In this way, the spring increase of phytoplankton is represented by a rightward response of biomass to an upward income in harvestable energy, with the increase in light harvesting capacity moving diagonally upward and rightward in consequence. By analogy, the autumnal decline in biomass is a leftward response to regain balance between now-high maintenance costs against a diminishing energy income. The excess of energy-harvesting capacity over the costs of its maintenance provides the existing crop with a second asset—a cushion of capacity that can absorb the impact of short-term variability in the energy

income without resorting to a restructuring of the biomass.

Because we have defined this cushion in terms of energy exchanges, it is appropriate to refer to it in thermodynamic terms—it is the exergy of the system (Jørgensen, 1992). We may see at once a presumption that a high income flux relative to biomass maintenance favors exploitation by species that can build the highest levels of harvesting capacity for the biomass—that is, those which contribute most to a high level of exergy. However, high net fluxes are not reliable and energy harvesting is ultimately sensitive to changing day length and solar elevation and to the variability of cloud cover and atmospheric absorption and scatter, all of which affect the daily flux of harvestable energy penetrating the water. Moreover, atmospheric variability also includes the work of wind and the impacts of varying dissipation rates on the mixed-layer depth and, hence, the mean exposure of entrained plant biomass to harvestable energy.

The reality of biomass assembly and its additional energy-harvesting capacity is that they have to be conducted against a background of a stochastically changing energy income. Intriguingly, the components of the relationship are provisionally quantifiable in energetic terms, at least probabilistically, if the energy harvest is solved as the mean biomass-specific photon capture by plants entrained in the mixed-layer circulation (Reynolds, 1997a). In this way, the temporal fluctuations in the harvestable income can be represented as an irregularly oscillating time track (the broken line in Fig. 4B). High values correspond to intense photon fluxes into clear or shallow mixed layers, low points to overcast skies or strong wind-forcing and mixed-layer deepening. We may also recognize that the externally forced depression of the harvestable energy flux may still be sufficient to meet the biomass maintenance cost, with a little in hand to maintain positive increase—that is, the fall in income is absorbed by the exergy cushion: the structure survives and has the resilience to recover promptly to its maximum capacity when the harvestable energy flux is restored to an optimum and the capacity to expand the biomass is fully restored. On other occasions, however, especially once a substantial standing biomass has been put in place, the flux of harvestable energy falls below the minimum maintenance requirement. The cushion of exergy is exceeded; the energy needs of the system are now out of balance with its supplies and are fundamentally unsustainable. This situation cannot persist for long before there has to be a reduction in the standing biomass, back to a level that is energetically sustainable. There is abundant resort



to minimal metabolism, including the production of resting spores and propagules, through to mass mortalities of vegetative cells. The biomass response is sharp and severe, conforming to all conventional appreciations of an externally forced disturbance reaction.

Following the course of irregular forcing episodes, represented in Fig. 4(B), it is easy to appreciate that planktonic systems are far more liable to restructuring disturbances than they are to achieve their autopoietic potential of a low-diversity, competitively excluded climax to the succession. This maturation process is simply overridden by the frequency of externally forced disturbances that it rarely, if ever, proceeds to its logical outcome. It may also be seen that the intervention of weather fluctuations and events, superimposed on predictable climatic cycles, is a, if not the, principal agent resisting exclusion and local extinction of planktonic species and contributing to Hutchinson's paradox of local species richness. Steady states are just so rarely achieved that species are retained, at least at the scale of plankton generations, in a nonequilibrium coexistence.

Here, at least, is a candidate mechanism contributing to the maintenance of a high biodiversity in the plankton. We should pursue its workings a little further.

### E. Resource Limitation and Diversity

Long before most pelagic systems can approach the attainment of a producer base at the capacity of the harvestable energy flux, biomass assembly will have been constrained by the capacity of the bioavailable nutrient resources. We can represent this constraint very simply by superimposing the vertical axis across the exergy cushion (Fig. 4C). Between the origin and the resource limit there remains an opportunity for biomass fluctuation but, against the vertical axis, the assembly opportunity remains strictly within the definition of exergy. Thus, the area of the plot corresponding to sustainability of growth (Fig. 4A) is finally shaped as in Figure 4D.

The new geometric figure is without a name—the best likeness I can think of is the sail of a windsurfer. Its periphery, however, readily corresponds to quantities established in sections III–V: the abundance of resources, relative to the consumptive demand, favors the advance of the most exploitative species, whose geometric range spreads from the origin along the line of the maximum biotic exergy flux. The closer this approaches the capacity of the resources, the greater is the stress of resource deficiency and the greater is the adaptation required to exploit it. Maintaining a high exergy flux remains the strongest driver of the plank-

tonic succession. Plainly, the curved upper surface corresponds to the C-S axis of increasing structural development, increasingly subject to resource competition and powerful selection. In contrast, the maintenance axis is the lower boundary of the adequacy of the exergy flux to drive the assembly of the community. Thus, the lower boundary of the “sail” represents the extremes of disturbance tolerance permitted by the attuning R-strategy. The two straight boundaries could equally be scaled in terms of  $K^{**}$  and  $I^{**}$ , the rightward trend representing increasing resource limitation, the downward trend corresponding to carbon-processing limitations. The axes serve just as well in separating the animal analogues of food supply and resource stress and “explerent” opportunism provided by externally imposed disturbance. Subject to current methodological uncertainties, there seems to be every probability that analogous axes describe the availability of organic carbon sources and processing opportunities for microbial plankton too. Note also that the fluctuating coordinates of stochastically variable environments continue to fall both within and beyond the sail area defining positive growth responses; only when the track is kept firmly and persistently in the S or R areas of the plot is there likely to be fierce and ongoing competition leading to the progressive installation of a single species, dominating a low-diversity, low-equitability community. On the other hand, for as long as externally driven variability keeps resetting the coordinates of the environmental conditions, at least with respect to the chosen axes, the time track keeps moving freely and extensively across and beyond the body of the shape, signifying that conditions rarely exist for sustained competition to last either for long enough or in one direction for any of the species present to have sufficient chance to deploy its superior adaptations to the competitive exclusion of others.

## VII. MECHANISMS PROMOTING AND MAINTAINING DIVERSITY IN THE PLANKTON

### A. Diversity within Habitats

The geometric representation of environmental variability and of its impact on species selection may properly be pursued in relation to the exergy model, but it is actually easier to bring the concept of fluctuation tracking back to the habitat template (of Fig. 3A), because it is easier to relate to the environments it seeks to represent (Figs. 1 and 2) and because sufficient pre-

liminary knowledge exists about how the tracking thus represented actually selects for the preferred traits of species, at least of the freshwater phytoplankton (Fig. 3B). To be clear, the resource- and energy-replete conditions, wherein resources fully meet present organismic demands, are represented in the upper right-hand corner of the template. Of those present, the species most advantaged are the opportunistic, invasive C-type species (like *Chlorella*; see section III.B and Fig. 3B). Biomass growth creates a strain on the readily available resources, more elaborate resource gathering is required, and autopoiesis favors a succession to more conservative, accumulative S-type dominants (like *Microcystis*) with a high tolerance of resource-supply stress and a resistance to disturbance. Large size becomes selectively valued but at the tradeoff in terms of surface-to-volume ratio (Fig. 3B) and at the price of slower metabolism and growth. In more nutrient-replete environments, the principal stress is imposed by having to operate on low or intermittent light doses. Survival prospects are enhanced by motility, combined with increased size, but the selective advantages of a high surface-to-volume ratio are not abandoned—the morphological attenuation among the acclimating R-type diatoms and *Planktothrix*-type filaments offers a high tolerance of mechanical disturbance and the resilience to recover from severe forcing events.

The organizational trends imposed by disturbance and stress press the selective bias toward the R or S apices of the triangular template, shown in Fig. 5A. We have already seen that seasonal trends track through the template matrix in broad, predictable ways (summarized as S-ward and R-ward trends in Fig. 5B), but

substantial within-season variability will lead to almost chaotic short-term time tracks, such as the fragment included in Figure 5C.

Our deduction is again that the effect of environmental variability is to move any selective advantage among species at a faster rate than autopoiesis narrows the opportunities or that competition can forge a low-diversity monoculture dominated by a single well-adapted species. We know that this theoretical outcome, correctly anticipated by Hutchinson (1961), certainly is achievable and good descriptions of local, competitively excluded community structures are described in the literature. These cover the sort of arrested “plagioclimaces” of overwhelming *Microcystis* dominance of tropical eutrophic Lake George (Ganf and Viner, 1973) and imitated in several field-scale enclosure experiments (summarized in Reynolds, 1988a) of year-round *Planktothrix* dominance of exposed, continuously mixed hypertrophic polder lakes in the Netherlands (first described by Berger, 1975), of the sustained dominance of nanoplankton in the continuously flushed, groundwater-flushed Montezuma’s Well in Arizona (Boucher *et al.*, 1984), and the unique dominance by *Chlorella* of a cooling gradient across the hot-spring fed Rotowhero, New Zealand (Jolly and Brown, 1975). No less impressive is the striking commonality of the autopoietic organization of vertically segregated layers (or “plates”) of algal and microbial producers on the stable physicochemical gradients in permanently ice-covered lakes in Antarctica (e.g., Vincent, 1981), in tropical forest lakes (Reynolds *et al.*, 1983), and in midlatitude karstic dolines (Vicente and Miracle, 1988).

In contrast, however, the great majority of assem-

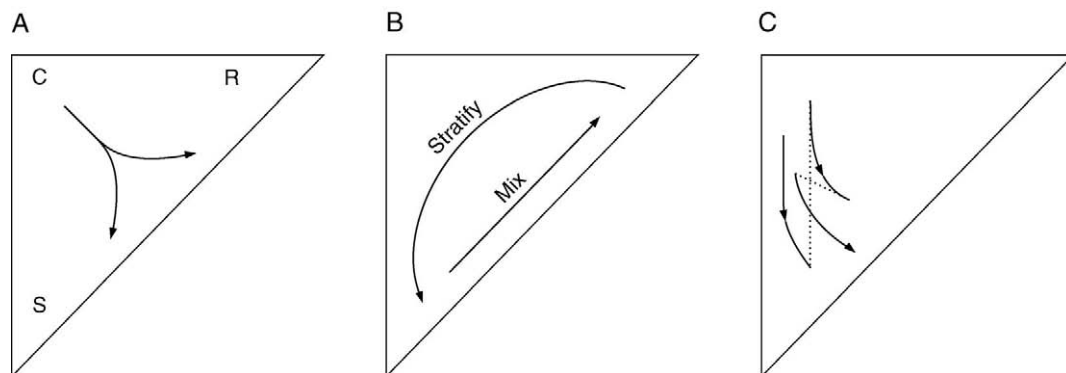


FIGURE 5 (A) From a starting condition, near the top left (C) corner of the triangle, environmental change moves to select increasingly for (S-) species tolerant of resource stress or for (R-) species tolerant of energy limitation (processing constraints); (B) major environmental restructuring alters the trajectory, changing the selection significantly; (C) under more frequent and perhaps less severe forcing, trajectories are redirected erratically, delaying progress to the competitive exclusion signified by the R and S apices.

blages that have been sampled, identified, and described in the literature are, simply, weakly organized. There need not be any paradox in this, once it is appreciated that in such accumulating, mid-successional assemblages most of the specific populations are not in any steady state but rather in a state of flux, either increasing or decreasing in response to events in the recent past. Thus, most so-called communities, in fact, comprise populations which, at the given point in time, are in some stage between being dismantled and reassembled.

The crucial questions about local diversity in the plankton should be "How long has it been since a restructuring was initiated by external forcing of a magnitude sufficient to create a disturbance?" and "At what rate is the restructuring taking place?" The field-scale experiments carried out in the large limnetic enclosures in Blelham Tarn, England, pointed to a progress to steady-state monocultural dominance requiring from 12 to 16 generations of the eventual dominant (Reynolds, 1988a). The time required for this could be as few as 35 to 60 days, provided growth rates could be sustained by water temperatures  $\geq 20^{\circ}\text{C}$  and a supply of carbon and nutrients of a capacity to support the last population doubling. Accordingly, limitations on growth rate imposed by low temperature, slow carbon renewal, or nutrient depletion extend the generation times reciprocally, delaying the onset of the climactic condition and slowing the exclusion of diversity. More likely, the intervention of external forcing (strong winds, heavy rain) would interrupt the successional maturation at an earlier date, and if of sufficient magnitude (exceeding the net structural exergy flux), selecting for alternative species with a new target outcome. It was found consistently that the community response, represented by two to four divisions of the newly selected species simultaneously with the dieback of the erstwhile-selected species, took some 5 to 15 days before it was clearly manifest. It was deduced that a disturbance frequency of this magnitude (two to four generations) is sufficient to maintain an optimal local species diversity (Reynolds, 1988a).

Analyses of temporal diversity fluctuations in small lakes in locations as far apart as Canada (Trimbee and Harris, 1983) and central Hungary (Padišák, 1993) confirm the validity of this deduction, showing diversities in phytoplankton composition peaking within 11 days of a recognized physical stimulus. These and numerous other case studies have been brought together to demonstrate this planktonic validation of Connell's (1978) intermediate disturbance hypothesis (Padišák *et al.*, 1993). Its articulation followed from a consideration of the celebrated diversity recognized in coral reefs and

tropical forests, which in some ways, posed analogous questions to those raised by Hutchinson (1961) about the plankton. The essence of Connell's hypothesis is that frequent disturbance excludes all but fast-maturing species; very infrequent disturbance allows competitive exclusion to reduce diversity; therefore, maximum diversity is maintained at intermediate frequencies. Recently, this simply stated insight has been the subject of a belated debate among ecologists. Setting aside those who contest whether others had not put forward the idea before Connell distilled it so elegantly and those who complain (mistakenly) of a lack of experimental evidence, the main line of argument has sought to distinguish the effects of disturbance intensity and of disturbance frequency. The debate is partly resolved by recalling the fact that disturbance is solely the response to an imposed forcing. The reference point of whether a disturbance has occurred at all and, if so, how intense it was, should be most usefully found in comparison of the forcing energy and the accumulated cushioning of the exergy flux. If structure has to be lost and re-grouped as a consequence of an external force and if the recovery depends on the restoration of opportunity, local Shannon diversity can be shown consistently to benefit during the reconstruction phase, when competition is least. At both high and low frequencies of forcing, the numbers of intermediate species are reduced so that the inocula are no longer readily available to take advantage of conditions under which their growth might be favored. In contrast, local stocks of mid-successional species are enhanced by the sorts of opportunities that might be provided every two or three generations by the (intermediate) rejuvenation of appropriate growing conditions and with a marked alleviation of severe interspecific competition in the early post-disturbance sequel.

## B. Species Richness among Habitats

It can be seen that the species representation contributing the local diversity remains heavily biased toward those that have been well represented in the recent past, and which, of those whose growth should be favored by the onset of the appropriate environmental characteristics, remain potentially capable of seeding the largest inocula from standing stocks of vegetative cells and resting propagules. With the same kinds of environmental variability affecting Shannon-type species diversity in each tangible locality, there is still no clear explanation as to why the collective representation of planktonic organisms is so relatively species rich. We need a proposition for separate species representation

in separated localities which recognizes that communication between localities is sufficient to maintain the high level of apparent cosmopolitanism among planktonic organisms. In other words, we need to be aware of the roles of perennation and dispersal in relation to the maintenance of biodiversity.

The survival of any species whose range of habitat suitability (as defined in section II) is discontinuous, in time as well as space, involves the separate development of separated populations, but with a measure of regenerative connection and gene renewal that resists their permanent divergence. The metapopulation ecology of plankton is not, formally, a well-rehearsed topic, though there have been numerous studies to contribute a general appreciation of the principles. Although the existence of liquid water on the planet has a very long geological history, individual bodies of fresh waters are extremely transient. With the exception of the basins formed by tectonic movements, a majority of these is less than 20,000 years in age. The idea of lakes as islands in a terrestrial sea is an easy one to assimilate, even if we wish to extend the analogy to suggest they are like volcanic eruptions that are shortly to be eroded back under the sea. The actual oceans are physically and temporally contiguous, but the patterns of global circulation allow significant habitat differentiation to be maintained at the scales at which planktonic organisms live their lives. The principles of island biogeography propounded by MacArthur and Wilson (1967) provide a good model for plankton ecology.

The temporal discontinuities are most appropriately bridged by the production of resistant, resting propagules. Even the most elementary biological texts on aquatic protists are memorably punctuated with references to the production "of cysts, to survive adverse conditions." Certainly, among the freshwater phytoplankton, almost all the major groups represented produce some kind of physiological resting stage, if not a discrete resting spore, cyst or akinete, some of which can remain dormant for many years and still be fully viable (see Reynolds, 1984a; over 60 years in the case of some *Anabaena* akinetes in a dated lake sediment). Eggs may fulfill a similar role for zooplankton, maintaining banks of inocula, pending the restoration of conditions favorable to growth and recruitment of successive generations (Hairston, 1996). Produced in adequate numbers, such dormant life-history stages provide a significant survival "hedge" through periods of hostile conditions (low temperature, low light, poor resources or food availability, drought, and so on). The inocula potentially supplied by subsequent spore germination and egg hatching provide a powerful mechanism

for maintaining diversity in seasonally fluctuating environments.

In many cases, the propagules facilitate spatial transfers as well. Dispersal in water droplets, or in dust, or in or on the bodies of animals have been shown to be effective pathways of planktonic organisms (review of Kristiansen, 1996) and to have been directly implicated in the establishment of populations in new or isolated bodies of water (Maguire, 1963, 1977). Planktonic species are not equally amenable to dispersal, the trait being related to other aspects of the life-history strategies; for instance, effective perennation and dispersion are essential properties of species of temporary or temporally variable habitats. In general, the dispersability, or invasiveness, of propagules (or, indeed, of vegetative cells) is dependent on a raft of such species-specific features as their sizes, their resistance to desiccation, and the numbers and frequencies with which they are produced. The probability ( $p$ ) of a given species ( $A$ ) being able to establish itself in another water body will be determined as a function, partly of its relevant species-specific traits ( $T_A$ ) and partly of the problems posed to all would-be invaders of the distance ( $d$ ) and size ( $a$ ) of the new (or "target") site from the existing (or "source": MacArthur and Wilson, 1967) population:

$$p = f(1/d, a, T_A)$$

An important converse of this relationship also manifests itself in relation to the ongoing suitability or effectiveness of source and target sites as a habitat for the given species. Either because the habitat changes in consequence of autogenic properties or because species  $A$  is prevented from completing its perennation in that habitat, through the intervention of a facultative predator or pathogen, it is ultimately likely that the survival of the species in that location is at risk. Repeated at several sites, this process would threaten the survival of the species in its entirety and bring about a diminution in total species richness (i.e., the biodiversity). The resistance to that comes, quite literally, from the patch dynamics of habitat distribution and their temporal suitability to species  $A$ . Thus, the status of species  $A$  is a function of the number of possible habitat patches ( $N$ ), the number that is occupied ( $O$ ), the rate at which they become excluded therefrom ( $e$ ), and the probability ( $p$ ) of colonizing the ( $N-O$ ) unoccupied habitats. The rate of expansion (or loss) of the metapopulation in time ( $dO/dt$ ) may be symbolized:

$$(dO/dt) = pO(N-O) - eO$$

For the more familiar and (supposedly) more cosmopolitan planktonic forms, relative ease of dispersal is, manifestly, the dominant factor upholding the widespread occupancy among suitable habitats. For the majority of known species that are relatively rare (that is, there is a relatively low occupancy among the total number of patches), effective dispersal may be resisted by specific traits ( $T_A$ ) or simply by low numbers. In this case, survival may be considered more tenuous. However, from the limited investigative evidence available, many of these rarer species have been numerous in the past and occasionally continue to perform relatively well on the inertia of persistent, viable propagules. While the length of this "ecological memory" (Padisák, 1992) is not certainly quantified, we may be aware of the fact that it is generally the rarer species that contribute most to the total biodiversity of planktonic organisms. Whereas the metabolically significant planktonic biomass relies on the productivity of a relatively small number of species in a series of variable and renewable habitats, the majority of species is limited to habitats in which temporal variability is less extreme or exclusion is long protracted.

## VIII. CONCLUSIONS AND IMPLICATIONS

There is scarcely a short answer to the question, "How is biodiversity maintained in the plankton?" The available evidence is that there is an operative blend of those mechanisms hypothesized to underpin diversity in other ecosystems. Global species richness is assisted by protracted isolation of populations and endemism of subsequently differentiated species to such locations as Lake Baykal and small lakes in Australasia. Progressive functional adaptations and niche specialism among, for example, the photobacteria, the cyanoprokaryotes, and the planktonic crustacea contribute to the pervasion of selected species into sites offering their preferred habitat conditions for long time periods, even when those sites are sometimes mutually remote. However, the majority of species seem to be reliant on the periodic rejuvenation of a broad range of habitat conditions and the opportunity to fill, mainly noncompetitively, the spare capacity that is thus regenerated. On the basis that opportunities are simultaneously closing, there is an implicit dependence on the maintenance of viable seed banks and the ability to combine this either with efficient dispersal or with high survivorship.

This last distinction permits the differentiation of invasive and accumulative life-history strategies. They have equal merit because they exploit different frequencies and intensities of environmental fluctuation. Indeed, the full range of conservative mechanisms may be considered to have evolved to maximize the exploitative opportunities provided by the variability within and between the available habitats. Consequently and conversely, it is the variety of dimensions and scales of variability that upholds the diversity of species. The nearest that we can get to a short answer to the biodiversity question is that it is the means by which living systems cope with a nonequilibrium world.

The feature distinguishing plankton-based ecosystems from terrestrial ones is that the relevant temporal scales are mostly much shorter in open waters than on the land. Not only are the mechanisms of planktonic biodiversity conveniently observed but, provided that appropriate scaling factors are applied, the study of planktonic systems may be rewarded with insights into current issues about the importance and the protection of biodiversity.

Of the many propositions that have been advanced about the role of diversity in upholding ecosystem processes, current attention focuses mainly on four broad postulates (Lawton, 1997). In various ways, they relate to Darwin's (1859) proposition that a large number of species imparts a higher level of functional stability than does a small one. Thus, to lose species inevitably impairs the functional integrity of the ecosystem. The "redundant species hypothesis" counters that, because species are not equally represented, some are contributing much more than others to ecosystem function. Indeed, some "keystone species" (Paine, 1969) drive the main functions and have a much stronger influence than "passenger" species (Walker, 1992). Then, logic suggests, a minimal diversity is essential to adequate ecosystem functioning and most of the species are really redundant in their roles (Lawton and Brown, 1993; Walker, 1992). Almost diametrically opposed to this view is the "rivet hypothesis" (Ehrlich and Ehrlich, 1981), which accords to each species an essential contributory role, and which, if lost from the whole, like rivets lost from the structure of an aircraft, quickly lead to serious functional impairment and failure. A third hypothesis takes a much looser view, suggesting that function is modified by changes in the richness of species composition but in unpredictable ways ("idiosyncratic"; Lawton, 1994) because the contributions of individual species to system function are unequal. The fourth possibility is the null hypothesis that ecosystems are insensitive to changes in species composition. This

seems increasingly to be implausible and will not be discussed.

One of the important lessons from the plankton scale is that a distinction should be made between structural stability and functional stability. Taking the basic function of the plankton to be the material cycling of matter in open water, driven by solar energy invested in carbon bonds, the photoautotrophic, heterotrophic and phagotrophic roles are fulfilled, respectively, by any or all of a large number of species of phytoplankton, bacterioplankton, and zooplankton. Despite having vastly different and frequently changing species compositions, studies of the productivity of planktonic systems demonstrate a remarkable level of functional coherence (Schindler, 1990). In this way, pelagic fish may continue to feed planktivorously and to respire photosynthetically generated oxygen almost without reference to the planktonic structure, just so long as similar functions are maintained.

At the level of planktonic communities, a decline in the population of a dominant species will carry fewer implications if a second species is poised to substitute quickly. Frost *et al.* (1995) made the case that species complementarity made for functional compensation of the smaller scales of structural variability. The case is analogous to the arguments about the flexibility of communicating information through networks rather than along a single pathway (Pahl-Wostl, 1990), or what has been called the "World Wide Web" explanation of biodiversity (Reynolds, 1997b). This recognizes the idea that some species may contribute less to overall ecosystem function than other, more functionally creative keystone species (following Paine, 1969), or "ecosystem engineers" (Lawton, 1997), and by providing opportunities to others may also serve to promote and maintain the species richness. It also values the role of the reserve (richness) of complementary species able to fulfill the compensatory function. The versatile airplane still relies on having most of its rivets in place.

What clues does planktonic diversity give us about the approaches to upholding species richness in other ecosystems? The advice "Keep up the disturbance" is a little trite and somewhat counterintuitive to traditional conservation attitudes. Only at the global scale of aquatic systems is it possible to conceive the simultaneous existence of sufficient a range of habitats to accommodate the entirety of species. No single location, no matter how variable, offers habitat or refuge for more than a minority representation of the aquatic biota. Aquatic biodiversity has to be pursued at the scale of fluvial catchments and regional seas—every lake, pond, evolving flood plain, river estuary, coastal shelf, and

ocean circulation is but a piece in the biospheric mosaic. The essential general aim should be to ensure that suitable habitat elements persist, each in a range of developmental states. Local extinctions are resisted by good between-locality communication of propagules. In this sense, managing biodiversity has to adopt the philosophies of patch dynamics, taking full account of the longevities and the invasiveness of species-specific populations in relation to the externally influenced availability and renewal of suitable patches.

### See Also the Following Articles

BACTERIAL BIODIVERSITY • ESTUARINE ECOSYSTEMS • LAKE AND POND ECOSYSTEMS • OCEAN ECOSYSTEMS • REEF ECOSYSTEMS • RIVER ECOSYSTEMS

### Bibliography

- Atlas, R. M., and Bartha, R. (1993). *Microbial Ecology—Fundamentals and applications*. 3rd ed. Benjamin Cummings, Redwood City.
- Azam, F., Fenchel, T., Field, J. G., Gray, J. S., Meyer-Reil, L. A., and Thingstad, F. (1983). The ecological role of water-column microbes in the sea. *Marine Ecology, Progress Series* 10, 257–263.
- Bannister, T. T., and Weidemann, A. D. (1984). The maximum quantum yield of plankton photosynthesis. *Journal of Plankton Research* 6, 275–294.
- Berger, C. (1975). Occurrence of *Oscillatoria agardhii* Gomont in some shallow eutrophic lakes. *Verhandlungen der internationale Vereinigung für theoretische und angewandte Limnologie* 26, 97–113.
- Boucher, P., Blinn, D. W., and Johnson, D. B. (1984). Phytoplankton ecology in an unusually stable environment (Montezuma Well, Arizona, USA). *Hydrobiologia* 119, 149–160.
- Brooks, J. L., and Dodson, S. I. (1965). Predation, body size and composition of the plankton. *Science* 150, 28–35.
- Connell, J. H. (1978). Diversity in tropical rain forests and coral reefs. *Science* 199, 1302–1310.
- Darwin, C. R. (1859). *The Origin of Species by Natural Selection* (Reprinted). Penguin Books, London.
- Dumont, H. (1999). Effects of reservoirs on faunal richness and dispersal. In *Theoretical Reservoir Ecology* (M. Straškraba and J. G. Tundisi, Eds.). 477–491. IIE, São Carlos.
- Ehrlich, P. R., and Ehrlich, A. H. (1981). *Extinction. The Causes and Consequences of the Disappearance of Species*. Random House, New York.
- Elliott, J. M. (1996). British freshwater Megaloptera and Neuroptera. *Scientific Publications of the Freshwater Biological Association* 54, 70pp.
- Fogg, G. E. (1965). *Algal Cultures and Phytoplankton Ecology*. Athlone Press, London.
- Frost, T. M., Carpenter, S. R., Ives, A. R., and Kratz, T. R. (1995). Species compensation and complementarity in ecosystem function. In *Linking Species and Ecosystems* (C. G. Jones and J. H. Lawton, Eds.), pp. 224–239. Chapman and Hall, New York.
- Ganf, G. G., and Viner, A. B. (1973). Ecological stability in a shallow equatorial lake (Lake George, Uganda). *Proceedings of the Royal Society of London B* 184, 321–346.

- George, D. G., and Reynolds, C. S. (1997). Zooplankton-phytoplankton interactions: The case for refining methods, measurements and models. *Aquatic Ecology* 31, 59–71.
- Grime, J. P. (1979). *Plant Strategies and Vegetation Processes*. Wiley Interscience, Chichester.
- Hairton, N. G., Jr. (1996). Zooplankton egg banks as biotic reservoirs in changing environments. *Limnology and Oceanography* 41, 987–1092.
- Hall, D. J., Threlkeld, S. T., Burns, C. W., and Crowley, P. H. (1976). The size-efficiency hypothesis and the structure of zooplankton communities. *Annual Review of Ecology and Systematics* 7, 177–208.
- Hardin, G. (1960). The competitive exclusion principle. *Science* 131, 1292–1297.
- Hart, R. C. (1996). Naupliar and copepodite growth and survival of two freshwater calanoids at various food levels: Demographic contrasts, similarities and food needs. *Limnology and Oceanography* 41, 648–658.
- Huntley, M. E., and Lopez, M. D. G. (1992). Temperature-dependent production of marine copepods: A global synthesis. *American Naturalist* 140, 201–242.
- Hutchinson, G. E. (1961). The paradox of the plankton. *American Naturalist* 95, 137–147.
- Jolly, V. H., and Brown, J. M. A. (1975). *New Zealand Lakes*. Auckland University Press, Auckland.
- Jørgensen, S. E. (1992). *Integration of Ecosystem Theory: A Pattern*. Kluwer, Dordrecht.
- Juhász-Nagy, P. (1993). Notes on compositional diversity. *Hydrobiologia* 249, 173–182.
- Kirk, J. T. O. (1994). *Light and Photosynthesis in Aquatic Ecosystems*, 2nd ed. Cambridge University Press, Cambridge.
- Kristiansen, J. (1996). Dispersal of freshwater algae: A review. *Hydrobiologia* 336, 151–157.
- Lawton, J. H. (1994). What do species do in ecosystems? *Oikos* 71, 367–374.
- Lawton, J. H. (1997). The role of species in ecosystems: Aspects of ecological complexity and biological diversity. In *Ecological Perspectives of Biodiversity* (T. Abe, S. R. Levin, and M. Higashi, Eds.), pp. 215–228. Springer, New York.
- Lawton, J. H., and Brown, V. K. (1993). Redundancy in ecosystems. In *Biodiversity and Ecosystem Function* (E. D. Schulze and H. A. Mooney, Eds.), pp. 255–270. Springer University Press, New York.
- MacArthur, R. H., and Wilson, E. O. (1967). *The Theory of Island Biogeography*. Princeton University Press, Princeton.
- Maguire, B. (1963). The passive dispersal of small aquatic organisms and their colonization of isolated bodies of water. *Ecological Monographs* 33, 161–185.
- Maguire, B. (1977). Community structure of protozoans and algae with particular emphasis on recently colonized bodies of water. In *Aquatic Microbial Communities* (J. Cairns, Ed.), pp. 355–397. Garland, New York.
- Mann, K. H., and Lazier, J. R. N. (1991). *Dynamics of Marine Systems*. Blackwell Scientific Publications, Oxford.
- Margalef, R. (1978). Life-forms of phytoplankton as survival alternatives in an unstable environment. *Oceanologica Acta* 1, 493–509.
- Margalef, R. (1997). *Our Biosphere*. ECI, Oldendorf.
- Margalef, R., Estrada, M., and Blasco, D. (1979). Functional morphology of organisms involved in red tides, as adapted to decaying turbulence. In *Toxic Dinoflagellate Blooms* (D. L. Taylor and H. H. Seliger, Eds.), pp. 89–94. Elsevier, Amsterdam.
- Nielsen, S. N. (1992). *Application of Maximum Energy in Structural Dynamic Models*. Miljøministeriet, København.
- Padisák, J. (1992). Seasonal succession of phytoplankton in a large, shallow lake (Balaton, Hungary): A dynamic approach to ecological memory, its possible role and mechanisms. *Journal of Ecology* 80, 217–230.
- Padisák, J. (1993). The influence of different disturbance frequencies on the species richness, diversity and equitability of phytoplankton in shallow lakes. *Hydrobiologia* 249, 135–156.
- Padisák, J., Reynolds, C. S., and Sommer, U. (1993). *The Intermediate Disturbance Hypothesis in Phytoplankton Ecology*. Kluwer, Dordrecht.
- Pahl-Wostl, C. (1990). Temporal organisation: A new perspective on the ecological network. *Oikos* 58, 293–305.
- Paine, R. T. (1969). A note on trophic complexity and community stability. *American Naturalist* 103, 91–93.
- Pourriot, R. (1977). Food and feeding habits of Rotifera. *Ergebnisse der Limnologie* 8, 243–260.
- Ragan, M. A., and Chapman, D. A. (1978). *A Biochemical Phylogeny of Protists*. Academic Press, New York.
- Redfield, A. C. (1958). The biological control of chemical factors in the environment. *American Scientist* 46, 205–221.
- Reynolds, C. S. (1984a). *The Ecology of Freshwater Phytoplankton*. Cambridge University Press, Cambridge.
- Reynolds, C. S. (1984b). Phytoplankton periodicity: The interaction of form, function and environmental variability. *Freshwater Biology* 14, 111–142.
- Reynolds, C. S. (1988a). The concept of ecological succession applied to the seasonal periodicity of freshwater phytoplankton. *Verhandlungen der internationale Vereinigung für theoretische und angewandte Limnologie* 22, 683–691.
- Reynolds, C. S. (1988b). Functional morphology and the adaptive strategies of freshwater phytoplankton. In *Growth and Reproductive Strategies of Freshwater Phytoplankton* (C. D. Sandgren, Ed.), pp. 388–433. Cambridge University Press, New York.
- Reynolds, C. S. (1995). Successional change in the planktonic vegetation: Species, structures, scales. In *The Molecular Ecology of Aquatic Microbes* (I. Joint, Ed.), pp. 115–132. Springer-Verlag, Berlin.
- Reynolds, C. S. (1996). The plant life of the pelagic. *Verhandlungen der internationale Vereinigung für theoretische und angewandte Limnologie* 26, 97–113.
- Reynolds, C. S. (1997a). Successional development, energetics and diversity in planktonic communities. In *Ecological Perspectives of Biodiversity* (T. Abe, S. R. Levin, and M. Higashi, Eds.), pp. 167–202. Springer, New York.
- Reynolds, C. S. (1997b). *Vegetation Processes in the Pelagic*. ECI, Oldendorf.
- Reynolds, C. S., Thompson, J. M., Ferguson, A. J. D., and Wiseman, S. W. (1982). Loss processes in the population dynamics of phytoplankton maintained in closed limnetic systems. *Journal of Plankton Research* 4, 561–600.
- Reynolds, C. S., Tundisi, J. G., and Hino, K. (1983). Observations on a metalimnetic Lyngbya population in a stably stratified tropical lake (Lagoa Carioca, Eastern Brasil). *Archiv für Hydrobiologie* 97, 7–17.
- Romanovsky, Yu. E. (1985). Food limitation and life-history strategies in cladoceran crustaceans. *Ergebnisse der Limnologie* 21, 363–372.
- Rothschild, B. J., and Osborn, T. R. (1988). Small-scale turbulence and plankton contact rates. *Journal of Plankton Research* 10, 465–474.
- Schindler, D. W. (1990). Experimental perturbations of whole lakes as tests of hypotheses concerning ecosystem structure and function. *Oikos* 57, 25–41.

- Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal* 27, 623–656.
- Sournia, A., Chrétiennot-Dinet, M. J., and Ricard, M. (1991). Marine plankton: How many species in the world oceans? *Journal of Plankton Research* 13, 1093–1099.
- Sommer, U. (1981). The role of *r*- and *K*- selection in the succession of phytoplankton in Lake Constance. *Acta Oecologia* 2, 327–342.
- Sommer, U., Gliwicz, Z. M., Lampert, W., and Duncan, A. (1986). The PEG model of seasonal succession of planktonic events in fresh waters. *Archiv für Hydrobiologie* 106, 433–471.
- Sverdrup, H. U., Johnson, M. W., and Fleming, R. H. (1942). *The Oceans: Their Physics, Chemistry and General Biology*. Prentice Hall, New York.
- Talling, J. F., and Lemoalle, J. (1998). *Ecological Dynamics of Tropical Inland Waters*. Cambridge University Press, Cambridge.
- Tappan, H. (1980). *The Palaeobiology of Plant Protists*. W. H. Freeman and Co., San Francisco.
- Thierstein, H. R. (1989). Inventory of palaeproductivity records. In *Productivity of the Ocean, Present and Past* (W. H. Berger, V. S. Smetacek, and G. Wefer, Eds.), pp. 355–375. John Wiley and Sons, Chichester.
- Trimbee, A. M., and Harris, G. P. (1983). Use of time-series analysis to demonstrate advection rates of different variables in a small lake. *Journal of Plankton Research* 5, 819–833.
- Tyler, P. A. (1996). Endemism in freshwater algae. *Hydrobiologia* 336, 127–135.
- Vicente, E., and Miracle, M. R. (1988). Physicochemical and microbial stratification in a meromictic karstic lake of Spain. *Verhandlungen der internationale Vereinigung für theoretische und angewandte Limnologie* 23, 522–529.
- Vincent, W. F. (1981). Production strategies in Antarctic inland waters: Phytoplankton ecophysiology in a permanently ice-covered lake. *Ecology* 62, 1215–1224.
- Viroux, L. (1997). Zooplankton development in two large lowland rivers, the Moselle (France) and the Meuse (Belgium) in 1993. *Journal of Plankton Research* 19, 1743–1762.
- Walker, B. H. (1992). Biodiversity and ecological redundancy. *Biological Conservation* 6, 18–23.
- Waterbury, J., Watson, S., Guillard, R. R. L., and Brand, L. (1979). Widespread occurrence of a unicellular, marine, planktonic cyanobacterium. *Nature* 277, 293–294.
- Wetzel, R. G. (1995). Death, detritus and energy flow in aquatic ecosystems. *Freshwater Biology* 33, 83–89.
- Wilson, E. O., and Peter, F. M. (1986). *BioDiversity*. National Academy Press, Washington, D.C.







# PLANT–ANIMAL INTERACTIONS

Ellen L. Simms

*University of California Botanical Garden*

---

- I. Types of Interactions
  - II. Antagonistic Interactions
  - III. Mutualistic Interactions
  - IV. Summary
- 

## GLOSSARY

**Allee effect** For population size to be regulated, it must exhibit a negative density dependence. That is, the population growth rate must decline as the population gets larger. However, under certain circumstances, some populations exhibit positive density dependence. This phenomenon, in which population growth rate increases as the population gets larger, is called an Allee effect. An Allee effect may generate a critical minimum population size, below which extinction will occur.

**angiosperm** Flowering plant. A lineage characterized by flowers, seeds enclosed in carpels, specialized conducting elements in the phloem (sieve tube members) and xylem (vessels), presence of endosperm, double fertilization, and tectate pollen.

**diaspore** A plant part distributed by dispersal, regardless of its developmental and morphological origins. A diaspore may be a naked seed, a seed enclosed in a fruit, or many seeds enclosed in a fruit. It may also mean bulbs or lengths of rhizomes. A good synonym is propagule.

**fitness** Relative contribution of offspring to the next generation. An individual, genotype, or phenotype

whose progeny constitutes a large proportion of the succeeding generation has high fitness.

**granivores** Animals that eat seeds or achenes (grass fruits).

**herbivores** Animals that eat plants. Usually excludes instances when a single animal eats an entire plant, which is categorized as predation.

**phylogeny** The evolutionary relationships among taxa (groups of related organisms), often portrayed with some kind of branching diagram, with branches representing speciation events. Typically, the true phylogeny of a group is hidden deep in the past and evolutionary biologists must infer relationships. Various types of data and methods of analysis are used in this effort and there is considerable contention among groups of scientists as to which are most likely to estimate the true phylogeny.

**symbiosis** Interaction in which two organisms live in close proximity. Symbiosis can be antagonistic or mutualistic. Often, the larger individual is called the *host*, and its inhabitant is called the *guest*.

**trophic level** Position of a species in the food web (Fig. 1). Plants—autotrophs that convert solar energy to chemical energy and utilize mineral nutrients—constitute the first trophic level. Primary consumers—animals that feed on living plants—constitute the second trophic level. The third trophic level is composed of secondary consumers—animals that feed on primary consumers. Predation, parasitism, grazing, and herbivory are intertrophic level interactions; competition occurs among species in the same trophic level.

---

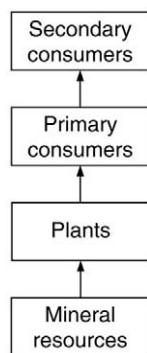


FIGURE 1 Trophic levels within a food web. Each trophic level is comprised of one or more species that consume individuals or resources at the next lower trophic level and are, in turn, consumed by species in the next higher trophic level. Arrows indicate the direction of flow of energy and resources in the food web.

**BECAUSE PLANTS CAN OBTAIN** nourishment and energy from inorganic sources, they are the foundation of most biotic communities. Animal consumption of plants is the primary conduit by which energy and resources enter the food web. However, the wide-ranging effects of plant-animal interactions on biotic diversity extend far beyond simple trophic links. This article will summarize the various types of plant-animal interactions and use a multidisciplinary approach to examine their implications for biotic diversity.

## I. TYPES OF INTERACTIONS

Biotic interactions can be categorized by their effects on the interacting parties (Table I). An interaction may not affect a species, or may be beneficial or detrimental. Antagonistic interactions negatively influence one or both species. Some interactions are clearly antagonistic: When a vole consumes an oak seedling, the rodent benefits while the plant dies. In other cases, it seems clear that both parties are benefiting, as when a hummingbird obtains nectar while transporting pollen from one plant to another. These interactions are termed mutualistic. The implications of other interactions may be more ambiguous. For example, Clark's nutcrackers consume pine nuts but are also important agents of seed dispersal, and it is unclear whether the interaction is mutually beneficial to both species or whether one species is benefiting at the expense of the other. The fitness costs and benefits to the parties involved in such an interaction may be conditional on the current environment. For example, during years of heavy seed production, birds may provide plants with a net benefit,

TABLE I  
Pairwise Ecological Interactions between Plants and Animals  
Categorized by Their Effects on the Fitness of Each Party

	Plant effect on animal	Animal effect on plant
Antagonisms		
Herbivory	+	-/0
Carnivory	-	+
Mutualisms	+	+

whereas during poor seed years, the net effect of birds on the plant may be negative.

Ecologists seek to categorize interactions because they have distinct ecological and evolutionary implications. Within biotic communities, certain kinds of interactions increase species diversity whereas others reduce it. Similarly, the nature of interactions determines their impact on genetic diversity within populations.

## II. ANTAGONISTIC INTERACTIONS

### A. Plant Consumers—Herbivores and Granivores

#### 1. Types of Plant Consumers

Animals consume plants in all kinds of habitats, including marine, terrestrial, and freshwater, and do so in wildly diverse ways. For some plants, there is at least one animal species devoted to consuming each type of organ. Some animals remove tissues by chewing; others suck plant sap. Grazing individuals feed from many different individual plants. Some animals live within plants, literally surrounded by food. These include borers, gallers, and leaf miners.

Many animals chew on plants, just as we do. However, another important mode of consumption is to use strawlike mouthparts to pierce and suck fluids from vascular structures such as xylem and phloem, which transport water, minerals, and other compounds throughout the plant. Aphids are common sap feeders. Spider mites are another type of plant-sucking arthropod that may be familiar to unhappy owners of house plants. The origin of this feeding mode probably extends back to the Carboniferous with the Paleodictyopteroidea, an assemblage of insects with sucking mouthparts.

Sap feeders share an interesting problem in common with many blood feeders: Neither blood nor plant sap provides a balanced or complete supply of vitamins

and amino acids. One solution to this problem among aphids has been to host intracellular microbial symbionts with the enzymes necessary to convert common nonessential amino acids into essential rare ones, much as Midas converted base metals to gold (Douglas, 1994). These intracellular guests inhabit special cells near the gut and oviducts and are transmitted by the mother to her eggs.

Perhaps the best known mode of plant consumption is grazing. All of us are familiar with picturesque scenes of cows grazing or deer browsing in verdant pastures. Far from peaceful, these are scenes of graphic violence. Hundreds, perhaps thousands, of plants are being eaten alive! Each grazer is eating photosynthetic organs from numerous living plant individuals. However, because the aboveground parts of most plants are constructed of repeating, renewable modules called shoots, a single episode of grazing or browsing rarely kills a plant.

Less well-known grazers include parrot fish, which maintain well-mown lawns of algae on coral reefs. When portions of the reef are protected from parrot fish, luxuriant algal growth can smother the coral. Sea urchins graze kelp forests and, when their natural predators are eliminated, may create vast barrens on the ocean floor. On land, many plant-eating insects, such as grasshoppers, katydids, and some beetles, are grazers. Some grazers feed on roots. These include tiny soil insects called springtails, relatives to silverfish, which move around in the air spaces between soil particles, nibbling on roots. Microscopic crustaceans in the water column of oceans and freshwater lakes and ponds graze on tiny photosynthetic single-celled organisms called phytoplankton. In this case, however, the animals function more like predators because they must kill their prey to eat it.

In contrast to grazers, which move from plant to plant and eat only a portion of each plant, some animals feed entirely on one plant during their lifetime. In some cases, many generations of the consumer will occupy the same tree, evolving greater and greater specialization to that one individual. The rate of evolutionary change in a lineage is a negative function of generation time. Organisms such as long-lived trees, which have very long generation times, are at a distinct evolutionary disadvantage relative to their short-lived pests, which may be able to evolve very quickly because they go through multiple generations every year. Indeed, an important unanswered question in evolutionary biology is how long-lived plants such as trees avoid being destroyed by rapidly evolving pests.

Some consumers live inside their food. Perhaps the most intriguing are gall-forming arthropods—wasps, flies, and mites that induce plants to form elaborate

structures within which the animal munches away on the host. Galls occur in many forms, including small red swellings on leaf blades, large globular swellings on branches, or bristly structures in odd places. In most cases, the female insect induces the gall when she injects her eggs into the plant tissues. Her young hatch and feed inside the gall, where they are relatively protected from enemies and the rigors of the physical environment. The relative frequency of gall-forming herbivores increases with increasing aridity of the environment, presumably because insects in galls are less vulnerable to desiccation. Plant-feeding nematodes, called vinegar eels, often live in galls on roots. The earliest known gall was produced by an insect feeding on tree fern fronds in the late Carboniferous.

Experiments show that the same species of galling insect produces very different looking galls on different plant species, suggesting that the plant determines gall form. Detailed work on the goldenrod ball galler, however, indicates that both insect and plant genes interact to determine gall size (Abrahamson and Weis, 1997). Although these insects appear to successfully manipulate plant development, which is a goal of many biotechnologists, relatively little research has been devoted to understanding the molecular mechanisms of this fascinating interaction.

Another group of consumers that lives inside plants are leaf miners. As their name suggests, the larvae of these flies and moths burrow within a leaf like a miner, eating the tissue they excavate and leaving behind their waste material, called frass. Some leaf miners snake along, leaving a long serpentine mine in their wake. Others remain in one place, creating a large blotch in the leaf. In many cases, leaf miners can be identified to species on the basis of host plant and the appearance of the mine. This has also made leaf miners easy to detect in the fossil record. Because they must live between the top and bottom surfaces of leaves, leaf miner larvae are very flattened from top to bottom. Some leaf miners pupate within the mine; others exit the mine and pupate in the soil beneath the host plant. In addition to leaf miners, some other insects mine plant stems, burrowing just beneath the stem surface and leaving similar-looking feeding galleries.

Bark beetles are economically very important plant consumers (Paine *et al.*, 1997). Adults attempt to bore into tree trunks, but healthy trees generally can “pitch” them out, literally flooding the bore holes with sap. However, the sap attracts more adults of these gregarious beetles and an ailing tree’s defenses can be quickly overcome. Female beetles then excavate extensive galleries just beneath the bark and lay their eggs in the termini of the tunnels. The larvae feed on both the host

wood and tree-feeding fungi that adults transport in special structures called mycangia. Often, bark beetles do not kill their host but are indirectly responsible for its death because they introduce deadly fungal pathogens.

## 2. Evolutionary Responses by Plants

Plants have evolved numerous responses to their consumers. Primary among these is resistance. Plants resist consumers in three main ways. They may defend themselves chemically or mechanically, or they may escape damage by being difficult for consumers to discover.

### a. Chemical Defenses

Plants are highly proficient chemists. There has been ongoing controversy as to why plants produce such a startlingly diverse array of chemicals, which, because they had no known function in primary metabolism, became known as secondary compounds. Early theories proposed that these compounds were waste products from “pathological overproduction of carbon.” But this explanation begged the question, “Why so many ways of throwing out the trash?”

By the 1950s phytochemists had begun speculating that secondary chemicals were important in plant defense, and this theory has dominated the latter half of the twentieth century. Nevertheless, competing hypotheses have held their own—notably that secondary chemicals are important in protecting plants from physical dangers such as ultraviolet light. Undoubtedly some plant chemicals do perform such functions, but it is certainly not clear why plants should have so many kinds of sun screen. Perhaps the most defensible alternative hypothesis is that secondary compounds are important in defense against microbial enemies such as fungal, bacterial, and viral pathogens. It is quite likely, in fact, that many compounds defend plants against both these agents and animal consumers.

Chemical defenses are divided into three categories: digestibility reducers, toxins, and repellents. Many plants produce digestibility reducers such as tannins, which are high molecular weight carbon-rich compounds that bind proteins and make them difficult to digest. Thus, these compounds do not directly kill the consumer, but make the plant less nutritious to eat. This indirect mechanism of defense leads to important questions about the selection pressures that might have led to their evolution. In particular, it is not always clear that individual consumers know the nutritional value of their host. Moreover, in many cases the individual that chooses the plant is not the one destined to feed on it. For example, when choosing among a population of the same plant species, female butterflies do

not necessarily lay their eggs on the individuals on which their larvae can develop best. Another question is, “If the chemical does not kill the consumer, why would it be evolutionarily advantageous to plants possessing it?” This question becomes even more vexing in light of considerable evidence that, when protected from other causes of mortality, consumers can compensate for poor food quality by consuming more (not less!) plant tissue. The leading hypothesis to explain this paradox is that consumers feeding on a plant with a digestibility-reducing compound will grow more slowly, making them more vulnerable to enemies, such as predators, parasites, or diseases, and therefore die before consuming much plant tissue.

Toxins, as their name implies, are poisons that have direct negative effects on animals that consume them. Some plants are so toxic that apparently no animals will eat them. However, in many cases certain animals have evolved mechanisms that detoxify extremely potent poisons. Often, these animals are specialists on that host plant. There may be strong evolutionary benefits to feeding on such a previously unexploited resource. The most obvious benefit is that no one else is using the plant, which reduces competition for food. However, because a toxic plant harbors so few consumers, it may also provide “enemy-free space” in which consumers are less likely to be discovered by predators or parasites. Enemy-free space could also arise by a slightly different mechanism, which is nicely illustrated by small consumers on seaweeds. Small crustaceans called amphipods live on the algae they eat. Parrot fish also eat algae, but when they discover an amphipod, they snap it up like candy. Thus, amphipods that feed on very toxic species of algae are less likely to be consumed by herbivorous fish that are opportunistically omnivorous.

Perhaps because of these kinds of benefits, many animals that have evolved the ability to consume a toxic plant have also evolved specialization to that host. For example, specialist herbivores may use the toxic compound as a cue to find their host. Additionally, the compounds may stimulate feeding or egg laying. This level of specialization sets up an important evolutionary trade-off for toxic plants. A compound that previously killed all consumers has become an attractant to at least one consumer. Any individuals that have higher levels of the compound may be better defended against most consumers but may be more attractive to the specialist. Conversely, individuals with lower toxin concentrations will be less attractive to the specialist but may become vulnerable to other generalist consumers. This type of situation may create stabilizing selection and cause the plant population to evolve an intermediate

concentration of the compound (Simms, 1992). Alternatively, the specialist could become a major cause of death and seriously reduce abundance of the host. Indeed, when plants introduced into new habitats become noxious weeds, specialist consumers are sometimes imported as biocontrol agents.

A final class of chemical defenses are repellents, which, as their name suggests, repel animals from laying eggs on or eating plants that possess them. Some evolutionary biologists have argued that herbivores will not quickly evolve mechanisms that overcome repellents. This argument is based on the understanding that the rate of evolution depends in part on the strength of selection. Toxins impose strong selection for detoxification mechanisms because they kill animals. Repellents simply cause animals to look elsewhere for food, which might impose weaker selection. The strength of selection imposed by repellents on consumers will, however, depend critically on the fitness cost to the animal of finding an acceptable alternate host.

#### *i. Animal Uses of Plant Defensive Compounds*

Whatever the evolutionary reasons for these compounds, plant secondary chemicals have enormous value to humans. For example, synthetic pyrethroids, which were originally extracted from a species of chrysanthemum, are valuable insecticides because they are effective but degrade quickly and so do not accumulate in the environment. Plant secondary chemicals are the source of virtually all of the herbs and spices that make food interesting to eat. In many cases, these compounds were first important in preserving foods and keeping them safe to eat. Finally, plant secondary compounds are an important source of pharmaceuticals. In fact, 25% of the modern medical drug prescriptions (119 different chemical substances) written between 1959 and 1980 in the United States were pharmaceuticals derived from 90 different plant species (Farnsworth *et al.*, 1985).

Other animals also use plant secondary compounds in interesting ways. Many consumers sequester plant compounds from their food and use them for their own defense, the best known example being the monarch butterfly, which harbors cardiac glycosides from its milkweed hosts. Some animals even self-medicate with secondary compounds (Rodriquez and Wrangham, 1993). For example, healthy woolly bear caterpillars avoid poison hemlock, the plant used to execute Socrates. But woolly bears infested with a lethal parasitoid will preferentially consume poison hemlock (*Conium maculatum*), which can kill the parasite and allow the caterpillar to survive to adulthood. European starlings

may protect their nestlings by lining their nests with fresh plant materials that inhibit arthropod hatching and bacterial growth. Animals may also use plant secondary compounds to preserve their food. Pikas, relatives of rabbits, live in burrows in talus slopes. To survive the winter in their alpine homes, they harvest and store enormous quantities of vegetation, which they store in haypiles. Pikas prefer to eat grass hay, but they often will harvest toxic herbs as well. Haypiles with these herbs are less likely to become moldy than those without them.

#### **b. Mechanical Defenses**

Mechanical resistance to consumers may be obvious, as in the case of spiny cacti and thorny shrubs, or more subtle, as in the case of silica bodies that render grass leaves less palatable to many consumers. Some plants combine both chemical and mechanical defenses. For example, Wright's datura possesses both toxic alkaloids and leaf hairs, called trichomes. Moreover, some genotypes possess simple hairs (mechanical defense), whereas others have a sticky surface provided by glandular hairs that excrete a sticky exudate (mechanical and chemical defense). The sticky hairs defend plants against whiteflies but not mirid bugs; plants with simple hairs are better defended against the latter pests. It is generally supposed that mechanical defenses are more difficult for consumers to overcome.

When animals consume other animals, they eat tissue that has a composition similar to their own. However, plant tissues are generally richer in carbon and poorer in nitrogen than animal tissues. In part because of mechanical defenses, a high proportion of the carbon in terrestrial plants is devoted to structural molecules such as cellulose and lignins. Animals lack the enzymes necessary to digest these compounds. Microbes do have such enzymes, however, and herbivores sometimes have elaborate modifications of their guts that house microbial symbionts that can digest the fibrous fare.

#### **c. Escape**

In addition to defending tissues, plants may also escape consumers in time or space. Ephemeral plants, especially those with annual life cycles, may obtain temporal escape with a life cycle that does not sustain consumers long enough for them to complete their life cycles. Even if short-lived plants do not starve out their consumers, they may limit them to only a single generation. This constraint prevents the buildup of dense pest populations which short-lived consumers can attain on long-lived plants. Escape may also be achieved by appearing only during seasons when consumers are rare. For ex-

ample, many cool-season plants are relatively unaffected by insect consumers, which are far more abundant during warm weather.

Highly dispersed plants may also escape consumption if the distances between them are greater than the average distance traveled by their consumers. This mechanism may be in part responsible for the astonishing diversity of certain tropical forests, in which only one individual of each tree species will be found in a large area. This concept is also embodied in the “resource-concentration hypothesis,” which states that dense concentrations of host plants will harbor the highest densities of consumers. If a plant species is both short-lived and rare, it may be so difficult to find that it can complete its life cycle before being found by consumers.

#### d. Tolerance and Compensatory Growth

Even if plants are damaged by consumers, they may evolve mechanisms that allow them to maintain fitness in the face of damage. Plants may tolerate damage through various compensatory mechanisms, including reallocating resources from undamaged plant parts to replace damaged tissues. Resources are usually allocated among plant parts in response to gradients between points of production (sources) and points of use (sinks). Thus, leaves, which produce photosynthate, typically function as sources, sending resources to meristematic sinks where new growth is occurring. However, if consumers damage actively photosynthesizing leaves, the area around the damaged leaves may shift from source to sink, thereby attracting resources for compensatory growth. Tolerance of consumers may be especially important among fast-growing plants living in resource-rich environments.

### 3. Coevolution of Plants and Herbivores

In 1964, Ehrlich and Raven conceived a coevolutionary hypothesis to explain the magnificent diversity of plant chemistry. They postulated that herbivorous insects are a strong mortality agent for plants and that if any trait arose that protected a plant from its herbivorous insects, that trait would quickly spread in the plant population. Further, they argued that escaping its herbivores would create for the plant the opportunity for a period of rapid speciation, called an adaptive radiation. They postulated a similar process for the insects. For the herbivores, the newly evolved plant species represent unused resources. Any insect trait that allowed an insect to exploit these plants would likewise result in an adaptive radiation of herbivores. As Janzen later argued, the crucial characteristic of this process that distinguishes it from

ordinary evolution by natural selection is its reciprocal nature. The plant evolves resistance to the herbivore, the herbivore then evolves a mechanism that negates the resistance, after which the plant evolves resistance to the herbivore, and so on, *ad infinitum*. With every crank of the coevolutionary process, new species arise through adaptive radiation.

This hypothesis has excited considerable controversy. Some authors argue that reciprocal coevolution is rare or nonexistent because insect herbivores do not impose sufficiently strong selection pressures on plants. Others argue that herbivores experience more selection from their natural enemies than from plants and that moving to new plants is driven by the adaptive advantage of enemy-free space. Further, Labandeira and Sepkoski have pointed out fossil evidence that indicates that the great radiation of modern insects began 245 million years ago and was not accelerated by the expansion of angiosperms during the Cretaceous period. However, neither insects nor plants have stopped evolving, and currently evolving systems provide the best tests of the coevolutionary hypothesis.

One way that scientists have tried to test the coevolutionary hypothesis is to compare the evolutionary lineages (phylogenies) of host plants and insect consumers that narrowly specialize on that group of hosts. Coevolution can lead to a pattern called cospeciation, in which the two phylogenies match, much like your fingers do when you place your palms together, fingers up. One hand represents the plant lineages, the other the insects. At the base of your palm is the ancestral species; the fingers represent various derivative lineages. An alternative phylogenetic pattern, in which lineages do not match, is produced by host switching. Host switching is the phenomenon whereby specialist consumers shift host species and then speciate on that new species, without any speciation by the host plant.

The alliance of Hawaiian silverswords and the plant hoppers that live and feed on them provides a particularly exciting pair of phylogenies with which to test the coevolutionary hypothesis. The Hawaiian silverswords are derived from a pair of ancestral species in the rather prosaic group of California plants called the tarweeds. Over the past 5 million years the Hawaiian descendents have rapidly radiated into a stunningly diverse array of species. Some of these Hawaiian descendents are magnificent rosette plants with life histories much like century plants—they live long lives terminating in the production of giant flowering stalks. Others are multi-branched perennial shrubs. Most are attacked by members of the plant hopper genus, *Nesosydne*, in the family Delphacidae. The plant hoppers are highly host specific;

each species feeds on one or a few closely related plant species. Phylogenetic analysis of molecular data from the plant hoppers and their hosts reveals a pattern of cospeciation: Each plant hopper species is most likely to use the host species that is most closely related to the host of its most close relative. This pattern is exciting, because it is expected to arise from reciprocal coevolution.

However, other mechanisms could produce the same pattern. For example, the plants could be evolving in response to some other selective pressure, with the insects following along behind. Matching phylogenies can also arise when both an insect and its host plant might speciate simultaneously in response to some external event as, for example, when they become geographically isolated from their main populations by a geologic event such as mountain building. Thus, phylogenetic comparisons alone cannot deduce whether coevolution has occurred. Other types of data must be examined. Careful observation of host use across hybrid zones suggests that the most likely explanation of the match between silversword and plant hopper phylogenies is that plant hosts are speciating in response to some external pressure and plant hoppers are tagging along behind plant host speciation.

#### 4. Community and Ecosystem Effects of Plant–Consumer Interactions

Considerable evidence suggests that consumers can reduce the fitness of individual plants and thereby impose selection pressures that produce evolutionary changes in plant traits. However, it is less clear whether consumers influence plant abundance in the landscape. This issue is at the heart of a controversy that has raged among community ecologists for the past two decades over the relative importance to community structure of top-down control by consumers of their resource populations versus bottom-up control of consumer populations by resources at the base of the food web. The current view is that both types of forces interact in complex ways to structure biotic communities. However, several competing hypotheses aim to explain how these forces interact. Three of the major models are summarized here.

Donor control models predict that while plants are food to their consumers, consumers have little effect on plant abundance. Thus, biomass of organisms at one trophic level is a function of the productivity of their resource base at lower trophic levels. This means that adding resources to the base of a food web will trickle up the web, increasing biomass at all trophic levels. In

contrast, consumer control models predict that each trophic level can be controlled by either its resources or its consumers, but not by both. They further predict that the direction control moves depends on the trophic level being examined and the number of trophic levels in the ecosystem. In particular, plants, at the base of the food web, are expected to dominate in ecosystems with odd numbers of trophic levels whereas herbivores will dominate in ecosystems with even numbers of trophic levels. Thus, increasing abundance in a particular trophic level will cascade up and down the food web, alternately expanding or shrinking trophic levels. Finally, keystone predation models predict that the species composition at each trophic level modifies the relative effects of resources and consumers.

The last model is more complex than the previous two, incorporating aspects of each. It is of particular interest here because it incorporates information about the relative vulnerability to consumers (i.e., resistance) of different resource (e.g., plant) species. The keystone model predicts that species diversity in resource populations can be maintained if resource species exhibit trade-offs between their relative competitive abilities and their relative resistance to consumers. Further, the model predicts that when resources are scarce, consumer populations will be small and the plant community will be dominated by a few fast-growing, strong competitors that are highly vulnerable to consumers, which are therefore consumer controlled. Under nutrient-rich conditions, consumer populations will be large and the plant community will consist primarily of a few slow-growing, well-defended species that are resource controlled because of their heavy investment in resistance. At intermediate levels of nutrient availability, both types of plants can coexist because of the trade-off between competitive ability and resistance to consumption (Leibold *et al.*, 1997). Consequently, species diversity will be greatest at intermediate levels of productivity. Other factors will determine whether these diverse communities are controlled by consumers (top down) or by resources (bottom up).

Leibold examined this prediction in planktonic communities of fishless ponds that varied in their level of mineral nutrient availability. These communities consist of photosynthetic planktonic algae (phytoplankton: single-celled green plants) that are grazed by herbivorous microarthropod zooplankton. Leibold found that algae in low-nutrient ponds consisted primarily of small, unprotected forms thought to be fast-growing but susceptible to grazing. Algae from more eutrophic (nutrient rich) ponds were larger and often sheathed



or gelatinous forms thought to be slow-growing but resistant to grazers.

In another recent analysis, Chase and colleagues reviewed studies of temperate terrestrial grasslands to determine whether the effects of consumers on plant biomass fit the keystone herbivore model, which predicts that consumer control should be strongest at high levels of resource availability and decline with declining productivity. They reviewed the results of experiments that manipulated the presence or absence of large grazers. Because most temperate grasslands are water-limited, they sought evidence for a correlation between consumer effect and precipitation. As predicted, the proportional effect of consumers on plant biomass declined significantly with increasing precipitation. Schmitz found a similar relationship between plant productivity and the effect of insect herbivores on plant biomass. Further, Chase and colleagues found a turnover in species composition among plants along the precipitation gradient, as predicted by the keystone model.

## B. Plants as Consumers— Carnivorous Plants

The most ubiquitous interaction between plants and animals is the use by animals of plants as sources of material resources and energy. However, there are a few plants that turn the tables. In a world of plant-eating animals, carnivorous plants eat animals. To be considered carnivorous, a plant must have some mechanism to attract, capture, and/or digest prey and must be able to absorb nutrients from those prey (Givnish, 1989). Over 500 species in nine plant families have evolved the carnivorous habit.

### 1. Mechanisms of Prey Capture

Carnivorous plants capture prey in several remarkable ways (Table II). The evolutionarily independent origin of carnivory is demonstrated by the many ontogenetic origins of the traps. A pitfall trap is a tubular structure, often containing liquid, which prey can enter but have difficulty leaving. Although five different plant families capture prey in some kind of pitfall trap, called tanks or pitchers, these traps may be comprised of leaf rosettes (e.g., *Brocchinia*), modified leaves (e.g., *Sarracenia*), or modified leaf tips (e.g., *Nepenthes*). Some plants have active mechanisms to trap prey. For example, the leaves of the Venus flytrap (*Dionea muscipula*) function like a

TABLE II  
Presence (+) or Absence (–) of Adaptations for Active Prey  
Attraction, Capture, and Digestion in Carnivorous Plant Genera  
and Species<sup>a</sup>

Genera (no. of species)	Attraction	Digestion	Type of trap
Bromeliaceae			
<i>Catopsis</i> (1)	+	–	Pitfall
<i>Brocchinia</i> (2)	+	–	Pitfall
Eriocaulaceae			
<i>Paepalanthus</i> (1)	+	–	Pitfall
Sarraceniaceae			
<i>Heliamphora</i> (6)	Chemical and visual	–/+	Pitfall
<i>Darlingtonia</i> (1)	+	–	Pitfall
<i>Sarracenia</i> (9)	+	+/-	Pitfall
Nepenthaceae			
<i>Nepenthes</i> (82)	+	+	Pitfall
Cephalotaceae			
<i>Cephalotus</i> (1)	+	+	Pitfall
Droseraceae			
<i>Drosera</i> (90)	–	+	Active flypaper
<i>Aldrovanda</i> (1)	–	+	Steel trap
<i>Dionaea</i> (1)	+	+	Steel trap
Dioncophyllaceae			
<i>Triphyophyllum</i> (1)	–	+	Passive flypaper
<i>Drosophyllum</i> (1)	+	+	Passive flypaper
Roridulaceae			
<i>Roridula</i> (2)	+	+	Passive flypaper <sup>b</sup>
Lentibulariaceae			
<i>Pinguicula</i> (35)	–	+	Active flypaper
<i>Utricularia</i> (280)	+/-	+	Bladder trap
<i>Genlisea</i> (35)	–	+	Lobster pot
<i>Biovularia</i> (1)	–	+	Bladder trap
<i>Polypompholyx</i> (2)	–	+	Bladder trap
Byblidaceae			
<i>Byblis</i> (2)	–	+	Passive flypaper

<sup>a</sup> Modified from Givnish *et al.*, 1984 (Givnish, 1989).

<sup>b</sup> Nutrient uptake apparently assisted by exudations of the kleptoparasitic bug *Pameridea roridulae*.

miniature steel-jawed trap when tripped by the hapless prey. There are even aquatic plants (e.g., *Utricularia*) with sophisticated underwater traps that slurp up prey unlucky enough to trip them.

### 2. Costs of Carnivory

Carnivorous plants are usually restricted to sunny, moist, nutrient-poor habitats, such as bogs and fens. Their slow growth and restricted distribution suggest

that there are fitness costs associated with carnivory. For example, it has been argued that leaves morphologically specialized for prey capture have compromised photosynthetic abilities, making carnivorous plants poor competitors. This hypothesis is supported by studies of the pitcher plant, *Sarracenia alata*, which produces two kinds of leaves, a “regular” leaf and one that is modified into a pitcher. The plant responds to competition from neighboring vegetation by diverting resource allocation from pitchers to regular leaves. Small stature and slow growth also make many carnivorous plants (e.g., *Pinguicula* and *Utricularia*) vulnerable to being buried by litter fall. Because of their compromised competitive ability, carnivorous plants generally respond poorly to addition of nutrients to their habitat, being easily outcompeted by plants that thrive under richer conditions. This characteristic creates important conservation concerns in the many parts of the world where atmospheric input of anthropogenic nitrogen is significantly increasing nitrogen availability in previously nitrogen-poor bogs and fens.

Another fitness cost of carnivory was identified by Zamora, who found that *Pinguicula vallisneriifolia* tends to trap its own pollinators. He also found that reproduction in the plant is limited by pollen availability, indicating that feeding on its pollinators reduces plant fitness.

### 3. Benefits of Carnivory

As with so many plant novelties, these were noticed by the inquiring mind of Charles Darwin (1874), who demonstrated that the sticky traps of *Drosera rotundifolia* do indeed capture and digest animals. Darwin's son, Francis (1878), first demonstrated experimentally that prey capture enhances the growth and reproduction of this species. Similar studies have subsequently found that in most circumstances, growth of carnivorous plants benefits from prey capture.

In many cases, the majority of a carnivorous plant's nitrogen and phosphorus is obtained from prey. However, carnivorous plants appear to obtain only a small proportion of other necessary nutrients from prey. Comparisons of greenhouse and field studies suggest that plant growth is generally restricted by the rate of prey capture and that plants could utilize many more prey than they are able to catch. Carnivorous plants also tend to be frugal with their nutrients, practicing particularly efficient internal recycling of nitrogen and phosphorus.

Many elegant methods have been used to examine in more detail nutrient uptake from prey in carnivorous plants (Adamcec, 1997). In particular, putatively carnivorous plants can be offered insects reared on media

enriched in the stable nitrogen isotope  $^{15}\text{N}$ . If the plant tissues subsequently become  $^{15}\text{N}$  enriched, this indicates that their nitrogen supply has been supplemented by insect proteins. Using this method, Hanslin and Karlsson found that *Drosera rotundifolia* and several species of *Pinguicula* in a subarctic environment took up 29–41% of the nitrogen available in insect prey they were offered. Further, root uptake of nitrogen was stimulated by prey capture, an unanticipated additional benefit of carnivory. Experiments performed in glasshouse or laboratory environments generally reveal even greater uptake efficiencies.

Carnivorous plants show interesting developmental changes during maturation of their carnivorous organs. For example, using fluorescent dye tracers, Owen and colleagues found developmentally regulated bidirectional transport by leaf glands in the pitcher vine, *Nepenthes alata*. In mature leaves, the glands transport fluids directly from the pitcher fluid to the plant vasculature (internal plumbing system), apparently functioning in nutrient uptake. However, in immature, closed leaves, the glands secrete fluid from the vascular tissues into the pitcher, building up a supply of fluid in which to eventually trap prey. Gallie and Chang examined developmental regulation of hydrolase expression in *Sarracenia purpurea* pitchers. Hydrolase is an enzyme involved in prey digestion. Hydrolase expression commenced when the pitcher first opened upon maturity, increased for several days, and would largely cease after 2 weeks without prey. However, adding prey-derived resources such as amino acids to the pitcher fluid could induce hydrolase expression in pitchers that had ceased expression due to lack of prey.

### 4. Nonprey Guests—Iniquiline Communities

Prey-digesting guests (iniquilines) are very common in pitfall traps and provide for some carnivorous plants the sole means of benefiting from prey. In fact, the food webs of iniquilines in pitcher plant (*Sarracenia* spp.) traps have been the subject of numerous highly informative community and population ecology studies.

In one of the most arcane modes of nutrient uptake, *Roridula gorgonias* hosts a bug, *Pameridea roridulae*, that feeds on insects trapped on its sticky leaves. Although the plant has no known method of digesting its prey, stable-isotope studies indicate that it acquires  $^{15}\text{N}$  label from prey. It apparently derives nutritional benefit via exudations from the bug guest that has dined on the labeled prey. This example, however, highlights the vulnerability of insectivorous plants to kleptoparasites

(animals that steal prey). Spiders, in particular, frequently compete with carnivorous plants for prey that have been attracted by the plant's attractive structures.

### III. MUTUALISTIC INTERACTIONS

In addition to antagonistic interactions in which one party feeds on the other, plants and animals may also interact in ways that can benefit both parties. Far from being pleasant affairs, however, mutualistic associations can be highly vulnerable to cheating, which often makes them evolutionarily uneasy truces between parties. The delicate evolutionary and ecological balances that can be achieved by these organisms are truly fascinating and lead to some common evolutionary issues.

One important question regards the degree of specialization between mutualists. Specialization is useful because it allows the evolutionary development of elaborate lock-and-key mechanisms that exclude cheaters and maintain the mutualism. However, specialization is an evolutionarily vulnerable position because extinction of one partner species can spell doom for the other. Further, in mutualistic interactions that must be reconstituted each generation, which is the case for all animal-plant interactions, specialization may doom individuals that cannot find the correct partner in the environment.

#### A. Plant-Protecting Ants and Plant-Feeding Ants

As described previously, plants have evolved a variety of mechanisms that defend them against herbivores. One of the strangest defenses, though, is provided when plants are guarded by ants. Once thought to be rare and unusual, myrmecophytes (ant plants) are now recognized as widespread and ecologically important. In many cases, these relationships appear to be quite casual. Visiting ants may rob nectar from flowers, but perform some guarding services in return. However, the term myrmecophyte is generally reserved for the more specialized case in which an ant colony resides in special structures provided by the plant.

##### 1. Benefits to Ants—Costs to Plants

Tropical myrmecophytes display the most sophisticated development of this type of interaction. In the neotropical regions alone, associations between plants and myrmecophytes have been described for about 250 plant species, from 19 families, and up to 180 ant species

from 5 subfamilies. These ant plants may provide ant housing in specially modified stems, hollow thorns, or specialized leaf pouches called leaf domatia. The trees or shrubs often feed ants with amino acid- or sugar-based solutions produced by extrafloral nectaries. The most developed myrmecophytes may also provide food in the form of specialized structures composed of lipids (Beccarian bodies), proteins (Beltian bodies), glycogen (Müllerian bodies), or some combination. In the most elaborate cases, plants may provide everything—room, board, and drink—to their ants.

Food provision for ants may be quite costly to myrmecophytes and studies show that plants regulate production of food structures. For example, in Central American *Cecropia* (Moraceae) trees grown at intermediate nutrient availability, removing Müllerian bodies stimulates their production. In contrast, when Müllerian bodies accumulate, which would happen if ants were not present, plants cease production.

In many ant-plant interactions, however, there is an important third partner through which the ants obtain benefit: sap-sucking homopterans tended by ants. Ants derive benefit from these homopterans in two ways. They may keep "milk herds" of homopterans from which they obtain honeydew, or they may be in the "beef business" and eat the homopterans they tend. Rather than simply supporting ants, the plants in these situations must also support homopteran consumers. Homopterans increase plant risks as well. They are often important vectors of plant diseases. Further, colonization by ant queens is a relatively rare event; presumably colonization by both ant queen and homopterans would be even rarer.

The homopteran mode of ant benefit appears to provide few options for control by the plant. This problem is illustrated with the African myrmecophyte *Leonardoxa africana*, on which the same ant species (*Aphomyrmex afer*) may tend one or both of two different homopteran species. Gaume and colleagues found that homopteran identity influenced the costs and benefits to the plant of ant patrol. One homopteran, the pseudococcid, could support larger colonies of ants, leading to better plant defense. This homopteran was also more efficient at producing ant biomass; ants tending pseudococcids did not use other plant resources. However, when ants tended coccids, the other homopteran, they also used plant resources from the extrafloral nectaries. Thus, when ants tended coccids, the only control that plants had over homopteran feeding was indirectly through nectary production. Plants that produced fewer nectaries supported fewer ants and fewer ants could tend fewer coccids. The plant could control pseudococ-

cid colony size more directly via domatia volume. Plants that provided a smaller total volume of the swollen stems used as domatia supported fewer pseudococcids and therefore fewer ants.

## 2. Benefits to Plants—What Motivates Ants?

Ants can benefit plants in three ways. First, they may patrol the plant and discourage or repel would-be herbivores. They also prune neighboring plants, thereby reducing plant competition for their host. Finally, some ants feed their host plant (myrmecotrophy).

Ants have frequently been observed killing and removing insect herbivores, and numerous experiments demonstrate the efficacy of this defense. For example, Fonseca observed four times as many herbivores on *Tachigali myrmecophila* plants from which he removed *Pseudomyrmex concolor* ants as on plants with intact ant colonies. Further, the daily rate of herbivory was about 10 times lower when ants were present, resulting in experimental plants without ants exhibiting about twice as much cumulative herbivore damage during the 18-month experiment. Leaf longevity was also substantially higher on plants with ants. It is interesting to note that these ants do not eat the herbivores they kill. Instead, they feed exclusively on catenococcid insects they tend inside the domatium, which is the hollow rachis of the compound leaf.

Ants are also effective deterrents of mammalian herbivory. For example, the African myrmecophyte *Acacia drepanolobium*, possesses two kinds of thorns. The swollen thorns are domatia in which *Crematogaster* ants live and rear their broods. Stapley has shown that the unswollen thorns slow plant damage by browsing mammals, but that browsers may compensate by feeding longer. Ants were far more effective defenses. When a browsing mammal encountered and was stung by ants, it stopped feeding immediately and could not be induced to feed further on that tree.

A second benefit that patrolling ants may provide is in competition with neighboring plants. Ants will prune vines (lianas) and branches of neighboring trees, effectively preventing their host tree from being overgrown. The result of this vigilance is that the host tree occupies a dramatically open cylinder of space amid otherwise densely packed tree canopies. Although such pruning of neighbors clearly benefits the host tree, it also benefits the ant colony by reducing the number of directions from which it may be attacked by competing or predatory ants.

In certain circumstances, ants may harm their own host by pruning it rather than its neighbors. For exam-

ple, Stanton and colleagues discovered a situation in which *Crematogaster nigriceps* so severely prunes its host tree, *Acacia drepanolobium*, that the tree cannot flower and is sterilized. In the habitat studied, four species of ants compete strongly for hosts, and *C. nigriceps* fares poorly in the violent conflicts over nest space. Instead of pruning neighboring trees, *C. nigriceps* prunes its own tree, apparently because it cannot prune neighboring trees occupied by competitively dominant ants. Indeed, careful observation of a large number of trees occupied by *C. nigriceps* revealed that these trees were always pruned in such a way as to avoid canopy contact with adjacent trees occupied by competing ant colonies. Canopy pruning of its own tree appears to be a defensive response by a *C. nigriceps* colony to competition with dominant ants that prevent it from pruning their trees.

## 3. Feeding Plants—Myrmecotrophy

Finally, a very different group of plants receives nutrients from the ants they house. These plants are also known as ant epiphytes. Epiphytes are plants that live on, but derive no nutritional benefit from, the branches of other plants. Myrmecotrophic epiphytes provide ant domatia in hollow or inflated roots, hollow rhizomes, or folded leaves. Ants then act as “mobile roots,” gathering food items for the nest, processing them, and then depositing the resulting waste and fecal matter within the plant. The best studied of these systems is *Myrmecodia tuberosa* (Rubiaceae), an epiphytic shrub of Southeast Asia and northern Australia. This species also has elaiosome-bearing seeds, a typical feature of ant-dispersed seeds. The ants feed on these food bodies and then “plant” the seeds along the walkways they create in the canopies of the trees their plant “homes” inhabit.

## B. Plant Pollinators

An important and obvious characteristic of plants is that, with few exceptions, they are rooted to the ground and cannot move. This poses a crucial problem for sexual reproduction: Immobile mates must exchange gametes, meaning that pollen must be moved to the ovule. Pollen can move passively with fluid flow in the physical environment. Wind pollination has been successful among the gymnosperms and water pollination is found among many aquatic angiosperms. However, as many allergy sufferers know too well, most of the pollen produced by wind-pollinated plants never reaches its intended target. Many land plants avoid this inefficiency by using animals as pollen vectors. Most

temperate angiosperms and almost all tropical angiosperms are animal pollinated.

Of course, animals will not move pollen around as a favor to the plant. An important evolutionary problem for plants, then, has been to attract pollen vectors. It seems quite clear that the need to attract pollinators has been a primary driver in the evolution of flower morphology. Animal-pollinated flowers often have large, brightly colored structures that function as “advertising” for the “goodies” available to the visitor. The rewards may be nectar, other more specialized chemicals, or even the pollen itself. As with any purveyor of delectables, the flower also must contend with thieves. For example, nectar-robbing bees may drill through the flower wall and gain access to the nectar without transporting pollen. Other visitors, such as ants, may be too small to trip the elaborate pollen application mechanisms of some flowers.

The potential for cheating selects for specificity in pollinator attraction. Another advantage of specificity is that it can promote pollinator fidelity. Simply attracting a pollinator once is not sufficient. To ensure fatherhood, the plant must attract an animal that will visit other flowers of the same species. Moreover, to avoid inbreeding, the flowers must be on different individuals.

Thus, the twin needs to avoid exploitation by cheaters and ensure pollinator fidelity create strong evolutionary pressure for plant traits that promote pollinator specificity. As might be expected when sex is involved, response to this selection pressure has led many plants and animal pollinators to exhibit fascinating and baroque relationships (Darwin, [1877] 1984; Grant and Grant, 1965). For example, male euglossine bees depend upon flowers of plants in the euphorb and orchid families for fragrances that they convert to pheromones with which to attract mates. In another instance, male insects are tricked by orchids into thinking they have found a mate, when in fact all they have discovered is a cleverly shaped and scented mimic.

### 1. Role of Plant Pollinators in Plant Diversification

An important aspect of these highly specialized relationships is that they can prevent mating between individuals of otherwise very closely related populations, a phenomenon known as reproductive isolation. Evolution in a trait that promotes reproductive isolation can quickly lead to speciation. Traits that create the opportunity for rapid speciation by exploiting novel resources such as new pollinators are considered “key innovations.”

Identifying such key innovations is a sticky problem in evolutionary biology. Circularity arises because the

characteristic that best defines a group, and therefore allows it to be identified as speciose (having lots of species), is frequently also the character postulated to be the key innovation responsible for the radiation. One way out of this thicket is to identify a causal link between the putative key innovation and one of the processes that determines diversity.

#### a. Floral Nectar Spurs as a Key Innovation

One particularly persuasive example of a key innovation is floral nectar spurs. These structures are critically involved in pollinator specialization because they hold nectar deep within the flower and make it available to only a narrow range of floral visitors. Animals must either be small enough to enter the spur or have sufficiently long and narrow mouthparts to sip nectar from the spur. Observing the 11½-in.-long nectaries of a Madagascar orchid prompted Charles Darwin (1877, pp. 162–163) to predict correctly the existence of a moth with a sufficiently long proboscis to pollinate this fantastic flower. Nectar spurs have evolved independently in several distantly related families and genera of flowering plants. They may be constructed from petals, sepals, or both, and genetic studies suggest that simple genetic differences can produce quite different shapes, which might favor different pollinators. Comparative studies suggest that, as would be expected were spurs driving diversification, spurred groups have significantly more species than closely related groups without spurs (Hodges, 1997).

Detailed studies within genera also corroborate the putative link between nectar spur morphology and pollinator fidelity. Within two groups of orchids, experimental manipulation of spur length significantly decreased both pollinia removal by pollinators and fruit set. This observation demonstrates that spur morphology directly influences pollinator-mediated reproductive success. Another study examined the effect of spur morphology on pollinator attraction across a hybrid zone between two species of columbines. Hybrids varied in floral characters, including spur length and orientation, and these morphologies differentially attracted either hummingbirds (primary pollinators of *Aquilegia formosa*) or hawkmoths (primary pollinators of *A. pubescens*), and thereby promoted reproductive isolation. On the other hand, the presence of a hybrid zone indicates that floral morphology has not prevented pollination “mistakes.”

#### b. Insect Pollination and the Angiosperm Radiation

The role of plant-pollinator interactions in reproductive isolation has also led to the much grander hypothe-

sis that insect pollination was a “key innovation” leading to the co-radiation of flowering plants (angiosperms) and anthophilous insects, which are those groups most involved in pollination, including certain bees and wasps (Hymenoptera), various families of flies (Diptera), and butterflies and moths (Lepidoptera). Fossil evidence suggests that the angiosperms diversified very rapidly, and many hypotheses have been advanced to explain this phenomenon. Like most hypotheses in the historical sciences, however, these have been very difficult to test. One necessary prediction of the insect-pollination angiosperm radiation hypothesis is concurrent diversification of the angiosperms and anthophilous insects.

Considerable controversy surrounds the dates of diversification of angiosperms and anthophilous insects. As early as the Carboniferous, seed ferns had large pollen that was probably too heavy for wind transport and may have been pollinated by paleodictyopteran insects found in the same formations. The first direct evidence associating insects with plant pollen appears in the Lower Permian. However, the radiation of the insect groups that today are most strongly associated with angiosperm pollination probably occurred in the late Middle to early Upper Cretaceous, the period most commonly thought to have witnessed the radiation of the flowering plants. While not proving the codiversification hypothesis, these estimates at least do not rule it out.

## 2. Pollination Syndromes

Many plant species that share animal pollen vectors also share similar suites of floral traits, such as color, shape, symmetry, and scent, which appear to attract those kinds of animals. For example, hummingbirds fly by day, have excellent color vision, possess long beaks, and visit flowers while in flight. Correspondingly, hummingbird-pollinated flowers are day-blooming and tend to be brightly colored, often red, bilaterally symmetrical and tubular in shape, and frequently pendant. Floral traits also may correspond to the physiological needs of pollinators, as with hummingbird-pollinated flowers, which tend to produce copious sucrose-rich nectar that helps fuel the notoriously high metabolic rates of their pollinators. A number of these trait combinations, which are called pollination syndromes, are summarized in Table III (Howe and Westley, 1988).

Implicit in the concept of pollination syndromes is the assumption that floral evolution has been strongly entrained by interactions with specific classes of pollinators, leading to strong specialization. Specialization, however, may be an evolutionary dead end. A plant

highly specialized and dependent upon one or a few species of pollinator is seriously vulnerable to extinction of its pollinators. Dependence upon specialized pollinators can also lead to Allee effects, which arise when populations become too sparse to persist. For example, individuals in small, isolated populations of the annual plant farewell-to-spring (*Clarkia coccinea*) were visited by so few pollinators that they could not produce enough seeds to replace themselves. If individual reproductive rates dip below replacement levels for very long, a population can dwindle to extinction.

Figs may provide a particularly impressive example of this problem. Fig trees are ecologically important components of tropical forests because the fruits they produce support frugivorous animals that are important seed dispersers for many other forest plants. Figs do not have large showy flowers to attract pollinators. Instead, fig reproduction is exquisitely dependent upon minute wasps that pollinate small flowers held tightly inside the closed fig. The wasp's end of the bargain is met when the eggs it lays inside the fig hatch and its larvae feed on a few of the many developing seeds. Individual fig trees flower synchronously for a relatively short time but fig wasp populations cycle constantly. Maintenance of the wasp population requires that female wasps emerging from a fig find and lay eggs in figs of other trees of the same species that are flowering at times other than their host individual. Genetic evidence shows that, despite their minute size and short life span, these wasps routinely move pollen between trees 5 to 14 km apart. This interdependence argues that viable fig populations may require a higher density of trees than is generally assumed necessary for plants not involved in such a tight mutualism. The only study to test this contention, however, found that despite heavy forest fragmentation by humans, banyan tree (a type of fig) populations on the Cook Islands have not yet suffered from Allee effects.

As demonstrated in a comprehensive study of a group of Hawaiian shrubs by the laboratory of Weller and Sakai, plants can sometimes escape the grips of their pollinator addiction. Here, species derived from insect-pollinated ancestors have evolved wind pollination, apparently by passing through a transitional stage characterized by a relatively generalized pollination system.

Indeed, several researchers have questioned the assumption that most plant-pollinator relationships are highly specific (Jordano, 1987; Waser *et al.*, 1996). First, they argue that even specialization in insect pollinators is usually defined at the level of plant genus or family, not species. Further, insect species confined to

TABLE III  
Pollination Syndromes: Putative Characteristics of Flowers Associated with Particular Groups of Pollinators<sup>a</sup>

Animal	Flower					
	Opening time	Color	Odor	Shape	Symmetry	Nectar
Entomophilous						
Beetles	Day/night	Dull or white	Fruity or aminoid	Flat or bowl-shaped	Radial	Often absent
Carrion or dung flies	Day/night	Brownish or greenish	Fetid	Flat or deep; often traps	Radial	Rich in amino acids, if present
Bee flies	Day/night	Variable	Variable	Moderately deep	Radial	Hexose-rich
Bees	Day/night	Variable, but not pure red	Sweet	Flat or broad tube	Radial or bilateral	Sucrose-rich for long-tongued bees; hexose-rich for short-tongued bees
Hawkmoths	Night	White or pale green	Sweet	Deep, often with spur	Radial	Ample and sucrose-rich
Butterflies	Day/night	Variable; often pink	Sweet	Deep or with spur	Radial	Often sucrose-rich
Vertebrate pollinated						
Bats	Night	Drab, pale; often green	Musty	Flat "shaving brush" or deep tube; often on branch or trunk; hanging; abundant pollen	Radial	Ample and hexose-rich
Birds	Day	Vivid; often red	None	Tube; often hanging	Radial or bilateral	Ample and sucrose-rich

<sup>a</sup> Adapted from Howe and Westley, 1988.

one plant species in one geographical area often visit other plant species in other parts of their range (Thompson, 1994). They also argue that specialization by pollinators may be more a function of which plants are available. Short-lived insects are more likely to visit only one or a few plant species with which they temporally co-occur whereas social insects, which have long-lived colonies and thus the opportunity to overlap with many plant species, often exhibit serial specialization on a large variety of plant species. For example, the only bee species on the Galapagos Islands is colonial and has been recorded visiting flowers of at least 60 plant species in 28 families. Finally, clustering of flowers into certain categories thought to be canalized through selection by pollinators may actually reflect physiological or morphological constraints in plants. For example, Chittka has argued that clustering of flower colors into particular narrow ranges may be a function of the physical and chemical constraints imposed by plant pigments rather than constraints im-

posed by the vision systems of different pollinators. If these arguments are true, the pollinator syndrome concept may be clouding our thinking about and study of plant-pollinator interactions. Researchers may be oblivious to, or fail to record, flower visitation by the "wrong" pollinators (Waser *et al.*, 1996).

### 3. Competition for Pollinators

Pollinators are an important resource that plants may compete over. Whether plants compete for pollinators is determined by the factors that limit seed production. In many situations, plant reproduction is limited by mineral resources, in which case plants are unable to increase fruit set with increased pollinator availability. However, there are many examples of plants in which reproduction is limited by pollinator availability. Pollinator limitation has important implications for plant conservation. For example, prolific flowering by invading plant species may negatively impact native plant communities by depriving native species of pollination.

## C. Fruit and Seed Dispersers

Many plants depend upon animal dispersal of seeds and fruits (diaspores). Unlike pollen dispersal, in which insects are major players, diaspore dispersal is dominated by vertebrates. Seed dispersal is important because plants generally cannot develop successfully in the shade of their mother. Seed dispersal may also be an important mechanism by which seeds escape the predator or pathogen populations that are well adapted to exploit them, having built upon and become adapted to their parents. Finally, seed dispersal is important for plants to colonize newly opened habitat.

### 1. Types of Dispersers

#### a. Vertebrates

*i. Fish* Perhaps the most unexpected and amazing diaspore dispersers are fish. This phenomenon has been most extensively studied along the Amazon River. However, it is likely to be important in many areas with extensive seasonal flooding. In the Amazon basin, where the timing of plant reproduction corresponds with the seasonal flooding, some fruits are adapted to passive dispersal by water (hydrochorous). However, even these fruits can derive facultative benefit from fish dispersal. Plants with heavy fruits or with seeds embedded within hard shells are apparently obligately fish dispersed. Fish differ in their efficiency as dispersers. Catfish are effective dispersers whereas characins are destructive and act largely as predators of all but the smallest seeds.

*ii. Mammals* Many mammals are important seed dispersal agents. Primates and bats are the most important mammalian dispersers in tropical areas. Both types of animals can move quickly across the landscape, thereby dispersing diaspores long distances. Diaspore dispersal by bats is particularly important for forest regeneration after land abandonment in the neotropics. In temperate regions, diaspores may be dispersed by ungulates (e.g., antelopes, elephants, and zebras) and by many supposed carnivores. For example, black bears consume prodigious quantities of fruit, sometimes competing with humans for delectable berries (McCloskey, 1948).

*iii. Birds and Reptiles* Birds are arguably the most important class of fruit and seed dispersers. The earliest known examples of animal-dispersed plant propagules include the fleshy seeds of cycad progenitors, which appear to have been consumed by ancient reptiles (Howe and Westley, 1988). Many dinosaurs were cer-

tainly important fruit and seed eaters and may have functioned as dispersal agents. However, except perhaps for birds, modern reptiles are only rarely important diaspore dispersers today.

Birds can be hard on seeds. Beaks may break up the seed coat, rendering the embryo vulnerable to digestive acids and enzymes. Seeds may be ground up in the gizzard. However, the guts of frugivorous birds tend to be short and gentle (not highly muscular). In fact, many seeds require a trip through a bird's digestive tract to germinate successfully.

Although large numbers of a broad range of birds feed on fruits, few depend solely on fruits. Even waxwings, perhaps the most specialized frugivores in the temperate region, also feed on insects when they are available. Nevertheless, fruits are an important resource with which many birds produce body fat prior to migration. Moreover, in the tropics, where seasonal constraints on fruit production may be weaker, several groups of birds depend almost exclusively on fruits (e.g., quetzals, toucans, and barbets).

#### b. Invertebrates

*i. Ants* The only major insect seed dispersers are ants. Myrmecorous seeds, those adapted to ant dispersal, often possess a starch- or lipid-rich body called an elaiosome attached to a tough and smooth seed coat that is difficult for ants to crack. Seed size is also constrained by selection by ants—large ants tend to carry larger seeds than small ants.

In comparison with vertebrates, ants do not carry seeds very far. Nevertheless, ants can be important dispersers for many plants. Ants may either store seeds in the nest or remove the elaiosome and then discard seeds at the nest entrance or colony waste pile. Both locations tend to have well-aerated, nutrient-rich soil that can improve plant growth. Seeds collected by ants may also gain some protection from other seed predators through their association with active ant nests, which are generally avoided by most other animals.

An important conservation issue in many areas has been loss of native seed-collecting ants to competition from invading Argentine ants (*Iridomyrmex humilis*), which are not strongly attracted to myrmecorous seeds. Red imported fire ants (*Solenopsis invicta*) do feed on seeds, though, and their arrival in an area may have a negative impact on myrmecorous plant populations by competitively excluding other more effective seed-dispersing ants.

*ii. Other Insects* Occasionally seeds are dispersed by insects other than ants. For example, scarab beetles



TABLE IV  
 Characteristics of Fruits and Seed Dispersed by Different Animals<sup>a</sup>

Animal	Fruit			
	Color	Odor	Form	Reward
Vertebrate dispersers				
Hoarding mammals	Brown	Weak or aromatic	Indehiscent thick-walled nuts	Seed itself
Hoarding birds	Green or brown	None	Rounded seeds or nuts	Seed itself
Arboreal mammals	Yellow, white, green, or brown	Aromatic	Arillate seeds or drupes; often compound and dehiscent	Pulp protein, sugar, or starch
Bats	Pale yellow or green	Musky	Various; often hanging	Pulp lipid- or starch-rich
Terrestrial mammals	Often green or brown	None	Indehiscent nuts, pods, or capsules	Pulp lipid- or starch-rich
Highly frugivorous birds	Black, blue, red, green, or purple	None	Large drupes or arillate seeds; often dehiscent; seeds >10 mm long	Pulp lipid- or starch-rich
Partly frugivorous birds	Black, blue, red, orange, or white	None	Small- or medium-sized drupes, arillate seeds, or berries; seeds <10 mm long	Pulp often sugar- or starch-rich
Feathers or fur	Undistinguished	None	Barbs, hooks, or sticky hairs	None
Insect dispersers				
Ants	Undistinguished	None to humans	Elaiosome on seed coat; seed <3 mm long	Oil or starch elaiosome with chemical attractant

<sup>a</sup> From Howe and Westley, 1988.

may bury seeds with dung. It is also likely that grassland termites aid in dispersal of grass seeds.

## 2. Plant Adaptations to Diaspore Dispersers

### a. Dispersal Syndromes

Animal dispersers impose selection on fruit and seed characters. As with the suites of floral traits ascribed to selection by particular groups of pollinators, biologists have also described "dispersal syndromes," suites of characters that appear to be shared by propagules sharing certain groups of animal dispersers or particular modes of transportation (Table IV). For example, most dog owners can describe the common characteristics of propagules dispersed on animal fur and feathers, which include barbs, hooks, and barbed hairs that cause these annoying passengers to attach firmly to socks as well as fur. However, as with pollination syndromes, considerable controversy exists over whether these suites of traits are the result of selection by particular groups of animal species or whether they instead reflect the evolutionary constraints imposed by the morphol-

ogy and physiology of plant ancestors (Jordano, 1987).

### b. Plant Adaptations to Frugivores

The ripe fruit, of course, functions as the attractant and reward for many seed dispersers. The problem for the plant in this case is to ensure survival of at least some of the seeds. First, the plant must ensure that the fruit remains on the plant long enough for the seed to develop and be provisioned by the mother. Consequently, immature fruits share with other plant organs various mechanisms that deter animal consumption. Unripe fruits often contain toxins and palatability-inhibiting compounds (Stiles, 1989). Perhaps for some readers the most memorable example of this phenomenon will be the inadvertent bite into an unripe persimmon.

A ripe fruit is advertising for dispersal of mature seeds. During ripening, fruits often change color, from an inconspicuous green that is poorly discernable amid the foliage to a contrasting color such as red, blue, yellow, or black, which is conspicuous to visually

searching frugivores. The ripe fruit may contain various nutritious goodies that attract prospective dispersers, including sugars, minerals, water, lipids, and proteins. These resources are costly for plants to provide, and some plants produce fruits or seeds that mimic more nutritious fruits but either lack any nutritional value or are much less nutritious than their model fruits.

Occupying an attractive and nutritious fruit, seeds must possess traits that promote survival of fruit consumption and digestion. Sometimes seeds can be easily separated from the pulp and are discarded by frugivores prior to digestion (remember watermelon seed spitting contests). Other seeds are relatively indigestible, due either to hard seed coats, toxins, or simply the incompetence of the frugivore gut, and pass through unharmed.

Frugivores vary in where and how they deposit seeds. Those that discard seeds as they are eating often deposit seeds individually, but may also fail to move seeds any distance from the parent. Dispersers that regurgitate seeds may move them further and often deposit seeds individually. When seeds are defecated, however, they may be concentrated at high densities. If no secondary dispersal process spreads these seeds, they may experience extremely competitive growing conditions upon germination. Such dense concentrations of seeds may also attract seed predators or secondary seed dispersers. Ants, for example, often remove seeds from dung piles. Nevertheless, some seeds require passage through a vertebrate gut before they will germinate (Traveset, 1998).

Frugivores function as seed dispersers only if they move away from the parent plant before defecating, regurgitating, or otherwise discarding the seed. Fruits often contain mild toxins, which are thought to improve dispersal by deterring foraging frugivores from finishing all the fruits at a bonanza tree. Specialization on particular fruits is rare among frugivores, perhaps because no one species of fruit provides a complete diet. This, too, may be a mechanism that benefits plants because it encourages frugivores to move about to achieve a complete diet.

### c. Seed Predators as Dispersers

Many diaspores lack a fleshy fruit or other enticement with which to attract frugivores. These seeds may benefit from dispersal by their predators. Many rodent and bird seed predators hoard seeds for future use. To the extent that they forget or lose buried seeds, they may be extraordinarily effective dispersal agents. Many members of the crow and jay family are important seed hoarders and dispersers. For example, Clark's nutcrack-

ers carry pinyon pine seeds up to 22 km and bury seeds in small clumps (scatter-hoards). Behavioral studies indicate that these birds are highly effective at finding hoarded seeds and will even dig into a meter of snow to find known hoards. However, individual birds will commonly store more than twice as many seeds as needed, perhaps to protect themselves against theft. Theft may be common, especially among social seed predators, such as pinyon jays. Pinyon jays that observe other individuals burying seeds can also find and exploit those hoards.

Caching may have significant effects on the local ecosystem. For example, Clark's nutcracker initiates forest succession after large fires by moving in limber pine seeds from long distances. Genetic structure of tree populations may also be influenced by the seed-caching behavior of Clark's nutcracker.

Seeds dispersed by predators have interesting adaptations that promote offspring survival. Oaks, for example, are dispersed by their seed predators, which include birds and rodents. Both types of animals crack open acorns and feed on the cotyledons of the embryo. During germination, the embryo root (radical) emerges from the apical (pointy) end of the acorn. Several studies by Steele and colleagues have discovered that digestibility-reducing astringent tannins are concentrated at the apical end of the acorn. Further, all acorn consumers studied, including insects, birds, and rodents, preferentially consume the basal (cap) end of the acorn, which leaves the embryo intact and viable. Partially eaten acorns can germinate, sometimes at higher rates than found for intact acorns. Finally, Frost and colleagues have found that in the European oak *Quercus robur*, experimentally removing cotyledons from seedlings does not reduce their germination success when compared with seedlings with intact cotyledons. These authors suggest that the primary function of the cotyledons is to attract seed dispersers.

### d. Evolutionary Dead Ends?

As with other forms of specialized dependence, specialization to a particular dispersal agent can be an evolutionary dead end. Perhaps the most famous example involves the fruits of the *Calvaria major* tree on the island of Mauritius in the Indian Ocean. This tree had not been observed to recruit young seedlings for over 300 years when Temple surmised that it was lacking its essential fruit disperser, which was the dodo, a large flightless pigeon that had gone extinct in the late 1600s. Temple was able to mimic dodo digestion by feeding the *Calvaria* seeds to turkeys.

Turkeys were not available to rescue *Calvaria* on Mauritius, but Janzen has argued that another tree has survived the loss of its dispersal agent in Central America through such a substitution. Huge, rare guanacaste (*Enterolobium cyclocarpum*) trees in Costa Rica produce fruits that are readily eaten by domesticated horses and cattle. Janzen speculated that these fruits were once dispersed by the native Pleistocene horse (*Equus fraternus*), which went extinct some 10,000 years ago. He demonstrated that most seeds pass unharmed through the guts of domesticated horses and went on to argue that these animals, which were introduced by Spanish conquistadors 500 years ago, have replaced the lost dispersal agent. Thus, adaptation to one seed dispersal agent preadapted the fruit to other similar agents.

#### IV. SUMMARY

Plant-animal interactions are ubiquitous and important. A common theme throughout the study of plant-animal interactions is the enormous effects that these interactions have on plant and animal evolution. There is strong evidence that the interaction between plants and insect pollinators is the primary driver of diversity in flowering plants and the groups of insects most involved in pollination. Selection by animal consumers has driven the evolution of numerous plant defense traits. These traits form the basis of many of the uses that we make of plants today. Plant-based fibers, pharmaceuticals, and flavorings all derive from plant evolutionary responses to consumers. The plant fitness trade-offs between these defensive traits and competitive ability also play an important role in determining the composition of biotic communities.

The primary benefit that plants obtain from animals is mobility. Many, perhaps most, plants depend upon animals to transport pollen and propagules. In many cases, the interactions between plants and their animal transportation providers are highly specialized and mutually beneficial. These specialized mutualisms can be quite vulnerable to extinction of either party, which is an important issue in both plant and animal conservation.

Finally, plant photosynthesis converts solar energy into chemical energy and thereby provides the energetic basis for most of the world's life. Plants are therefore the foundation of the global ecosystem. Aside from the decomposition of plant litter through the microbial food chain, this energy flows into the global ecosystem via

animal consumers of plants. Understanding how plant-animal interactions influence this process is crucial to understanding how intact ecosystems provide the goods and services upon which human endeavor, and indeed all life, depends.

#### See Also the Following Articles

ADAPTIVE RADIATION • COEVOLUTION • DEFENSES, ECOLOGY OF • FOOD WEBS • GRAZING, EFFECTS OF • PLANT-SOIL INTERACTIONS

#### Bibliography

- Abrahamson, W. G., and Weis, A. E. (1997). *Evolutionary Ecology across Three Trophic Levels: Goldenrods, Gallmakers, and Natural Enemies*. Princeton Univ. Press, Princeton, NJ.
- Adamcewicz, L. (1997). Mineral nutrition of carnivorous plants: A review. *Bot. Rev.* **63**, 273-299.
- Darwin, C. [1877] (1984). *The Various Contrivances by Which British and Foreign Orchids Are Fertilised by Insects*. Univ. of Chicago Press, Chicago.
- Douglas, A. E. (1994). *Symbiotic Interactions*. Oxford Univ. Press, New York.
- Farnsworth, N. R., Akerele, O., Bingel, A. S., Soejarto, D. D., and Guo, Z. (1985). Medicinal plants in therapy. *Bull. World Health Org.* **63**, 965-981.
- Givnish, T. J. (1989). Ecology and evolution of carnivorous plants. In *Plant-Animal Interactions* (W. G. Abrahamson, Ed.), pp. 243-290. McGraw-Hill, New York.
- Grant, V., and Grant, K. A. (1965). *Flower Pollination in the Phlox Family*. Columbia Univ. Press, New York.
- Herrera, C. M. (1995). Plant-vertebrate seed dispersal systems in the Mediterranean: Ecological, evolutionary, and historical determinants. *Annu. Rev. Ecol. Syst.* **26**, 705-727.
- Hodges, S. A. (1997). Floral nectar spurs and diversification. *Int. J. Plant Sci.* **158**(6, Suppl.), S81-S88.
- Howe, H. F., and Westley, L. C. (1988). *Ecological Relationships of Plants and Animals*. Oxford Univ. Press, New York.
- Jordano, P. (1987). Patterns of mutualistic interactions in pollination and seed dispersal: Connectance, dependence asymmetries, and coevolution. *Am. Nat.* **129**, 657-677.
- Leibold, M. A., Chase, J. M., Shurin, J. B., and Downing, A. L. (1997). Species turnover and the regulation of trophic structure. *Annu. Rev. Ecol. Syst.* **28**, 467-494.
- McCloskey, R. (1948). *Blueberries for Sal*. Viking, New York.
- Paine, T. D., Raffa, K. F., and Harrington, T. C. (1997). Interactions among scolytid bark beetles, their associated fungi, and live host conifers. *Annu. Rev. Entomol.* **42**, 179-206.
- Rodriguez, E., and Wrangham, R. (1993). Zoopharmacognosy: The use of medicinal plants by animals. In *Phytochemical Potential of Tropical Plants*. (K. R. Downum, J. T. Romeo, and H. A. Stafford, Eds.). Plenum, New York.
- Simms, E. L. (1992). Costs of plant resistance to herbivory. In *Plant Resistance to Herbivores and Pathogens: Ecology, Evolution, and*

- Genetics* (R. S. Fritz and E. L. Simms, Eds.), pp. 392–425. Univ. of Chicago Press, Chicago.
- Stiles, E. W. (1989). Fruits, seeds, and dispersal agents. In *Plant–Animal Interactions* (W. G. Abrahamson, Ed.), pp. 87–122. McGraw-Hill, New York.
- Stowe, K. A., Marquis, R. J., Hochwender, C. G., and Simms, E. L. (2000). The evolutionary ecology of tolerance to consumer damage. *Annu. Rev. Ecol. Syst.* (in press).
- Thompson, J. N. (1994). *The Coevolutionary Process*. Univ. of Chicago Press, Chicago.
- Traveset, A. (1998). Effects of seed passage through vertebrate frugivores' guts on germination: A review. *Perspect. Plant Ecol. Evol. Syst.* **1**, 151–190.
- Waser, N. M., Chittka, L., Price, M. V., Williams, N. M., and Ollerton, J. (1996). Generalization in pollination systems, and why it matters. *Ecology* **77**, 1043–1060.





# PLANT BIODIVERSITY, OVERVIEW

Jeannette Whitton and Nishanta Rajakaruna  
*The University of British Columbia*

---

- I. Introduction
  - II. Plants and People
  - III. Plant Biodiversity Described
  - IV. Patterns of Plant Biodiversity
  - V. Threats to Plant Biodiversity
  - VI. The Need for Further Study
- 

## GLOSSARY

- alternation of generations** In plants, a reproductive cycle in which a multicellular gametophytic phase alternates with a multicellular sporophytic phase to complete the life cycle.
- antheridium** A unicellular or multicellular structure that produces male gametes (sperm).
- archegonium** A multicellular structure that produces the female gamete (egg).
- autotroph** An organism that obtains its nutrition by synthesizing organic substances from inorganic substances acquired from its environment.
- gametophyte** In an organism with an alternation of generations, the haploid, gamete-producing phase.
- sporophyte** In an organism with an alternation of generations, the diploid, spore-producing phase.
- 

**PLANTS, ALSO KNOWN AS EMBRYOPHYTES**, are members of the kingdom Plantae and include a wide

array of organisms that vary greatly in many characteristics, including growth form, reproduction, and ecology. They range in size from the smallest of mosses, a few millimeters in height, to the giant redwoods, over 117 m tall, and occupy all continents and virtually every terrestrial, many freshwater, and a few marine ecosystems. This article begins with a discussion of the various roles that plants play in the global ecosystem and in the lives of humans. This is followed by a description of the 12 phyla included in the plant kingdom, along with a brief overview of their characteristics, especially as relates to issues in plant biodiversity.

## I. INTRODUCTION

### A. Defining Plant Biodiversity

Traditionally, the study of plants (kingdom Plantae), fungi (kingdom Mycota), and algae (kingdom Protista) is included under the umbrella of botany (the branch of biology dealing with plants), though this article focuses singly on the plant kingdom, including mosses, ferns, conifers, flowering plants, and related lineages.

According to this circumscription, plants are characterized as autotrophs (including some derived heterotrophs), with complex multicellular structures. Their life cycles include an alternation of generations, with multicellular diploid and haploid phases, the sporophyte and gametophyte, respectively; although the rela-

tive duration of these phases and their degree of autonomy differ among groups of plants.

The kingdom Plantae is variously divided into formal and informal groupings. Informally, plant diversity is divided into four groups that formally include 12 living phyla, as used in Raven *et al.* (1999). These are the nonvascular plants (mosses, liverworts, and hornworts), the seedless vascular plants (ferns, horsetails, club mosses, and wisk ferns), gymnosperms (conifers, cycads, Ginkgo, and gnetophytes), and angiosperms, or flowering plants. Although the informal groupings, for the most part, reflect nonnatural assemblages (i.e., the set of phyla described do not share a unique common ancestor), life history and ecological characteristics unite their component phyla, and thus, we use these designations to structure the presentation of characteristics of plant biodiversity outlined below (see Section III).

Estimates for the number of species included in the plant kingdom hover around 270,000 to over 300,000, with the great majority of this diversity occurring within flowering plants (approximately 235,000 species), followed by about 11,000 species of ferns and 9000 species of mosses. The degree of our taxonomic knowledge of various groups is not equal, so while the estimate for the number of conifer species may be quite good, the estimate for flowering plants may require substantial revision as areas of high diversity become better understood.

## B. Plants as Units of Biodiversity, as Resources, and as Habitat

In nearly all terrestrial habitats, plants form the dominant features of the landscape. In satellite images, most of what covers the land is plants. Their dominance hints at the critical role that they play in the global ecosystem; yet as living organisms, plants also constitute units of biodiversity, and thus consideration of their biological characteristics, habitat requirements, and conservation status should be an integral component of research, discussion, and policy affecting local, regional, and global biodiversity issues. In the context of a reference on biodiversity, this may seem an unnecessary statement, yet as vertebrate species continue to be the dominant focus of much of our conservation effort, plants are often considered habitat or food resources.

The complex web of interactions that constitutes our global ecosystem critically depends on marine and terrestrial photosynthetic organisms, including the members of the plant kingdom. Photosynthetic organisms are the planet's primary producers, harvesting light

energy for their own growth, fixing carbon from the atmosphere into organic molecules, releasing oxygen to the environment, and ultimately sustaining life on earth. Globally, photosynthetic organisms fix approximately 100 billion metric tons of carbon annually and play a key role in maintaining the balance between fixed and atmospheric carbon. Human activities, most critically the use of fossil fuels, are liberating fixed carbon into the atmosphere at a rate in excess of the global carbon fixation rate, contributing to global warming. In light of this trend, it is increasingly important that the potential for carbon fixation by photosynthetic organisms be maintained at the very least. Many human activities that threaten plant biodiversity, such as deforestation and habitat destruction or deterioration, also may decrease plant biomass, further exacerbating the global warming trend. In consideration of global carbon cycles, vegetation is a central focus, yet the emphasis tends to be on biomass rather than biodiversity. While this perspective is reasonable given the role that plants play in these contexts, it is important to encompass all groups of taxa within the framework of biodiversity policy and practice. Therefore in the presentation that follows, the biological properties of each group of plants are presented, with emphasis on those features that may impact issues relevant to discussions of biodiversity.

## II. PLANTS AND PEOPLE

As members of the global ecosystem, human life ultimately depends on plant life. In addition, we have dependencies on plants as sources of food, medicine, clothing, shelter, and fuel that predate the origin of the human species.

The history of human civilization is inextricably linked to the history of agriculture, including the development of crops and livestock as food sources. It is estimated that only about 3000 of 235,000 flowering plant species have been used by people as food, and only 150 or so of these have been cultivated to any extent. Today, only 6 crops (wheat, rice, corn, potatoes, sweet potatoes, and manioc) directly or indirectly provide over 80% of the human population's calories. The world's more highly industrialized regions depend on high energy input agriculture, in which crops are largely grown as genetically uniform monocultures. Such cultivation practices render these pure stands susceptible to pest and pathogen attack, and it is estimated that one-third of all crops are destroyed by pests.

Most of the global human population, however, lives in tropical and subtropical regions where relatively low

input agriculture is most important. Further improvements in agricultural yields in these regions will have the greatest impact on global human welfare, where improving yields, nutritional quality, and disease resistance could greatly increase overall productivity.

Sources of wild germplasm remain of critical importance for future crop development and improvements required to feed the growing human population. Conserving the wild relatives of our crop plants has become a primary concern of plant biologists the world over. Most plants are capable of interbreeding to some degree with wild relatives, and thus, traditional breeding can be used to introduce beneficial traits into cultivars. With the advent of genetic engineering tools, the ability to interbreed is no longer required to introduce traits into cultivars, thus removing limitations on species that can potentially be used as sources of genetic material. In addition, given the small fraction of plants that have been exploited as food sources by humans, the totality of plant biodiversity can be viewed as a potential source of new crops.

In addition to the major food crops described above, plants also make their way into our diets and our cultures as sources of beverages such as coffee and tea and in the form of spices, condiments, and oils.

Another use of plants that has and will likely continue to affect human lives is as sources of medicinal compounds. It is estimated that between 35,000 and 70,000 different species have been used as medicines by various peoples of the world and 80% of the global population derives almost all of their medicines exclusively from plants. Plants are extraordinary chemical factories: they produce a wide array of alkaloids, glycosides, and saponins that are the principal source of drugs used to prevent and cure illness in both traditional and Western medicines. Modern Western pharmacopoeia have derived approximately 7000 different medical compounds from botanical sources, and 120 or so of these are presently widely prescribed. Included among these are antimicrobial agents such as quinine, used in the treatment of malaria, and analgesics, including the alkaloids morphine and codeine, both derived from the opium poppy, *Papaver somniferum*. Glycosides produced by species of *Digitalis* have enormous value for the treatment of heart ailments, both in improving the action of a failing heart and in reducing a dangerously fast heartbeat. Plants are also sources of anticancer agents. Among these is the well-known rosy periwinkle (*Catharanthus roseus*) of Madagascar, which is the source of alkaloids used in the treatment of both Hodgkin's disease and acute lymphocytic leukemia. More recently, taxol, a drug made from extracts of the Pacific

yew (*Taxus brevifolius*), has been introduced as a promising treatment for ovarian and breast cancer. These are just a few of the thousands of species that have been or are presently used to cure human diseases. There is no doubt that more species have immense untapped potential. Since tests for determining pharmaceutical potential are both time-consuming and expensive, and synthetic drugs continue to be unaffordable to a large proportion of the global population, it is likely that the dependence on local plant-based cures will continue to rise and spread throughout the world.

The qualities of flexibility, durability, and strength, along with ready availability, make plant material such as wood an ideal source of building material. Wood and related products, including paper, are of enormous commercial importance. In industrialized nations these commodities account for a significant proportion of the total national consumption of goods. Plant species such as pine, spruce, hemlock, fir, redwoods, birch, beech, and oak are used as building materials in the temperate regions while mahogany, species of the family Dipterocarpaceae, bamboo, and rattan are used in the tropics.

A wide variety of other materials are also derived from plants, including the fibers used to weave cotton, linen, and jute, as well as latexes, resins, perfumes, and dyes. Finally, plants are also used by humans for their aesthetic qualities. Ornamental plants in all their color and shape add beauty to our surroundings and are used in both public and private gardens, along roadsides in our cities, and as houseplants. In an increasingly urban environment, ornamental plants often constitute a rare link with the natural world.

It is critical that while we exploit plants to satisfy our needs we continue to conserve the natural habitats from which they came, as such areas are undoubtedly the homes for a wealth of other species with the potential to serve needs of a growing human population.

### III. PLANT BIODIVERSITY DESCRIBED

#### A. Nonvascular Plants

Nonvascular plants (often referred to collectively as the bryophytes) include three phyla: the mosses (Bryophyta), approximately 9000 species; liverworts (Hepatophyta), approximately 8000 species; and hornworts (Anthocerophyta), approximately 330 species (Table I). These three groups are characterized by their small stature, by the absence of specialized conducting tissues (i.e., xylem and phloem), which occur in other groups of plants, and by their life cycle. The conspicuous, dom-



inant phase of the life cycle of these plants is the gametophyte (haploid) generation. Gametes are formed in specialized structures on the gametophyte called antheridia (containing male gametes) and archegonia (containing female gametes). Antheridia and archegonia may occur on the same gametophyte or on separate male and female gametophytes. Male gametes in the nonvascular plants have motile sperm and are dependent on water for dispersal to the female gametes, which are nonmotile eggs, retained in the archegonia on the gametophyte. Dispersal of gametes is thought to be extremely localized (less than 1 m). After fertilization, the diploid phase of the life cycle, the sporophyte, grows out of the archegonium, where it remains attached throughout its development, dependent on the gametophyte for most or all of its nutrition throughout its existence. Meiosis occurs in the sporangia, specialized structures of the sporophyte, resulting in the formation of haploid spores that are dispersed primarily by air currents. Spores are extremely small, are tolerant of environmental extremes, and thus have the capacity to disperse over much greater distances than gametes. Many nonvascular plants also have the capacity to reproduce asexually as gametophytes, either by fragmentation or via specialized propagules called gemmae.

Ecologically, the three phyla of nonvascular plants share their dependency on water for sperm dispersal, and thus bryophytes tend to occur where water is at least seasonally available and/or tends to accumulate. Among the richest habitats for bryophyte diversity are temperate and tropical cool, moist forests and arctic and alpine habitats. Nonvascular plants grow on a variety of substrates, including soil and bare rock, and on other plants as epiphytes. On bare rock, they play an important role as initial colonizers. Peatlands are dominated by mosses in the genus *Sphagnum*. These unique and important ecosystems cover about 1% of the Earth's surface and represent a significant global carbon sink, containing over 100 billion metric tons of unavailable carbon.

More than 60% of bryophyte families have worldwide distributions, occurring on all continents. Bryophyte species also tend to have wide geographic distributions: species that occur on more than one continent are common. Two processes have contributed to establishing present-day distributions. First, much of the diversity that we see today may be ancient, and present-day distributions may be the result of vicariance, or the splitting of a larger, once continuous range into smaller isolates. Second, especially in taxa that have originated more recently, present distributions are likely the result of long-distance spore dispersal. The role of ongoing

dispersal in bryophytes is relevant to biodiversity research, because it has broad implications for structuring of genetic variation in these species. As suitable habitats become fragmented or patchily distributed as a result, for example, of clear-cut logging on small and large scales, the potential for recolonization by extirpated bryophytes depends critically on dispersal and establishment, which are poorly characterized in these groups. Although we know that spores are easily carried by air currents, only a handful of studies have examined patterns of gene flow in bryophytes. Another consideration in bryophytes that may be relevant to biodiversity is the relationship of taxonomic species described primarily on morphological and ecological criteria to biological species—units that are united by gene flow. The reduced morphology and absence of variation in traits related to gene flow are reasons to suspect greater decoupling of taxonomic species and biological species among bryophytes. As a consequence of our lack of understanding of evolutionary processes in these groups, a particular taxonomic species that may be considered common could comprise several cryptic, genetically isolated units, some of which may be critically endangered. Compounding this problem is the relative rarity of expertise in identification of bryophytes. While both trained professionals and amateur botanists contribute to our knowledge of flowering plant biodiversity, in many areas, no local experts in bryophyte identification, professional or amateur, are found. Although the bryoflora of a number of regions is well characterized, the bryophytes are often far less thoroughly studied than vascular plants in the same region. This shortage of expertise is reflected in the status of conservation initiatives in the bryophytes: to date, no comprehensive global listing of threatened bryophytes exists, although the International Union for the Conservation of Nature and Natural Resources (IUCN) has produced a listing for vascular plants. At the present time, the Red List for bryophytes consists of a list of 92 exemplary globally threatened bryophytes.

## B. Seedless Vascular Plants

Seedless vascular plants include four phyla with living representatives: phylum Psilophyta (including only two genera, *Psilotum* and *Tmesipteris*), phylum Lycophta (about 1000 species, including the club mosses and the genera *Selaginella* and *Isoetes*), phylum Sphenophyta (the horsetails in the genus *Equisetum*, represented by about 15 species), and the phylum Pterophyta (the ferns, represented by about 11,000 species; Table I). All of these groups share a number of characteristics.

They are the earliest set of lineages to have evolved vascular tissue (xylem and phloem). These tissues provide structural stability that allows vascular plants (seedless vascular plants, gymnosperms, and angiosperms) to grow much larger than the bryophytes. The seedless vascular plants also lack seed formation, which distinguishes them from gymnosperms and angiosperms considered below.

Like the nonvascular plants, the seedless vascular plants have a free-living gametophyte generation that forms antheridia and archegonia, in which gametes are produced. Gametophytes in some groups are subterranean, nonphotosynthetic, and associated with mycorrhizae (fungal symbionts that occupy the roots). The gametophytes of most seedless vascular plants are inconspicuous and ephemeral relative to the sporophytes, although in some lycophytes, it may take 6–15 years for gametophytes to become fertile. Male gametes are flagellate sperm, and therefore, as with the nonvascular plants, they are dependent on water for transport to the stationary egg. After fertilization, the diploid sporophyte grows out of the archegonium on the gametophyte, but unlike the nonvascular plants, in these groups the sporophyte grows much larger than the gametophyte and eventually becomes the free-living, dominant form of the life cycle. Meiosis occurs in sporangia on the sporophyte, producing large numbers of wind-dispersed spores that settle and germinate to form the next generation of gametophytes.

The seedless vascular plants display a diversity of ecological tolerances. Even though the motile sperm remain dependent on moisture for transport to the egg, the dominance and perennial habit of the sporophytes appear to have somewhat lessened the strength of selection for regular water availability that appears to limit the range of habitats suitable for bryophytes. Living representatives of the LycopHYTA are all herbaceous and mostly tropical. The two living genera of Psilophyta are tropical and subtropical. Approximately 75% of fern species are tropical, and about a third of these are epiphytes. A few species of ferns are known only as gametophytes, while others do not appear to form sporophytes near their range limits. It may be that when ecological conditions are no longer favorable for some ferns, sexual reproduction is not possible. This suggests that rates of sporophyte production could act as an indication of habitat quality in some species.

Among the phyla of seedless vascular plants, a number of families contain globally threatened species, according to the 1997 IUCN Red List for vascular plants. Within the lycophytes, 39 of 79 species (49.7%) of the quillwort family, Isoetaceae, are globally threatened.

TABLE I  
The Phyla of Plants

Phylum	Estimated number of species	Percentage of globally threatened species
AnthoceroPHYTA	330	Not available
Hepaticophyta	8,000	Not available
Bryophyta	9,000	Not available
Psilophyta	7	30.8
LycopHYTA	1,000	6.5
Shenophyta	15	Not available
Pterophyta	11,000	7.5
Cycadophyta	140	82.8
Gnetophyta	70	2.5
Ginkgophyta	1	100
Coniferophyta	550	55.8
Anthophyta	235,000	13.9

These small plants mostly occur in aquatic or wet terrestrial habitats in tropical and subtropical regions. They represent the nearest living relatives of ancient tree lycophytes that dominated the coal-forming swamps of the Carboniferous.

Among fern families, the Cyatheaceae, Ophioglossaceae, Loxomataceae, and Parkeriaceae all have in excess of 25% of their species listed as globally threatened. The Cyatheaceae is the family that includes the tree fern genus *Cyathea*. Tree ferns are conspicuous elements of tropical and subtropical montane forests. The Parkeriaceae includes only four species in the genus *Ceratopteris*, an aquatic genus of mostly tropical, edible floating ferns, some of which are cultivated for human consumption. The Ophioglossaceae includes three genera, among them the genus *Botrychium*. Members of the genus *Botrychium* have unusual population structures that remain poorly characterized. Species of *Botrychium*, most of which are of conservation concern, occur as rare, widely scattered sporophytes. These ferns are perennials and may not produce aboveground structures every year. Gametophytes, which are subterranean in this genus, are rarely found, and thus the dynamics of sexual reproduction have not been explored.

### C. Gymnosperms

The gymnosperms are conspicuous and important components of many terrestrial ecosystems. They include four phyla with living representatives, including the conifers (phylum Coniferophyta; about 550 species), cycads (phylum Cycadophyta; about 140 species in 11 genera), the maidenhair tree, *Ginkgo biloba* (the only

species in the phylum Ginkgophyta), and the gnetophytes (phylum Gnetophyta, including three genera, *Gnetum*, *Welwitschia*, and *Ephedra*, totaling about 70 species; Table I). Together, the gymnosperms and angiosperms constitute a uniquely derived group, the seed plants. The four phyla of gymnosperms (as well as the angiosperms) all form seeds, defined as mature ovules that contain embryos. As such, they represent a further shift in the importance of the sporophyte relative to the gametophyte generation. In all seed plants, the gametophyte is enclosed within sporophytic tissue and, thus, is no longer free-living at any time during its existence. Male gametophytes are packaged and dispersed as pollen grains, while female gametophytes are retained on sporophytes in ovaries. Pollen may be transferred to the vicinity of the ovules by wind or insects (e.g., some gnetophytes and cycads are likely beetle pollinated). Pollen tubes are formed, through which male nuclei are transported to the egg cells. After fertilization, the embryo develops inside the megasporangium wall, resulting in a seed that may be dispersed by wind or animal vectors. For example, the seeds of pinyon pine are dispersed by nutcrackers.

Although the cycads and *Ginkgo* have flagellate sperm, these are released from pollen grains after they reach the vicinity of the ovule, and thus, the seed plants no longer have the requirement for water for gamete movement. This opens up yet another set of habitats for these plants. Among gymnosperm lineages, ecological requirements are highly varied. The Coniferophyta, the most species-rich phylum of gymnosperms, are also the most diverse ecologically. Conifers play an especially important role in temperate and boreal forest ecosystems, where they are often the dominant tree species. In these regions, conifer species are also of significant economic importance and are managed for timber and paper pulp production. Overexploitation of this resource has impacted these ecosystems in numerous ways, impacting not only narrow endemics such as the burrowing owl of the Pacific northwestern United States but also other resources such as native fisheries habitats, including the spawning grounds of Pacific salmon.

The 1997 IUCN Red List for vascular plants indicates that all four phyla of gymnosperms include families with globally threatened species. The conifers include seven families with more than 49% globally threatened species. The single species in the Ginkgophyta, the maidenhair tree, *Ginkgo biloba*, is a relictual Asian species that was preserved in temple grounds in China and Japan and is not known to occur in the wild. A very large percentage of cycads (82.8%) are globally threatened (Table I). The cycads occur in a variety of tropical and

subtropical habitats. These relictual plants are cultivated as ornamentals and are threatened in part due to over-collecting.

## D. Angiosperms

With somewhere over 235,000 extant species, flowering plants represent over 80% of plant biodiversity. They essentially share the life cycle common to all seed plants, with the innovation of having reproductive structures contained in flowers and with further reduction in the gametophytic phase of the life cycle. The flowering plants have evolved a number of features that have been described as important in their radiation, including relationships with pollinators, which opened the way for diversification of floral structures that likely contributed to reproductive isolation among diverging lineages.

Flowering plants display a wide array of variation in ecology, life history, growth form, breeding system, and other traits that make them impossible to characterize as a group. Ecologically, flowering plants are more diverse than all other groups of plants described. They occur in virtually all terrestrial habitats, and in many of these, they are the dominant elements in the landscape. A number of flowering plant species have become adapted to aquatic habitats, while others have evolved tolerance to arctic, extreme desert, and even marine conditions.

The 1997 IUCN Red List for vascular plants lists nearly 300 families that include globally threatened species. Some generalizations about broad-scale patterns of diversity in flowering plants can be made, though regional, local, and taxon-specific patterns are also important in discussions of biodiversity. Because discussions of patterns of diversity most often consider all vascular plants (including seedless vascular plants and seed plants), these patterns will be considered in subsequent sections.

A number of features of flowering plants impact the way in which genetic variation is apportioned in nature, and thus these features impact not only the susceptibility of plant species to becoming threatened but also their potential for recovery. For example, the breeding systems of plants range from highly self-fertilizing to obligately outcrossing. Among outcrossing species, a variety of mechanisms exist to minimize the potential for self-fertilization. Chief among these are genetic self-incompatibility mechanisms, which are estimated to occur in as many as 50% of angiosperms. These mechanisms operate via the action of alleles at a single locus that must differ in paternal and maternal parents in order for matings to be compatible. Genetic self-incom-

patibility can have major effects on species conservation and recovery plans, because the number of self-incompatibility alleles that occur in a population can become depleted to the point of severely limiting the potential for successful matings. Highly selfing species tend to display much higher levels of differentiation among populations than many outcrossing species and thus may show a greater tendency for adaptation to locally varying conditions. While inbreeding is generally viewed as having a negative impact on threatened species, plants that habitually inbreed are well adapted to do so. Thus it is important to understand breeding systems in order to make well-informed conservation and management decisions.

#### IV. PATTERNS OF PLANT BIODIVERSITY

##### A. Measures of Biodiversity

Various measures are used to determine plant diversity. A simple measure of diversity is species richness. This measure refers to the number of different species in a given region. Although there are other more robust measures of diversity, plant diversity studies are often based on the measure of species richness. The most species-rich regions of the world are in the tropics and subtropics. Of the 270,000 or so currently known vascular plant species, approximately two-thirds are restricted to these regions.

Species endemism is often used in discussions of species richness. Species can be categorized as broad or narrow endemics. Broad endemics are restricted to a specific region (which may be rather large) but are seldom rare. Narrow endemics, species that best fit the colloquial notion of rarity, are species that are highly localized and are found in small numbers in only a few localities. Measures of species richness and endemism are important criteria for estimating plant diversity as well as prioritizing conservation of a given region.

The extent of species endemism is also a useful measure in describing biodiversity hot spots (Myers, 1990). These are regions where exceptional concentrations of endemic species are facing exceptional threats of destruction of habitat. To qualify as a hot spot, a region must contain at least 0.5% or 1500 of the world's 270,000 vascular plant species as endemics. Secondly, a hot spot should have lost 70% or more of its primary vegetation, described as the form of habitat that usually contains the most species, especially endemics.

##### B. Patterns of Endemism

As many as 133,149 species (44% of all species of vascular plants) are confined to 25 hot spots comprising only 1.4% of Earth's land surface. Fifteen of these hot spots contain at least 2500 endemic plant species. The remaining 10 contain over 5000 endemic species (Myers *et al.*, 2000). Note that nonvascular plants are not included in these determinations because of the lack of the understanding we have of their taxonomy and global patterns of distribution. It is believed that even greater numbers of nonvascular species may be confined to these hot spots.

The 25 hot spots can be grouped into several habitat types. Predominant are tropical forests, appearing in 15 hot spots, and Mediterranean-type zones, in five. Nine comprise islands and almost all tropical islands fall into one or another hot spot. The five of the "hottest hot spots" are predominantly tropical—the tropical Andes, Sunderland (Indonesia, Malaysia, and Brunei), Madagascar, Brazil's Atlantic forest, and the Caribbean. These five hot spots along with the Philippines, Mesoamerica, the Mediterranean Basin, and Indo-Burma contain endemics amounting to 30% of global totals for vascular plant species and represent a mere 0.7% of the Earth's land surface.

The species-level analyses are complemented by an assessment of endemism among higher taxa such as plant families and genera. Madagascar and the surrounding islands possess 11 endemic families and 310 endemic genera, Cape Floristic Province has 6 endemic families and 198 endemic genera, and New Caledonia has 5 endemic families and 112 endemic genera. Such unique patterns of species richness and endemism make these areas a high priority for conservation.

#### V. THREATS TO PLANT BIODIVERSITY

##### A. Factors Affecting Susceptibility of Plant Populations to Threats

Human-induced disturbances relating to agriculture and urbanization, in addition to natural catastrophes such as droughts, floods, fires, disease, and hurricanes, have negative effects on plant diversity. Among the factors that affect the degree of susceptibility of a species to such threats are local population size, geographic range, habitat specificity, and the extent of dependence on other organisms.

Population size can be an important determinant of the fate of a species, especially when considered in

combination with the extent of a species geographic range. A species with a narrow geographic range and small population sizes will have an increased chance of all of its populations experiencing the same threat, and each small population will be at greater risk of extinction than in a species with a broader range and larger populations.

Specific biotic interactions can also be important for species persistence. Plants do not exist in isolation: in fact their survival often depends on other organisms. Many species of plants have evolved intimate relationships with a diverse array of organisms, without which they are unable to complete their life cycles. Symbiotic or mutualistic relationships with fungi, bacteria, insects, and birds are obligatory for the successful germination, growth, reproduction, and dispersal of propagules in many species of plants. For example, flowering plants that depend on one or a few species of animals for pollination and seed dispersal may become threatened with extinction in the face of a rapid decline in the population of these animals. Reduced seed set is the most direct effect of pollinator decline, yet indirect effects may also contribute to a decline in the plant population. For example, in the absence of pollinators, a higher percentage of seed may be set through self-pollination, decreasing heterozygosity and increasing the expression of deleterious traits often associated with a shift to inbreeding. Ultimately, loss of pollinators or disruption of pollination systems can cause plant extinction. A number of factors, including habitat alteration or fragmentation, grazing, introduction of alien pollinators, and the use of pesticides, can have drastic effects on pollinator populations. The plants most at risk from loss of pollinators are dioecious (species where male and female flowers occur on separate plants) or self-incompatible, those that have a single pollinator, and those that propagate only by seed. Once again, a plant species with a restricted range can be especially vulnerable to population fluctuations among its pollinators and seed dispersers.

## B. Anthropogenic Threats to Plant Biodiversity

Anthropogenic phenomena pose the greatest threat to plant biodiversity. These threats include pressures such as habitat fragmentation resulting from agriculture, logging, and development as well as climate change, grazing, invasions by exotic species, and overharvesting of individual species.

Agriculture has had exhaustive impacts on the planet. The search for a food supply has done more

to decrease plant diversity and to physically alter our surroundings than any other human activity. Habitat fragmentation, often a result of agriculture, is a serious threat to plant diversity and is the primary cause of the present extinction crisis. Livestock grazing can also have drastic impacts on plant communities. The selective nature of grazing, combined with the limited tolerance of some plant species to grazing, can result in substantial shifts in species composition.

One of the most visible aspects of habitat degradation in natural landscapes is the spread of exotic, occasionally native, invasive vegetation that displaces native communities. Disturbances such as those described above increase opportunities for these invasions. The introduction and often widespread dissemination of alien species, such as Norway maple (*Acer platanoides*), kudzu (*Pueraria lobata*), purple loosestrife (*Lythrum salicaria*), and Japanese honeysuckle (*Lonicera japonica*), planted in environments where there are no natural controls or defenses, have been devastating.

Another rising threat to present-day plant diversity is global warming. The predicted warming of the atmosphere is a result of the increase in the levels of carbon compounds, especially carbon dioxide (CO<sub>2</sub>), produced as a result of many anthropogenic activities. A rise of 1–2% of atmospheric CO<sub>2</sub> can double the present atmospheric levels (350 ppm) in about 50 years, with the resulting temperature increasing by 1–5°C. This increase in temperature can eliminate species confined to mountain peaks and northern temperate regions while some species could presumably move north, changing global patterns of plant diversity. Global warming can also induce other natural changes—changes in precipitation, evaporation rates, sea level, atmospheric chemistry (rising CO<sub>2</sub>), UV penetration of the atmosphere, and soil and water chemistry—which, singly or in combination, may differentially affect plant species and their communities, leading to drastic changes in patterns of global plant diversity.

The current wave of extinctions resulting from systematic pressures imposed by human activities is eliminating 27,000 species each year, making this the sixth greatest mass extinction in the history of our planet. The principal factor determining the rate at which species are becoming endangered and extinct is habitat destruction and deterioration, especially in the species-rich tropics. Of the 270,000 or so currently known vascular plant species, 170,000 are tropical or subtropical endemics. The rain forests, which harbor this amazing diversity of species, are subject to destruction by exploitation and land clearing on an unprecedented scale. It is suggested that 60,000 tropical plant species

will be at risk of extinction within the next 50 years. On a similar scale, of the 80,000 species of the temperate zone, about 8000 are threatened and several hundred endangered. At least 217 species of vascular plants and undoubtedly many species of nonvascular plants have gone extinct in North America over the past 500 years. Many extinctions have gone unnoticed because the species were not known to science. In the United States, old-growth forests of the Pacific Northwest, longleaf pine forests of the southeastern coastal plains, rangelands, grasslands and savannas, and wetlands—33% of the species listed under the Endangered Species Act are dependent on wetlands—are all considered endangered or threatened ecosystems due to one or more reasons ranging from logging, to conversion to agriculture, and secondarily, to fire suppression or overgrazing and subsequent invasion by exotics.

Eleven hot spots have already lost at least 90% of their primary habitats while three have lost 95%, making conservation a high and immediate priority. Some hot spots have their endemic species concentrated in exceptionally small areas, making these endemics highly susceptible to extinction. The Eastern arc, New Caledonia, and the Philippines are especially significant in this regard. The criteria used to define hot spots exclude some of the most species-rich areas of New Guinea, Amazonia, and the Congo Basin because these regions contain rich endemic floras (New Guinea has 15,000 endemic species while the other two regions amount to 30,000); they also retain at least 75% of their primary vegetation, disqualifying them for hot spot status. Further, regions such as the Ethiopian Highlands, southeastern China, and northern Rwanda, where exceptionally rich floras face exceptional threats, are not sufficiently documented to meet the hot spot criteria. Thus, while it can be argued that hot spot analysis may be used to prioritize conservation efforts to areas where efforts will have the greatest impact, other poorly characterized regions and areas of high endemism must also be considered.

### C. Natural Threats to Plant Biodiversity

In addition to the devastating effects on diversity by anthropogenic activities, natural catastrophes can also have negative impacts on plant diversity. Natural catastrophes on a global scale are extremely rare but natural fires, floods, hurricanes, landslides, and droughts occur at various times in all parts of the world. However, regularly or sporadically occurring events such as fire and hurricanes are critical for the maintenance of some plant communities. Many plant species are not only

adapted to such disturbances but also depend on them for successful growth and reproduction. In such cases, human interference that hinders or eliminates natural disturbances is likely to pose a threat to plant diversity. Such interference can often intensify these natural disturbances, as in the case of forest fires, resulting in drastic reductions in plant diversity. Hence, a knowledge of natural disturbance regimes is important in the conservation of plant diversity. Prairies, other grasslands, and ponderosa and longleaf pine forests often depend on frequent, yet low-intensity ground fires. Without these fires, these communities gradually change into other community types that are often less diverse.

Diseases also pose threats to plant diversity. The recent rapid decline in *Chamaecyparis lawsoniana* (Port Orford/Oregon Cedar) across its range in northern California and southern Oregon has been attributed to a root rot disease caused by a fungus. Apparently, the spores of this fungus are transported mainly in the mud on tires of logging trucks. Other pests and pathogens that disperse along road systems include black stain root disease fungus and gypsy moth, all having had drastic impacts on plant diversity.

### D. Conservation of Plant Biodiversity

The consequences of loss of plant biodiversity for human welfare and global ecology are unpredictable and wholly irreversible. Hence, identification and immediate protection of sites of high conservation value (i.e., biodiversity hot spots) must be a highest priority. This underlines the vital necessity of increasing inventory and ecological and biogeographical information as a prerequisite to developing plans for conservation. Species are not distributed randomly across the planet nor do they occur in a uniform pattern. Rather, they respond to environmental gradients, including climate, topography, and substrate. They further reflect a long history of species colonizations and extinctions, plate tectonics, and other global processes. By paying attention to these processes, conservationists can identify areas of greatest importance for protection. There are several criteria used in assessing the conservation value of natural areas. Species richness (or diversity), endemism, naturalness, rarity (extent of habitat), threat of human interference, amenity value, educational value, scientific value, and representativeness are some of those criteria often considered when declaring areas as protected.

Decreases in and fragmentation of natural areas is certain to lead to substantial increases in extinction rates. This is therefore a case for some form of selective

program of *ex situ* (off-site) conservation although it is universally agreed that the most effective and efficient mechanism for the conservation of plant diversity is habitat protection or *in situ* conservation. It is also acknowledged that *ex situ* facilities can be critical in a comprehensive conservation program. *Ex situ* conservation programs supplement *in situ* conservation by providing for the long-term storage, analysis, testing, and propagation of threatened and rare species of plants and their propagules. They are particularly important for wild species whose populations are highly reduced in numbers, serving as an alternative to *in situ* conservation, as a source of material for reintroductions, and as a major repository of genetic material for future breeding programs of domesticated species. Methods of *ex situ* conservation can be classified according to the part of plant that is conserved—the whole plant, seed, tissues, or genetic material in culture. All these methods of *ex situ* preservation of live plant material require periodic regeneration and sexual reproduction of the stock. The latter requires knowledge of the breeding system and pattern of genetic variability of the species concerned.

Botanic gardens and arboreta have always played key roles in *ex situ* conservation. The most widely known function of these institutions is to assemble and maintain a diversity of plant species. They also conduct and facilitate botanical research, especially in plant taxonomy and systematics. Botanic gardens make their most important contribution to the conservation of plant diversity through education as well as their influence on public opinion.

## VI. THE NEED FOR FURTHER STUDY

Representatives of all groups of plants contribute to the health of the global ecosystem and to human welfare. As such, any loss of biodiversity has the potential to negatively impact the future of the human population. Among the top priorities for efforts in plant diversity research are training of amateur and professional experts in the study and identification of poorly understood plant groups, including nonvascular plants as well as taxa that have restricted distributions. Special

emphasis should be placed on characterizing species that are thought to have potential for medicinal use. In addition, regions of the world that are highest in diversity, especially those regions that are most critically threatened, must be the target of a significant proportion of resources available for conservation. This can best be accomplished by training of additional highly qualified personnel and increasing funding to conservation agencies, including *ex situ* conservation programs. All of these initiatives will be most effective if they are coordinated by an international body that relies heavily on recommendations made by scientists and regional experts. It is only through concerted efforts such as these that we will be able to slow the loss of plant diversity around the world. As plants play such a vital role in the health of our planet and our species, it is of the utmost importance that we value and protect this diversity for future generations.

### See Also the Following Articles

PLANT-ANIMAL INTERACTIONS • PLANT COMMUNITIES, EVOLUTION OF • PLANT CONSERVATION, OVERVIEW

### Bibliography

- Given, D. R. (1994). *Principles and Practice of Plant Conservation*. Timber Press, Portland, OR.
- Klein, R. M. (1987). *The Green World. An Introduction to Plants and People*, 2nd ed. Harper & Row, New York.
- McNeely, J. A. (1990). *Conserving the World's Biological Diversity*. IUCN, Gland, Switzerland; World Bank, Washington, D.C.
- Myers, N. (1990). The biodiversity challenge: Expanded hot-spots analysis. *Environmentalist* 10, 243–256.
- Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature* 403, 853–858.
- Noss, R. F., and Cooperrider, A. Y. (1994). *Saving Nature's Legacy: Protecting and Restoring Biodiversity*. Island Press, Washington, D.C.
- Raven, P. H., Evert, R. F., and Eichhorn, S. E. (1999). *Biology of Plants*, 6th ed. Freeman/Worth Publishers, New York.
- Simpson, B. B., and Ogorzaly, M. C. (1995). *Economic Botany: Plants in Our World*. McGraw-Hill, New York.
- Wilson, E. O. (1992). *The Diversity of Life*. The Belknap Press of Harvard University Press, Cambridge, MA.
- Wilson, E. O., and Peters, F. M. (Eds.) (1988). *Biodiversity*. National Academy Press, Washington, D.C.
- Web sites: <http://www.wcmc.org.uk>; <http://www.dha.slu.se>.



# PLANT COMMUNITIES, EVOLUTION OF

Brian J. Enquist,<sup>\*</sup> John Haskell,<sup>†</sup> Bruce H. Tiffney,<sup>\*</sup> and  
Karl J. Niklas<sup>‡</sup>

<sup>\*</sup>University of California at Santa Barbara; <sup>†</sup>University of New Mexico; and  
<sup>‡</sup>Cornell University

---

- I. Introduction
  - II. Brief Overview of the Evolution of Terrestrial Plants
  - III. Niche Diversification and Competition
  - IV. Body Size and Plant Form
  - V. Allocation and Reproductive Strategies
  - VI. Plants and Climate: Physiognomy and Physiology
  - VII. Coevolution
  - VIII. Succession
  - IX. Stability and Resilience
  - X. Diversity Equilibria
  - XI. Latitudinal Gradients
  - XII. Abundance
  - XIII. Summary
- 

## GLOSSARY

- community** A group of coexisting species characterized by taxonomic composition and/or structural type, possessing repeatable associations of dominance, diversity, and trophic interaction.
- competition** Interactions between two or more organisms for biotic or abiotic resources, resulting in decreased fitness of one or more of the participants.
- diversity** A quantitative measure of the taxonomic or physiognomic composition of coexisting organisms at one of three scales: alpha (local diversity), beta

- (diversity change between adjacent local sites), and gamma (regional diversity).
- niche** An organism's biotic and abiotic resource requirements, collectively enabling the organism's trophic and behavioral role in the community.
- succession** Changes within ecological time spans in the taxonomic or physiognomic composition of a specific community in response to biotic or abiotic perturbation.
- turnover** Changes in the taxonomic and possibly physiognomic composition of a community over evolutionary timescales in response to environmental or evolutionary change.
- 

*THE PATTERNS AND PROCESSES WITHIN PLANT COMMUNITIES* that underlie the fundamental mechanisms regulating and maintaining biological diversity are reviewed and evaluated. In particular, we focus on general processes or "rules" influencing the evolution of diversity within the framework of ecological communities. We do so by comparing ecological theory with patterns observed in the fossil record and in contemporary communities.

## I. INTRODUCTION

In the past few decades, questions about the creation, maintenance, and role of biological diversity have stim-



ulated social, political, and scientific inquiry. However, the fundamental mechanisms responsible for biodiversity remain poorly understood. We review the patterns in contemporary and historical plant communities that we believe are critical to achieving a better understanding of the processes that regulate diversity. In doing so, we take an actualist (or pure uniformitarian) stance: We assume that the basic physical and biological processes that regulate diversity today are manifestations of natural "laws" that have been in operation throughout the existence of the earth.

Clearly, any realistic juxtaposition of the concepts derived from neo- and paleoecological studies requires a clear exposition of the limitations and evidence on which they rest. Paleo- and neontological research programs each bring different strengths to bear on the question of how biological diversity originates and is maintained. The greatest strength of neocology is the ability to achieve fine resolution. Much of what is known about the ecological dynamics of plant and animal communities is based on detailed, short-term, small-scale, neontological studies. These studies emphasize the role of local processes such as succession, predation, competition, mutualism, abiotic stress, and a host of other ecological phenomena in the regulation of biodiversity. On the other hand, the paleontological research program tends to emphasize the influence of evolutionary and physical changes on the species composition of communities. Despite their different strengths, the two disciplines often reach similar or complementary conclusions. For example, both paleontology and neontology concur that community composition is highly variable over space and time (Ricklefs and Schluter, 1993; Brown, 1995). Communities do not consist of a "tight-knit" assemblage of co-occurring species; instead, species respond individually to changing biotic and abiotic conditions.

Differences between the disciplines in the level of taxonomic, spatial, and temporal resolution have important consequences on how phenomena are defined and perceived. Within neontology, the definition of coexisting species can vary from very small sets to regional assemblages. In most cases, neo- and paleoecologists use the same vocabulary (e.g., community, taxonomic diversity, and stability), but the meanings of these words change as a function of the spatial and temporal scales used in the two fields. Despite these potential pitfalls, we believe that we can improve our knowledge of the important community processes by focusing on the complementary strengths of each discipline.

## II. BRIEF OVERVIEW OF THE EVOLUTION OF TERRESTRIAL PLANTS

The fossil record clearly shows that the number of plant species has increased over geological time (Fig. 1; Niklas, 1997). The first land plants were diminutive, nonvascular, bryophyte-like organisms that inhabited the margins of the terrestrial landscape and provided resources and shelter to the first land-dwelling animals. These plants were restricted in their geographic coverage owing to their dependence on liquid water for vegetative survival and reproductive success (Bateman *et al.*, 1998).

The earliest vascular plants possessed better vegetative adaptations to life on land but were also tied to moist habitats for reproduction. They quickly diversified in morphological complexity, primarily during the Devonian, a period that saw the first plants with leaves, roots, lateral meristems, and an erect habit. These plants were larger than their nonvascular progenitors. By Late Devonian times, vascular plants evolved the capacity for secondary growth and with it the ability to form forests. These changes increased the complexity of the physical structure of plants and the diversity of resources available within terrestrial communities, creating opportunities that were subsequently exploited by the evolution of plants with vining, epiphytic, and understory lifestyles. Initial stages of the seed habit also evolved by the Late Devonian, allowing new ecological solutions within the swamp community and laying the groundwork for the subsequent establishment of communities in drier habitats in the Carboniferous. The basic physiognomies of modern plant communities were well established approximately 300 million years ago.

From the late Carboniferous through the early Mesozoic, gradual environmental changes associated with the formation of Pangea resulted in extinctions of lowland taxa and their replacement by increasingly drought-tolerant taxa migrating from more mesic sites. This established an early mid-Mesozoic flora dominated by a host of seed-bearing clades collectively known as the gymnosperms. Pteridophytes survived but were a secondary element of Mesozoic vegetation.

Angiosperms became an important ecological element approximately 115 million years ago, in the Early Cretaceous. They initially evolved as disturbance-tolerant herbaceous and shrubby plants occupying early successional environments, but they quickly radiated. By the Late Cretaceous, angiosperms started to appear as forest trees in increasingly more stable sites, and they

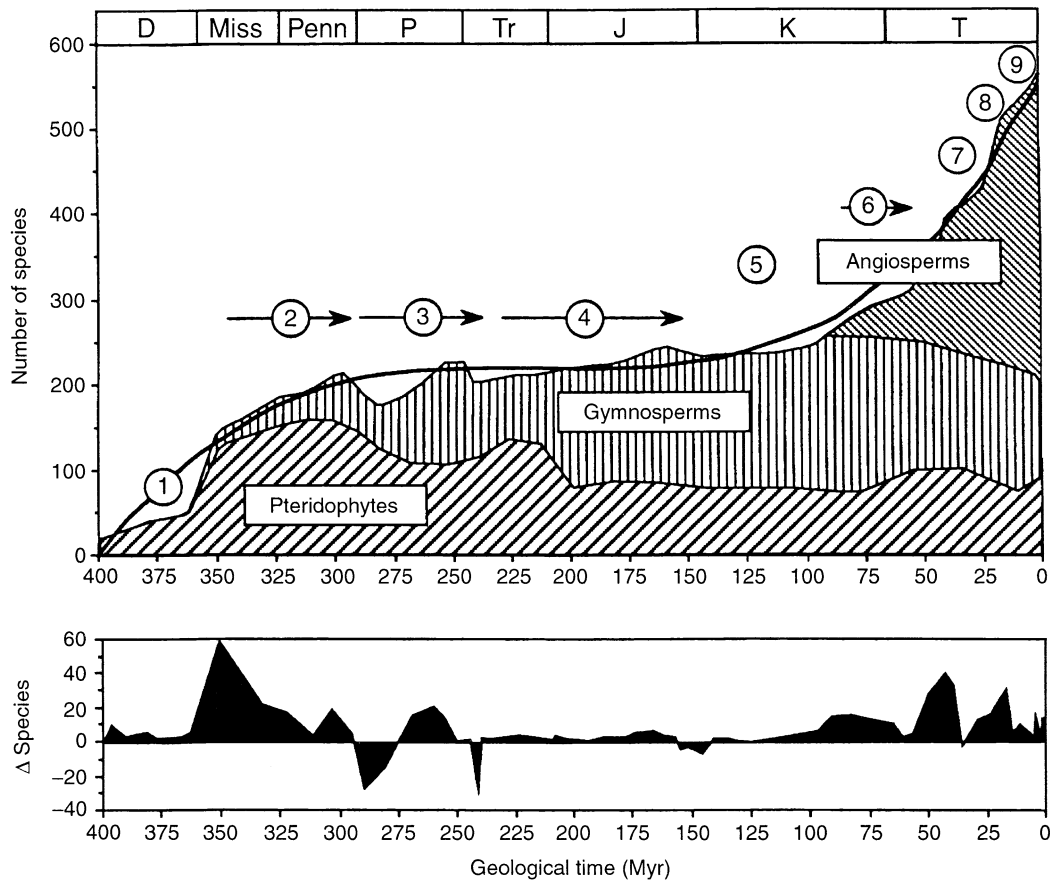


FIGURE 1 Total number of species and species turnover. Number of vascular land plants and change in species number are plotted as a function of geological time. Data for vascular land plant species are segregated into three reproductive groups (pteridophytes, gymnosperms, and angiosperms). Solid curved line denotes ordinary least squares regression curve for the cumulative data. Change in species number (turnover) is computed on the basis of the appearance and disappearance of species in each geological stage (data from Niklas 1997). 1, Origin of seed, origin of arborescent habit; 2, diverse lowland swamp communities, high atmospheric oxygen; 3, sequential decline in diversity and occurrence of pteridophyte-dominated communities, coalescence of Pangea; 4, gymnosperm-dominated communities common, Pangea slowly breaks up; 5, angiosperms first appear in the fossil record and begin to spread, becoming the dominant floristic element; 6, Cretaceous–Tertiary boundary, angiosperms diversify and become the dominant vegetational element, including closed forest communities; 7, Eocene–Oligocene climatic shift leads to dramatic Earth cooling; 8, herbaceous angiosperm communities radiate, including grasslands; 9, taiga and tundra biomes fully developed.

were the most diverse terrestrial plant clade. However, it is not clear if they formed the dominant vegetation at this time. By the early Tertiary, angiosperms were represented by many living families and genera forming closed-canopy forests in temperate and tropical climates, and they dominated Earth's vegetation. Cooling and drying climates in the later Tertiary resulted in the diversification of midcontinental grassland communities and the rearrangement of individual distribution patterns that created the biomes of the present. The rise of angiosperms established the full comple-

ment of structural and reproductive features (e.g., complex canopies, diverse life history, and reproductive strategies) observed today as well as created the most diverse terrestrial communities in Earth's history. In their radiation, angiosperms did not displace equally diverse nonangiospermous seed plant communities. Rather, angiosperms both inserted themselves into existing communities and created totally new communities, often creating new opportunities for existing clades. The result has been the most speciose plant communities yet seen on Earth (Fig. 1).

### III. NICHE DIVERSIFICATION AND COMPETITION

Understanding the evolution of plants in terms of niches provides the basis by which one can begin to assess the evolution of diversity within plant communities. The concept of the niche is useful because it provides a framework linking individual traits, such as morphology, physiology, and behavior, with their interactions with the biotic and abiotic environment. These features in turn influence population level traits such as growth, abundance, and distribution (Brown, 1995).

Both ecological theory and empirical evidence stress that competition is one of the important forces driving niche diversification and thus community diversity [Knoll (1986) as cited in Rosenzweig, 1995]. The increase in plant species diversity through time is paralleled by evolutionary shifts in resource requirements, body size, allocation strategies, and life history traits suggestive of niche diversification and thus of the action of competition. Indeed, the degree of difference or similarity that is required for coexistence has not been quantified, and it has been hypothesized that two or more plant species with very similar resource requirements can coexist indefinitely in the same community as long as their reproductive units (e.g., spores, seeds, or fruits) are dispersed randomly such that adults do not consistently grow in close proximity (Hubbell and Foster as cited in Ricklefs and Schluter, 1993; Brown, 1995; Rosenzweig, 1995; Crawley, 1997).

Competition, however, is difficult to effectively demonstrate in the fossil record. At large temporal scales, such as those encountered in the fossil record, the evidence suggests that competition may be supplanted by other features such as shifts in the physical environment [Niklas *et al.* (1985) as cited in Ricklefs and Schluter, 1993]. The shift from a Pteridophyte-dominated late Paleozoic flora to a seed plant-dominated Mesozoic flora appears to have been caused by global climate change that selected against the reproductive modes of many pteridophyte lineages, resulting in extinctions in existing lowland communities. Opportunistically, seed plants migrated in from more mesic sites to replace the lowland pteridophytes. Similarly, gymnosperm diversity was not drastically affected by the evolutionary appearance and subsequent diversification of angiosperms, despite the observation that contemporary gymnosperms are ecologically displaced by the more rapid vegetative growth and reproduction of flowering plant species [Niklas *et al.* (1985) as cited in Ricklefs and Schluter, 1993]. The evidence suggests that al-

though niche diversification is driven largely by competition, major changes in community composition at evolutionary timescales have been initiated by larger scale physical changes.

### IV. BODY SIZE AND PLANT FORM

Body size influences nearly all the characteristics of organisms and ultimately the composition of the communities they inhabit (Brown, 1995). The ecological and evolutionary ramifications of body size appear to have been an important basis for differentiation in vascular plants. Neocological observations indicate that mass-specific metabolism, age to reproductive maturity, relative growth rate, and population density are all inversely correlated with body size (Niklas, 1997; Enquist *et al.*, 1999). Thus, although the total resource needs of larger individuals increase, their use per-unit-volume or per-unit-mass decreases. Body size also affects biological timing. For example, larger species have longer life spans and longer times until reproduction than do smaller species. Increases in size allow species to partition resources over time: Small species tend to be more sensitive to short-term environmental variation, whereas large species apparently cue to longer term fluctuations.

Individual clades within pteridophytes, gymnosperms, and angiosperms often display Cope's rule: Derived taxa tend to become larger. An increase in size confers a competitive advantage in the struggle for the resources, such as light and nutrients, that are necessary for growth and reproduction [Harper (1977) as cited in Crawley, 1997]. Paleocological data indicate that Cope's rule may also reflect the tendency for new clades to make their evolutionary debut in disturbed habitats [DiMichele *et al.*, 1987; Niklas *et al.* (1985) as cited in Ricklefs and Schluter, 1993]. Such species often possess a small body size and its associated rapid vegetative growth and onset of sexual maturity (i.e., the classic r-selected species). Conversely, larger species within many plant clades appear to occupy more ecologically stable habitats.

Over long time spans there may be selection against larger sizes. Larger species tend to be more prone to extinction during periods of environmental change than their smaller counterparts [Bakker (1977), Stanley (1979), and Tiffney and Niklas (1985) as cited in Niklas, 1997; Brown, 1995]. Presumably this is because larger species take more time to reach reproductive maturity, tend to produce fewer progeny per reproductive cycle, and have lower population densities per unit

area. This suggests that evolutionary processes operating on microevolutionary and macroevolutionary timescales may have conflicting outcomes. Under the limitations of physical and biological constraints, organisms benefit from competitive advantage by becoming larger until an unpredictable environmental change forces them to extinction.

A plant's form, or shape, is as important as its size in partitioning ecological opportunities. Many aspects of plant form and architecture can be interpreted as the result of natural selection acting to optimize biomechanical stability and the ability to procure resources from the environment while minimizing hydrodynamic costs of resource transport [Niklas and Kerchner (1984) as cited in Niklas, 1997; West *et al.* (1997, 1999) as cited in Enquist *et al.*, 1999]. Only a relatively small number of "plant designs" appear to be functionally viable over the range of sizes within which vascular plants operate. It can be shown via mathematical modeling and computer simulation that most of these viable options have been exploited [Niklas and Kerchner (1984) as cited in Niklas, 1997] and that the evolution of plant size has been strongly limited by biomechanical constraints (Niklas, 1997; Bateman *et al.*, 1998). In summary, although selection has apparently operated to increase plant size within clades, physical and biomechanical constraints have restricted maximum size (Fig. 2).

## V. ALLOCATION AND REPRODUCTIVE STRATEGIES

Biomechanical limitations restrict niche partitioning involving body form. However, plants with different forms may vary the allocation or timing of energetic output, thus increasing the potential for coexistence. Separate species attain their "adult" size at different times by altering growth rates. Unless the total amount of energy available to the individual plant changes, such changes in growth rates must be accompanied by a trade-off in other activities requiring energy. For some species, changes in growth rate are compensated for by changes in wood density: Fast-growing arborescent species generally produce wood of a much lower density than do slower growing species (e.g., balsa versus mahogany; Enquist *et al.*, 1999). This has obvious implications for relative mechanical stability, susceptibility to disease, and species longevity. Similarly, an energetic trade-off exists with regard to the ability to tolerate herbivory, parasitism, and disease (Crawley, 1997). By investing energy in the formation of secondary chemicals, plants can often avoid enemies, albeit at the cost of slower growth and delayed sexual maturity.

Perhaps the area in which different allocation strategies have the strongest effect is within reproduction.

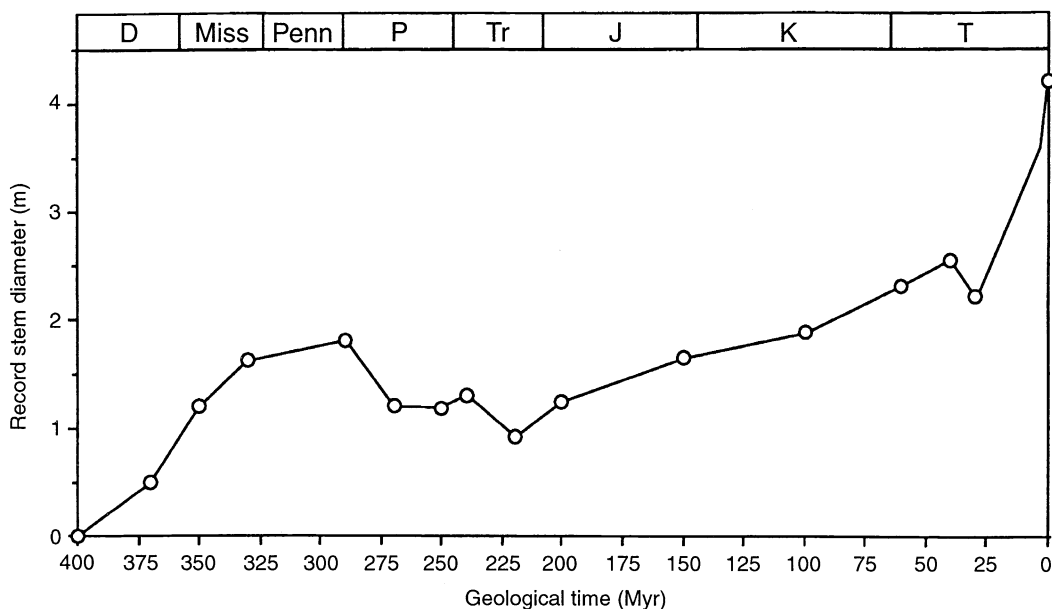


FIGURE 2 Record stem diameters. Largest reported stem diameters plotted as a function of geological time (data from Niklas, 1997).

Plants can alter both the proportion of resources used in reproduction and the amount of resources allocated to individual disseminules. Such adaptations are often linked to community structure. Large seeds, for example, tend to occur in forest communities, whereas smaller seeds are more common in open or disturbed habitats (Crawley, 1997). Variation in reproductive strategy provides plants with additional ways to partition environmental opportunities and coexist.

The fossil record suggests that reproductive adaptations have permitted the ecological exploration of previously poorly colonized terrestrial environments, starting with the invasion of land, the evolution of the seed, and finally the suite of adaptations involved in the origin of angiosperms (Niklas, 1997). Early land plants possessed pteridophytic reproduction, requiring freestanding water. With evolution of advanced forms of gymnospermy, this dependence was lost, but at the cost of vegetative reproduction, because all nonangiospermous seed-bearing plants were apparently either shrubs or trees capable of only limited vegetative reproduction.

The evolution of the seed habit, specialized dispersal mechanisms, and the capacity for seed dormancy opened up many new habitats in the late Paleozoic and Mesozoic. Together, specialized dispersal and seed dormancy permitted further niche diversification within plant communities. The morphology of dispersal units through the Late Cretaceous suggests that seed dispersal was primarily abiotic. The evolution of large seeds in many angiosperm lineages during the early Cenozoic indicates the increasing importance of animal dispersal. Additionally, the ability of seeds to remain dormant has many important ecological implications, potentially enhancing local coexistence by enabling potential competitors to partition niches in time and space. Furthermore, dormancy enables plants to "wait" until specific environmental conditions are present, such as a rare rain in the desert or a high nutrient defecation following passage through an animal's digestive tract. Dormancy holds the potential for greater dispersal distances and the formation of seed banks, enhancing the geographic ranges of species and their capacity to deal with potentially protracted inclement environmental conditions. Dormancy, however, may also increase the probability of seed predation.

It is only with the angiosperms that asexual vegetative reproduction and the seed habit are found to commonly co-occur within an individual with either an herbaceous or woody growth habit. The possession of rapid sexual life cycles, the potential for an herbaceous habit, and in many cases the potential for vegetative reproduction predisposed angiosperms to outdiversify and potentially outcompete other lineages in multiple

ways in response to environmental changes especially in the Cenozoic. These traits allowed for further invasion of new habitats, such as desert and arctic/alpine environments.

## VI. PLANTS AND CLIMATE: PHYSIOGNOMY AND PHYSIOLOGY

The patterns of plant diversity and distribution are influenced by climate. Most biomes show a remarkable degree of convergence in plant architectures, size structure, and physiology, despite drastically different biogeographic and evolutionary histories of their component taxa [Leigh (1975) and Mooney (1977) as cited in Brown, 1995; Milewski (1983) as cited in Ricklefs and Schluter, 1993]. Many of the physiognomic features associated with plants found in communities within each of these biomes, such as the relative proportion of compound and simple leaves, root to shoot ratios, plant heights, leaf cuticle thickness, and leaf indument, all show consistent trends with variation in temperature, precipitation, humidity, and rainfall [Crawley, 1997; Raunkiaer (1934) as cited in Crawley, 1997; Wolfe (1978, 1979) as cited in Ricklefs and Schluter, 1993]. Indeed, because of the tight association between vegetative physiognomy and climate, physiognomy is often used as a proxy for reconstructing past climates.

The tight link between climate and physiognomy is due mainly to physiological and stoichiometric trade-offs (Crawley, 1997). One of the most important is the trade-off between water and carbon balance. In order to assimilate atmospheric CO<sub>2</sub>, plants must be able to exchange gasses through their stomata, but open stomata cause water loss via transpiration. In many terrestrial environments water stress can limit rates of production by reducing photosynthesis. The evolutionary history of terrestrial plants is marked by the evolution of adaptations that enhance the ability to increase productivity in the face of water stress. The initial radiation of vascular plants was most likely due to the evolution of a cuticle pierced by stomata whose aperture could be regulated (Niklas, 1997). These adaptations were honed through the subsequent evolution of sunken stomata, leaf pubescence, physiologically dormancy, and the ability to mechanically position leaves in response to environmental change. Such morphological adaptations allowed for the colonization of more seasonal environments [Hutton *et al.* (1998) as cited in Bateman *et al.*, 1998].

Although it appears that changes in morphology allowed for the continued exploitation of new resources

and habitats, physiological processes appear to be highly conserved over geologic time [Raven (1995) as cited in Bateman *et al.*, 1998]. However, interpreting these features is difficult from fossils because the process of preservation alters chemical fossils. When physiological processes are mirrored in morphology, we infer physiological evolution. For example, changes in the density of leaf-borne stomata appear to correlate with geochemically predicted changes in atmospheric CO<sub>2</sub> content, at least back through the Mesozoic. Other morphological features suggest that basic plant physiology has changed little in 400 million years. C<sub>3</sub> photosynthesis appears the norm; CAM photosynthesis may have evolved by the Pennsylvanian. C<sub>4</sub> photosynthesis is, at the oldest, Late Cretaceous and probably younger, matching the radiation of angiosperms into hot and dry environments. The timing of observed increases in plant diversity does not appear to be associated with physiological changes.

## VII. COEVOLUTION

The diversification of vascular plants is associated with an increasing number of coevolutionary relationships. Coevolution can be subdivided into two extreme forms: "tight" coevolution, in which the interactions between two closely linked organisms are intimately interdependent resulting in mutual evolutionary influences, and "loose" coevolution, in which the interdependence of two or more species is facultative and transient. Coevolution includes interactions in which all parties benefit and interactions in which one party benefits at the expense of the other.

The fossil record suggests tight coevolution is rare or short lived. For example, Abert's squirrel currently feeds almost exclusively on ponderosa pine, which currently has a very large geographic range; the modern distributions of Abert's squirrel and ponderosa pine closely overlap. However, Betancourt and Van Dender (as cited in Brown, 1995) suggest that this is a very recent feature since the geographic range of ponderosa pine was reduced to a handful of relict populations in the Pleistocene that probably could not have supported a viable squirrel population. Likewise, there are many angiosperm genera whose fruits or seeds are currently dispersed by specific mammals or birds. Although many of these genera had similar seed and fruit morphologies more than 30 million years ago, none of their modern dispersal agents existed during this time. In fact, these plant lineages passed through two major turnovers of mammals during the Tertiary, indicating that dispersal coevolution must have been either loose

or capable of rapid adaptation. Additionally, recent analyses indicate that the radiation of the angiosperms had little apparent influence on the diversification of insect families, although it is possible that coevolutionary response occurred below the family level. In fact, insects began radiating more than 100 million years before the ascendancy of angiosperms [Labandeira and Sepkoski (1993) as cited in Rosenzweig, 1995]. These observations suggest that many of the tight coevolutionary couplings seen in modern communities may be relatively transient phenomena.

Nonetheless, both neo- and paleontologists have shown that loose coevolutionary relationships have been an important part of plant evolution. Insect pollination is a basal feature among angiosperms and their closest relatives. Many floral features, including bilateral symmetry, prolonged calyx tubes, nectar and resin rewards, and sympetaly, appeared by the Turonian and suggest specialization for bee pollination (Crepet, 1996). The obligate mutualisms between yuccas and yucca moths, which have been well documented by molecular data, indicate that the general relationship was likely established as long as 40 million years ago. Similarly, fossils indicate that the close association between figs and wasps is at least as old (Pellmyr, 1999). Coevolutionary relationships with pollinators and seed dispersers probably enabled the angiosperms to maintain broadly dispersed but rare or patchy populations by promoting genetic outcrossing and minimizing local inbreeding. This created an opportunity to colonize habitats that were inaccessible to older plant clades that were more reliant on wind and water for the dispersal of seeds and gametes. Further, the presence of mycorrhizal fungi in the earliest plant communities supports the hypothesis that fungi played a key role in allowing plants to spread over terrestrial habitats [Selose and LeTacon (1998) as cited in Bateman *et al.*, 1998].

As suggested in the Red Queen hypothesis, coevolution can result in an apparent "evolutionary arms race" in which both participants are evolving "as fast as they can" only to maintain their relationship relative to each other [Van Valen (1973) as cited in Brown, 1995]. Ehrlich and Raven (as cited in Ricklefs and Schluter, 1993), for example, suggested that advances in antiherbivory compounds by plants are matched by the evolution of more efficient detoxification mechanisms by herbivores. Fossil evidence for this kind of coevolution is difficult to obtain. One possible example involves seeds of *Zanthoxylon* (Rutaceae), which possess oils commonly interpreted as deterrents. Such seeds are commonly found pierced by a hole of consistent morphology from approximately 40 million years ago. In extant *Zanthoxylon* seeds, bruchid beetle larvae create

similar holes. This suggests a pest–host relationship that has persisted for approximately 40 million years. Nevertheless, it is unknown if such an arms race of evolving strategy and counterstrategy occurred during this period.

The paleoecological perspective also provides evidence for a very broad coevolution between vertebrate herbivores and plants. Vertebrate herbivory appeared in the Pennsylvanian and became well established by the Permian. During this time a terrestrial food pyramid of plants and synapsid herbivores was established, lasting to the Early to mid-Triassic time. These herbivores became extinct during the Middle Triassic and, after a brief hiatus, were replaced by dinosaurs. From the Late Triassic through to the end of the Cretaceous, dinosaurian herbivores dominated. Some of these herbivores were small (on the order of 30 kg), but the dominant ones were nearly two orders of magnitude larger. At the Cretaceous–Tertiary boundary, average herbivore size plummeted with the extinction of terrestrial dinosaurs and the radiation of birds and mammals.

Although the forces driving turnovers in vertebrates are hotly debated, it is evident that changes in the composition and structure of floras accompanied changes in the composition of the vertebrate faunas. The Permo–Triassic marked a very broad global transition from floras dominated by pteridophytes and seed ferns to floras dominated by cycadophytes and conifers. This transition resulted in less digestible forage for herbivores, which the fossil record suggests generally favors the evolution of larger herbivores. The radiation of herbaceous and shrubby angiosperms and low-feeding ornithischian dinosaurs during the later Cretaceous again suggests a very broad-scale coevolutionary process. Dense angiosperm communities, spatially well suited to exploitation by small mammalian and avian herbivores, apparently became common only after the extinction of the terrestrial dinosaurs at the Cretaceous–Tertiary boundary. Together, these observations suggest a coupling between community composition and structure and the evolutionary history of large herbivores.

## VIII. SUCCESSION

Succession is one of the most conspicuous processes in the change of community composition and structure over ecological time scales and is one of the most widely written about topics in plant ecology. Because succession involves local immigration and extinction coupled with changes in species relative abundance, it is concep-

tually tied to the notion of community stability and equilibrium.

As a neoecological process, succession occurs during timescales defined by the lifetimes of individuals in populations, which vary across species and the communities they form. It is also dependent on the frequency, duration, predictability, and magnitude of environmental change, the proximity of source areas, and the diversity of life history traits present in the species pool. These features are believed to dictate the apparent orderliness of succession and thus the abundance and diversity of species within a community at any given time (Crawley, 1997).

Succession was minimally important in the earliest land plant communities because they were patchy and composed of taxa with similar growth habits and physiological requirements. By the Late Devonian a variety of clades had evolved shrubs and trees, creating multilevel communities and ushering in light-controlled succession. Other features also affected succession, including nutrient availability, and physical disturbance. Fire-driven succession has been demonstrated for the arborescent lycopod *Sigillaria*, which occurs immediately above charcoal-rich layers in Pennsylvanian swamps.

Succession in Mesozoic gymnosperm communities is more difficult to demonstrate. Retallack and Dilcher (as cited in Niklas, 1981) note that angiosperms and cycadeoids in the Late Cretaceous Dakota Formation tend to be associated with sedimentary features indicative of disturbance, whereas conifers and cycads are generally absent from such environments. Similarly, Hickey and Doyle draw attention to the fact that angiosperms in the Early Cretaceous Potomac Formation appear in otherwise conifer-dominated environments either directly above layers of charcoal (created by fire) or in sediments indicative of disturbance by water [Niklas *et al.* (1985) as cited in Ricklefs and Schluter, 1993]. Such angiosperm-dominated localities subsequently revert to conifer-dominated communities, suggesting that some or all of the earliest flowering plant species were early successional specialists.

## IX. STABILITY AND RESILIENCE

The stability and long-term diversity of living communities are features of considerable contemporary interest and debate. Discussion of these features, however, is often confused by matters of precise definition. Although recent neoecological debate has focused on the effects of diversity on ecosystem processes, such as nutrient cycling and biomass production, some studies have examined the persistence of taxonomic composi-

tion. With reference to taxonomic stability, some ecologists maintain that the more species that are in the community, the more interactions will occur and the more resilient the community becomes to perturbation. A similar argument is made for ecosystem processes. However, a similar number of interactions could occur within a community in which continuous species turnover was occurring, changing the species composition but maintaining species richness and ecosystem processes. Thus, two different measures of the community, taxonomic composition and ecosystem properties such as species richness, could demonstrate conflicting patterns (Haskell as cited in Brown *et al.*, 2000). These features of ecosystems are complex and require careful identification of the appropriate models, definitions, and scales of investigation necessary to define stability and the methods used to measure it.

Within communities, species diversity ultimately reflects the dynamics of local colonization and extinction as influenced by the prevailing physical and climatic conditions, species life histories, and evolutionary clades present. As such, diversity (or species richness) is an emergent property of both ecosystems and evolution. It is the result of a myriad of complex interactions of abiotic and biotic factors. Species diversity per se does not dictate whether or not a community displays stability. However, increased biological diversity may confer an increased array of potential responses to varying environmental conditions. Thus, as diversity increases through increased niche diversification, the probability of including species with tolerances for new sets of environmental conditions also increases. During periods of drastic environmental change, the different capacities for productivity, mortality, reproduction, etc. may allow some community members to survive, or change their relative abundance within the community under the new environmental regime, and thus to produce a similar community under the new regime.

Community stability in the fossil record is primarily gauged in terms of taxonomic composition. However, as data are summed over increasingly long time periods, the level of taxonomic resolution generally decreases. Still, stability may be estimated by rates of extinction and origination, with high taxonomic turnover rates suggesting community instability. Using this gauge, late Paleozoic swamp communities appear to be very stable, changing slowly in the face of increasing global cooling leading to the threshold of the Permo–Triassic boundary [DiMichele and Phillips (1990) as cited in Ricklefs and Schluter, 1993]. These communities contained large numbers of pteridophytic and early seed plant species, each of which offered a different “solution” to environmental change. With the transition to the

generally continental climates of the early mid-Mesozoic, conifers, cycadophytes, and several other less diverse seed plant clades characterized by slow reproduction and growth rates replaced the Carboniferous swamp communities, but at a lower level of species diversity. However, despite their comparatively low diversities, these floras never experienced substantial extinction events, suggesting that taxonomic diversity and stability are not inevitably linked to one another. Similarly, even though the evolution of angiosperms led to arguably the most diverse communities in Earth history, there is no indication from origination/extinction rates that angiosperms form more or less stable communities than their predecessors.

Instead, it is possible that rates of compositional turnover may be more heavily influenced by rates of environmental or evolutionary change and independent of ecosystem function, taxonomic composition, or species richness. For example, climatically driven migrations observed in the late Pleistocene raise the possibility that species are constantly shifting in response to stimuli and may repeatedly reassert themselves into similar associations. The players may be the same over long periods of time, but the communities are not stable in the sense that we would discuss them in a neoecological context. The relative stability of communities observed from the Paleozoic to the present may suggest that most communities are relatively stable in the face of “normal” Earth change. It may take catastrophic change to destabilize the community, regardless of the diversity.

From the perspective of individual lineages, the fossil record suggests that those clades including species with a range of growth habits appear to be more long lived than clades possessed of but one growth habit. Additionally, large organisms tend to be more insulated from rapid changes in the environment and to respond to longer temporal oscillations in the abiotic environment that may be unavailable to smaller plants with shorter life spans. However, small organisms can respond rapidly to short-term environmental changes. If a clade is represented in a community by species with large and small body sizes, then that clade could persist in a community, despite short- or long-term changes in the environment, thus giving the community the appearance of taxonomic stability when gauged at the higher taxonomic levels.

## X. DIVERSITY EQUILIBRIA

Many experiments have apparently demonstrated shortcomings in the original theory of island biogeography,



especially regarding the idea that island species richness is maintained at consistent equilibrium values by colonization and extinction (MacArthur and Wilson, 1967; Rosenzweig, 1995). However, there is increasing evidence that colonization and extinction processes may be important in maintaining an equilibrium in systems in which species composition is not severely limited by low colonization probabilities (Brown *et al.*, 2000). For example, European pollen data show that there is no overriding directional trend in plant family diversity throughout the past 10,000 years at 24 European sites located between 40 and 70°N latitude. Eighteen of the 24 sites examined show remarkably consistent diversity despite high rates of familial turnover at each site of 30–50% per 1000-year interval (Haskell as cited in Brown *et al.*, 2000).

Testing theories about diversity equilibria over the longer time spans of the pre-Quaternary record is fraught with potential sampling and taphonomic pitfalls [Alroy (1998) as in McKinney and Drake, 1998]. Nonetheless, diversity equilibria models have been extended to evolutionary timescales (Sepkoski, 1991). In these models, diversity is hypothesized to be ultimately lim-

ited by rates of speciation and extinction. Equilibrium in this case is hypothesized to be due to “niche saturation” and occurs when a clade has exploited all possible opportunities.

For several animal clades, at intermediate timescales, species origination rates appear to be richness dependent, whereas extinction rates are not. Thus, as diversity increases, origination rates decline and diversity reaches an equilibrium number [Rosenzweig, 1995; Alroy (1998), and McKinney (1998) as in McKinney and Drake, 1998]. A similar pattern is also observed in the plant fossil record. During the initial debut of each major land plant clade, speciation rates tend to be high and species longevity is short, but subsequent rates of origination tend to decrease while species longevities increase (Fig 3). Furthermore, extinction rates have remained essentially independent of standing diversity (Fig 4). Species diversity equilibria, however, are attained only for short and intermediate intervals of geological time. Over longer periods, diversity has progressively increased (Fig 3). It is not difficult to understand why: Species equilibria require stable environmental conditions and an unchanging field of players. Over

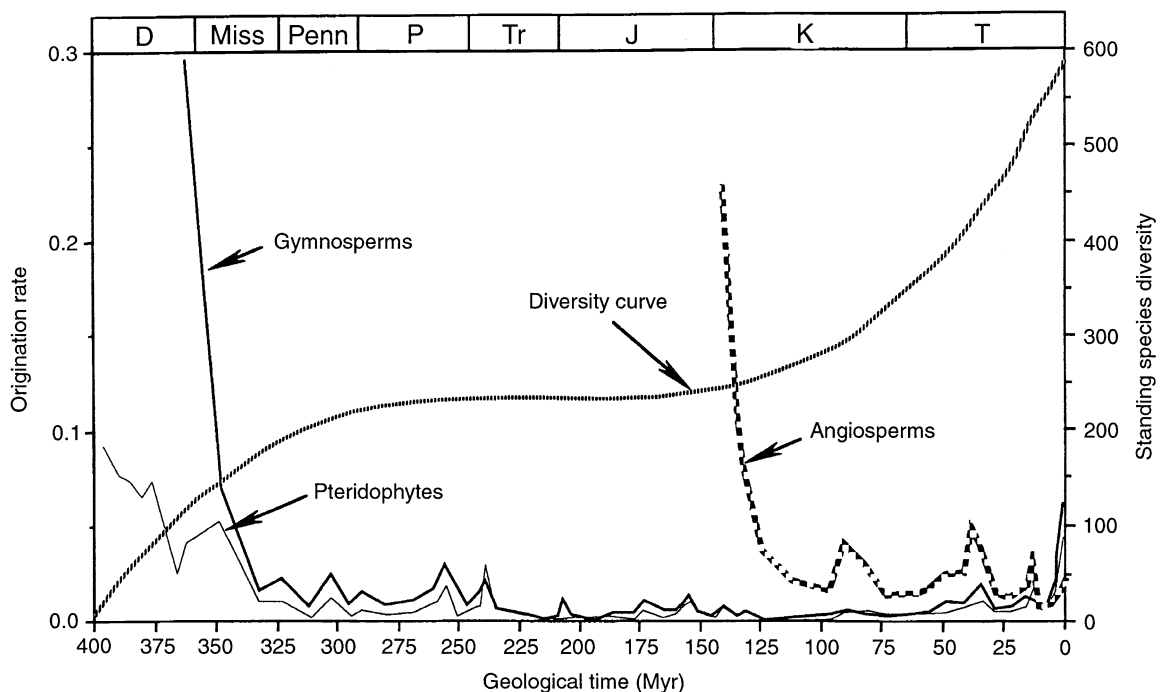


FIGURE 3 Origination rates of vascular plants. Origination rates of pteridophytes, gymnosperms, and angiosperms are plotted as a function of geological time. Curved line denotes ordinary least squares regression for total standing diversity. Rates are computed as the appearance of new species per unit time (relevant geological stage) per standing local diversity. Exceptionally high rates at the appearance of each group reflect low standing diversity of the group at first appearance (data from Niklas, 1997).

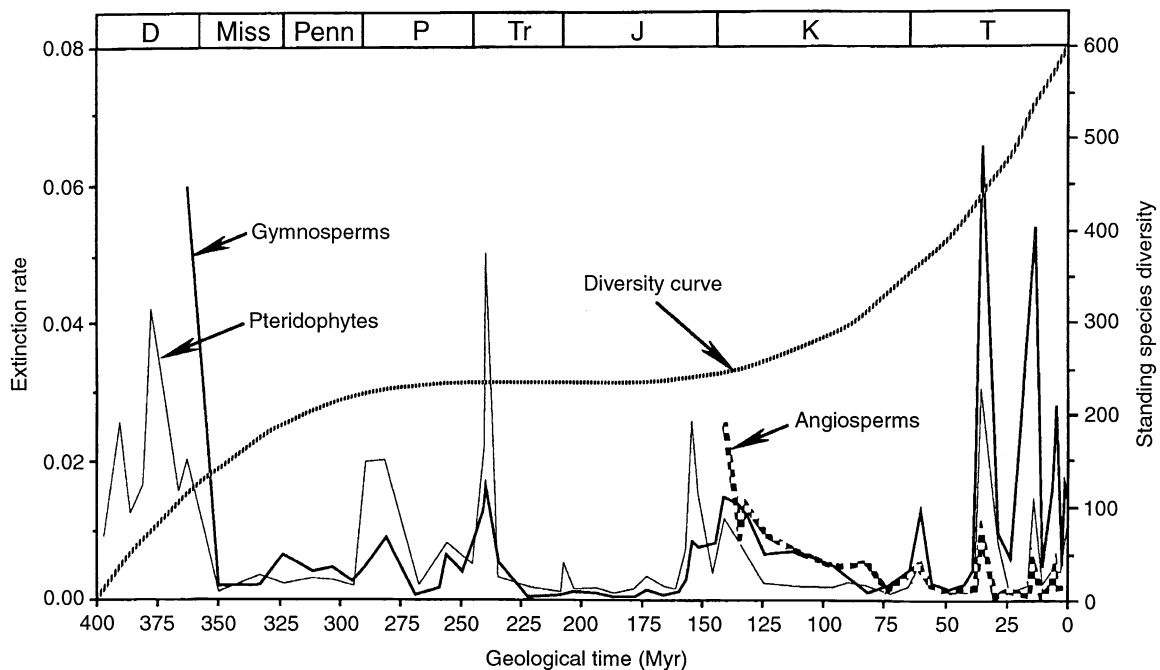


FIGURE 4 Extinction rates of vascular plants. Extinction rates of pteridophytes, gymnosperms, and angiosperms are plotted as a function of geological time. Curved line denotes ordinary least squares regression for total standing diversity. Rates are computed as the disappearance of new species per unit time (relevant geological stage) per standing local diversity. Exceptionally high rates at the appearance of each group reflect low standing diversity of the group at first appearance (data from Niklas, 1997).

geologic time, neither occurs for a variety of abiotic and biotic reasons, such as plate tectonics evolution.

## XI. LATITUDINAL GRADIENTS

An important macroecological pattern of diversity is the latitudinal diversity gradient, which is the increase in alpha and beta diversity from polar to tropical regions. This gradient is reflected by both the number of higher taxa and the number of species nested within those taxa [Ricklefs (1989) as cited in Ricklefs and Schluter, 1993]. For example, typically there are approximately 1.4–2.3 species per plant family and 12 plant families per 0.1-ha area plots within temperate tree communities. In contrast, tropical tree communities of the same size have between 2 and 4 species per family and as many as 58 families [Gentry (1988) as cited in Ricklefs and Schluter, 1993]. Evidence from the fossil record suggests that a latitudinal gradient in angiosperm diversity has existed since the origin of this clade. The Eocene floras of 40–50°N latitude were more diverse than those at 70°N latitude during this relatively warm geological interval. The same pattern is seen when

comparisons are made among 20- to 16-million-year-old floras found in eastern or western North America and the Canadian Arctic.

Many hypotheses have been advanced to explain the latitudinal gradient in species diversity. The most promising propose multiple processes and consider both local- and global-scale patterns, invoking gradients of productivity and energy availability that correlate with changes in latitude (Brown, 1995). Productivity ultimately controls biodiversity, and conditions that favor high productivity, such as warm temperatures and ample precipitation, are often associated with high diversity [Currie and Paquin (1987) as cited in Rosenzweig, 1995].

Differential origination rates from the tropics to the poles may also play an important role in creating and maintaining the latitudinal diversity gradient, although a fundamental mechanism for this pattern has yet to be identified. In general, it appears that major evolutionary innovations and speciation events occur more frequently in lower rather than higher latitudes and that these innovations subsequently spread toward the poles over comparatively long periods of time. The hypothesis of the equatorial region as a source of novelty is sup-

ported by trends in the invasion of land, the origin of seeds and angiosperms, and even the invasion of land by tetrapods, all of which appear to comply with a “tropics to the poles” trend (Jablonski, 1993).

## XII. ABUNDANCE

As ubiquitous as the latitudinal diversity gradient but possibly more mysterious are the patterns of abundance that are found in both modern and historical communities. Although it is generally agreed that increasing resource availability can increase diversity, very little is understood about how these resources are divided among members of a community. Abundance patterns

in extant communities tend to conform to a “canonical” pattern: Most species are rare, whereas very few are common [Preston (1948, 1962) as cited in Brown, 1995; Rosenzweig, 1995; Rosenzweig (1998) as cited in McKinney and Drake, 1998].

Unfortunately, examination of past abundance patterns is confounded by preservational features. The plant fossil record consists of plant parts (dispersed leaves, fruits, wood, etc.) that are preserved in a manner that imperfectly reflects the abundance of species in communities. A deciduous tree, for example, within an otherwise evergreen forest might appear to dominate due to excessive production and shedding of organs. However, several Tertiary floras and a few Carboniferous ones are both geographically extensive and widely

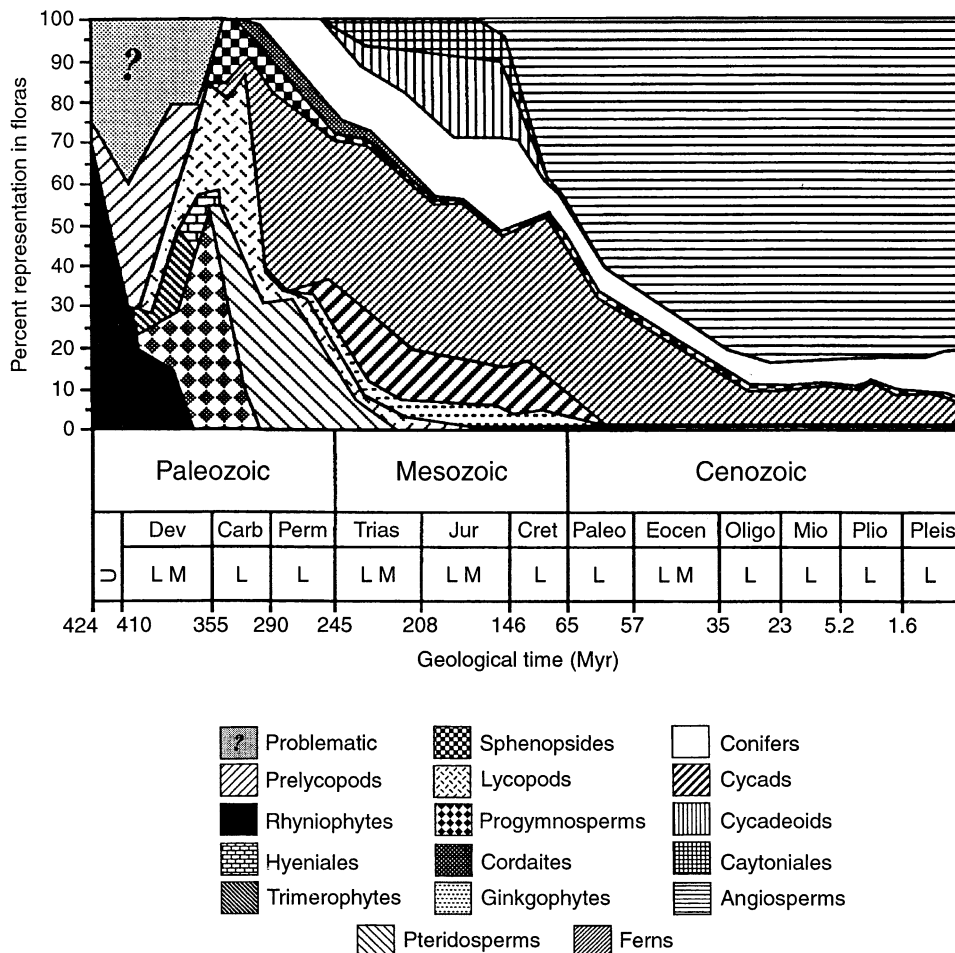


FIGURE 5 Changes in floristic composition. Taxonomic composition of representative floras (expressed as the percentage of each major group of vascular land plants) are plotted against geological time. Rhyniophytes and taxonomically “problematic” groups of plants dominate the first land plant floras; angiosperms dominate the most recent fossil plant floras.

sampled, and they suggest that canonical abundance distributions may be a common feature in historical communities.

Dominance within a community, as reflected by abundance, appears to be a transient feature for an individual species. Indeed, the fossil record indicates that most species that were once common typically become rare with time [Knoll (1986) as cited in Rosenzweig, 1995]. At a higher taxonomic level, however, certain groups do exhibit long-term dominance (e.g., the Mesozoic is the age of conifers and cycadophytes). This dominance likely reflects the increased genetic, developmental, morphological, and reproductive diversity inherent in pooling larger numbers of species within higher taxa. When members of a clade breach a developmental barrier, radically new solutions to life may arise that may allow for increased abundance via the colonization of new environments and/or the ability to outcompete with older clades. This dynamic may account for major transitions in dominance such as that between gymnosperms and angiosperms (Fig. 5).

### XIII. SUMMARY

Examination of the major patterns of plant diversity from both contemporary and historical communities holds the promise of illuminating the fundamental mechanisms responsible for the origin and maintenance of biological diversity. Here we have outlined several "rules" and patterns that appear to have been important in the evolution of plant communities. In many situations, the contemporary and historical evidence clearly complement each other. In other cases, such as the relative importance of competition, they appear to conflict. The latter cases may reflect the inherent research limitations of paleontology and neoecology, the real effects of temporal scale, or the operation of different processes.

Of particular note to ecologists interested in diversity should be identification of those factors that cause diversity to change at both short and long timescales. Dramatic climate change and evolutionary innovation appear to be the most important features in the historical record. It is interesting to note that the current diversity–stability debate has considered these variables not as cause but as effect, examining the influence of diversity on community and ecosystem processes rather than vice versa. We live in the most diverse terrestrial communities in the history of Earth, but they function as the result of the same basic rules of interaction,

energetics, and stoichiometry as did their Paleozoic predecessors. When events, human or otherwise, create an extraordinary disturbance, biodiversity may plummet and modern communities, as we recognize them, may cease to exist. However, the evidence of 400 million years strongly suggests that the survivors will reorganize into communities and continue to function following the same basic physical, chemical, and ecological rules that created the world around us. They have done so before, and there is good reason to assume that they will continue to do so as long as this planet harbors life.

### Acknowledgments

This article was completed as part of the Body-size Working Group supported by the National Center for Ecological Analysis and Synthesis, a center funded by NSF (Grant DEB-94–21535), the University of California at Santa Barbara (UCSB), and the state of California. We thank J. Damuth, W. DiMichele, F. A. Smith, and J. H. Brown for commenting on initial versions of this article. BJE was supported by a NSF Postdoctoral Fellowship. JPH was supported by NSF Grant DEB-9707406 to J. H. Brown and by the National Center for Ecological Analysis and Synthesis. BHT acknowledges research support provided by the College of Creative Studies, UCSB.

### See Also the Following Articles

COEVOLUTION • DIVERSITY, COMMUNITY/REGIONAL LEVEL • HABITAT AND NICHE, CONCEPT OF • LATITUDE, COMMON TRENDS WITHIN • PALEOECOLOGY • PLANT BIODIVERSITY, OVERVIEW

### Bibliography

- Bateman, R. M., Crane, P. R., DiMichele, W. A., Kenrick, P. R., Rowe, N. P., Speck, T., and Stein, W. E. (1998). Early evolution of land plants: Phylogeny, Physiology, and Ecology of the primary terrestrial radiation. *Annu. Rev. Ecol. Syst.* 29, 263–292.
- Brown, J. H. (1995). *Macroecology*. Univ. of Chicago Press, Chicago.
- Brown, J. H., Ernest, S. K. M., Parody, J. M., and Haskell, J. P. (2000). Regulation of diversity: Maintenance of species richness in changing environments. *Ecology*. (in press).
- Crawley, M. J. (Ed.) (1997). *Plant Ecology*. Blackwell, Cambridge, MA.
- Crepet, W. L. (1996). Timing in the evolution of derived floral characteristics: Upper Cretaceous (Turonian) taxa with Tricolpate and Tricolpate-derived pollen. *Rev. Paleobot. Palynol.* 90, 339–359.
- DiMichele, W. A., Phillips, T. L., and Olmstead, R. G. (1987). Opportunistic evolution, abiotic environmental stress, and the fossil record of plants. *Rev. Paleobot. Palynol.* 50, 151–178.
- Enquist, B. J., West, G. B., Charnov, E. L., and Brown, J. H. (1999). Allometric scaling of production and life-history variation in vascular plants. *Nature* 401, 907–911.
- Jablonski, D. (1993). The tropics as a source of evolutionary novelty through geological time. *Nature* 364, 142–144.

- McKinney, M. L., and Drake, J. A. (Eds.) (1998). *Biodiversity Dynamics: Turnover of Populations, Taxa, and Communities*. Columbia Univ. Press, New York.
- Niklas, K. J. (Ed.) (1981). *Paleobotany, Paleocology and Evolution*, Vol. 2. Praeger, New York.
- Niklas, K. J. (1997). *The Evolutionary Biology of Plants*. Univ. of Chicago Press, Chicago.
- Pellmyr, O. (1999). Systematic revision of the yucca moths in the *Tegeticula yuccasella* complex (Lepidoptera: Prodoxidae) north of Mexico. *Syst. Entomol.* 24, 243–271.
- Ricklefs, R. E., and Schluter, D. (Eds.) (1993). *Species Diversity in Ecological Communities*. Univ. of Chicago Press, Chicago.
- Rosenzweig, M. L. (1995). *Species Diversity in Space and Time*. Cambridge Univ. Press, New York.
- Sepkoski, J. J. (1991). A model of onshore–offshore change in faunal diversity. *Paleobiology* 17, 58–77.



# PLANT CONSERVATION

Mike Maunder  
*Royal Botanic Gardens*

---

- I. A History of Plant Conservation
  - II. Distribution and Loss of Plant Diversity
  - III. Human Influences on Plant Diversity
  - IV. Plant Extinctions: How Many and Where?
  - V. Responses to Biodiversity Loss
  - VI. *Ex Situ* Conservation
  - VII. Toward an Integrated Approach
  - VIII. Facilities and Skills for Plant Conservation
- 

## GLOSSARY

- ark paradigm** The concept that threatened species can be preserved in or at a special facility; a term derived from the Biblical account of Noah's Ark saving every extant species on earth during a great flood.
- ex situ** Literally, away from the site or location; in this context referring to conservation efforts elsewhere than the natural habitat; e.g., in botanical gardens.
- hot-spot** An area identified as having an unusually large number of plant species, a high proportion of endemic (localized) species, and a threat of habitat destruction.
- in situ** Literally, in or at the site or location; in this context referring to conservation efforts within the natural habitat.
- in vitro** Literally, in glass; i.e., in a test tube; more broadly, in a laboratory or other artificial setting rather than in nature.
- living dead** A term for a species in which scattered adult individuals still survive but the capacity for

reproduction has been lost; thus extinction will occur when the present generation dies out.

---

*PLANT CONSERVATION* is the management of plant resources to maintain current levels of plant diversity and to avoid population and taxonomic extinctions. Historically, plant conservation has been overshadowed by the politically more pressing issues of wildlife conservation. This article provides a historical review of plant conservation issues and provides an overview of present approaches and activities.

## I. A HISTORY OF PLANT CONSERVATION

Human societies have historically managed plant resources to ensure dependable access to timber, fruits, roots, and medicinals. However, for both rural communities and larger political dynasties, this management has been driven largely by utility. The cedars of Lebanon, *Cedrus libani*, now surviving as scattered and isolated groves in Lebanon, Syria, and Turkey and covering probably less than 5% of their historical area, have been subject to fluctuating periods of management and overexploitation for over 2000 years. Indeed the marker stones to Roman forest reserves can still be found on Lebanese hillsides. The concept of human-caused habitat degradation and species loss was established only

during the eighteenth century, as a reaction to catastrophic habitat changes following colonial occupation of oceanic islands such as Mauritius (Grove, 1996). One of the earliest *ex situ* conservation attempts can be attributed to Governor Byfield of St. Helena, who in the early eighteenth century collected two seedlings of the then very scarce St. Helena redwood, *Trochetiopsis erythroxylon*, for cultivation in his garden on the island.

The recorded extinction of *Franklinia alatamaha* in the United States established the concept of plant extinctions. During the late nineteenth century and early twentieth century, a small number of species were assumed to have become extinct in the wild, most notably *Ginkgo biloba* and *Amherstia nobilis*. Both species were considered to be extinct in the wild and only surviving in Asian temple gardens. However, they appear to have been treated as isolated novelties and did not prompt any general concern from the botanical community. In contrast, animal extinctions attracted significant attention; as early as the 1880s the National Zoo, Washington, DC, was proposed as a "home and a city of refuge for the vanishing races of the continent."

The broader environmental movement can be traced, in part, to the late nineteenth century protected area movement in the United States (the Yellowstone National Park Act of 1872) and the wilderness-inspired literature of Aldo Leopold. The 1960s saw the start of a public concern for the environment and the beginning of international and national structures for conservation. This concern led to a public and political environmental awareness of the 1970s as manifested through the establishment of Earth Day on April 22, 1970, in the United States and the Endangered Species Act of 1973. The Russian agricultural botanist N. I. Vavilov, during the 1920s and 1930s, promoted the value of crop relatives and wild species in supporting agriculture. The present concerns about plant conservation, as both a political and scientific issue, can be traced to a series of international conferences on plant genetic resources during the late 1960s and 1970s. The erosion of these resources was recognized as an urgent problem at a Food and Agriculture Organization (FAO) conference in 1961 and subsequently discussed by a joint FAO and International Biological Programme (IBP) meeting in 1967, where the term "genetic resources" was introduced. The FAO Panel of Experts on Plant Exploration and Introduction led the debate on plant resources from the mid-sixties to seventies, with the creation in 1974 of the International Board for Plant Genetic Resources (IBPGR, now International Plant Genetic Resources Institute (IPGRI)).

A widely recognized initial impetus to conservation

was derived from the United Nations Conference on the Human Environment held at Stockholm in 1972. The Threatened Plants Committee (TPC) of the International Union for the Conservation of Nature (IUCN) was established in 1974 as a direct product of the initial *Red Data Book for Threatened Plant Species*. This network was designed to deal specifically with plant species outside of the remit of the FAO. The TPC acted as the template for the subsequent evolution of the IUCN's Species Survival Commission (SSC).

The discussions in the early and mid-1970s set the stage for botanic gardens to take up plant conservation as a serious responsibility. The "Ark Paradigm," the idea that *ex situ* facilities would hold stocks of threatened plants during a period of habitat degradation, the "demographic winter" *sensu* Soulé (1991), was established as a working objective by botanic gardens in the 1970s. This attitude is manifest in the 1978 *Red Data Book* (Lucas and Synge, 1978); for instance, in the data sheet for *Dracaena ombet*, it was stated that "it seems too late for such a proposal [*in situ* conservation] to be worthwhile. Great efforts must now be made to bring the ombet into cultivation and maintain it safely in the botanic gardens of the world." A notable conference in the late 1970s was "Extinction is Forever" (Prance and Elias, 1977); it provided a regional overview of American plant and habitat conservation issues with both a geographical and taxonomic focus. This conference attempted to assess levels of species decline and loss in the region and subsequently set the scene for later developments in tropical forest inventory.

Two conferences hosted by the Royal Botanic Gardens, Kew, in 1976 and 1978, established the agenda for botanic garden activities over the next two decades (Simmons *et al.*, 1976; Synge and Townsend, 1979). As a result of the 1976 conference, the Threatened Plants Committee of the SSC of the IUCN was requested to establish a Botanic Gardens Conservation Co-ordinating Body "to find out which threatened plants are in cultivation and where, and to keep botanic gardens informed of current conservation activities." This need was reiterated at the First International Botanic Gardens Conservation Congress in Las Palmas, Canary Islands, in 1985 with the recommendation of the establishment by IUCN of the Botanic Gardens Conservation Secretariat (BGCS), later to become Botanic Gardens Conservation International (BGCI). In parallel with discussions on botanic gardens, a number of broader conservation references were published in the early 1980s that paved the way for the 1992 UNCED meeting. These included The Brandt Report (Independent Commission on International Development Issues, 1980), the Global 2000



FIGURE 1 *Echium wildpretii* endemic to Tenerife and Las Palmas, Canary Islands, Spain. Careful management of the wild populations has recovered wild populations on Tenerife. However, the extensive cultivated populations in botanic gardens show evidence of inbreeding and hybridization.

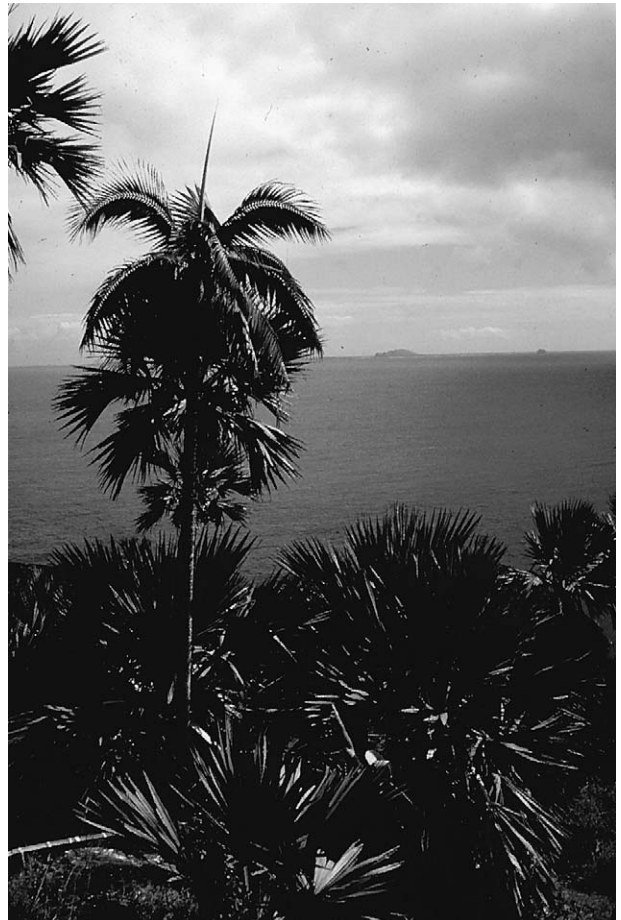


FIGURE 2 *Dictyosperma album* var. *conjugatum*, endemic to Round Island, Mauritius. This tree is the last wild tree; however, extensive, but undocumented, cultivated stocks survive on Mauritius. The potential exists to use these for reintroductions onto Round Island.

Report (U.S. Council on Environmental Quality and U.S. Department of State, 1980), and the World Conservation Strategy (IUCN/UNEP/WWF, 1980). All three reports served to emphasize ongoing loss of species and habitats and the linkages to human welfare and development.

The Convention on Biological Diversity (CBD) is the major legislative influence on the conservation of global biodiversity; directly influencing national activities through the requirement to produce and implement Biodiversity Action Plans.

The main objectives of the CBD are

- The conservation of biological diversity (Articles 6–9, 11, and 14)

- The sustainable use of its components (Articles 6, 10, and 14)
- The fair and equitable sharing of the benefits arising from the use of genetic resources, including by appropriate (1) access to genetic resources (Article 15), taking into account all rights over those resources; (2) transfer of relevant technologies (Articles 16 and 19), taking into account all rights to technologies; and (3) funding (Articles 20 and 21).

Outlined under Article 6 (General Measures for Conservation and Sustainable Use) is the requirement for each Contracting Party to develop a national biodiversity strategy, action plans, or program. The primary function is to make specific recommendations for national action through identifying areas for action; iden-





FIGURE 3 *Encephalartos woodii*. This cycad is extinct in the wild and only survives as a male clone in botanic garden collections. It is an example of a cultivated member of the "living dead."

tifying obstacles, such as national capacity, finances, technology, inadequate legal mechanisms, etc.; identifying relevant government sectors and affected constituencies; identifying cost-effective solutions; and assigning tasks. In general, this study will be undertaken through a National Biodiversity Unit.

The past 20 years has seen a revolution in the available technology for supporting plant conservation. Fundamental to this has been the explosion in information technology allowing the collection, exchange, and analysis of data pertaining to the wild status of species. The past two decades has seen the move from record keeping in a ledger and the largely verbal transmission of associated information toward the increased use of computerized records and geographical information systems (GIS). Based upon the scientific advances attained in the fields of crop genetic resources, there have been major advances in seed storage, cryopreservation, and



FIGURE 4 Restoration of a degraded island habitat on Round Island, Mauritius. A collaborative exercise between the government of Mauritius and the Mauritian Wildlife Fund with support from the Durrell Wildlife Conservation Trust, Royal Botanic Gardens, Kew, and Fauna and Flora International is restoring the original forest after control of introduced goats and rabbits.

*in vitro* propagation. The FAO Panel of Experts on Plant Exploration and Introduction drafted standards and procedures for the long-term storage of seed. Recent advances have included protocols for the storage of *in vitro* cultures. An increasing number of technical manuals are available to guide professionals (e.g., Given, 1994). The development of polymerase chain reaction (PCR) based molecular technologies has opened the door to identification and management of both evolutionary lineages and conservation units. The utility of these techniques is increasing as both levels of resolution increase and costs decrease. In addition, as the understanding of major phylogenetic relationships advances, it will become increasingly feasible to focus conservation activities on evolutionarily significant clades.

## II. DISTRIBUTION AND LOSS OF PLANT DIVERSITY

Botanical diversity is concentrated into areas of unusual richness and exhibits gradients on both a global and regional scale. The taxonomic diversity of higher plants broadly correlates to latitude and is generally highest in the tropical and subtropical regions. Species richness and endemism were commonly used attributes for the identification of biodiversity priorities since they presumably reflected both the complexity and uniqueness of ecosystems. Accordingly, early discussions identified tropical rain forests and oceanic islands as conservation priorities.

In 1986, the International Union for the Conservation of Nature (IUCN) and the Worldwide Fund for Nature (WWF) initiated a project to identify major centers of plant diversity. The objectives of the project were to identify which areas around the world, if conserved, would safeguard the greatest number of plant species; to document the many benefits, economic and scientific, that conservation of those areas would bring to society and to outline the potential value of each for sustainable development; and to outline a strategy for the conservation of the areas selected. The project (Davis *et al.*, 1994, 1996, 1997) has identified 234 areas of especially high species diversity, and correlates strongly with the listed 12 megadiversity countries (*sensu* Mittermeier *et al.*, 1998) (Table I).

Myers (1988, 1990) identified conservation "hot spot" areas, defined as areas having an exceptional concentration of plant species, having high endemism, and threatened with destruction. Myers initially identified 10 hot spots, all in tropical rain forest areas. The analysis was subsequently expanded to include four additional rain forest areas and four Mediterranean climate regions, including the Cape Floristic region.

The "megadiversity countries" concept, developed by Mittermeier (1988) and expanded in Mittermeier *et al.* (1998), is based on three premises: (1) biodiversity is not evenly distributed among the world's countries; (2) although international conservation priorities should be based on scientific information on biodiversity, it is governments that develop conservation policies and programs; and (3) a very small number of countries, lying wholly or partly within the tropics, contain a high percentage of the world's species. The first phase of analyses identified 12 centers of diversity to be mostly in the tropical regions. This initial focus on the megadiverse tropical countries overlooked the diversity of Mediterranean and subtropical floras; for instance, the five Mediterranean climate regions of the planet contain about 20% of the world's plant species. The 24 hot spots are estimated to hold about 124,000 endemic plant species, equating to 40% of the world's plant diversity (Mittermeier *et al.*, 1998).

### III. HUMAN INFLUENCES ON PLANT DIVERSITY

Historically, about half of the ice-free surface of the globe has been transformed, managed, or utilized by man (Turner *et al.*, 1990); nearly 40% of the potential terrestrial net primary productivity of the Earth is used

by mankind or lost as a result of land change (Vitousek, 1994). Hannah *et al.*, (1994) concluded that 73% of the Earth's land surface, other than rock, ice, and barren land, is either human dominated (36.3%) or partially disturbed (36.7%), with only 27% undisturbed. Mackinnon and Mackinnon (1986a, 1986b) estimate that 65% of the original ecosystems south of the Sahara have been subject to major ecological disturbance and that 68% of the natural habitat has been lost in the Indo-Malayan nations. This trend will continue as human populations and their demands grow.

The extensive conversion of wild habitats has resulted in changes to both the ecological and taxonomic composition of these areas. In many parts of the world the change is almost complete; taxonomically and ecologically diverse wildlands have been converted to agricultural landscapes dominated by a small range of domesticates, e.g., cereals. In addition, the modernization of agriculture in traditional landscapes has left scattered fragments of original habitats. This fragmentation of habitats is causing both high levels of population loss (Hughes *et al.*, 1997) and regional species loss (Harris and Silva-Lopez, 1992) and modifications in patterns of genetic diversity (Young *et al.*, 1996), often with a reduction in the level of genetic diversity in persisting populations (Hall *et al.*, 1996). On the other hand, these surviving fragments of habitat can act as refuges for threatened species and act as the focus for later habitat restoration programs (Saunders *et al.*, 1993; Packard and Mutel, 1996) (Fig. 4).

The Mediterranean Basin has been subject to the near-complete transformation of its original ecology, a process that can be traced back over 6000 years, through the classical civilizations of Greece and Rome, the establishment of pastoralism, and the contemporary pressures of intensive agriculture, urbanization, tourism developments, and invasive species. The Mediterranean Basin is recognized as a globally important hot spot of botanical diversity, with ca. 24,000 plant species and, surprisingly, only 31 recorded species extinctions (Greuter, 1994). It is likely that the Mediterranean Basin has suffered unrecorded plant extinctions, as the initial forest ecosystems were replaced with an anthropogenic mosaic of scrub and forest under the increasing influences of domestic stock and fire.

Spectacular patterns of degradation have occurred on oceanic islands. Oceanic islands are particularly susceptible to damage by invasive species (MacDonald and Cooper, 1995). Colonized Pacific islands were subject to profound environmental changes and high levels of avian, and presumably plant, extinctions following Polynesian settlement. One example, the Pacific island

TABLE I  
Major Plant Diversity Countries<sup>a</sup>

Country	Megadiversity country ranking	Number of centers of plant diversity	Number of vascular plant taxa	Biodiversity hot spots with number of endemic plant species
1. Brazil	1	12	56,000	Brazilian cerrado, 6000 Atlantic Forest, 4400
2. Colombia	2	8	51,000	Tropical Andes, 20,000 (in part) Choco, 2500 (in part)
3. China	3	8	27,100	Indo-Burma, 7000 (in part) Eastern Himalayas, 5000 (in part)
4. Mexico	4	12	>20,000	Mesoamerican Forest (in part)
5. Australia	5	10	15,600	Southwestern Australia, 3724
6. Indonesia	6	18	>12,000	Wallacea, 1500 Sundaland, 5000 (in part)
7. Peru	7	8	18,200	Tropical Andes, 20,000 (in part)
8. Ecuador	8	6	>20,000	Choco, 2500 (in part)
9. Malaysia	9	13	>14,000	Sundaland, 5000 (in part) Indo-Burma, 7000 (in part)
10. India	10	6	17,000	Western Ghats, 2182 (in part) Eastern Himalayas, 5000 (in part) Indo-Burma, 7000 (in part)
11. Zaire	11	12	11,000	
12. Madagascar	12	1	10,000	Madagascar, 9700 (in part)
13. United States	n/a	8	20,000	California, 2125 (in part) Caribbean islands, 7000 (in part)
14. Philippines	n/a	6	8,931	Philippines, 5832
15. Republic of South Africa	n/a	6	23,420	Karoo, 1940 Cape, 5682
16. Turkey	n/a	6	8,650	Mediterranean Basin, 13,000 (in part)

<sup>a</sup> This table illustrates the concentration of plant diversity in 16 countries holding half of the world's centers of plant diversity. The 12 megadiversity countries, are ranked in column two. Total number of centers of plant diversity in the 16 countries is 140/234 = c. 60%. The 16 countries contain (entirely or in part) 18/24 of the most important plant biodiversity hot spots. Data are derived from McNeely *et al.* (1990), Myers (1990), and Davis *et al.* (1994, 1996, 1997).

of Rapa Nui (Easter Island), underwent massive environmental degradation as long ago as 1200 to 800 B.P. In the Caribbean, Atlantic islands, and Mascarenes, environmental degradation can be attributed to European colonial administration and the development of unstable and fragile economies based upon plantation products. Colonial settlement would often focus on the clearance and agricultural development of the lowlands, with the retention, planned or otherwise, of upland habitats, albeit in a modified and fragmented form. Examples of islands with severely modified lowlands include Mauritius, the Gulf of Guinea islands, and New Caledonia. Although the island of Sao Tome, Republica Democratica de Sao Tome and Principe, has over 90%

forest cover, the remaining primary forest is restricted to high ground and steep slopes. While the lowlands of Gran Canaria, Islas Canarias, were converted to plantations, the uplands were decimated as a source of both firewood and topsoil. On Gran Canaria less than 1% of the original laurel forest survives.

Some islands have suffered the virtual complete loss of original habitats (Box 1). This can result from periods of over-exploitation over centuries. On Rapa Nui, no original plant communities have survived the Polynesian settlement and later colonial sheep ranching. On St. Helena, only fragments of gumwood forest and montane tree fern thickets have survived, covering less than 1% of the island's land area (Mauder *et al.*, 1995). The



FIGURE 5 Ascension Island: View over colonized lava fields where introduced plants have transformed the ecology and landscape from a bare lava field to a vegetated ecosystem. The lava fields are dominated by African forage grasses and the Madagascar endemic *Catharanthus roseus*.

total transformation of an oceanic island can result from modern industrial exploitation, for instance, phosphate mining on Nauru in the Pacific.

As a result of deliberate introductions and as artifacts of the agricultural and horticultural industries, an increasing number of exotic species are growing in new territories and habitats (Box 1; Fig. 5). While such introductions may in the short term increase local biodiversity, in the long term such introductions are having disastrous impacts on both species diversity and ecological processes. Disastrous invasions include the impact of guava in Mauritius, miconia on the Polynesian Islands, and water hyacinth in the East African Rift Valley



FIGURE 6 The retention and conservation of some tropical habitats will be dependent upon balancing the needs of resident large mammal populations with plant conservation. In the Mwaluganje Elephant Sanctuary, Kenya, increasing elephant populations are damaging the mosaic of the coastal forest.

Lakes. In addition, introduced diseases, such as chestnut blight in the United States, and pathogens, such as the New Zealand flatworm in the United Kingdom, are impacting upon plant diversity.

Human activities, notably the combustion of fossil fuels and the subsequent release of carbon dioxide, are altering the composition of the atmosphere. It is predicted that the historic concentration of carbon diox-

#### Box 1

##### Case Study: The South Atlantic Island of Ascension—Too Late for Habitat Restoration?

On discovery in 1501, the island was a bare volcanic landscape, with vast seabird colonies supporting only 25 vascular plant species; of these 11 are endemic. However, as a result of nineteenth century garrison settlement, a large number of exotic plants were introduced for both agricultural and horticultural purposes (Fig. 5). The effects of exotic species have transformed the original bare volcanic landscape of Ascension into a semiforested one. Successive importations of plants from British and South African botanic gardens represented massive introductions of both plant species and associated invertebrate and fungal diversity. Consequently, the original upland *Marattia* fern thicket has been transformed into an exotic woodland dominated by an introduced bamboo. In the middle altitudes, introduced *Casuarina*, *Araucaria*, and *Juniperus bermudiana* are forming open woodlands. The xeric lowland lava fields have been transformed into open woodland of *Prosopis juliflora* with introduced African grass species. The *Prosopis* is now encroaching on the turtle (*Chelonia mydas*) nesting beaches and threatening to expand onto the unique breeding colonies of Sooty Tern (*Sterna fuscata*). The transformation of terrestrial ecology has been massive and habitat restoration is no longer feasible; the priority should be to maintain and promote specific species and communities, e.g., seabird colonies, through the control of introduced predators (cats and rats) and herbivores (sheep and donkeys). The net effect has been an increase in total botanical diversity and primary productivity and the creation of a new synthetic ecosystem.



FIGURE 7 Botanic gardens in the tropics are playing a real role in the conservation and sustainable use of plant resources. The Entebbe Botanic Garden, Uganda, is working with traditional healers to catalog and propagate medicinal plants.

ide in the Earth's atmosphere may double in the next century and change global climate. This could result in an increase in mean global temperature of approximately 1.5–4.5°C during the twenty-first century (IPCC, 1992). Botanical diversity is expected to change as species react through both local extinctions and the colonization of new areas. Ecological processes will be modified as fire and hydrological systems alter. It is expected that these changes will cause the extinctions of plant species as populations die out and are unable to migrate between isolated habitats (Morse *et al.*, 1995). It is likely that invasive weed species will dominate many of these new and unstable plant communities.

All of these damaging impacts are driven by one fundamental force, the increase in human population. By the year 2025 the world population is projected to total about 8.3 billion people, compared with the current estimate of 5.7 billion. The patterns of increase will vary regionally; European populations may decline slightly, Asia could grow from a current 3.46 billion to 4.96 billion, while the population of Africa will possibly double between 1995 and 2025, from 728 million to 1.49 billion (WRI/UNEP/UNDP/World Bank, 1996). As populations increase, there will be continued demand for both new and traditional plant products; for instance, between 4000 and 6000 products of medicinal plants are traded internationally. One new wild product is derived from *Prunus africana*, harvested from montane African forests for its bark, from which alkaloids are extracted for the treatment of prostate disorders. The annual over the counter trade is worth over \$200 million per year.

#### IV. PLANT EXTINCTIONS: HOW MANY AND WHERE?

Plant resources are being exploited to the point of both economic and biological extinction. The exhaustion of the Caribbean mahogany (*Swietenia* spp.) and the Mauritian ebony (*Diospyros* spp.) are examples of island resources taken to economic extinction, while the Juan Fernandez endemic sandalwood (*Santalum fernandezianum*) has been exploited to biological extinction. Horticultural exploitation has resulted in a small number of species being eradicated in the wild and surviving in cultivation, an example being the Chilean crocus *Tecophilaea cyanocrocus*.

Since 1600, it is estimated that 654 flowering plant species have become extinct (WCMC, 1992). This is certainly an underestimate as many areas of the world (e.g., the Mediterranean Basin) have undergone massive levels of habitat destruction, with presumed species loss prior to scientific inventory. The most comprehensive survey of species decline undertaken on behalf of the IUCN indicates that about 33,400 plant species are threatened with extinction, equating to about 10% of the world's 250,000–300,000 plant species (Walter and Gillett, 1998). For the largest part of the planet, there is no clear consensus on the rate of species and population loss. It could be argued that the plant extinctions logged by the IUCN and World Conservation Monitoring Centre (WCMC) reflect, in part, the geographical distribution of botanical knowledge and monitoring, rather than actual rates of species loss; for instance, all the recorded sub-Saharan plant extinctions (45) are from the Republic of South Africa alone.

Koopowitz *et al.* (1994), estimating both historical and prevailing rates of extinction based upon rates of habitat conversion and distribution of restricted endemic plant species for tropical Latin America, produced figures exhibiting a dramatic lack of congruity with the WCMC/IUCN records. This is particularly notable for Brazil, where WCMC/IUCN has recorded only 5 extinctions and Koopowitz estimates a loss since 1950 of 2261 species. It is evident that the data pertaining to plant extinctions are not sufficient to identify in advance which plant species are at greatest risk of extinction; the paucity of field survey work and rapidity of habitat loss mean that many species' extinction will be identified only in retrospect. A few attempts have been made to identify extinction-prone tropical forest species in advance (Martini *et al.*, 1994).

Apart from the well-publicized estimates of loss from the tropics, it is evident that the European nations are

## Box 2

## Case Study: The "Living Dead"

Plant conservationists face a dilemma in dealing with the living dead, species that are surviving as scattered, nonreproductive, individuals isolated from their original habitats (Fig. 3). Oceanic islands having suffered massive levels of habitat loss contain examples of the living dead. Both *Kokia cookei* from Hawaii and *Ramosmania rodriguesiana* from Rodrigues survive only as sterile single clones. The endemic palm, *Hyophorbe amaricaulis*, survives only as a single tree in the Curepipe Botanic Gardens, southwest Mauritius. The single specimen has persisted within the botanic garden for over 50 years with no regeneration recorded. The specimen produces flowers and fruits regularly, but attempts to grow these using both micropropagation and conventional horticultural techniques have not been successful.

losing species as a result of changes in both habitat cover and agricultural practice. For example, Britain and Ireland have lost 19 nonendemic species over the past 250 years. In contrast to areas in the tropics, European plant conservation frequently focuses on the retention of traditional agricultural practices such as hay meadow management or transhumance.

In many parts of the world it is difficult to differentiate clearly between wild and domesticated plant diversity. Traditional farmers often maintain an agriculturally sophisticated and dynamic mix of local cultivars and landraces in close proximity with wild species (Nabhan, 1989; Casas and Caballero, 1996). These systems are threatened by modern agricultural practices and economic changes encouraging a more industrial, and less diverse, style of agriculture (Burgess, 1994). The number of traditional cultivars used in agriculture in Europe and North America is declining dramatically. A study by Rural Advancement Fund International found that of the 7098 apple varieties in use between 1804 and 1904 in the United States, 86% have been lost.

## V. RESPONSES TO BIODIVERSITY LOSS

Conservation actions are implemented to mitigate or reverse the damaging impacts of human social, demographic, and economic change on plant diversity. Manage-

ment actions aimed at the conservation of biodiversity will take place at various levels of the biodiversity hierarchy, a nested hierarchy of spatially defined units with often ill-defined boundaries (Soulé, 1991). These efforts vary in the scale of spatial, capital, and scientific investment, encompassing single-species management through to wilderness retention. Increasingly, the traditional schism between on-site activities (*in situ* conservation) and off-site activities with a species or genetic focus (*ex situ* conservation) is eroding. Conservation activities focused on any one layer of the biospatial hierarchy should take into account the linkages to other levels. An integrated approach to plant conservation is being promoted encompassing species recovery programs and habitat management (Falk *et al.*, 1996). A parallel development has been an increasing trend toward ecosystem-level conservation in both the temperate and tropical regions (Weeks, 1996). In addition, the linkages between biodiversity management and sustainable development have been given due recognition (Barzeti, 1993).

The majority of the world's species will be retained through the "coarse filter" approach of habitat conservation; this potentially could conserve all levels in the biodiversity hierarchy and their interactions. However, many protected areas will require increasing management because of external influences impacting on ecological processes and promoting changes in both community structure and composition. Protected area borders are permeable to disease, invasive species, poaching, civil unrest, climate change, etc. Protected areas have been established with the assumption that environmental conditions and community patterns/composition have been relatively stable for long periods in the past and will continue to be stable into the future. It could be argued that habitats are loosely organized collections of species whose coexistence is dependent on their individual limits and subsequent distribution along environmental gradients. On a geological time scale, they could be viewed as relatively transient assemblages. Accordingly, a "fine filter" approach will be required as a backup to catch those species not secured through the priority action of habitat conservation.

The surviving major wilderness areas (Hannah *et al.*, 1994; Mittermeier *et al.*, 1998), large relatively undisturbed natural areas, offer the best opportunities for retaining ecosystem and evolutionary processes. Beyond these areas, plant conservation will depend upon a number of core skills: (1) the protection and active management of habitats to maintain plant diversity and ecological processes, (2) the management of individual plant populations to retain viable populations, and (3) dealing with the human context of plant conservation,

including sustainable use, cultural value, and public access.

Single-species management for a threatened species can take a number of forms: (1) management of harvesting (both illegal and legal), (2) protection from invasive organisms and pathogens, (3) habitat modification and management, e.g., prescribed burning, (4) reintroduction or translocation, (5) artificial propagation (preferably on-site and in-country). Species as the compositional unit of a community or ecosystem are a convenient and discrete unit of management, particularly where that taxa is threatened and requires species-specific management. Species recovery, the management activities required to halt or reverse the decline in a threatened species' population, can be best achieved through cross-disciplinary recovery teams applying the recommendations of a Species Recovery Plan (Bowles and Whelan, 1994). Such a program should have clear, numerically based objectives utilizing, where appropriate, information from ecological, taxonomic, genetic, and social studies. A Species Recovery Plan provides a forum to bring all required expertise together to ensure a balanced integrated approach to species conservation.

## VI. EX SITU CONSERVATION

Increasingly, *ex situ* collections are promoted as integral components of conservation programs for both wild taxa and crop genetic resources. The world's *ex situ* plant resources encompass the traditional garden plots of the tropics, the intensive allotments, farms and gardens of Europe, and the global networks of seed banks and botanic gardens. The world's network of plant genetic resource collections (e.g., seed banks and field gene banks) contain relatively few species, but with high levels of infraspecific sampling. The network maintains some 6 million accessions, representing mostly agricultural crop cultivars and landraces. The most important 150 major crop gene banks are maintained, as part of the global network of the International Agricultural Research Centers (IARCs), under the umbrella of the Consultative Group on International Agricultural Research (CGIAR). Important centers include the International Rice Research Institute in the Philippines and the International Wheat and Maize Improvement Centre (CIMMYT) in Mexico. The total annual cost of maintaining all accessions currently in gene banks is estimated at \$300 million per annum, with many facilities suffering a chronic lack of resources.

The largest collections of noncrop wild species are maintained by the world's botanic garden networks; cur-

rently they are estimated to maintain 4 million accessions, representing 80,000 taxa. The majority of these collections are maintained as living plants in mixed collections serving a wide range of purposes (Maunder, 1994). There are approximately 1700 botanic gardens in 148 countries; however, over 40% of these are concentrated in western Europe and North America. These collections cultivate a skewed representation of the world's wild diversity, but can be relatively comprehensive at the generic level for attractive and horticulturally amenable families such as the palms, cacti, bromeliads, and orchids. The levels of infraspecific genetic diversity are low since most species are represented by relatively few individuals, often from a limited number of introductions. These collections contain specimens of threatened species, including species now extinct in the wild, e.g., *Sophora toromiro* (Maunder *et al.*, 1999) (Fig. 2). Botanic gardens should be recognized as vitally important resources of horticultural skills and conservation biology research that can support the imperatives of *in situ* management of wild populations and conservation of economically useful plants (Fig. 7), and as readily accessible venues for public education. Botanic gardens have a fundamental role in public education as accessible and often urban venues for introducing the public to plant diversity. Large collections of horticultural plants, mostly cultivars, are maintained by commercial nurserymen and amateur horticulturists. In Europe, the collections closely overlap with botanic gardens at the taxonomic level for hardy herbaceous and woody taxa, but in contrast with botanic gardens, the level of documentation is often poor. Some countries have established national collections of garden plants; such networks exist in Australia, France, and the UK. The National Council for the Conservation of Plants and Gardens (NCCPG) in the UK coordinates over 600 collections, maintained by professional and amateur horticulturists, containing 13,000 species and 39,000 cultivars. Such networks can also serve noncommercial crop resources; the Seed Savers Exchange in the United States exchanges over 5000 crop varieties.

## VII. TOWARD AN INTEGRATED APPROACH

Current trends of both habitat loss and conversion have promoted discussion on the most effective responses to retaining both the taxonomic and ecological components of biodiversity. Conservation biologists have realized that plant conservation management needs can be targeted at distinct, but interlinked, levels of the

biodiversity hierarchy, namely, whole systems at the landscape or ecosystem level, habitats, species, populations, and genes. Superimposed upon these patterns will be an increasing knowledge about plant phylogeny, allowing conservation issues to be guided by evolutionary perspectives.

Traditionally, conservation professionals have operated within isolated and distinct hierarchies; for instance, the ecosystem-level conservationists (protected areas) did not interact greatly with *ex situ* practitioners with a focus on the population/gene level (plant genetic resources and botanic gardens). Accordingly, plant conservation practice and debate have traditionally been polarized between *in situ* activities and *ex situ* activities. Recently, a number of authors have recognized that plant conservation strategies need to utilize a variety of complementary techniques, "integrated strategies" *sensu* Falk *et al.* (1996). Recovery Planning for threatened plant species is a relatively recent development, with the first United States plant recovery plan initiated in 1979. The later production of regional plant conservation strategies reflects this integrated approach, for instance, the Federal Native Plant Conservation Initiative of the United States and the Australian Network for Plant Conservation.

Habitat conversion as the major threat to biodiversity produces an "extinction debt," a pool of species destined for extinction unless the habitat is repaired or restored (Tilman *et al.*, 1994). While stocks of such taxa could be maintained *ex situ* for a future reintroduction, the physical and technical capacity does not currently exist to hold large numbers of taxa or individuals over time. In addition, the longer material is held in cultivation *ex situ*, the more likely it is that genetic modification will take place (Fig. 1). The restoration of those threatened habitats, particularly areas adjacent to protected areas, will provide a vital means of enabling a significant number of species to recover.

## VIII. FACILITIES AND SKILLS FOR PLANT CONSERVATION

Plant conservation has often been marginalized as a peripheral activity for government agencies involved in protected area or forestry management, often as a subset of "wildlife" management. The establishment of national plant conservation networks is playing an important role through (1) promoting an integrated approach to plant conservation, utilizing and promoting the available professional skills; (2) developing collaborative re-

### Box 3

#### Case Study: East Usambara Mountains, Tanzania—The Management of Diversity in a Global Hot Spot

The East Usambara Mountains are part of the Eastern Arc Mountains, a chain of nonvolcanic mountains running from the Taita Hills in Kenya south to the Uzungwa massifs in Tanzania. The East Usambara are a group of low mountains close to the northeastern Tanzanian coast, with the main range only ca. 40 km long by 10 km wide and bound by steep escarpments. The East Usambaras are a matrix of lowland semideciduous and evergreen submontane tropical forest and a recognized center of endemism for birds, vascular plants, bryophytes, reptiles, amphibians, and invertebrates (Burgess *et al.*, 1998). Specific botanical studies have shown that the Usambara Mountains contain an important concentration of endemic vascular plant species. The flora of the East Usambaras consist of 1921 indigenous vascular plant taxa, of which 64 (3.3%) are strict endemics. The numbers of endemic plant species, including endemic African violet species, has led to the East Usambara Mountains being designated an Afromontane Regional Centre of Endemism (CPD Site AF71). Forest loss and degradation in the Usambaras has been severe. The area of natural closed high forest has declined as a result of colonial timber extraction and plantation establishment and more recently through clearance for village agricultural plots. The conservation of the East Usambara forests is dependent upon their retention as valued watershed areas, as resources of local resources, e.g., timber and medicinal plants, and as internationally valued biodiversity resources, e.g., habitat for wild African violets. The East Usambara Catchment Forest Project and Tanzanian Department of Forestry and Bee Keeping are working on local initiatives but with the backing of international finances. International funding for conservation can only succeed if complementing locally supported initiatives.

lationships with protected area networks, government agencies, parastatals, and NGOs; (3) establishing a network of research/conservation facilities in different climatic/vegetation zones; (4) developing in-country



taxonomic expertise; (5) initiating a national system for identifying plant and habitat conservation priorities based on information gathering and monitoring; and (6) fostering institutional strengthening through locally focused professional training.

A number of international networks have established for plant conservation the plant genetic resources network under the International Plant Genetic Resources Institute (IPGRI) and the wild species network under the Species Survival Commission of the World Conservation Union (SSC/IUCN). While there is continued need to further consolidate international networks, a more urgent need is the establishment of local and regional networks to support and direct local frontline initiatives. Examples include the Korean Plant Specialist Group of the SSC, a voluntary network that links government, NGO, and academic interests in plant conservation.

The Center for Plant Conservation (CPC) in the United States can be viewed as one model for a national network. The center is based at the Missouri Botanical Garden, St. Louis, MO, and was established in 1984 as a national network of collaborating botanic gardens with the clear focus on conserving threatened native flora. The CPC has promoted the utilization of both *in situ* and *ex situ* techniques. Fundamental to the effectiveness of the CPC has been the availability of quality data on distribution and status of botanical diversity; this has been heavily dependent upon the activities of the Nature Conservancy's Heritage Programs. In addition, CPC has produced two keystone references for plant conservation, namely, *Genetics and Conservation of Rare Plants* (Falk and Holsinger, 1991) and *Restoring Diversity: Strategies for Reintroduction of Endangered Plants* (Falk *et al.*, 1996). This has resulted in the development of working collaborations between protected area authorities, government agencies, and the adoption of population genetics as a working tool for botanic garden conservation activities. National networks for plant conservation are now established in a number of countries, including Australia, Indonesia, and Canada.

There is an urgent need to incorporate plant conservation into national conservation infrastructures and in particular foster integral links with protected area agencies. The recent implementation of Biodiversity Action Plans is a great opportunity to consolidate the plant conservationist's national role. Plant conservation needs to identify specific national and local incentives and sanctions that can be used as linkages between the central government institutions (the legal infrastructure) and the public sector (the financial infrastructure). Wild plants will continue to be an integral component

of rural life for millions of people. Traditional resource management practices will need to be protected and given a contemporary economic and legislative relevance, e.g., the cooperative government and NGO management of habitats in Brazil, Belize, and India.

However, these activities must be built upon sound science and, particularly, an understanding of how plant populations respond to change. The future challenge is to maintain plant populations, as evolutionary lineages, as ecological components of functioning landscapes, and as valued economic resources, within a changing ecological and political climate. For instance, the 1980s saw the collapse of one of the world's most effective protected area and plant genetic resource networks during Russia's transition from communism to a free-market economy. The increased need for species and habitat management will require greater continuity of information between successive management regimes.

Plant conservation, encompassing the needs for managing increasingly fragmented habitats and populations within human-dominated landscapes (Box 3; Fig. 6), will face a number of challenges: (1) Sound conservation solutions must include strong elements of social science, resource economics, and commercial practice. (2) An "theoretical" awareness of the need for sustainable-use practices alone does not necessarily change practices. (3) Many important plant habitats are subject to increasing levels of encroachment by settlement and cultivation as pressure for land forces people further into protected areas. These areas are increasingly becoming ecological islands, and many of their enclosed species' populations will increasingly face issues of viability. (4) As land between protected areas is progressively modified by human activities and its suitability for sustaining original levels of biodiversity will decline, the opportunities for linkage between prime habitats in parks will vary. However, it is evident that modified habitats can contribute to maintaining connectivity between protected areas, and increased research is urgently needed on this topic. (5) While there have been efforts to "soften" the edge between parks and the lands outside them, there is a continued need to develop schemes by which protected area resources and revenues are shared with neighboring communities.

Plant conservation could easily become focused on salvaging the lost (Box 2). Future investment in scientific and practical activity should focus increasingly on the retention of viable habitat areas and the management of socially and economically important plant resources. While a proportion of the world's plant diversity can be saved as "threatened species," funded as abstracted and emotionally charged conservation projects, the vast ma-

jority of plant diversity will depend on a pragmatic balance between conservation and local utilization. This will be dependent upon establishing effective monitoring and conservation tools that can be locally adopted and modified.

## Bibliography

- Barzeti, V. (1993). *Parks and Progress: Protected Areas and Economic Development in Latin America and the Caribbean*. IUCN (Protected Areas Programme), Gland, Switzerland.
- Bowles, M. L., and Whelan, C. J. (1994). *Restoration of Endangered Species: Conceptual Issues, Planning and Implementation*. Cambridge Univ. Press, Cambridge, UK.
- Burgess, M. A. (1994). Cultural responsibility in the preservation of local economic plant resources. *Biodiversity Conserv.* 3, 126–136.
- Burgess, N. D., Clarke, G. P., and Rodgers, W. A. (1998). Coastal forests of Eastern Africa: Status, endemism patterns and their potential causes. *Biol. J. Linnean Soc.* 64(3), 337–367.
- Casas, A., and Caballero, J. (1996). Traditional management and morphological variation in *Leucaena esculenta* (Fabaceae: Mimosoideae) in the Mixtec region of Guerrero, Mexico. *Econ. Bot.* 50(2), 167–181.
- Davis, S. D., Heywood, V. H., and Hamilton, A. C. (Eds.) (1994). *Centres of Plant Diversity: A Strategy for Their Conservation. Europe, Africa, South West Asia and the Middle East, Vol. 1*. IUCN/WWF, Gland, Switzerland.
- Davis, S. D., Heywood, V. H., and Hamilton, A. C. (Eds.) (1996). *Centres of Plant Diversity: A Strategy for Their Conservation. Asia, Australia and the Pacific, Vol. 2*. IUCN/WWF, Gland, Switzerland.
- Davis, S. D., Heywood, V. H., and Hamilton, A. C. (Eds.) (1997). *Centres of Plant Diversity: A Strategy for Their Conservation. The Americas, Vol. 3*. IUCN/WWF, Gland, Switzerland.
- Falk, D., Millar, C. I., and Olwell, M. (Eds.) (1996). *Restoring Diversity: Strategies for Reintroduction of Endangered Plants*. Island Press, Washington, D.C.
- Given, D. R. (1994). *Principles and Practices of Plant Conservation*. Timber Press, Portland, OR, and Chapman & Hall, London/New York.
- Glowka, L., Burhenne-Guilman, F., Syngé, H., McNeely, J. A., and Gündling, L. (1994). *A Guide to the Convention on Biological Diversity*. Environment Policy and Law Paper No. 30. IUCN, Gland, Switzerland.
- Greuter, W. (1994). Extinctions in Mediterranean areas. *Philos. Trans. R. Soc. London*, B 344, 41–46.
- Grove, R. (1996). *Green Imperialism: Colonial Expansion, Tropical Island Edens and the Origins of Environmentalism, 1600–1860*. Cambridge Univ. Press, Cambridge, UK.
- Hall, P., Walker, S., and Bawa, K. (1996). Effect of forest fragmentation on genetic diversity and mating system in a tropical tree, *Pithecellobium elegans*. *Conserv. Biol.* 10(3), 757–768.
- Hannah, L., Lohse, D., Hutchinson, C., Carr, J. L., and Lankerani, A. (1994). A preliminary inventory of human disturbance of world ecosystems. *Ambio* 23, 246–250.
- Harris, L. D., and Silva-Lopez, G. (1992). Forest fragmentation and the conservation of biological diversity. In *Conservation Biology: The Theory and Practice of Nature Conservation, Preservation and Management* (P. L. Fiedler and S. K. Jain, Eds.), pp. 197–237. Chapman & Hall, London/New York.
- Hughes, J. B., Daily, G. C., and Ehrlich, P. R. (1997). Population diversity: Its extent and extinction. *Science* 278, 689–692.
- Koopowitz, H., Thornhill, A., and Anderson, M. (1994). A general model for the prediction of biodiversity losses based on habitat conversion. *Conserv. Biol.* 8(2), 425–438.
- Lucas, G. Ll., and Syngé, H. (1978). *The IUCN Plant Red Data Book*. IUCN, Morges, Switzerland.
- MacDonald, I. A. W., and Cooper, J. (1995). Insular lessons for global biodiversity conservation with particular reference to alien invasions. In *Islands: Biological Diversity and Ecosystem Function* (P. M. Vitousek, L. L. Loope, and H. Adersen, Eds.), pp. 189–203. Springer-Verlag, Berlin/New York.
- Mackinnon, J., and Mackinnon, K. (1986a). *Review of the Protected Areas System in the Afrotropical Realm*. IUCN, Gland, Switzerland.
- Mackinnon, J., and Mackinnon, K. (1986b). *Review of the Protected Areas System in the Indo-Malaysian Realm*. IUCN, Gland, Switzerland.
- Martini, A. M. Z., Rosa, N. A., and Uhl, C. (1994). An attempt to predict which Amazonian tree species may be threatened by logging activities. *Environ. Conserv.* 21(2), 152–162.
- Maunder, M. (1994). Botanic gardens: Future challenges and responsibilities. *Biodiversity Conserv.* 3, 97–103.
- Maunder, M., Culham, A., Bordeu, A., Allanguillame, J., and Wilkinson, M. (1999). Genetic diversity and pedigree for *Sophora toromiro* (Leguminosae): A tree extinct in the wild. *Mol. Ecol.* 8, 725–738.
- Maunder, M., Upson, T., Spooner, B., and Kendle, T. (1995). Saint Helena: Sustainable development and conservation of a highly degraded island ecosystem. In *Islands: Biological Diversity and Ecosystem Function* (P. M. Vitousek, L. L. Loope, and H. Adersen, Eds.), pp. 205–217. Springer-Verlag, Berlin/New York.
- Mittermeier, R. A. (1988). Primate diversity and tropical forest: Case studies from Brazil, Madagascar and the importance of megadiversity countries. In *Biodiversity* (E. O. Wilson, Ed.), pp. 145–154. National Academy Press, Washington, D.C.
- Mittermeier, R. A., Myers, N., Thomsen, J. B., da Fonseca, G. A. B., and Olivieri, S. (1998). Biodiversity hotspots and major tropical wilderness areas: Approaches to setting conservation priorities. *Conserv. Biol.* 12(3), 516–519.
- Morse, L. E., Kutner, L. S., and Kartesz, J. T. (1995). Potential impacts of climate change on North American flora. In *Our Living Resources: A Report to the Nation on the Distribution, Abundance, and Health of U.S. Plants, Animals and Ecosystems* (E. T. LaRoe, G. S. Farris, C. E. Plunkett, P. D. Doran, and M. J. Mac, Eds.). U.S. Department of the Interior, National Biological Service, Washington, D.C.
- Myers, N. (1988). Threatened biotas: Hotspots in tropical forests. *Environmentalist* 8, 1–20.
- Myers, N. (1990). The biodiversity challenge: Expanded hotspot analysis. *Environmentalist* 10, 243–255.
- Nabhan, G. (1989). *Enduring Seeds*. North Point Press, San Francisco.
- Packard, S., and Mutel, C. F. (Eds.) (1996). *The Tallgrass Restoration Handbook for Prairies, Savannas and Woodlands*. Island Press, Washington, D.C.
- Prance, G. T., and Elias, T. S. (Eds.) (1977). *Extinction Is Forever*. New York Botanical Garden, New York.
- Saunders, D. A., Hobbs, R. J., and Ehrlich, P. R. (Eds.) (1993). *The Reconstruction of Fragmented Habitats: Global and Regional Perspectives*. Surrey Beatty & Sons, Ltd., Chipping Norton, Australia.

- Simmons, J. B., Beyer, R. I., Brandham, P. E., Lucas, G. L., and Parry, V. (Eds.) (1976). *Conservation of Threatened Plants*. Plenum, London.
- Soulé, M. E. (1991). Conservation: Tactics for a constant crisis. *Science* 253, 744–750.
- Synge, H., and Townsend, H. (Eds.) (1979). *Survival or Extinction: The Practical Role of Botanic Gardens in the Conservation of Rare and Threatened Plants*. The Bentham–Moxon Trust, Royal Botanic Gardens, Kew.
- Tilman, D., May, R. M., Lehman, C. L., and Nowak, M. A. (1994). Habitat destruction and the extinction debt. *Nature* 371, 65–66.
- Turner, B. L., Clark, W. C., Kates, R. W., Richards, J. F., Mathews, J. T., and Meyer, W. B. (Eds.) (1990). *The Earth as Transformed by Human Action*. Cambridge Univ. Press, Cambridge, UK.
- Vitousek, P. M. (1994). Beyond global warming: Ecology and global change. *Ecology* 75(7), 1861–1876.
- Walter, K. S., and Gillett, H. J. (1998). *1997 IUCN Red List of Threatened Plants*. World Conservation Monitoring Centre and IUCN–World Conservation Union, Gland, Switzerland.
- Weeks, W. W. (1996). *Beyond the Ark: Tools for an Ecosystem Approach to Conservation*. Island Press, Washington, D.C.
- Wilson, E. O. (1988). The current state of biological diversity. In *Biodiversity* (E. O. Wilson, Ed.). National Academy Press, Washington, D.C.
- Wilson, E. O. (1992). *The Diversity of Life*. Belknap Press, Cambridge, MA.
- World Conservation Monitoring Centre. (1992). *Global Biodiversity: Status of the Earth's Living Resources*. Chapman & Hall, London/New York.
- WRI/UNEP/UNDP/World Bank. (1996). *World Resources 1996–1997*. Oxford Univ. Press, Oxford, UK.
- Young, A., Boyle, T., and Brown, T. (1996). The population genetic consequences of habitat fragmentation for plants. *Trends Ecol. Evol.* 11(10), 413–418.



# PLANT HYBRIDS

Robert S. Fritz  
Vassar College

---

- I. Plant Hybridization
  - II. Plant Hybrid Zones
  - III. Genetic Basis of Hybrid Plant Resistance
  - IV. Communities on Plant Hybrids
  - V. Population and Evolutionary Consequences
  - VI. Conservation of Hybrid Plants
  - VII. Conclusions
- 

## GLOSSARY

**allopolyploid (amphiploid)** Species with chromosome sets derived from interspecific hybridization and doubling of chromosomes of the sterile hybrid, which restores fertility.

**dependent community** The group of species that requires a particular host plant to complete some or all of their life cycle.

**diploid hybrids** Species formed from hybridization where chromosomes from species of equal chromosome numbers become stabilized through recombination.

**hybrid** An individual that is heterozygous (intermediate) for one or more heritable characters that distinguish two or more populations, including individuals that are  $F_1$ s,  $F_2$ s, and the set of all backcrosses. Hybrids are initially formed when gametes from two species, subspecies, or races combine to form  $F_1$  plants.

**hybridization** Interbreeding of individuals from two or more populations of species, subspecies, or races,

which are distinguishable by one or more heritable characters.

**hybrid swarm** Mixture of hybrid genotypes, including  $F_1$ s,  $F_2$ s, and backcrosses, due to hybridization between two or more species that co-occur in a locality.

**hybrid zone** Areas of overlap or points of contact between two populations that are distinguishable based on one or more heritable characters where viable or partially fertile hybrids are formed.

**introgression** Permanent transfer of genes from one or more species, subspecies, or race into another species, subspecies, or race via hybridization and backcrossing.

**reticulate evolution** Pattern of speciation where new species arise from interspecific hybridization between two species coupled with establishment of reproductive isolation.

---

**PLANTS HAVE PERVASIVE** effects on biological diversity. The diversity and structure of plants are used to define all ecological communities. But it is the biological diversity of organisms that live on plants, their herbivores, pathogens, pollinators, dispersers, root and leaf mutualists, and the natural enemies of these organisms that contributes the most to total biological diversity. Interspecific hybridization of plants contributes to biological diversity in at least four fundamental ways. First, new species or subspecies of plants are formed from hybridization via several mechanisms. Second, hybrid-

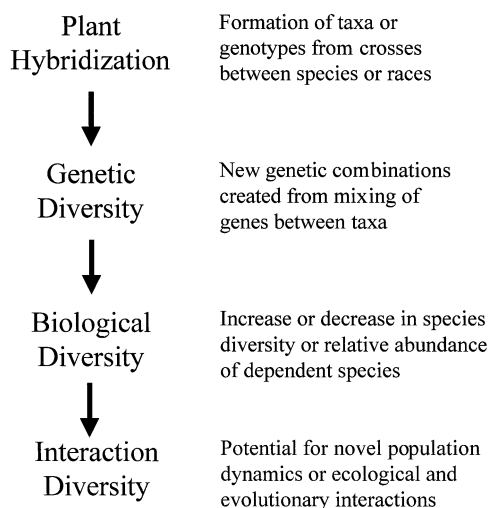


FIGURE 1 Conceptual model describing the cascade of effects of plant hybridization on genetic, biological, and interaction diversity.

ization leads to persistent hybrid populations that exist in a variety of hybrid zones. Third, introgression increases genetic diversity in populations, but complete introgression could lead to the loss of diversity. Fourth, biological diversity is increased by plant hybridization and introgression through its influence on the abundance and diversity of herbivores and their natural enemies, and their pathogens. This article will consider the formation of plant hybrids, the structure of hybrid zones, the effects of introgression on genetic diversity, and the distribution of hybrid plants in nature. However, the primary focus of the article will be on the effects of plant hybrids on the biological diversity and community structure of higher trophic levels, herbivores, their natural enemies, and pathogens. The conceptual model proposed in this article suggests that plant hybridization contributes to genetic diversity of plant populations, that genetic diversity increases biological diversity, and that biological diversity creates new opportunities for interaction diversity (Fig. 1). The proposed cascade of effects flowing from plant hybridization is discussed in the following sections.

## I. PLANT HYBRIDIZATION

### A. Occurrence of Hybridization in Plants

Interspecific hybridization is a common and important evolutionary mechanism in plants. Allopolyploid or homoploid (diploid) speciation gives rise to reticulate phylogenies. Estimates of the proportion of angiosperm

species derived from interspecific hybridization via polyploidy, of which allopolyploidy far predominates, are between 47 and 52% for angiosperms, 43% of ferns, and up to 95% of all pteridophytes (Grant, 1981). Species formed in this way occupy new habitats and can have expanded geographical ranges compared to the parental species. For example, *Tragopogon mirus* and *T. miscellus*, tetraploid species, have expanded ranges compared to their diploid parents.

Hybrid species may originate multiple times. Among polyploid hybrid species investigated using molecular means, the minimum number of recurrent origins of hybrid species ranges from 2 to 13. Thus, multiple origins appear to be the rule rather than the exception in polyploid species. This generates substantial genetic diversity, thereby providing greater opportunity for subsequent evolutionary change. Genetic mechanisms of gene silencing (sometimes leading to diploidization of polyploid genomes), gene diversification (divergence of duplicate genes), and genome diversification (repat- terning of chromosomes) operate in polyploid species, but knowledge of how widespread these mechanisms are in polyploid-derived species is unknown.

The proportion of families and genera where interspecific hybrids have been documented has been estimated for the floras of the British Isles, Scandinavia, the Great Plains, the Intermountain West of North America, and Hawaii. The proportions of families and genera with hybrids ranged from 16 to 34% and from 6 to 16%, respectively, and the numbers of hybrids identified in these floras ranged from 134 to 642. These values indicate that hybridization is common and naturally occurring hybrid zones are numerous.

Within these floras, certain families and genera have disproportionate numbers of hybrids. The families Asteraceae, Cyperaceae, Onagraceae, Poaceae, Rosaceae, Salicaceae, and Scophulareaceae and the genera *Bidens*, *Carex*, *Cyrtandra*, *Euphrasia*, *Rosa*, and *Salix*, among others, have large numbers of hybrids between various species relative to their representation in the flora. Traits that appear to promote the occurrence of hybrids in families and genera are perenniality, outcrossing, and vegetative or clonal reproduction. Outcrossing increases the likelihood of interspecific crossing and the other traits ensure that hybrids persist long enough to reproduce.

### B. Hybrid Complexes

Groups of species within a region that are connected by degrees of hybridization are called syngameons. Syngameons have been identified in *Betula*, *Geum*, *Iris*,

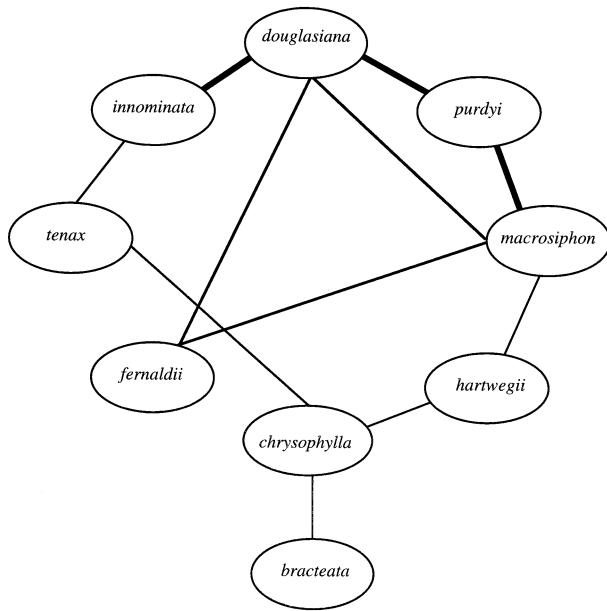


FIGURE 2 Syngameon of Pacific coast *Iris* species. The width of lines connecting species pairs is proportional to the amount of natural hybridization. (After Lenz, L. W. (1959). Hybridization and speciation in the Pacific coast irises. *Aliso* 4, 237–309.)

*Juniperus*, *Melandrium*, *Nothofagus*, *Pinus*, *Quercus*, and *Salix*. The extent of hybridization between pairs of species in syngameons varies, with some species having abundant gene flow and other pairs having limited gene flow (Grant, 1981). The web of hybridization provides the opportunity for species that do not hybridize directly to share genes via hybridization with a third species. A syngameon between irises along the Pacific coast of the United States is illustrated in Fig. 2.

### C. Types of Hybrids

Hybrids are formed when gametes from two species, subspecies, or races combine to form  $F_1$  plants.  $F_1$ s, if they are sterile, may be the only hybrid genotypes found in some hybrid zones (e.g., the sedges *Carex canescens*  $\times$  *C. mackenziei* in Scandinavia). They may persist by establishment of clones or by recurrent formation. In other populations where hybrids are common, formation of  $F_1$ s may be a rare event. In Louisiana irises, formation of  $F_1$ s between *I. fulva* and *I. hexagona* is rare (none are found in nature, though  $F_1$ s can be readily made artificially) (Arnold, 1997), but the rare  $F_1$ s serve to produce backcross hybrids that have high fitness and are common in some habitats. Often,  $F_1$  hybrids are intermediate in morphological characteristics to parent species.

$F_2$  hybrids have greater phenotypic variability than do  $F_1$  hybrids due to recombination that occurs in the formation of gametes. Combinations of parental traits present in  $F_1$  hybrids are dissociated in formation of  $F_2$  hybrids. The resulting phenotypic and genotypic variation spans the range between each parent's phenotype, with many  $F_2$ s resembling intermediate forms. This variation influences the fitness of these hybrids and the responses of herbivores and pathogens.

Progeny formed from crosses between  $F_1$ s and a parent species (P) are known as first backcrosses ( $BC_1$ ) (crosses of  $BC_1 \times P$  gives rise to  $BC_2$ , etc.). Backcrosses may be formed between hybrids and one or both of the parental species, resulting in unidirectional or bidirectional backcrossing, respectively. Unidirectional backcrossing is known in hybridization of *Populus fremontii* and *P. angustifolia*, with *P. angustifolia* being the recurrent parent (the parent species to which the  $F_1$  crosses). Preferential gene flow occurs from *Iris fulva* to *I. brevicaulis* through backcrossing, although backcrosses in both directions occur. In the willows *Salix sericea* and *S. eriocephala*, backcrosses occur to both parents. Backcross hybrids contain more of the recurrent parent's genetic material and thus tend to resemble that parent. Recombination creates substantial phenotypic and genetic variation among backcross progeny.

Hybrid swarms occur when extensive hybridization occurs within and among hybrid genotypes and parent species, creating a broad range of hybrid genotypes and phenotypes. Location of a putative hybrid within the range of variation between parents is usually made using a hybrid index based on plant morphology, and often this shows continuous variation in plant traits between parental species. Hybrid index methods based on morphology assume that parental traits are independent of each other (there is linkage equilibrium and lack of pleiotropy) and that traits are additively inherited. However, traits are not always independent and dominance of parental traits and transgressive morphologies make this method questionable. Molecular methods provide many more markers that are independent for creating hybrid indices that are useful for ecological studies (Rieseberg and Ellstrand, 1993).

Repeated backcrossing of hybrids with a parental species can permanently transfer genes from one species into the genome of another species, a process called introgression. Introgression may be localized or dispersed (Rieseberg and Brunfeldt, 1992). In localized introgression, gene transfer between species is restricted to the area of hybridization between the two species (e.g., the thistles *Carduus nutans*  $\times$  *C. acanthoides*). In dispersed introgression, gene flow occurs beyond the

range of hybridization. Evidence of dispersed introgression has been found for *Pinus banksiana* × *P. contorta* and *Iris fulva* × *I. hexagona*. However, localized introgression appears to be more common than dispersed introgression (Rieseberg and Brunsfeld, 1992). One outcome of introgression is the formation of new species or races. Molecular evidence has revealed that several plant species and races have originated from stabilized introgressants. The Fort Davis race of *Salix taxifolia* has cpDNA of *S. interior* and nuclear markers of *S. taxifolia*, *S. interior*, and as many as three other willow species. Rieseberg and Brunsfeld (1992) document numerous other cases of introgression.

The transfer of genes between plant species can generate substantial genetic variation that can be acted on by natural selection. In the evolutionary novelty model (Arnold, 1997), hybridization coupled with selection is hypothesized to favor transfer of adaptive traits with two possible outcomes. The first is the formation of new species via diploid speciation, which incorporates novel adaptations derived from both parental species. Transfer of similar chromosome segments between species has been demonstrated in three different pedigrees of hybrids between the sunflowers *Helianthus annuus* and *H. petiolaris*, and the resulting chromosome structures resemble the diploid hybrid species, *H. anomalus*. This strongly suggested that selection governed the formation of *H. anomalus* and that coadapted parental gene combinations were maintained by natural selection. The second outcome is adaptive transfer of traits between species, which increases genetic diversity. Thus, hybridization coupled with selection by abiotic and biotic factors is a creative evolutionary process that generates genetic diversity in plant populations.

## II. PLANT HYBRID ZONES

### A. Types of Hybrid Zones

Hybrid zones are locations where hybrids between species, subspecies, or races are found. Typically, hybrid zones are described as clines, spatial gradients in traits or alleles across a geographic transect where two taxa meet. The balance between selection on hybrids and dispersal of genes determines cline width. Clines are expected to be narrow if there is strong selection against hybrids, if gene flow is limited, or if there are steep environmental gradients. Clines will be wider if selection is weak, gene flow is more extensive, or environmental gradients are gradual. Cline shape is predicted to be a smooth, sigmoid curve if selection acts on single

genes or on quantitative traits, but linkage disequilibrium combined with dispersal of genes can distort the smooth shape of the cline, creating a stepped cline. In stepped clines most of the change in allele frequency or trait expression occurs in a narrow range in the middle of a cline. Allele frequency, linkage disequilibrium with other alleles or traits, and gene flow can be used to infer the form by which selection acts in these clines. A clinal hybrid zone is one type of spatial pattern of plant genetic diversity that can influence biodiversity and spatial distribution of animal and pathogen species.

Narrow hybrid zones 10 m wide are found between the sedges *Carex canescens* and *C. mackenziei* at the edge of water along the Bothnian coast in northern Sweden. Oak hybrid zones approximately 30 m wide between the oaks *Quercus depressipes* and *Q. rugosa* are found on steep slopes in northern Mexico. A much wider hybrid zone extending about 20 km occurs in the same area between *Q. coccolobifolia* and *Q. rugosa*.

In contrast to the traditional clinal model of hybrid zones, mosaic hybrid zones occur when habitats of the hybridizing taxa are patchily distributed. Hybridization may occur where the different habitat patches abut, leading to the patchy distribution of hybrids. In contrast to the more or less discrete location of hybrids geographically in the clinal hybrid zones, mosaic hybrid zones can be as widely distributed as the distribution of habitats and parental species. Therefore, the impact of mosaic hybrid zones on the distribution of plant genetic variation and its effects on diversity of communities of herbivores and pathogens are geographically more widespread.

Louisiana irises *Iris fulva*, *I. brevicaulis*, and *I. hexagona* fit the mosaic model since species and hybrid genotypes are associated with different, interspersed habitats. For example, *I. fulva* is associated with maple-dominated forest habitats, *I. brevicaulis* is associated with black oak forests, and *I. hexagona* is found at the edge of freshwater marshes. Hybrid genotypes are either not strongly associated with specific habitats or occupy intermediate or novel habitats. More dispersed mosaic hybrid zones occur where broadly sympatric species hybridize, either occasionally or extensively, where their habitats mix. The sunflowers *Helianthus annuus* and *H. petiolaris* exemplify this type of hybrid zone. These sunflower species have overlapping ranges and form local hybrid swarms in the western United States. Hybrids between *Salix sericea* and *S. eriocephala* also fit the mosaic model, with hybrids being found throughout the sympatric range of these species in eastern North America.

## B. Models of Hybrid Zone Dynamics

Clinal hybrid zones have been of particular interest to evolutionary biologists and a number of models have been proposed to explain them. These models differ in the causes of selection on hybrids and how it varies across the cline. The *hybrid disadvantage* or *tension zone* model proposes that a balance exists between gene flow across a cline and endogenous selection against hybrids due to genetic incompatibilities in hybrid individuals. The *environmental gradient* model proposes that a balance of gene flow exists across a gradient of environmental conditions (exogenous selection), favoring different alleles (from the two species) at opposite ends of the cline. The *hybrid advantage* or *bounded hybrid superiority* model predicts that within a narrow ecotone separating the parental species, hybrids have higher fitness. The *advancing wave* model proposes that selection favors one parental species over the other, so that hybrids have higher fitness than one parent but lower fitness than the other parent species. As the name implies, the zone of hybridization will shift under this model until the inferior species is eliminated or endogenous or exogenous selection results in a balance suggested by one of the previous models. Finally, the *neutral diffusion* model suggests that neutral alleles (alleles with equal fitness effects) mix after secondary contact between parental species but that endogenous or exogenous selection does not act on these alleles in the cline.

## C. Fitness in Hybrid Zones

Analysis of fitness of hybrids across parental and hybrid zones can reveal which of the models of hybrid zone dynamics is operating. The traditional view, derived mostly from studies of animal hybrids, is that hybrids are unfit compared to parental species. The presence of some degree of unfitness of naturally occurring or artificially created hybrids and the presence of coincident and concordant clines of traits are used to conclude that endogenous or exogenous selection acts against hybrids (Arnold, 1997). Lower viability or fertility of some hybrid genotypes is commonly found, suggesting endogenous selection against some hybrids. Higher susceptibility of hybrids to herbivores and pathogens or inferior ability to cope with environmental stress suggests exogenous selection against hybrids (Fritz *et al.*, 1999).

Support for the *hybrid superiority* model was found for intersubspecific hybrids of sagebrush, *Artemisia tridentata* spp. *tridentata* and *A. t.* spp. *vaseyana*. Reciprocal transplants of seeds and seedlings across the narrow

environmental gradient that separates the subspecies in Utah showed that hybrids had higher total fitness in the "hybrid" habitat and that parental subspecies had highest fitness in the parental habitats. Likewise, analysis of fitness of hybrid and parental *Iris* partially supported this model (Arnold, 1997). Fitness of *Ipomopsis aggregata* was greater than that of *I. tenuituba* and their interspecific hybrids across a narrow hybrid zone in the Rocky Mountains of Colorado due to strong hummingbird pollinator preference for *I. aggregata* floral traits, supporting the *advancing wave* model.

A review of fitness of hybrids by Arnold (1997) showed that frequently hybrids do not have lower fitness than parent species, rather that hybrid fitness may be equal to or greater than that of their parents. The genotype of hybrids affects their fitness, but the conclusion that hybrids are uniformly unfit is not generally supported by empirical studies.

## III. GENETIC BASIS OF HYBRID PLANT RESISTANCE

The influence of hybrid plants on biological diversity depends on the resistance of hybrids to phytophages. Resistance is the ability of a plant genotype to avoid attack or prevent the development of a specific herbivore or pathogen relative to other genotypes. Resistance is specific to herbivore or pathogen species and resistance to different herbivores or pathogens are separate traits that may or may not be correlated. Resistance determines the abundance of each species of phytophagous organism on hybrid plants; resistant plants have lower amounts of damage resulting from lower densities of herbivores, whereas susceptible plants experience higher amounts of damage resulting from higher densities of herbivores. Consequently, the resistance of hybrid plants to their phytophages contributes to the biodiversity, community structure, and interactions among phytophages.

### A. Theory

Several hypotheses concerning the resistance of hybrid plants compared to parental species assume that resistance is a polygenic trait, influenced by many genes, and that quantitative genetic models best describe the patterns of herbivore and pathogen response to hybrid plants. Some models specify that hybrids are F<sub>1</sub>s, but other models do not specify the genetic composition



of the hybrids. Box 1 describes the hypotheses that would be supported by comparisons of  $F_1$  hybrids to one or both parents. Providing that tests of resistance are performed in a uniform environment, the outcome of these comparisons is used to infer the genetic basis of resistance in hybrid plants. The additive pattern suggests that resistance genes act in a dosage-dependent

manner (Fig. 3A). Dominance of either susceptibility or resistance suggests dominant effects of genes from one parent species (Fig. 3B). Dominance of susceptibility implies that hybrids have similar attractant or performance traits as the equally susceptible parent. Dominance of resistance implies that hybrids have similar repellent or antibiosis traits as the resistant parent. Hybrid susceptibility (Fig. 3C) indicates that traits that condition plant resistance fall below a level that deters herbivores or pathogens, a response called hybrid breakdown. Hybrid resistance (Fig. 3D) occurs when hybrids are more resistant than either parent species, a response called heterosis. For this pattern, different resistance genes from each parent may be dominantly inherited and act together in their effects on herbivores.

Formation of  $F_2$  or backcross plants can have further effects on plant resistance. Backcrosses may be intermediate between the  $F_1$  and recurrent parent (Figs. 3A, 3C, and 3D). This would suggest that resistance traits act additively. Genetic recombination may lead to hybrid breakdown in these hybrid classes (higher susceptibility) compared to  $F_1$  and parents, as shown in backcross to Species A in Fig. 3B. Such patterns imply the breakup of coadapted gene complexes (several resistance genes evolved to work together in a species). Breakup of coadapted gene complexes may be one cause of hybrid susceptibility in  $F_2$  or BC plants. Backcrosses could resemble the recurrent parent (Species B in Figs. 3C and 3D), suggesting recovery of resistance or susceptibility factors above a threshold level. As with backcrosses,  $F_2$  resistance needs to be compared to parent and  $F_1$  progeny to document hybrid breakdown or some other pattern.

Regression analyses are useful in examining the patterns of resistance where continuous variation in hybrid genotypes exists, such as in hybrid swarms. Measures of herbivore abundance are plotted against hybrid indices and the best fitting regression models are determined. Significant linear, quadratic, and cubic regressions support the additive, susceptibility and resistance, and dominance hypotheses, respectively.

When resistance is due to single genes that have a major effect, segregation of discrete resistance phenotypes may occur. Gene-for-gene theory of the disease resistance of plants was based on the recognition of discrete infection types in the flax-flax rust (*Linum usitatissimum*-*Melampsora lini*) pathosystem, but discrete resistance phenotypes also frequently characterize plant parasites in natural plant-pathogen interactions.

Mendelian hypotheses of simple inheritance can be tested when distinct phenotypic classes of resistance are present in controlled crosses.  $F_1$  progeny of a controlled

## Box 1

### Predicted Patterns of Plant Hybrid Resistance to Herbivores and Pathogens

Pattern	Explanation
Additive	Abundance of an herbivore species on hybrids is intermediate between herbivore abundances on the two parental species.
Dominance	Abundance of an herbivore species on hybrids is equal to that of one parental species and is significantly different from the other parental species. If the abundance on hybrids is similar to that of the more susceptible, parent, the pattern is <i>dominance of susceptibility</i> . If the abundance on hybrids is similar to that of the more resistant parent, the pattern is <i>dominance of resistance</i> .
Susceptibility	Abundance of an herbivore species on hybrids is significantly higher than on each individual parental species.
Resistance	Abundance of an herbivore species on hybrids is significantly less than on each individual parental species.
Partial dominance	Abundance of an herbivore on hybrids is intermediate compared to both parents, but there is a significant deviation toward one of the parents.
No difference	Abundance of herbivores on hybrids and parents do not differ significantly.

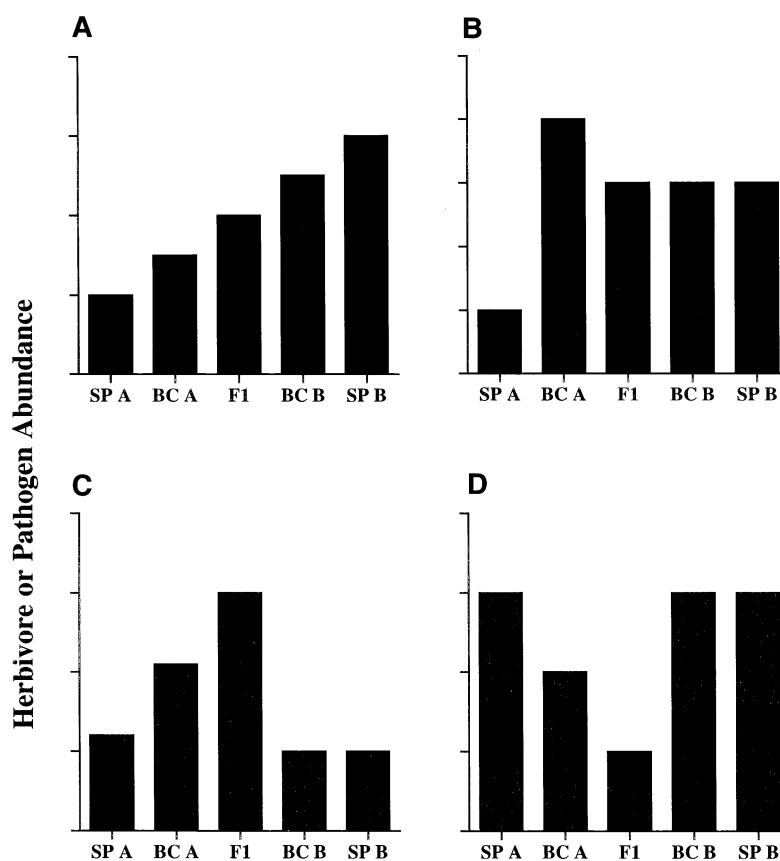


FIGURE 3 Four hypothetical patterns of susceptibility of  $F_1$  hybrid plants compared to their parental species (SP) and backcrosses (BC): (A) additive; (B) dominance; (C) susceptibility; (D) resistance.

cross may be either all susceptible or all resistant, matching the “dominance to susceptible” and “dominance to resistance” hypotheses, respectively. Segregation for resistance to *Venturia populina* and of resistance to the rust *Melampsora medusae* has been reported in  $F_1$  progenies of *Populus trichocarpa*  $\times$  *P. deltoides*. Likewise, segregation of recessive resistance can occur in the  $F_2$  progeny of two susceptible parents. Resistance of *Populus trichocarpa*  $\times$  *P. deltoides* to stem canker caused by *Septoria musiva* segregates in  $F_2$  progenies. *P. deltoides* is resistant, but its  $F_1$  progeny are always susceptible. However, when two susceptible  $F_1$  clones are crossed, resistant  $F_2$  individuals are seen. Conversely, resistance to *Melampsora occidentalis* is dominant, so those interspecific hybrids  $F_1$ s are all resistant. Yet,  $F_2$ s from crosses between these resistant  $F_1$ s produce some susceptible progeny (Fritz *et al.*, 1999). If intermediate resistance phenotypes can be identified, then crosses of  $F_1$ s and analysis of segregation ratios can be used to determine the number of genes responsi-

ble for resistance. Alternatively, correlation of resistance phenotypes with molecular markers can be used to identify the numbers and importance of resistance genes.

## B. Evidence

### 1. Resistance Traits

Inheritance of resistance traits in hybrids will determine the effects of hybridization on susceptibility to herbivores and pathogens. Chemical traits of hybrids are usually found in one or both parental species. Concentrations of phenolic glycosides (derived from *Salix sericea*) and condensed tannins (predominant in *S. eriocephala*) were found in intermediate concentrations in hybrids between these species (Fig. 4). Some chemical traits in hybrids are inherited as dominant traits, but other resistance mechanisms are recessive in hybrids. Pod abscission, a defensive trait of palo verde (*Cercidium microphyllum*) appears to be recessive in hybrids

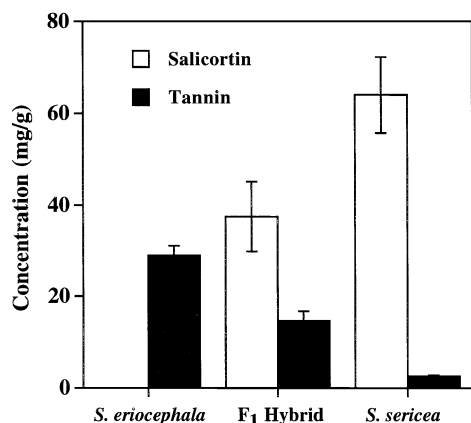


FIGURE 4 Concentrations (mean  $\pm$  1 SE) of salicortin (from *S. sericea*) and condensed tannins (primarily from *S. eriocephala*) in clones of natural hybrids. (From Orians, C. M., and Fritz, R. S. (1995). Secondary chemistry of hybrid and parental willows: Phenolic glycosides and condensed tannins in *Salix sericea*, *S. eriocephala*, and their hybrids. *J. Chem. Ecol.* 21, 1245–1253.)

with *C. floridum*, whereas seed coats of *C. floridum* are chemically defended against bruchid weevil attack and this trait is also recessive in hybrids. However, chemical traits of parents may be missing in hybrids, and hybrid plants may produce novel chemical traits (chemicals found in neither parent species nor parental chemical traits expressed in novel plant tissues). Thus, hybrids and hybrid-derived species can have unique or intermediate resistance characteristics (Rieseberg and Ellstrand, 1993).

Herbivores may show additive, dominant, threshold, or overdominant responses to hybrid traits. Additive responses suggest parallels between herbivore response and variation in a chemical or physical defensive or attractant trait (e.g., gall-inducing sawfly, *Pontania* sp., on willow hybrids), whereas dominant responses suggest inverse response of herbivores to dominant resistance traits (e.g., Japanese beetle (*Popillia japonica*) on willow hybrids) or positive responses of parasites to dominant attractants (e.g., the cynipid gall wasp *Andricus californicus* to *Quercus dumosa*  $\times$  *Q. engelmannii* hybrids).

Increased trichome density, a resistance trait of the alder, *Alnus incana*, may cause aphids to fall from leaves. In the F<sub>1</sub> hybrid, trichome density was intermediate between the two parents, but aphid susceptibility showed a dominance deviation toward *A. glutinosa*, which does not have trichomes, suggesting that a threshold density of trichomes exists to confer resistance that exceeds the density of F<sub>1</sub> hybrids. A threshold model could explain why hybrid willows are susceptible to the imported willow leaf beetle, *Plagioderia versicolora*, when concentrations of phenolic glycosides and tannins are inherited additively.

## 2. Resistance of Hybrid Plants

A review of 156 herbivores and pathogens from 38 plant hybrid systems revealed patterns in how these phytophages responded to plant hybrids (Table I). No difference was a pattern that accounted for 30.4% of the cases in the field and 34.7% of the cases in common

TABLE I  
Numbers (N) and Percentages (P) of Tests That Support Hybrid Resistance Hypotheses from Plant–Herbivore and Plant–Pathogen Studies<sup>a</sup>

Hypothesis	Plants/herbivores							
	Field census		Common garden		Laboratory		Plants/fungi	
	N	P (%)	N	P (%)	N	P (%)	N	P (%)
No difference	17	30.4	17	34.7	3	14.3	0	0
Additive	8	14.3	8	16.3	12	57.1	6	20
Dominance of								
Susceptibility	7	12.5	16	32.7	4	19.0	10	33.3
Resistance	7	12.5	0	0	0	0	4	14.3
Partial dominance								
Susceptible	11	19.6	8	16.3	2	9.5	8	26.7
Resistant	5	8.9	0	0	0	0	2	6.7
Total	56		49		21		30	

<sup>a</sup> From Fritz *et al.*, 1999

garden studies of herbivores, but in laboratory tests, only 14.3% of the cases fit this pattern. The additive pattern was found in 14.3 and 16.3% of the field and common garden studies, respectively, but a much larger proportion of the laboratory studies (57%) found this pattern. The next most common pattern was dominance of susceptibility, which accounted for 12.5, 32.7, and 19% of the three types of studies. This pattern indicates that an herbivore has equal abundance to the more susceptible parent species. Susceptibility, where the herbivore is more abundant than on either parent, accounted for 19.6, 16.3, and 9.5% of the three types of studies. Dominance of resistance was equal to dominance of susceptibility in frequency in the field studies (12.5%), and the resistance pattern occurred in 8.9% of the field studies. Cases of resistance and dominance of resistance did not occur in the common garden and laboratory studies.

The predominant patterns found for plant–pathogen studies of hybrids are dominance of susceptibility, susceptibility, and additive. Dominance of resistance and hybrid resistance seem to be more common in plant–pathogen systems, compared to common garden and laboratory studies of herbivores.

Studies performed in controlled environments point to the fundamental influence of genetics on hybrid plant resistance. No difference, additive, dominance of susceptibility, and hybrid susceptibility to insects are common patterns seen in these studies. Among these common patterns, only hybrid susceptibility would suggest that hybrid plants would have lower fitness compared to both parents, assuming the herbivores lower fitness of their host plants. The absence of dominance of resistance or hybrid resistance patterns indicates that resistance to herbivores is not inherited as a dominant trait in most systems. In contrast, plant–pathogen systems demonstrate several cases of hybrid resistance or dominance of resistance (20% of all studies).

Hybrid susceptibility is illustrated by dramatic increases in herbivore abundance. The bud gall mite, *Aceria parapopuli*, was over 800 times more abundant on F<sub>1</sub>-type hybrids of *Populus fremontii* and *P. angustifolia* than on either parent species. *Pemphigus betae*, a gall-inducing aphid of *Populus*, is found at much higher abundances (8- to 119-fold) in six poplar hybrid zones in river drainages throughout western North America. Severity of infection by the smut pathogen *Anthrocoidea fischeri* was 30- to 80-fold higher on F<sub>1</sub> *Carex canescens* × *C. mackenziei* hybrids than on pure species. Usually, differences in herbivore abundance on hybrids and parents are much less pronounced. Hybrid resistance effects on upper surface leaf and blotch miners in two oak hybrid zones (*Quercus depressipes* × *Q. rugosa*; *Q.*

*coccolobifolia* × *Q. emoryi*) resulted in 3–8 times as many miners on parents compared to the hybrid oaks, and less than twofold higher populations of *Phyllonorycter salicifoliella* occur on willow hybrids.

Even for the same plant hybrid system, several different hypotheses are supported. On the *Salix sericea* × *S. eriocephala* hybrid system, five different hypotheses were supported by different herbivore species. Figure 5 shows the different responses of three herbivores and a leaf pathogen on *Salix sericea*, *S. eriocephala*, and F<sub>1</sub> hybrid willows. Moreover, a guild of leaf gall inducers, three in the same genus, supported four different hypotheses in this system in a single year. This highlights the point that herbivore or pathogen resistance traits are independent traits of the plant that require separate consideration in determining patterns of plant resistance to particular herbivores.

### 3. Genetic and Environmental Effects on Plant Resistance

Studies of plants in the field combine the effects of plant genotype with the range of environmental factors experienced by the plant. Together, genetic and environmental factors can affect plant resistance. The role of environmental variation in determining plant resistance is important since some hybrid zones occur across steep environmental gradients, hybrids may occur at the limits of ranges of each parent species, and some hybrids are found in unique habitats where abiotic conditions differ. Plant stress, induced by environmental extremes, is known to make plants more susceptible to herbivores; thus the *stress hypothesis* proposes that elevated levels of herbivory on hybrid plants are a direct result of stress rather than due to direct genetic effects on plant resistance traits (Whitham *et al.*, 1994; Fritz, 1999). Mortality of pines, regardless of whether they were hybrid or parental, in the hybrid and overlap zones between *Pinus edulis* and *P. californianus* was greater than in the allopatric parental populations, suggesting stress as a factor. F<sub>1</sub> hybrids of *Salix sericea* and *S. eriocephala* are less tolerant to drought than are the pure species. The response of hybrid plants to stress is likely a consequence of their hybridity, and the consequent effects on resistance to herbivores or pathogens is therefore a genotype-by-environment interaction.

### 4. Generalist and Specialist Herbivores

If specialists are defined as herbivores that use only one parent species in a hybrid zone and generalists are defined as using both parent species, then these categories of herbivores are predicted to vary in how they respond to hybrid plants. Hybrid plants, which have some of the traits of the nonhost, should be less

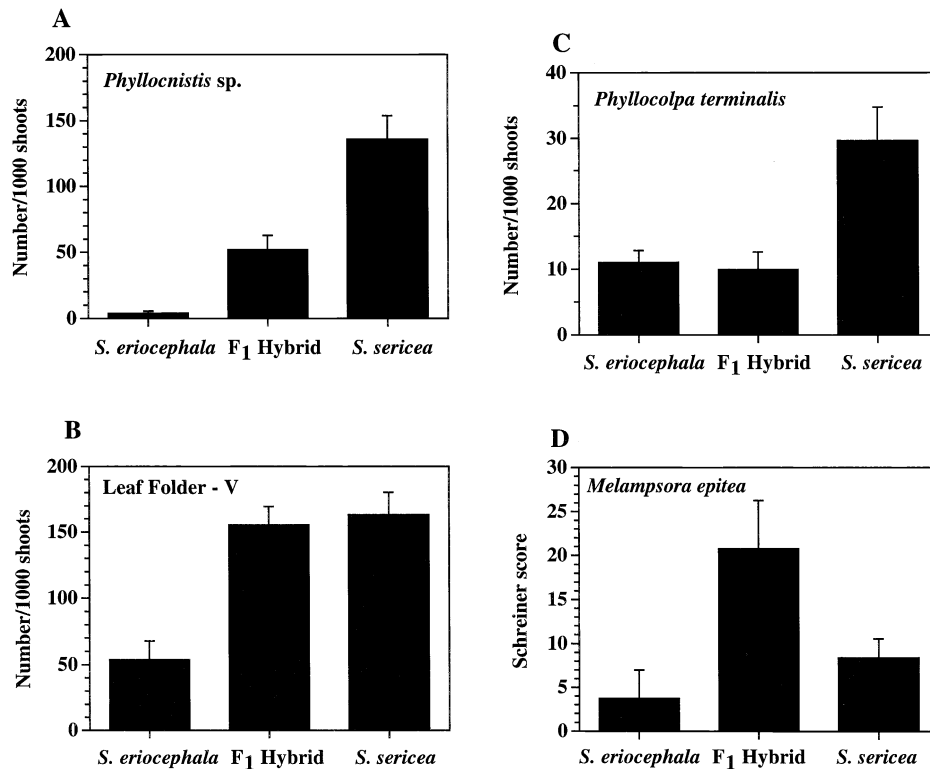


FIGURE 5 Abundances of three herbivores and infection score of a pathogen on 1-year-old seedling willows of *S. sericea*, *S. eriocephala*, and their F<sub>1</sub> hybrids growing in a common garden: (A) *Phyllocnistis* sp.; (B) leaf folder-V; (C) *Phyllocolpa terminalis*; (D) *Melampsora epitea*. (A–C from Fritz *et al.*, 1998; D from Roche, B. M., and Fritz, R. S. (1998). Effects of host plant hybridization on resistance to willow leaf rust caused by *Melampsora* sp. *Eur. J. For. Pathol.* 28, 259–270.)

suitable for specialists, whereas hybrids should be more suitable for generalists, which are already adapted to traits of both species. Fitting these predictions, 92% of generalists (11 of 12 taxa), but only 50% of specialists (14 of 28 taxa), were more abundant on hybrid than parental taxa in the *Eucalyptus amygdalina* × *E. risdonii* hybrid zone. However, there is less consistency among three other generalist species in other systems: dominance of susceptibility was found for a bagworm on sagebrush, dominance of resistance was found for the Japanese beetle on willow, and an additive pattern was found for the spruce budworm on spruce.

Specialist herbivores are predicted to have higher abundance on backcross hybrids that are genetically more similar to their host species when backcrosses are present in populations. Among the 28 specialists of the *E. amygdalina* × *E. risdonii* hybrid zone, 68% (19) were most abundant on the class of hybrids most similar to their host species, 28% (8) were equally abundant on all hybrid classes, and 4% (1) was most abundant on the least similar hybrid category. A gall-forming cynipid, *Andricus californicus*, occurred with equal abundance

on parental (*Quercus dumosa*) and all hybrid genotypes with *Q. engelmannii*, but not on pure *Q. engelmannii*, contrary to the prediction.

## 5. Conclusions

Several conclusions are apparent from these studies. First, resistance traits are inherited in several ways in hybrid plants. Chemical traits may be dominant, additive, or resistant. Moreover, novel plant chemicals or expression of chemicals in new plant tissues can occur in hybrid plants. Second, there is not a single predominant effect of hybridization on plant resistance to herbivores and pathogens. Diversity in the resistance mechanisms of plants to different herbivores or pathogens exists, with dominance of susceptibility, susceptibility, additive, and the no difference being the patterns most frequently reported. Third, genetic effects of hybridization on resistance are well documented and widespread, but the role of environmental variation in modifying the genetic effects is poorly understood. Finally, specialist and generalist insects vary in predictable ways in their abundance on hybrid plants.

## IV. COMMUNITIES ON PLANT HYBRIDS

### A. Diversity of Herbivore Species

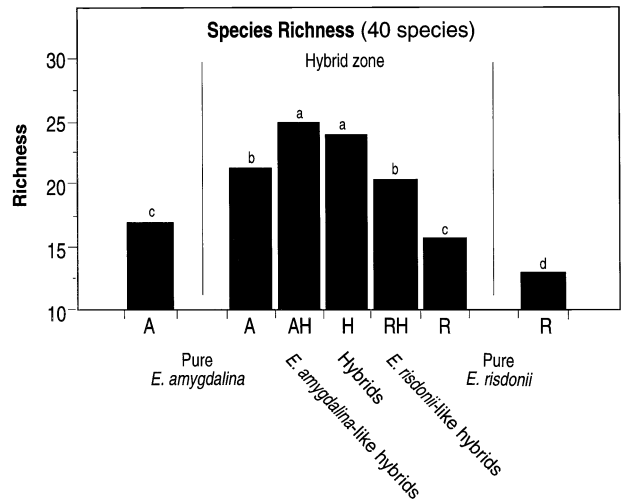
Species richness on hybrids will be greater than on either parent under two conditions that should commonly apply: parent species have species-specific herbivores and hybrids are not completely resistant to these herbivores. Given these conditions, hybrids will have herbivore communities derived from both parent species.

The effects of hybridization on species richness will be small if parental species share a large proportion of herbivores. This is the case for hybrids between *S. sericea* and *S. eriocephala*. Most of the gall formers, leaf miners, and leaf tiers, occur on both plant species and on hybrids. Several herbivores occur on one species and hybrids but not on the other parent. For example, *Pontania gracilis* occurs on *Salix sericea* and hybrids and *P. pomum* occurs on *S. eriocephala* and hybrids. Likewise, the stem gallers *Rabdophaga rigidae* and *R. salicisbrassicoides* occur on *S. sericea*, but *R. strobiloides* occurs on *S. eriocephala*. These herbivores are responsible for the higher diversity on hybrid willows compared to the two parents.

Extensive surveys of herbivore diversity on *Eucalyptus* hybrid systems have included comparisons of F<sub>1</sub>-like hybrids and putative backcrosses to each parent species. Species richness on F<sub>1</sub> hybrids and backcrosses to *E. amygdalina* were significantly greater than on either parent or backcrosses to *E. risdonii* (Fig. 6A). Additional species numbered 7–8 more than on *E. amygdalina* and 11–12 more than *E. risdonii* in their pure zones, respectively. Backcrosses to *E. risdonii* had higher diversity than on pure *E. risdonii* in the hybrid zone or in the pure zone. Parent species in pure zones had even lower diversity than parent species in the hybrid zone. Analysis of another *Eucalyptus* hybrid zone found greatest species richness on putative backcross hybrids than on either parents or F<sub>1</sub>-type hybrids (Fig. 6B). In both of these examples, relative abundances of herbivores were higher on hybrids than on parent species. This suggests that hybrid breakdown may create a more permissive environment for herbivores on backcross plants.

A significant number of species occurring on hybrids may be unique in that they are not found on either parent species. On spruce hybrids in two locations in Michigan, 12.2 and 15% of herbivore species were unique to hybrids. About half of species are restricted to one or the other parent species and 35 and 41.5% of herbivores on hybrids were found on both parent species (Fig. 7).

### A



### B

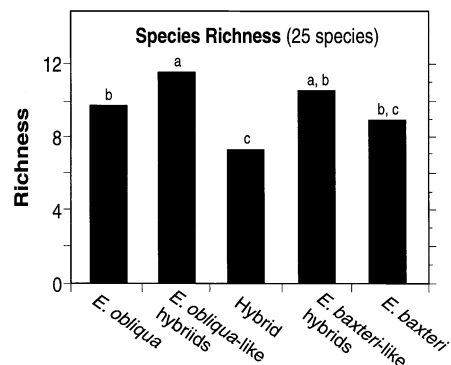


FIGURE 6 Species richness in a *Eucalyptus amygdalina* × *E. risdonii* hybrid zone (40 insect and fungal taxa species) (A) and in a *Eucalyptus obliqua* × *E. baxteri* hybrid zone (25 insect and fungal taxa species) (B). Bars with different letters differ significantly at  $P < 0.05$ . (A from Whitham *et al.* (1994); B from Morrow, P. A., Whitham, T. G., Potts, B. M., Ladiges, P., Ashton, D. H., and Williams, J. B. (1994). Gall-forming insects concentrate on hybrid phenotypes of *Eucalyptus*. In *The Ecology and Evolution of Gall-Forming Insects* (P. Price, W. Mattson, and Y. Baranchikov, Eds.), pp. 121–134. North Central Forest Experiment Station, Forest Service, USDA, St. Paul, MN.)

## B. Community Structure

### 1. Shape of Communities

The combination of altered resistance of hybrid plants compared to parents, species-specific responses of herbivores to hybrid plants, and new combinations of coexisting phytophages on hybrid plants dictates that community shape (the relative representation of different species) will differ between hybrids and parent species

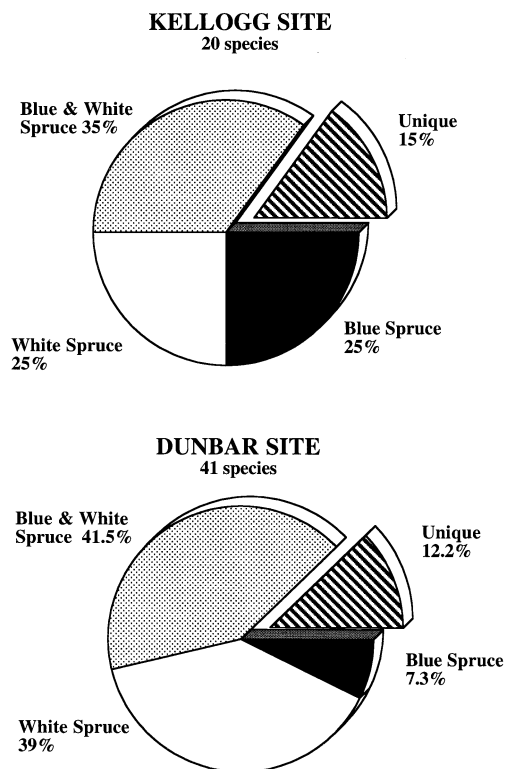


FIGURE 7 Proportions of herbivore species on F<sub>1</sub> hybrids between blue spruce (*Picea pungens*) and white spruce (*P. glauca*) derived from blue spruce specialists, white spruce specialists, species that utilize both hosts, and unique herbivores not found on either parent. (From Mattson, W. J., Haack, R. K., and Birr, B. A. (1996). F<sub>1</sub> hybrid spruces inherit the phytophagous insects of their parents. In *Dynamics of Forest Herbivory: Quest for Pattern and Principle*. (W. Mattson, P. Niemela, and M. Rousi, Eds.), No. 183, pp. 142–149. North Central Forest Experiment Station, Forest Service, USDA, St. Paul, MN.)

and among genetic classes of hybrids. The ecological consequences of the novel communities of phytophages will depend on species abundances and how species interact with each other. However, novel interactions (competition, mutualism, predation, and parasitism) are likely to result.

Community shape differs on *S. sericea*, *S. eriocephala*, and their hybrids. A discriminant function analysis suggests that the relative abundances of herbivore species varies among parents and hybrid field plants and among full-sib families derived from crosses between pure female *S. sericea* and male *S. eriocephala* (Figs. 8A and 8B). However, herbivore communities are not completely distinct between these taxa. For both analyses, some plants from each taxon are closely spaced, indicating that communities do not differ between some plants. A more distinct pattern of community structure

was seen between *Populus fremontii*, F<sub>1</sub> hybrids, and backcross hybrids (Fig. 8C). But even for this example, some individual plants fall within the space occupied by other taxa. The use of herbivore species to categorize hybrid and parental genotypes provides similar levels of discrimination of hybrids as the use of plant morphology.

## 2. Keystone Species

Species that have large effects on populations of other species in a community are called keystone species. Removing them from the community and measuring the responses of other species can reveal their indirect effects on community structure. Two such cases of indirect effects of keystone species on diversity and relative abundance of other species on hybrids were found for the gall aphid, *Pemphigus betae*, and the leaf beetle, *Chrysomela confluenta*, in the *Populus angustifolia*–*P. fremontii* hybrid zone (Whitham *et al.*, 1999). Removal of *P. betae* from hybrid trees reduced species richness by 32% and relative abundance by 55% of other herbivores in the system. The negative effects on diversity were due to elimination of the parasites and predators of the aphids and reduction of other herbivores, which benefit from aphid modification of the host plant. In contrast, removal of the beetle, which preferentially feeds on leaves of young plants, increased species richness of other herbivores by 120% and relative abundance by 75%. Defoliation by the beetle reduces the available foliage for several other herbivores, an intense competitive effect. Thus, when plant hybridization strongly alters resistance to even one herbivore, there may be a cascade of indirect effects on other species.

## 3. Third Trophic Level

The third trophic level is the group of parasitoids, predators, parasites, and pathogens of herbivores or plant pathogens. These species typically outnumber herbivores severalfold, and therefore their diversity is an important part of the biological diversity supported by hybrid plants. Few studies have considered the effects of parasitoids or predators on herbivores occurring on hybrid and parental plants, but variation in abundance and fitness of herbivores on hybrid plants may be explained in part by the impact of the third trophic level. Decreased levels of parasitism or predation could explain the higher abundance of some herbivores on hybrid plants compared to parental species. Higher parasitism of herbivores on hybrids may reduce the density and fitness impact of herbivores on hybrids. Parasitism by chalcid parasitoids did not differ among galls of *Andricus californicus* on *Quercus dumosa* and a wide

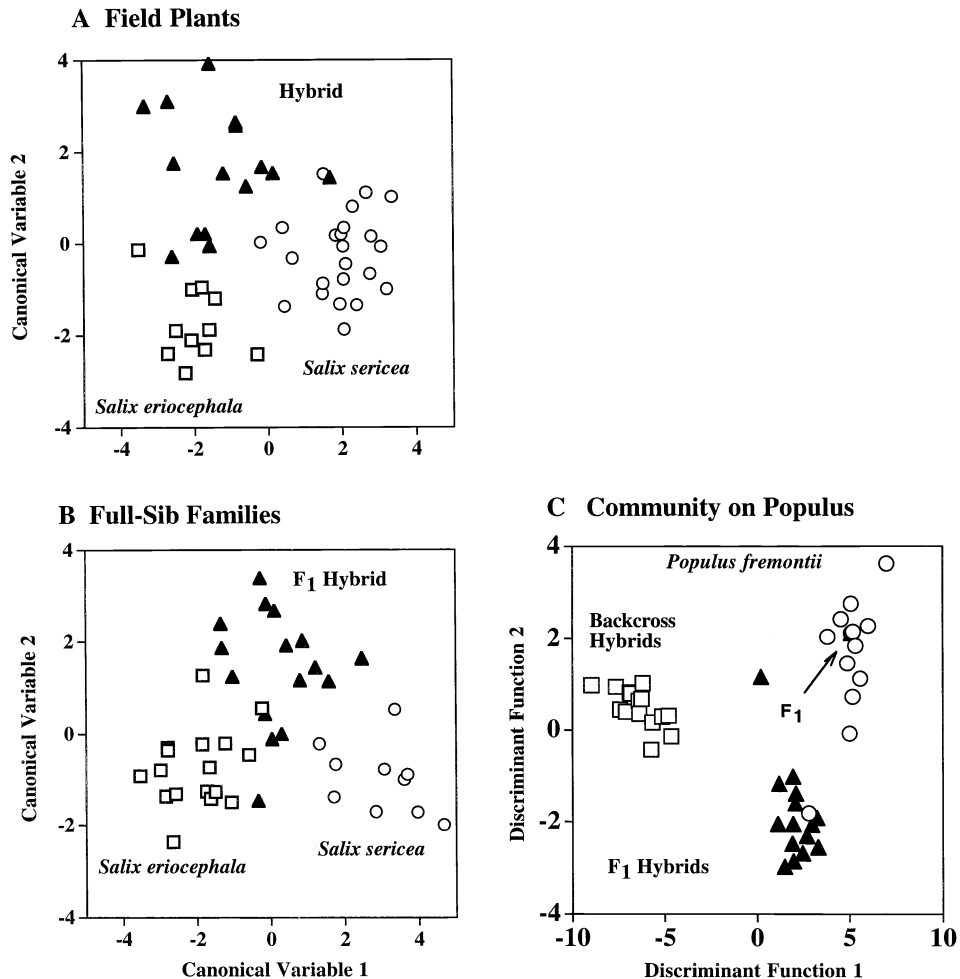


FIGURE 8 Community shape based on densities of herbivore and pathogen species on parents and hybrids. Plots of the first two canonical variables from discriminant function analyses using densities of 12 gall-inducing, leaf mining, and leaf-tying herbivores of *S. sericea*, *S. eriocephala*, and their hybrids. (A, B) *S. sericea*, *S. eriocephala*, and clones in the field (A), and F<sub>1</sub> hybrids and parents in a common garden (B). (C) *Populus fremontii*, and hybrids with *P. angustifolia* in the field. (From Floate, K. D., and Whitham, T. G. (1995). Insects as traits in plant systematics: Their use in discriminating between hybrid cottonwoods. *Can. J. Bot.* 73, 1–13.)

range of hybrid genotypes. Mortality of aphids on hybrids of *Alnus incana* and *A. glutinosa* was intermediate between that of the two parents (Table II). However, parasitism rates of a seed-eating herbivore of two cattails (*Typha latifolia* and *T. angustifolia*) and their F<sub>1</sub> hybrid were significantly greater on the hybrid, although densities of the seed predator were significantly lower on the hybrids. In this case, hybridization seems to directly affect herbivore susceptibility to parasites. Parasitism of the leaf miner *Phyllonorycter* sp. on oak hybrids was lower than on parental species, and lower parasitism of *Phyllonorycter salicifoliella* on willow hybrids was

found in the field, but was not repeated the next year or on F<sub>1</sub> hybrid plants in a common garden experiment (Table II) (Fritz *et al.*, 1999). Predation on beetle larvae on willow hybrids was intermediate between that on the parental species.

There is a wide range of effects of parasitoids and predators on herbivores of hybrid and parental plant species. The outcome of these interactions is likely to influence the population dynamics of herbivores of hybrids and therefore their effects on hybrid plant fitness. One consequence of unique communities on hybrid plants is that related herbivores that share natural ene-



TABLE II  
Summary of the Effects of the Third Trophic Level on Herbivores on Hybrid and Parental Species

Taxa	Herbivore	Enemy	Mortality relative to parent(s)
<i>Quercus dumosa</i> × <i>Q. engelmannii</i>	<i>Andricus californicus</i>	Chalcid	No difference
<i>Quercus depressipes</i> × <i>Q. rugosa</i>	<i>Phyllonorycter</i> sp.		Lower
<i>Salix sericea</i> × <i>S. eriocephala</i>	<i>Phyllonorycter salicifoliella</i>	Parasitoids	No difference/intermediate
<i>Salix viminalis</i> × <i>S. dasyclados</i>	<i>Galerucella lineola</i>	Various predators	Intermediate
<i>Alnus incana</i> × <i>A. glutinosa</i>	<i>Pterocallis alni</i>	Predators and parasitoids	Intermediate
<i>Typha latifolia</i> × <i>T. angustifolia</i>	<i>Lymnaecia phragmitella</i>	Four parasitoids	Higher

mies may interact more intensely. Apparent competition, where differential parasitism causes negative density relationships between species, would seem to be more likely on hybrid plants. Coexistence of related species on hybrids might facilitate shifts in host utilization by parasitoids.

#### 4. Other Indirect Effects

The interactions of plants with other species extend beyond herbivores and their natural enemies. Hybridization often results in novel plant architecture. F<sub>1</sub>-type hybrids in the cottonwood hybrid zone between *P. fremontii* and *P. angustifolia* have greater numbers of lateral branches than parent species. Compared to parent species and backcrosses to *P. angustifolia*, these hybrids supported significantly greater numbers of bird nests. Northern orioles (*Icterus galbula*) and black-billed magpies (*Pica pica*) were the most common species using hybrid cottonwoods as nest sites, but other common species using these trees were the American robin (*Turdus migratorius*) and the warbling vireo (*Vireo gilvus*) (Whitham *et al.*, 1999).

Plants have mutualist or commensalistic interactions with endophytic fungi, which may have detrimental effects on insect and mammalian herbivores. In the *Quercus grisea* and *Q. gambelii* hybrid zone, total endophyte frequency and frequency of *Gnomonia cerastis* was highest on *Q. gambelii* and was positively associated with hosts resembling *Q. gambelii*, but the frequency of *Coccochorella quercifolia* was most frequent on *Q. grisea* and on hybrids resembling *Q. grisea*. While direct effects on the density of a leaf mining moth (*Phyllonorycter* sp.) were not found, mortality of the tissue-feeding stage of *Phyllonorycter* sp. was positively associated with the presence of *G. cerastis*. This study suggests the possible interaction between endophytes and an insect herbivore mediated by hybrid genotype.

Mycorrhizal fungi are ubiquitous mutualists of

plants that commonly facilitate nutrient uptake. How hybridization affects these mutualists is unknown, but the effects of high levels of herbivory that occur on some hybrid plants could be detrimental to mycorrhizae. Herbivory on pinyon pine, *Pinus edulis*, by a stem-boring moth reduced mycorrhizal levels by as much as 33%, but when moths were removed, mycorrhizae recovered. Thus, interactions between herbivory on hybrids and root mutualists may be antagonistic, resulting in more harmful effects on plant fitness.

A rich diversity of ecological interactions (interaction diversity) is likely to occur on hybrid plants. Some interactions on hybrids will vary on themes present on parental species; other interactions will be unique due to unique species combinations on hybrids (e.g., mutualists and competitors). Still other interactions will be unique because novel or transgressive (traits that lie outside the parental range) traits of the hybrids themselves (e.g., morphology, phenology, or chemistry) create novel conditions for evolution between herbivores, pathogens, and their natural enemies.

## V. POPULATION AND EVOLUTIONARY CONSEQUENCES

### A. Population Dynamics of Phytophages on Hybrid Plants

The observation that most of the population of the gall-forming aphid (*P. betae*) was found on hybrid poplars rather than on pure parental species suggested the possibility that susceptible hybrids could limit the abundance of this parasite on parental poplars by attracting all of aphids (sink hypothesis). An alternative hypothesis is that hybrids are a reservoir for parasites that disperse and colonize parental species at higher rates than would occur in the absence of hybrids (source hypothe-

sis). These alternative population consequences for hybrid hosts in populations could profoundly affect the dynamics and epidemiology of natural parasite populations in hybrid zones. Where parental and hybrid individuals coexist, unless adaptation of the parasite to the hybrid host has occurred, parasite populations ought to be maintained at higher levels on parents than would occur in pure single-parent zones or in contact zones.

Several field studies of hybrid and pure parental zones compare herbivore abundances on parental species. No difference was found between parents in pure zones and parents in hybrid zones in 18 of 20 comparisons in a *Eucalyptus* system, but in 2 of 20 comparisons lower densities occurred on parents in pure zones than in the hybrid zone, supporting the source hypothesis. Environmental variation between pure and hybrid zones confounds the interpretation of these results, but it does not suggest broad support for either the sink or the source hypothesis. The smut fungus *Anthracoidea fischeri* on *Carex canescens* and *C. mackenziei* is influenced by the presence of hybrids. Only in populations where hybrids were present was the smut fungus found on parent species; here hybrids clearly act as a source.

## B. Evolutionary Outcomes

### 1. Effects of Phytophages on Plant Fitness

Herbivores and pathogens affect the fitness of plants. High levels of herbivore or pathogen attack typically result in slower growth, diminished reproduction, or even death of plants. Effects of herbivores and pathogens on hybrid plants can affect the fitness of hybrids, the distribution of hybrid zones, and selection among hybrids. Studies of *Eucalyptus* hybrids have revealed two effects of phytophages on fitness of hybrids. Hybrids of *E. melanophloia* × *E. crebra* have lower fitness, measured by seed production, than either parent. Insect predispersal seed predators and the seed pathogen *Ramularia* sp. cause greater losses on these hybrids compared to parents. In contrast, fitness of hybrids between *E. populnea* × *E. crebra* was intermediate between that of the parents in response to the same phytophages. In a third *Eucalyptus* hybrid system, fitness of hybrids and *E. amygdalina* was consistently lower than that of *E. risdonii*. Fitness of the F<sub>1</sub>-type hybrids was lower than that of either parent in 1980, but in 1990 and 1992, hybrid fitness was similar to that of *E. amygdalina*. In all years, fitness was inversely related to herbivore species richness and abundance, suggesting a direct role of phytophages on fitness.

A number of studies have found that absolute and relative abundances of herbivores and pathogens are

higher on hybrid genotypes than on parental species (Whitham *et al.*, 1994). If the higher load of herbivores causes more damage and more damage reduces plant fitness, then hybrid plants may be selected against. Two outcomes could result given this scenario. First, selection against hybrids could maintain narrow hybrid zones and favor reinforcement of species isolating mechanisms. Second, recombination of resistance traits can create genetic variation among hybrid progeny. Selection by herbivores or pathogens among the recombinant hybrids can favor certain hybrid genotypes over others, setting the stage for evolution of adaptive hybrid combinations and the formation of hybrid taxa.

### 2. Introgression of Plant Resistance

Backcrossing and selection can lead to introgression of plant resistance between plant species. Selection by grazers favored introgression of defensive traits from unpalatable *Cowania stansburyana* into populations of *Purshia tridentata* (bitterbrush), a plant heavily browsed by mammals in western North America. Introgressed *Purshia* populations are also resistant to browsers compared to nonintrogressed *Purshia* populations. In another example, introgression of jack pine (*Pinus banksiana*) resistance traits appears to explain the geographical gradient in resistance of Lodgepole pine (*Pinus contorta*) to several insects and pathogens. Likewise, introgression of black spruce (*Picea mariana*) genes into the red spruce (*Picea rubra*) genome resulted in greater resistance to spruce budworm. These studies suggest that adaptive transfer of plant traits that influence diversity of phytophage communities is a consequence of hybridization and selection imposed by plant parasites.

## VI. CONSERVATION OF HYBRID PLANTS

The value of hybrid plants to biological diversity depends on the attributes of the hybridizing taxa and on the effects on dependent communities. Hybridization of widespread, common plant species can produce genetically diverse hybrids that can contribute to evolution of plant taxa and support unique herbivore communities with novel interactions. On the other hand, interspecific hybridization threatens plant taxa that are rare, have isolated populations, or are endemic to islands. Decisions about conserving plant hybrid populations depend on the effects that hybridization are likely to have on the persistence of the parental species and on their dependent communities. Taxonomic, genetic,

and ecological information is required to answer these questions.

### A. Threats to Diversity from Plant Hybridization

Hybridization can threaten rare, endemic, or isolated plant species. Genetic assimilation via hybridization can threaten small populations of species occurring on islands or in isolated, relict populations when larger populations of related taxa are present. Most offspring of these rare taxa may be hybrid individuals, rather than pure conspecifics. A further threat is that these hybrid individuals may suffer outbreeding depression, reduced fitness relative to nonhybrid individuals, leading to a loss of their alleles from the population. Finally, inasmuch as hybrid individuals may not be recognized taxonomically, protection afforded to the threatened parent may not include hybrid individuals that possess a significant portion of the remaining genetic diversity.

The rarest plant in California, *Cercocarpus traskiae*, is found in a single canyon on Santa Catalina Island. A more common species, *C. betuloides*, also occurs on the island and hybridizes with *C. traskiae*. Five pure *C. traskiae* and two hybrid trees were discovered, and all of the seedlings found in the area were F<sub>1</sub> hybrids of just two of the pure *C. traskiae* individuals (Rieseberg, 1991). Thus, genetic assimilation and the loss of pure individuals threaten the remaining *C. traskiae* population. Rieseberg (1991) recommended that the *C. betuloides* individual near the remnant population of *C. traskiae* be destroyed, that isolated populations of *C. traskiae* be established in new locations, and that non-native mammals be eliminated. In other systems, hybrid species are difficult to identify, being confused with hybrid swarms or with putative species. Furthermore, hybrid species are sometimes found in novel and restricted habitats (e.g., *Helianthus paradoxus*) where they are threatened by habitat destruction.

Plants threatened with extinction by hybridization also have their dependent communities of herbivores and pathogens threatened. Hybridization can eliminate the only habitat of specialist herbivores or pathogens.

### B. Benefits of Hybrid Plants to Biological and Interaction Diversity

Hybridization between widespread and common species commonly creates stable hybrid zones or hybrid swarms that do not threaten the population or genetic integrity of parent species. Rather, these hybrid popula-

tions are sites of substantial genetic diversity where plant evolution occurs and to which communities of herbivores, pathogens, mutualists, and commensals respond.

Hybrid populations are centers of biological diversity of dependent phytophages. Not only do more species of herbivores and pathogens coexist on hybrid plants but some have much higher densities on hybrids and may be largely restricted to hybrid zones. This is the case for *Pemphigus betae* and *Chrysomela confluens* on poplar hybrids, but other hybrid zones have similar patterns. While a higher density of herbivores is not the most common response to hybrid plants, higher species richness on hybrids and in hybrid zones may be ubiquitous.

Independent herbivore response leads to unique communities on hybrid plants. These communities support greater interaction diversity and provide the environment for evolution of interactions between species. Competitive, parasitic, and mutualistic interactions in these unique species assemblages can influence the adaptations of herbivores to plants. Little experimental evidence exists for this assertion, but community patterns are suggestive.

The combinations of higher plant genetic diversity, higher herbivore and pathogen diversity, and more opportunities for interaction diversity strongly suggest that hybrid zones deserve conservation protection. The conservation status of hybrid plants is ambiguous at present. Since hybrids are not taxonomically distinct or may not even be recognized in nature when they occur, they usually do not receive consideration for protection. Because hybrid plants may be the major hosts of rare or threatened insect species, some hybrid populations may deserve special protection.

### C. Hybridization of Crops and Native Plants

Interspecific hybridization can transfer traits from crop plants to their native relatives. Gene flow is known for radish (*Raphanus*), squash (*Cucurbita*), sorghum (*Sorghum*), sunflower (*Helianthus*), and many other crop-native plant species pairs. Concerns over gene transfer are that crop genes may irreparably contaminate smaller populations of native plant species, that crop-native plant hybrids may make more aggressive weeds that disrupt native plant communities, and that selected or genetically altered traits may disrupt interactions with communities of dependent herbivores and pathogens. For gene flow to be evolutionarily important, F<sub>1</sub>s produced between crops and native species must be par-

tially fertile so that later generation backcrosses can occur. However, even production of  $F_1$ s can result in outbreeding depression, perhaps leading to reduced populations of native species.

Transgenic crop plants have specific traits, such as insect resistance and herbicide tolerance, introduced using molecular genetic methods. When these crops are grown for agriculture, hybridization with wild conspecific or heterospecific relatives can lead to formation of hybrids possessing the genetically engineered trait. Subsequent backcrossing can carry the transgenic trait into populations of the wild relative. Transgenic traits that provide a fitness advantage may spread throughout the population. Evidence of  $BC_2$  *Brassica campestris* (= *B. rapa*)-like plants with the transgenic herbicide-tolerance trait BASTA (glufosinate)-tolerant were produced experimentally and were found to occur in the field. Since these plants had high fertility, the rapid spread of herbicide tolerance is possible. The spread of transgenic herbicide tolerance or insect and pathogen resistance via hybridization could be a threat to biological diversity.

## VII. CONCLUSIONS

Hybridization is a naturally occurring process that contributes to biological and genetic diversity of plant species. Hybridization affects the resistance of plants to herbivorous insects and plant pathogens, typically affecting resistance differently for different phytophages. Consequently, dependent communities of herbivores and pathogens are unique assemblages on hybrid hosts compared to parent species. This presents the opportunity for novel interspecific interactions among the phytophages, creating interaction diversity on hybrid plants. The novel community can affect the population

dynamics and the evolution of hybrid plants. Conservation of naturally occurring hybrids and hybrid zones can maintain biological diversity of herbivore and pathogen species.

## See Also the Following Articles

CONSERVATION EFFORTS, CONTEMPORARY • EX SITU, IN SITU CONSERVATION • LATENT EXTINCTIONS: THE LIVING DEAD • PLANT BIODIVERSITY, OVERVIEW • PLANT INVASIONS

## Bibliography

- Arnold, M. L. (1997). *Natural Hybridization and Evolution*. Oxford Univ. Press, New York.
- Fritz, R. S. (1999). Resistance of hybrid plants to herbivores: Genes, environment, or both? *Ecology* 80, 382–391.
- Fritz, R. S., Roche, B. M., and Brunsfeld, S. J. (1998). Genetic variation in herbivore resistance of hybrid willows. *Oikos* 83, 117–128.
- Fritz, R. S., Moulia, C., and Newcombe, G. (1999). Resistance of hybrid plants and animals to herbivores, pathogens, and parasites. *Annu. Rev. Ecol. Syst.* 30, 365–391.
- Grant, V. (1981). *Plant Speciation*. Columbia Univ. Press, New York.
- Harrison, R. G. (Ed.) (1993). *Hybrid Zones and the Evolutionary Process*. Oxford Univ. Press, New York.
- Rieseberg, L. R. (1991). Hybridization in rare plants: Insights from case studies of *Cercocarpus* and *Helianthus*. In *Genetics and Conservation of Rare Plants* (D. A. Falk and K. E. Holsinger, Eds.), pp. 171–181. Oxford Univ. Press, New York.
- Rieseberg, L. R., and Brunsfeld, S. J. (1992). Molecular evidence and plant introgression. In *Molecular Systematics of Plants* (P. S. Soltis, D. E. Soltis, and J. J. Doyle, Eds.), pp. 151–176. Chapman & Hall, New York.
- Rieseberg, L. H., and Ellstrand, N. C. (1993). What can molecular and morphological markers tell us about plant hybridization? *Crit. Rev. Plant Sci.* 12, 213–241.
- Whitham, T. G., Morrow, P. A., and Potts, B. M. (1994). Plant hybrid zones as centers of biodiversity: The herbivore community of two endemic Tasmanian eucalypts. *Oecologia* 97, 481–490.
- Whitham, T. G., Martinsen, G. D., Floate, K. D., Dungey, H. S., Potts, B. M., and Keim, P. (1999). Plant hybrid zones affect biodiversity: Tools for a genetic-based understanding of community structure. *Ecology* 80, 416–428.





# PLANT INVASIONS

David M. Richardson  
*University of Cape Town*

---

- I. Concepts and Terminology
  - II. The History of Alien Plant Invasions
  - III. The Current Extent of Alien Plant Invasions
  - IV. Invasion Processes
  - V. Invasiveness and Invasibility
  - VI. Modeling Plant Invasions
  - VII. Managing Plant Invasions
- 

## GLOSSARY

**alien plants** Plant taxa in a given area whose presence there is due to intentional or accidental introduction as a result of human activities. Synonyms include “exotic plants,” “nonindigenous plants,” and “nonnative plants.”

**environmental weeds** Invasive alien plants that impact on natural or seminatural ecosystems, for example, by eliminating native organisms or altering ecosystem functioning (also known as “wildland weeds”). Native species can be environmental weeds, especially when the disturbance regime or resource levels have been altered.

**invasibility** The properties of a community or ecosystem that render it susceptible (or resistant) to invasion by alien plants.

**invasive alien plants** Naturalized plants that produce reproductive offspring, often in very large numbers, at considerable distances from parent plants, and thus have the potential to spread over a considerable area.

**invasiveness** The delimitation of features of an organism (e.g., life-history traits) that enable it to invade (i.e., to overcome various barriers to invasion).

**naturalized plants** Alien plants that reproduce consistently and sustain populations over many life cycles without the input of resources or direct intervention by humans; they often recruit offspring freely, but mainly very near adult plants, and do not necessarily invade natural ecosystems (*cf.* invasive alien plants).

---

**TECHNOLOGICAL INNOVATIONS**, driven by rapid increases in the global human population and requirements for diverse products and services from nonnative plants, have facilitated the large-scale (in terms of numbers of individuals, taxa, and foci of introduction) movement of plants to regions far beyond their natural dispersal range. This, together with widespread changes to disturbance regimes and the development of novel anthropogenic plant communities, has led to many plants (many, but not all, of which are “colonizers” in their natural habitats) becoming “weeds” (in agricultural systems, inside or outside their natural range) or “invaders” (in natural and seminatural systems outside their natural ranges). Plants carried to new habitats far from their natural ranges are very often introduced, intentionally or accidentally, as seeds and arrive without key natural enemies that limit their performance in their natural range. For many reasons, most introduced species do not spread from planting sites, and they

cause no damage in the receiving environment. Many alien plants provide us with food, fuel, and timber, beautify our cities, and serve many other useful purposes without invading. Those species that do become invasive are a major component of habitat transformation in many parts of the world. This chapter focuses primarily on the ecology of those species that have invaded natural and seminatural systems.

## I. CONCEPTS AND TERMINOLOGY

Terms such as *colonization* (or *colonizer*) and *invasion* (or *invader*) are frequently applied with reference to plants that reclaim their distributions rapidly following disturbance, or that quickly occupy new sites adjoining their original range after alterations to factors that previously limited their ranges. Within any assemblage, species vary in their ability to persist under conditions of environmental change, and also with respect to their ability to invade new sites. Most species occupy a reasonably fixed position on a continuum from “colonizers” or “pioneers” to late-seral species. The former typically undergo large fluctuations in population size whereas the latter are specialized for more stable environments where rapid population growth is unimportant. Such strategies are associated with distinctive suites of life-history traits.

Episodes of range expansion and contraction mediated by natural disturbances, interactions with co-occurring organisms, and restricted by natural barriers have driven the diversification of floras and the delimitation of the world’s phytogeographic zones. Humans have influenced the distribution of plants for millennia by, among other things, moving plants to new habitats (breaching geographic barriers) and altering opportunities for persistence, regeneration, and spread. Such actions have resulted in thousands of plant species spreading into areas which would, without human intervention, have been either unfavorable for the persistence of such species, or outside their normal dispersal range. There is some confusion in the literature regarding concepts relating to the different categories of “weedy” plants (Fig. 1).

The study of “weeds” (taxa in areas 1, 2, 4, and 5 in Fig. 1) has for centuries been directed at solving problems (how can weeds be killed or how can their effects on crops be reduced?). Colonization dynamics have been systematically explored by plant ecologists for about a century. Large-scale problems with the invasive spread of plants introduced to natural and semina-

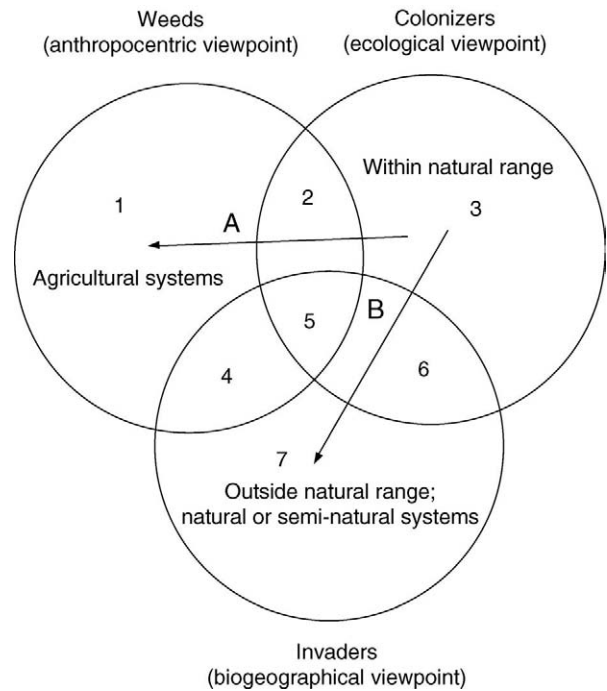


FIGURE 1 Weeds, colonizers, and invaders are overlapping (but not interchangeable) concepts that largely reflect perspectives in agriculture and range management (weeds), plant ecology and vegetation science (colonizers), and biogeography and invasion ecology (invaders), respectively. Weeds are predominantly herbaceous plants that reproduce rapidly and that do not rely on animals (other than humans) for dispersal (this applies especially to areas 1 and 2 in the diagram). Colonizers include a broader array of plant types. Similarly, invaders include a diverse group of taxa, with an overrepresentation of woody species. Many global lists of “weeds” or “invaders” (variously defined) do not distinguish adequately between these categories; the biggest problem is the lack of definition of which “weeds” (*sensu stricto*) are also “invaders” (*sensu stricto*); Daehler (1988) suggests an overlap of less than 25% of taxa from a global list of invasive angiosperms in areas 4 and 5 of the diagram. Arrows A (representing increasing disturbance associated with habitat transformation accompanying the development of agriculture) and B (depicting dispersal by humans to distant habitats, often resulting in release from natural enemies) show the origin of “weed” and “invader” floras from a pool of “colonizers” (taxa with inherent traits [sometimes masked by biotic interactions] that facilitate rapid population growth and spread in suitable habitats. In many regions, increasing fragmentation of natural systems and massive changes to disturbance regimes means that the distinction between “weeds” and “invaders” is becoming less well defined.

tural ecosystems outside their natural ranges have become a recent phenomenon, and “invasion ecology” has only recently begun to emerge as a field of specialization in ecology. Milestones in the study of such invasions have been the publication, in 1958, of Charles Elton’s book titled *The Ecology of Invasion by Animals and*

Plants, and the international program on "The Ecology of Biological Invasions" (1982–1988) under the auspices of SCOPE (Scientific Committee on the Problems of the Environment) which culminated in the volume *Biological Invasions: A Global Perspective* (Drake *et al.*, 1989).

Humans have influenced the geographic ranges of plants for millennia, by shuffling their distributions and altering their abundances, both within and outside their natural ranges. The extent to which such movements prior to the age of European colonialism shaped current distributions in Europe and Asia is often overlooked. The extent to which these range changes represent "invasions" can be debated since many changes were within or adjoining the natural ranges of such species. This is one reason for the blurred distinction between the categories "colonizers" and "invaders" (Fig. 1) in the ecological literature. Examples of human-induced changes in the distribution of pines (*Pinus* spp.; Pinaceae) serve to illustrate the different categories of range expansion. The distributions of species such as *Pinus brutia*, *P. halepensis*, *P. pinaster*, and *P. pinea* have changed dramatically due to human activities over many centuries. For example, *P. brutia* subsp. *eldarica* currently occurs in scattered locations from Azerbaijan to Pakistan. Genetic studies have shown that it occurs "naturally" only on a single mountain in Azerbaijan; the rest of its current range is due to humans who spread and managed this pine as a semidomesticated landrace. Most of its known locations follow ancient trade routes along the path taken by Alexander the Great (356–323 B.C.). In North America and other parts of the natural range of *Pinus*, many native species have invaded meadows, abandoned lands, grazed grasslands, and other habitats where they did not occur before humans arrived. In some cases such range expansions have been augmented by human-mediated dispersal of seeds beyond their normal dispersal ranges. Most pines are clearly "colonizers" as defined in Fig. 1; when moved by humans to areas outside their natural range, many of these species become "invaders," as shown for example by the invasive behavior of *P. radiata* and other pines in Australia, New Zealand, and South Africa.

## II. THE HISTORY OF ALIEN PLANT INVASIONS

Humans moved many plant species beyond their natural ranges long before the dawn of the Age of Discovery

and European colonialism. The frequency and pervasiveness of such early translocations were too limited to have resulted in widespread, damaging invasions. The voyages of exploration and the gradual settlement of the East and the New World by European powers (Britain, France, Germany, the Netherlands, Portugal, and Spain) heralded the start of a huge wave of translocations of plants and animals. This phase of "ecological imperialism" shaped the alien floras of many parts of the world that currently face the most severe problems with invasive alien plants. To cite but one instructive example: Plant introductions from Europe and Asia to South Africa between 1650 and 1806 reflected the need of Dutch colonists to cultivate a wide variety of agricultural and horticultural crops, mainly from their homeland in Europe and from the Dutch possessions in the East. Fifty or more crop plants were introduced within the first few years of European settlement, including many present-day weeds of agriculture. Subsequent waves of human immigrants, especially between 1800 and 1870, led to a surge in plant and animal introductions. Most of the plant species that now cause the greatest problems as invaders in fynbos and other South African biomes arrived between 1825 and 1860.

The magnitude and composition of the influx of alien plants to a given region during the colonial era was determined by many factors, including the geographical location of the area, its cultural links with colonial powers and other colonies, and special features of the region that influenced decisions on what types of plants to introduce. For example, the four regions of the world with a mediterranean-type climate outside the Mediterranean Basin (California, Central Chile, southern and southwestern Australia, and the southwestern tip of South Africa) all received the bulk of their alien floras from the Mediterranean Basin. Smaller components came from other regions with which they had colonial ties, and even fewer from other regions. Current problems with plant invasions in these regions clearly reflect the dynamics of biological exchange during the colonial era.

The rapid globalization of European trade routes starting in the eighteenth century led to a dramatic increase in the magnitude and diversity of the transcontinental movement in plants (Fig. 2). Trends affecting plant invasions in the nineteenth and twentieth centuries are complex and region specific. There are important differences in patterns between Northern and Southern Hemispheres, between eastern and western parts of the Northern Hemisphere, and between what are now known as "industrialized" and "developing" countries. The numbers of alien plant species that are



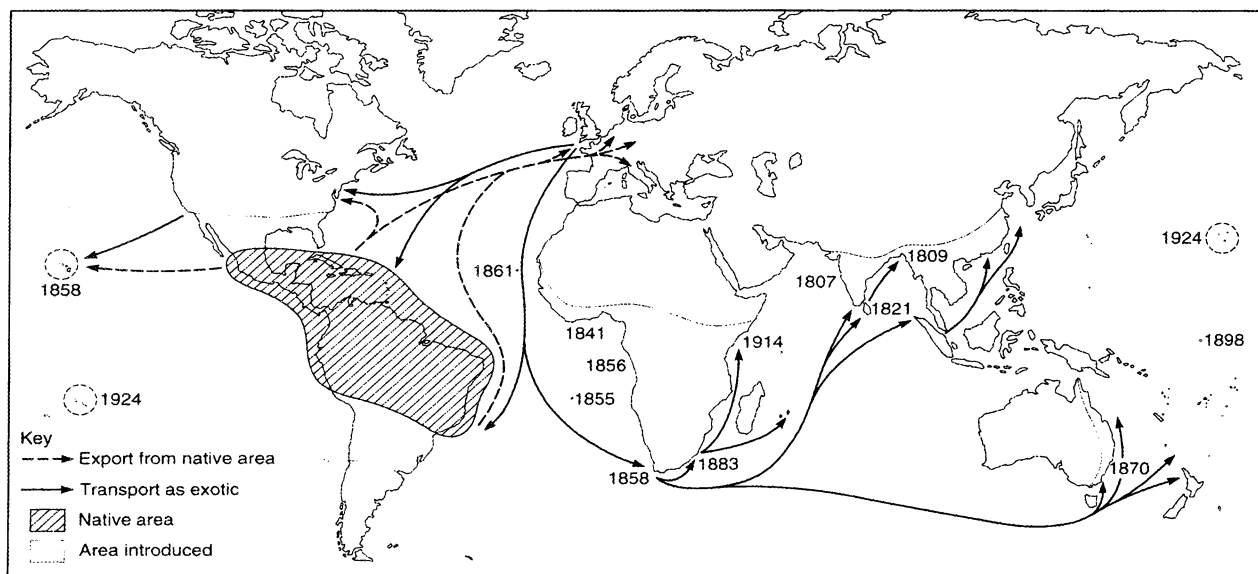


FIGURE 2 The human-orchestrated movement of plants throughout the world often involved movement from the native area as well as further dispersal from initial foci of establishment. A good example of this pattern is the history of human-aided dispersal of the *Lantana camara* (Verbenaceae) complex, now one of the most widespread and damaging plant invaders in the tropics and subtropics. The pattern reflects the major importance of European trade routes in disseminating plants.

becoming established is still increasing in most regions. Although new arrivals originate from a wider source area than the early introductions, they tend to arrive via more direct routes than in the past.

Among the fundamental factors that have affected invasions globally, especially in the twentieth century, have been (a) the improvement of transportation systems, both at an intercontinental scale, with the emergence of aircraft transportation and massive growth in shipping, and within regions (roads, railways, internal navigation canals, etc.); (b) massive habitat transformations and changes to prevailing disturbance regimes, greatly increasing the extent of ruderal habitats; (c) the introduction of large numbers of livestock to temperate grasslands dominated by caespitose grasses that lack the ability to withstand grazing pressure; (d) widespread afforestation using alien tree species (the area of tree plantations—comprising mainly alien species of *Pinus*, *Eucalyptus*, *Tectona* and *Gmelina*—doubled between 1980 and 1995 to cover 180 million ha; and most developing countries with large forest estates plan to double their plantation area by 2010); and (e) the increasing demand for alien plants for use in horticulture (North American seed and nursery catalogs offer more 59,000 plant taxa for sale to national and international markets). Developments in the twentieth century (especially the last quarter) that are particularly pertinent to

plant invasions have included the increasing attention given to scouring floras of previously underexplored areas for “new” useful plants; and the growth of agencies to facilitate the rapid transfer of germplasm of many “multipurpose” tree species for agroforestry and other purposes.

Most of the world’s worst invasive plants reached the stage where they cause major damage only in the past four decades. It is now well established that plant and animal invasions are a significant component of human-caused global change.

### III. THE CURRENT EXTENT OF ALIEN PLANT INVASIONS

Invasive alien floras include representatives of all main growth forms of higher plants (annuals, perennials, aquatic plants, epiphytes, geophytes, grasses, herbs, shrubs, trees, etc.). Alien plant invasions have affected aquatic ecosystems (estuarine, freshwater, marine, wetlands) and most types of terrestrial ecosystems throughout the world, including arid systems, coastal systems, forests (boreal, temperate, tropical), grasslands, savannas and woodlands (including mangroves) and shrublands.

Though alien plant invasions are now a global phenomenon, they have been much better studied in some parts of the world than others. This is partly because of the history and magnitude of invasions, and partly because of the global distribution of ecologists and the recent history of international scientific programs, notably the SCOPE program on biological invasions in the 1980s (Drake *et al.*, 1989). Most notable contributions to the recent literature on plant invasion ecology have come from Australia, the Mediterranean Basin (especially France), New Zealand, the Pacific islands (especially the Galapagos and Hawaiian archipelagos), South Africa (especially the fynbos biome), the United Kingdom, and the mainland United States (especially California and Florida). Comparative studies have been particularly valuable, especially those involving mediterranean-climate regions (different native floras, histories of introductions and human-mediated disturbance, but similar environmental features) and nature reserves (reflecting the only communities in many parts of the world where processes approximate natural conditions).

Among the most striking and disruptive of alien plant invasions worldwide have been the following: (a) The rapid spread of alien trees and shrubs of the genera *Acacia*, *Hakea*, and *Pinus* over large areas in fire-prone fynbos shrublands in South Africa. These invasions threaten hundreds of native plant species with extinction, change fire and nutrient-cycling regimes, and greatly reduce streamflow from watersheds. (b) The rapid invasion of mesic and wet montane forests on the Society and Hawaiian Islands by the alien tree *Miconia calvescens* (Melastomataceae), which transforms the forests, threatening many species with extinction. (c) The invasion of shrublands in North America's Great Basin by *Bromus tectorum* (cheatgrass; Poaceae), resulting in a huge increase in fire frequency (from 60–110 years on average to 3–5 years). (d) The rapid spread of *Mimosa pigra* (Fabaceae) to form dense thickets over about a million hectares of wetlands in Australia's Northern Territory. These thickets transform sedgeland into monotonous tall shrublands, and form an impenetrable understory to *Melaleuca* swamp forests, with profound consequences for biodiversity and for traditional Aboriginal food gathering. (e) The explosive spread of *Eichhornia crassipes* (water hyacinth; Pontederiaceae) in river and lagoons of west Africa and the Great Lakes of East Africa, especially in the past two decades, causing major disruption of fishing and navigation.

Which alien plant species are the most widespread or cause the most damage? There are numerous regional

or national lists of invasive alien plants, some of which provide data on areas invaded or rankings of the severity of invasions. However, since there are no generally accepted ways of quantifying when an area is "invaded" or when a species is "invasive" (as opposed to "naturalized"), attempts at compiling global lists of major plant invaders have generally been less than satisfactory.

## IV. INVASION PROCESSES

Plant invasions in natural and seminatural ecosystems involve the following fundamental phases: introduction to the region by humans, establishment, population growth (sometimes accompanied by genetic adjustment), spread to new areas within the region (and often also outside the region via further dispersal by humans; see Fig. 2), interaction with the local biota and disturbance regime, and displacement of native elements. There are many ways of conceptualizing the various processes involved in invasion and interactions with biotic and abiotic features of the new environment. One may depict the various potentially limiting factors as a series of "barriers." The simplest representation of such a model shows (a) a *geographic barrier*, which must be overcome by dispersal; (b) a *habitat barrier*, which requires preadaptation or genetic adjustment to the conditions of the new environment; and (c) a *biotic barrier*, which integrates the forces of predation, herbivory, competition, and interference that must be overcome in the new habitat, or the new mutualistic relationships that must develop. Additional complexity can be added by, for example, splitting the geographic barrier into two components (to account for factors limiting introduction to the region and dispersal within the region, respectively), by adding a *reproductive barrier* (to account specifically for factors that potentially limit seed set), or by splitting the biotic barrier into components (e.g., to isolate the role of mutualisms) (Fig. 3).

### A. Stages of Invasion

#### 1. Dispersal to a New Area

Important reasons for the intentional widespread translocation of plants include agriculture and forestry and agroforestry, botanical gardens, horticulture (including the commercial trade in seeds, bulbs, and cuttings and urban gardeners using seed exchanges), and soil stabilization. Many plants have been moved around the world inadvertently, notably in ship ballast, with military transport, and as contaminants in fertilizers, hay and straw, grains, wool, and cotton.

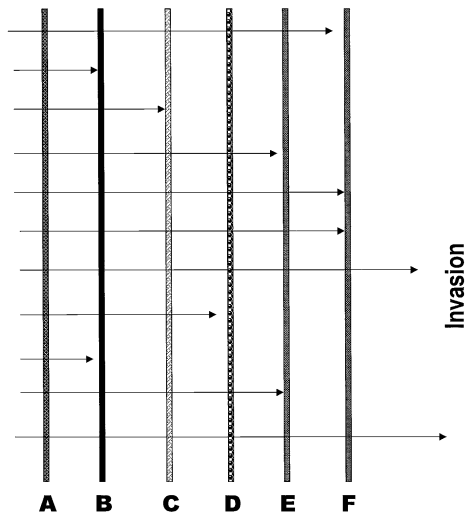


FIGURE 3 A schematic representation of major barriers limiting the invasion of introduced plants. The barriers are (A) geographic barrier I: Intercontinental or international and/or intracontinental; (B) reproductive barrier; (C) physical and chemical environmental barrier(s); (D) geographic barrier II (within the new region); (E) biotic barrier(s) I (general composition of fauna and flora); (F) biotic barriers(s) II (successionally mature, undisturbed plant communities). In many cases, alien plants establish mutualisms with other organisms to enable them to overcome barriers. For example, pollination (barrier B) and seed dispersal (D) by animals are often essential for invasion. Ants bury the seeds of some introduced plants, thus protecting them against predation and fire (barriers E and F). Many introduced plants fail in the absence of mycorrhizal fungi or nitrogen-fixing bacteria (barriers C, E, and F).

## 2. Establishment and Naturalization

Empirical evidence shows that the chance of becoming established, naturalized, and later invasive increases markedly with an increase in the number of propagules introduced, and with multiple introductions (including introductions at different times and from different source populations). More propagules reduce the likelihood of extinction and increase the chance of long-distance dispersal. Multiple introductions allow the incipient invader to sample a greater range of sites over space and time in the new environment and (since different introductions often originate from different source populations) improve the likelihood of introducing a genotype closely suited to local conditions. Also, multiple introductions increase the likelihood of forming novel genotypes that facilitate invasion. Successful establishment entails dealing with numerous physical, chemical, and biotic barriers (Fig. 3). Many introduced plants are initially grown in small populations that are inherently susceptible to extinction due to chance events. To establish and persist, a population must ex-

hibit  $dN/dt > 0$  when  $N$  is small (the “invasion criterion”). If an introduced plant can deal with various reproductive barriers, it becomes naturalized.

## 3. Spread

Invasion involves dispersal within the new area and population growth. Invasive alien floras show a wide range of adaptations for dispersal. Many species are dispersed by “passive” agents such as water or wind. A large proportion of the world’s most widespread and damaging invaders are dispersed by birds and mammals (both native and introduced). The rapidity with which these mutualistic seed-dispersal interactions establish suggests that vertebrate-dispersed plants have converged into generalized dispersal syndromes regardless of phylogenetic and geographical origins. Epizoochorous dispersal, mainly by cattle and sheep, facilitates the spread of many (mainly herbaceous) invaders. Vegetative reproduction is also important.

The dynamics of range expansion and population growth of an invasive alien plant typically follow the pattern shown in Fig. 4. There is frequently a time lag between the arrival of an alien plant in a new habitat and the start of widespread invasion. Examples include *Thlaspi caerulescens* (Brassicaceae), which was cultivated at Oslo Botanical Garden in Norway since 1814, was first collected as an escapee in 1874, spread slowly until 1900, then expanded rapidly until it reached most of its present range in about 1945; *Mimosa pigra* (Fabaceae), which was virtually confined to small areas around Darwin in Australia for 80 years before exploding; and *Melaleuca quinquenervia* (Myrtaceae), which showed no invasive tendencies for its first 50 years in Florida (United States). A recent analysis of the history of woody species introduced to Brandenburg in Germany revealed an average time lag between introduction

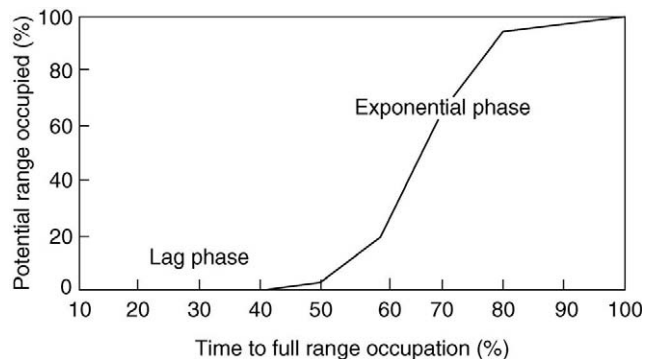
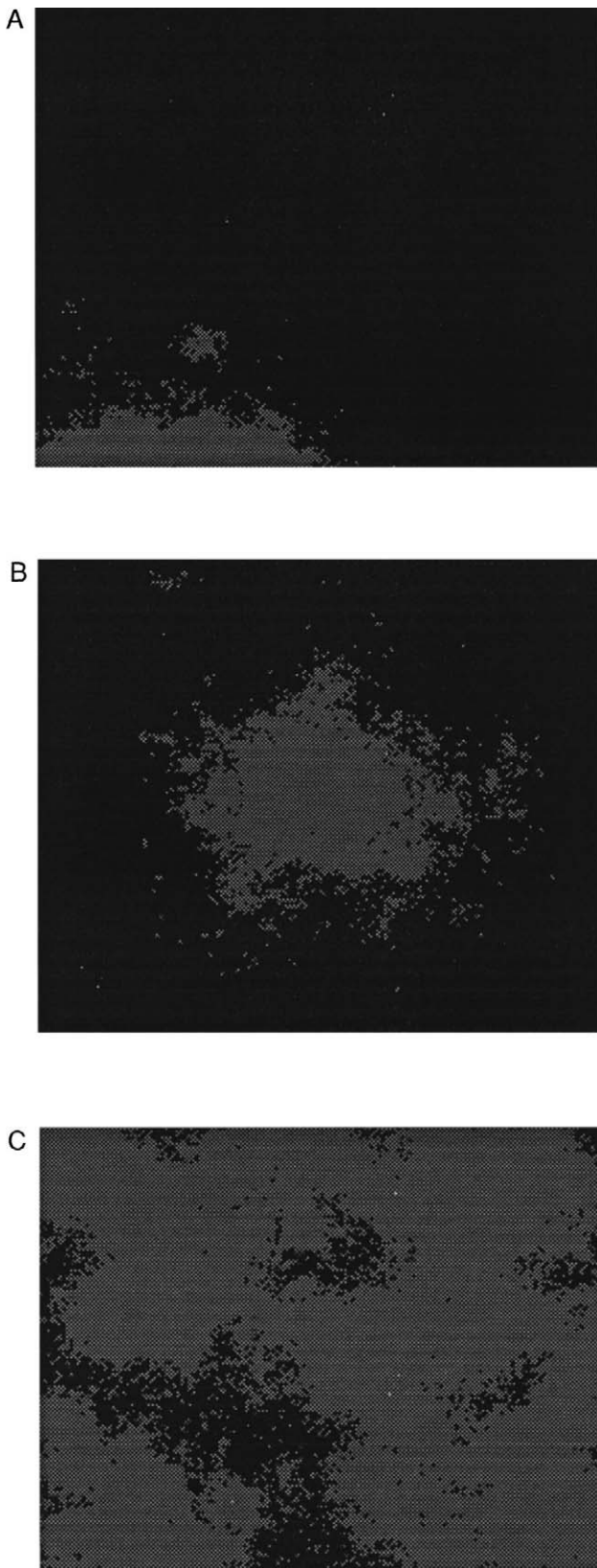


FIGURE 4 Model of the spread of an invasive alien plant over time. From Hobbs and Humphries (1995).



and invasive spread of 147 years (170 years for trees; 131 for shrubs)! Such lags, frequently alluded to but very seldom adequately explained in the invasion literature, are probably usually due to one or more of the following factors: (a) the founder population may maintain a stable, small population until genetic adjustment occurs or essential mutualists (seed dispersers, pollinators, mycorrhizal fungi) arrive (the extent of genetic alteration preceding or accompanying invasion has been studied for a few widespread invaders, e.g. *Ailanthus altissima* in the United States); (b) some lags are probably explained by an initial shortage of “safe sites,” which have become more abundant as human-induced disruption of ecosystems has increased (increasing time also improves the chance of a potential invader encountering a “safe site” formed by a rare event such as a flood); (c) populations spread slowly at their periphery but only show accelerated growth rates when there are many foci of growth.

The lag phase is followed by a phase of sudden growth during which populations increase at an exponential rate (and generally become noticed as invaders). One reason for the increased growth rate (besides those mentioned earlier that may prevent its realization) is the typical two-phase pattern of spatial expansion. This involves the densest recruitment of offspring close to founder populations (“neighborhood diffusion”) and the establishment of isolated colonists through long-distance dispersal. As satellite foci grow in size through diffusion, often coalescing with each other and the founder metapopulation, more propagules become available for additional jump dispersal. With increasing numbers of growth foci, population growth and range increase rapidly. The rate of spread is frequently augmented by intentional or accidental movement of plants within the invasion arena by humans, thus creating additional nascent foci. This process, termed “stratified diffusion,” has been documented for many plant invasions involving disparate plant taxa and environments and is evident at scales ranging from global (Fig. 2),

FIGURE 5 The simulated spread of alien pine trees in South African fynbos over 100 years in an area of 1.5 km × 1.5 km to illustrate the process of stratified diffusion. All simulations start with 75 plants; A shows spread from the edge of a plantation (at bottom left); B shows spread from a clump in the center; and C shows spread from plants randomly arranged across the invasion arena. White pixels show areas occupied by invading pines; black areas are free of pines. Details of the modeling procedure are described in Higgins and Richardson (1998).

regional, to landscape (Fig. 5). The end of the exponential phase usually occurs when most optimum sites for invasion are occupied (or when successful control is instituted). Alien plant invasions may also follow linear trajectories (e.g., in the case of spread along rivers or coastal foredunes). Invasions in these habitats usually proceed as a wave front.

Spread rates reported for invasive alien plants vary greatly. Some examples are 5000 m yr<sup>-1</sup> for *Bromus tectorum* (Poaceae) in North America; 4000 to 13,000 m yr<sup>-1</sup> for *Heterotheca latifolia* (Asteraceae) in the Georgia piedmont; 970 m yr<sup>-1</sup> for *Fraxinus ornus* (Oleaceae) along a river in France; 76 m yr<sup>-1</sup> for *Mimosa pigra* (Fabaceae) in wetlands in northern Australia; 21 to 31 m yr<sup>-1</sup> for *Acacia cyclops* (Fabaceae) and *Pinus pinaster* (Pinaceae) in South African fynbos; and 14 m yr<sup>-1</sup> for *Ammophila arenaria* (marram grass; Poaceae) on coastal dunes in California.

## V. INVASIVENESS AND INVASIBILITY

A major thrust of studies on plant invasions in recent decades has been the search for traits that separate invasive species from noninvaders and features that distinguish invaded systems from those that are apparently able to resist invasion. Although the two concepts are intimately linked, many studies have addressed the issues separately.

### A. Invasiveness

Numerous studies have described the features of individual alien plant species that have enabled them to invade given areas. For example, the "recipe for success" for *Lantana camara* (Verbenaceae) includes the following ingredients: (a) dispersal of seeds for long distances by many native and introduced birds; (b) toxicity of its fruit for many mammals, which limits damage by herbivory; (c) its ability to resprout vigorously following damage (e.g., by trampling); (d) its ability to invade a wide range of habitats, including sites disturbed by alien mammals on islands; (e) production of allelopathic substances, which improves its competitive ability; and (f) its ability to flower profusely for long periods, thus attracting pollinators and ensuring copious seed set.

Attempts have been made to define invaders for particular regions (usually biomes or countries) or the whole world and to partition groups of plants into in-

vaders and noninvaders. A notable early attempt to delimit traits of plants that enable them to become weeds or invaders was Herbert Baker's description of "the ideal weed." Briefly stated, Baker's hypothetical superweed is a plastic perennial that germinates in a wide range of physical conditions, grows quickly, flowers early, is self-compatible, produces many seeds that disperse widely, reproduces vegetatively, and is a good competitor. Many of the world's most widespread invaders and weeds possess many "Baker characters," but none has all, and many notable invaders possess very few, if any, of these traits. Are there an optimum number of Baker's traits that confers invasiveness? The work of Mark Williamson and coworkers on British annual plants suggests that optimum weediness is realized in plants with an *intermediate* number of Baker characters (4 or 5 out of 7).

There is evidence that plants that have performed best as alien invaders are not evenly distributed among taxonomic groups. Among the angiosperms, two families that are clearly overrepresented among the most widespread and damaging invaders of natural communities are Fabaceae and Poaceae. Other families that seem particularly prone to contribute invaders of natural areas are Myrtaceae, Rosaceae, Salicaceae, and Tamaricaceae. The Asteraceae, the largest family of flowering plants, contributes many species to alien floras in most regions, but the family is not significantly over represented.

Many studies have attempted to gain an understanding of the life-history attributes of plants that makes them successful invaders by comparing traits of native and alien species or invasive and less invasive alien plants. Rejmánek (1998) considered perspectives from many of these studies and added many new insights to arrive at a general model of seed plant invasiveness based on recognized and assumed causal and correlative relationships between a wide range of traits. This model is probably near the limit of what can be achieved in defining invasiveness at a global scale, and further advances will require the development of models based on plant-environment interactions (discussed later). The most consistent, and practical, predictor of invasiveness is still whether the species has become invasive elsewhere (discussed later).

### B. Invasibility

Elton's *The Ecology of Invasions by Animals and Plants* (which, it should be noted, is heavily biased in favor of evidence on animal invasions) put forward the view

that undisturbed native communities are not susceptible to invasion by introduced species. Recent debate on this topic has been rather futile since so few communities on earth are unaffected by human activities. What has become abundantly clear over the past few decades is that disturbance, be it naturally occurring or human induced, is a fundamental driver of plant invasions. Irrespective of other factors that facilitate or limit invasions, the susceptibility of communities to invasion by alien plants increases with increasing disturbance up to a threshold after which disturbance acts as a barrier. Useful insights on the types and magnitude of disturbance required to initiate and sustain invasions emerge from examining cases of invasion of pine trees (*Pinus* species) in many parts of the Southern Hemisphere. Most records document pine invasions in areas where the natural disturbance regime has been noticeably altered by humans. Changes to disturbance regimes have involved mainly increased or decreased herbivore pressure (grazing, browsing, and trampling), altered fire frequency, or mechanical clearing of vegetation. Different vegetation types differ in the type and magnitude of disturbance required to facilitate invasions. All records of pine invasions into native forests showed that severe disturbance is required to facilitate seedling recruitment and population growth. Invaded grasslands in Australia, New Zealand, and South Africa *invariably* have markedly changed disturbance regimes. The same applies for most invaded shrublands. The ultimate cause of increased invasibility caused by disturbance in these cases is clearly the reduced competition from resident plants through the reduction in ground cover. Several pine species have spread into South African fynbos where the only disturbance is the naturally occurring fire; intense fires and the dearth of vigorous plants immediately postfire combine to create opportunities for invasion. In general, intermediate levels of disturbance clearly make sites more vulnerable to invasion by pines, whereas severe disturbance usually prevents invasion. Naturally occurring disturbances where vegetation cover is removed is caused by factors such as volcanicity, slope instability, wind, and flooding; these are also important determinants of invasibility for pines. "Correlative" evidence from natural experiments such as this is very useful for defining the main factors that affect invasibility (Table I). This example shows that "invasibility" needs to be context specific; this pattern of invasibility relates to pines and probably not to other growth forms or to alien species in general.

Empirical evidence shows that certain communities are more susceptible to invasion by a wide range of

alien plants than others. For example, the greater susceptibility to invasion of island communities (the more isolated, the more invulnerable) compared to mainland ones has long been noted. Islands where established alien species make up 40% or more of the flora include Bermuda, Hawaii, Ascension, Rodrigues, Tristan da Cunha, Lord Howe, and New Zealand (figures for continental areas are much lower). Among the many reasons advanced to explain this are (a) the species richness of islands is low (theoretically reducing the likelihood of an introduced species encountering a close competitor); (b) island floras have evolved in isolation, often without adaptation to high levels of competition, grazing, trampling, or regular burning; (c) the small size of islands means that the history of human intervention is "concentrated" (more intense disturbance); and (d) many islands were colonized very early and have been the crossroads of intercontinental trade (and thus were exposed to greater numbers of potential invaders).

Among other systems that seem particularly susceptible to alien plant invasions are riparian zones. In virtually every region, these zones are among the most severely invaded habitats. An important reason for this is the frequent disturbance from water-level fluctuation, including floods, which disperses propagules and provides favorable sites for establishment.

There has been much debate on the link between the number of species in a community and its susceptibility to invasion. Some "correlative" studies report a negative correlation between plant species richness and invasibility, whereas others report a positive correlation. In many cases, such confusion is undoubtedly attributable to the overriding importance of disturbance (e.g., several studies point out that species-rich riparian zones are highly prone to invasion). Systematic experiments to explore the role of species numbers and identity have only begun in the past decade. Results to date emphasize the importance of considering spatial scale and productivity. For productive, small-scale grassland communities, for example, the identity of species is more important than the number of species.

### C. Linking Plant Traits to Environmental Features

There have been important recent advances in our understanding of the determinants of invasiveness and invasibility. There is now a reasonable understanding, at least for some severely invaded regions, of the plant traits that are correlated with success as invaders and also the environmental features that mediate invasions.

TABLE I  
Determinants of Invasibility for Pines (*Pinus* spp.; Pinaceae) Introduced to the Southern Hemisphere

Factors	Facilitating feature(s)	Limiting feature(s)
<b>Species attributes</b>		
Seed mass	Small seeds with large wings	Large seeds with small wings
Juvenile period	Short (<10 years)	Long (>10 years)
Interval between large seed crops	Short (<3 years)	Long (>5 years)
Ability to survive moderate browsing levels	Good	Poor
<b>Residence time</b>	Long (>50 years)	Short (<50 years)
<b>Extent of planting</b>		
Total area	Large	Small
Boundary: total area ratio	Large	Small
<b>Ground-cover characteristics</b>		
Basic vegetation structure	Bare or sparsely vegetated ground, shrubland, grassland	Forest
Vegetation cover	None-low (<50%)	High (>80%)
Latitude (°S)	30–45	<30
<b>Disturbance</b>		
Frequency	Low-moderate	Very low/very high
Human-induced types	Moderately increased herbivore pressure (grazing, browsing, trampling) or equivalent	Greatly increased or greatly reduced herbivore pressure or equivalent
Contributing factor	Decreased competition from ground-cover	Increased competition from ground-cover or physical elimination of pines (e.g., by mechanical clearing)
Natural types	Slope instability, wind, flooding, fires from volcanoes	Frequent fires (e.g., in grasslands)
<b>Resident biota</b>		
Composition of plant community	Naturally invadable community (e.g., <i>Dracophyllum subulatum</i> shrubland, <i>Chionochloa</i> tussock grassland in New Zealand)	Naturally resistant community e.g., <i>Eucalyptus blakelyi</i> forest in NE Victoria, <i>Protea nitida</i> fynbos in South Africa)
Indicators of invasibility	Conditions unsuitable for C <sub>4</sub> photosynthetic pathway and nutrient-poor soils; paucity of vigorous herbs	Conditions suitable for C <sub>4</sub> photosynthetic pathway and nutrient-rich soils: abundance of vigorous herbs
Role of mammals other than humans	Removal of competing vegetation (e.g., through grazing) Dispersing pine seeds (birds and mammals)	Destroying pine seedlings (browsing, trampling)
Role of fungi	Presence of appropriate mycorrhizal symbionts	Absence of mycorrhizal symbionts (no longer limiting following intentional widespread dissemination of spores) Influence of pathogenic fungi

Derived from the assessment of invasion histories of 19 *Pinus* species in Argentina, Australia, Brazil, Chile, Madagascar, Malawi, New Caledonia, New Zealand, South Africa, and Uruguay. Factors should be viewed "loadings"—the greater the number of facilitating features, the greater the chance of invasion for any taxon/site combination. Modified from Richardson *et al.* (1994).

Further progress, certainly at scales of resolution that have value in management, will rely on the development of models that link plant traits with critical environmental features. Such models show that *interactions* between the various determinants of invasive success are some-

times at least as important as the main effects. For example, in the case of pines invading forests, shrublands, and grasslands in the Southern Hemisphere, interactions between basic features of the environment, the features of the disturbance regime, and plant traits

explain the different spread rates observed in different areas.

## VI. MODELING PLANT INVASIONS

The use of models has greatly improved our understanding of invasions and how to deal with them. The many types of models that have been applied to plant invasions may be crudely grouped in three categories. *Simple demographic models* include exponential, logistic, logistic-difference, and stochastic models; they predict the future number of individuals in a population by making assumptions about the nature of population growth and by estimating demographic parameters regarded as being important in determining population dynamics. *Spatial-phenomenological models* describe plant-environment interactions using empirical data (invoking no ecological mechanisms); they include regression models, geometric models and Markov models. *Spatial-mechanistic models* are based on independent estimates of ecological parameters affecting invasions; they include reaction-diffusion models, population dynamic metapopulation models, and individual-based cellular automata models.

A particularly influential model in the study of invasions was that published by J. G. Skellam in 1951. He used a diffusion equation, combined with estimates of population growth, to show that invasion fronts advance at a constant velocity. Skellam's model, on which most subsequent invasion models have been based, did not incorporate long-distance dispersal, now known to be critically important for modeling the rate and variability of an expanding population front. Recent advances in mathematical modeling have facilitated the development of spatial-mechanical models (notably individual-based, cellular automata mixture models), which allow much more accurate modeling of all stages of plant invasions. Besides offering exciting opportunities for exploring critical processes in invasion, thus contributing to invasion theory, these models have important applications in management.

## VII. MANAGING PLANT INVASIONS

There are three basic options for managing invasive alien plants: (a) prevention, (b) removal ("eradication"), and (c) ecological management. The appropri-

ateness of different control approaches depends on the ecology of the invader and the invaded system, and on a myriad of socioeconomic issues.

### A. Prevention

Despite pessimistic prognoses in the 1980s regarding the ability of ecology to predict plant invasiveness and invasibility at scales useful for management, important advances have been made. Various methods have been developed for screening introduced plants to assess the risk of them invading particular habitats. As mentioned earlier, a good predictor of invasiveness is whether a species has invaded other (similar) areas where it has a longer history as an alien. For example, 90% of invasive plant species in Australia are also invasive in other locations. A major problem with applying this concept to management or regulation is the time lag inherent in many (most) invasions—at what stage should the "performance" of a species be scored with regard to invasiveness? Also, many species are currently being introduced directly from their native ranges to many different areas simultaneously.

### B. Removal

The number of options for controlling an invasive plant decline rapidly when the phase of exponential increase (Fig. 4) is reached. Management is clearly most effective when directed at removing invaders while they still occupy a small range and occur in small populations. Many potentially widespread invaders are effectively kept in check by various means of control at this stage. Removal is much less feasible once the invaders have spread over large areas, but is still attempted in some cases, usually when the damage the invaders cause overrides economic constraints and when other control measures are unavailable or impractical. For example, expensive, large-scale clearing of invasive *Hakea* and *Pinus* spp. in South African fynbos is warranted by estimated costs of reduced water production in the absence of control.

### C. Ecological Management

Experience has shown that satisfactory control of plant invaders is usually only achieved when several complementary methods, including biological control, improved land management practices (e.g., through pre-



scribed burning and modified stocking densities), herbicides, and mechanical methods are carefully integrated.

### D. Confounding Factors—New Trajectories and Conflicting Human Perceptions

Invasive plants are being effectively managed in many parts of the world, albeit at a great, and increasing, cost. The dimensions of the problem are changing rapidly, and management strategies need to be flexible enough to deal with exigencies. Among the main reasons for changing trajectories of invasions are the following:

- The increasing magnitude and pervasiveness of global trade and free trade agreements.
- The increasing availability and demand for alien plants for a wide range of purposes, especially for agroforestry and horticulture.
- Global change (including climate change, changed nutrient regimes, elevated CO<sub>2</sub>, fragmentation); although little is known of how all these factors affect invasions, especially in concert, there is consensus among ecologists that these changes greatly exacerbate problems with invading alien plants.
- Genetic engineering (GMOs, transgenic plants); genetically modified plants could acquire novel traits that may make them (better) invaders.
- Conflicts of interest; many alien plants are essential crops in some parts of the landscape/region/country and damaging invaders in other parts. Good examples are many tree species used in forestry and agroforestry, and *Echium plantagineum* (Boraginaceae), an important dry-season forage and a honey plant in South Australia, but an important weed of pastures in New South Wales. A major challenge is to develop objective methods for assessing the costs and benefits associated with our increasing dependence on alien species.

### See Also the Following Articles

DISTURBANCE, MECHANISMS OF • INTRODUCED SPECIES, EFFECT AND DISTRIBUTION • MIGRATION • PLANT-SOIL INTERACTIONS

### Bibliography

- Baker, H. G., and Stebbins, G. L. (Eds.) (1965). *The Genetics of Colonizing species*. Academic Press, New York.
- Daehler, C. C. (1998). The taxonomic distribution of invasive angiosperm plants: Ecological insights and comparison to agricultural weeds. *Biological Conservation* 84, 167–180.
- Di Castri, F., Hansen, A. J., and Debussche, M. (Eds.) (1990). *Biological Invasions in Europe and the Mediterranean Basin*. Kluwer, Dordrecht.
- Drake, J., Mooney, H. A., Di Castri, F., Groves, R., Kruger, F. J., Rejmánek, M., and Williamson, M. (Eds.) (1989). *Biological Invasions: A Global Perspective*. Wiley, Chichester.
- Elton, C. S. (1958). *The Ecology of Invasions by Animals and Plants*. Methuen, London.
- Higgins, S. I., and Richardson, D. M. (1998). Pine invasions in the southern hemisphere: modelling interactions between organism, environment and disturbance. *Plant Ecology* 135, 79–93.
- Hobbs, R. J., and Humphries, S. E. (1995). An integrated approach to the ecology and management of plant invasions. *Conservation Biology* 9, 761–770.
- Macdonald, I. A. W., Kruger, F. J., and Ferrar, A. A. (Eds.) (1986). *The Ecology and Management of Biological Invasions in Southern Africa*. Oxford University Press, Cape Town.
- Mooney, H. A., and Drake, J. A. (Eds.) (1986). *Ecology of Biological Invasions in North America and Hawaii*. Springer Verlag, New York.
- Rejmánek, M. (1998). Invasive plant species and invulnerable ecosystems. In *Biodiversity and the Management of Invasive Species* (O. T. Sandlund, J. Schei, and L. Viken, Eds.). Kluwer, Dordrecht.
- Richardson, D. M., Williams, P. A., and Hobbs, R. J. (1994). Pine invasions in the southern hemisphere: determinants of spread and invadability. *J. Biogeography* 21, 511–527.
- Richardson, D. M., Pyšek, P., Rejmánek, M., Barbour, M. G., Panetta, F. D., and West, C. J. (2000). Naturalization and invasion of alien planets: concepts and definitions. *Diversity and Distributions* 6 (in press).
- Sandlund, O. T., Schei, J., and Viken, L. (Eds.) (1998). *Biodiversity and the Management of Invasive Species*. Kluwer, Dordrecht.
- Simberloff, D., Schmitz, D. C., and Brown, T. C. (Eds.) (1997). *Strangers in Paradise: Impact and Management of Nonindigenous Species in Florida*. Island Press, Washington, D.C.
- Williamson, M. (1996). *Biological Invasions*. Chapman and Hall, London.



# PLANT-SOIL INTERACTIONS

Joan G. Ehrenfeld  
Rutgers University

---

- I. Introduction
  - II. Small-Scale Processes: Micron to Millimeter Scales
  - III. Mesoscale Processes: Centimeters to Meters
  - IV. Large-Scale Interactions: Stands and Biomes
  - V. Implications for the Management and Conservation of Ecosystems
- 

## GLOSSARY

**aerenchyma** Porous root tissue, especially well developed in wetland plants, that allows diffusive flux of oxygen from above-ground tissues to root tips. This tissue both supports the respiratory demand of the root tissues and allows oxygen to leak into the surrounding soil.

**mycorrhizae** A root tip that is infected with fungi in a mutually beneficial partnership. The fungal hyphae explore large volumes of bulk soil, absorbing nutrients and transferring them to the plant; the plant supplies the organic carbon necessary for growth and energy production to the fungus. Different groups of fungi form vesicular-arbuscular mycorrhizae (fungal hyphae invaginate into the plant root cells) and ectomycorrhizae (fungal hyphae grow between plant root cells and form a thick sheath over the root tip, but they do not invaginate). Several other forms are specific to particular plant families (Ericaceae, Orchidaceae).

**net mineralization/immobilization** The net result of microbial decomposition of organic matter is either the incorporation of nutrient elements (particularly nitrogen) into the microbial biomass, rendering it unavailable for plant uptake (*immobilization*), or their release into the soil solution (mineralization) after microbial demand for each element has been satisfied.

**nutrient uptake capacity** The instantaneous rate of nutrient acquisition, usually measured in brief (1–2 hr) incubations. Uptake capacity reflects the abundance of transport sites on the root cell membranes and their affinity for nutrient ions.

**rhizodeposition** The mixture of sloughed cells, mucilages, and small-molecular-weight sugars, amino acids, and other compounds leaked from root cells, which are deposited in the soil adjacent to the surface of fine roots. Exudation takes place from the root tip back to the zone of suberization. The chemical quality and quantity of the exudate is altered by the presence of mycorrhizae.

**rhizosphere** Volume of soil adjacent to, and strongly influenced by, a plant root. The rhizosphere is usually considered to extend about 2 mm from the root surface, and includes the “rhizoplane,” or soil directly in contact with the root surface.

**plasticity** Ability of a plant to respond to temporal changes or spatial variation in environmental conditions by altering the size or the distribution of plant parts. These are phenotypic, rather than genetic changes.

**siderophore** Chemicals secreted by roots (primarily non-protein-forming amino acids), which complex with insoluble metal ions bringing them into solution and permitting their transport to and uptake into the root.

**soil aggregate** A crumb-sized unit of soil, composed of aggregated soil minerals, microbes, and soil microfauna, which are cemented together by a combination of biological materials such as polysaccharide secretions, fungal hyphae, and chemical substances such as precipitated carbonates or silicates. Aggregates are classified by size and stability in water (disintegrating versus retaining their structure and integrity).

**soil organic matter** Organic substances, including a wide variety of carbohydrates, proteins, lipids, waxes, phenolic, and humic compounds, which accumulate in soil as a result of both plant and microbial growth. These compounds include small-molecular weight materials, which are rapidly decomposed to carbon dioxide; larger compounds, which may be slowly decomposed over years to decades; and large, complex, aromatic substances, which may be stable within the soil for millennia. Soil organic matter affects all aspects of the soil's biology, chemistry, and physics.

**soil texture** The relative abundance of sand ( $50 \mu\text{m} < \phi < 2 \text{mm}$ ), silt ( $2 \mu\text{m} < \phi < 50 \mu\text{m}$ ), and clay ( $\phi < 2 \mu\text{m}$ ) particles in the soil (USDA criteria). Particle size distribution determines the distribution of pore sizes, which in turn strongly affects the behavior of water in the soil.

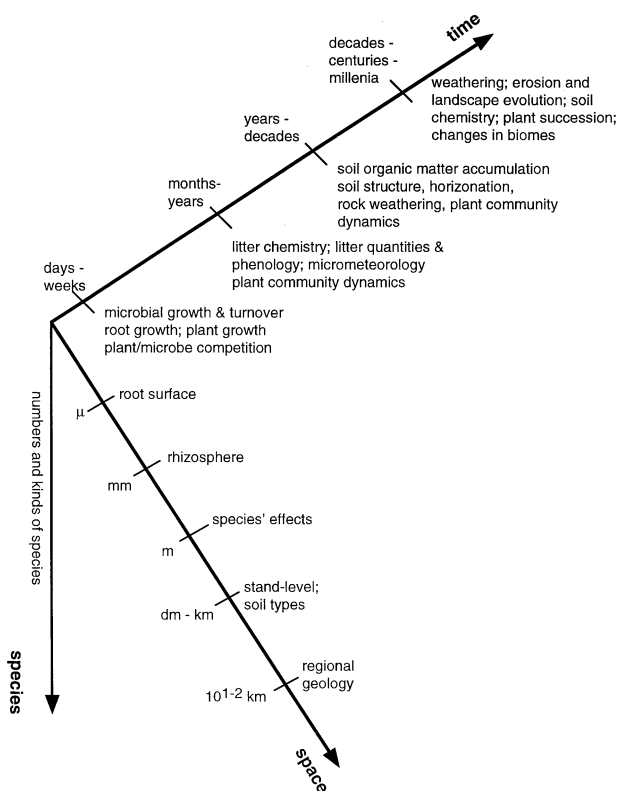
**PLANT-SOIL INTERACTIONS** include a wide range of biological, chemical, and physical effects exerted by soil on plant survival, growth, and reproduction, and reciprocal effects of plants on soil formation, soil physical structure, and the activities of the soil biota.

## I. INTRODUCTION

Plants are often assumed to be passive recipients of the environmental conditions that happen to exist at the location in which they grow. Soils supply mineral nutrients and water, and they provide anchorage, but they also set limits on plant growth. Plants, however, provide the organic substrate that forms the basis for all biological activity in the soil and that forms the soil organic matter; their growth drives the formation of soil from bedrock and alters its chemistry and physics. This chapter reviews

the two-way street that constitutes the plant-soil system and examines the implications of these mutual effects for the management and conservation of biodiversity.

Soil-plant interactions take place over a wide range of spatial and temporal scales (Fig. 1). At a scale of microns, roots etch microscopic channels of weathering within mineral grains and provide a variety of different habitats for microbial growth. These microscale phenomena cause millimeter-scale differences in the abundance and diversity of microbes and the soil micro- and mesofauna. Nutrients flow toward roots over distances of millimeters to centimeters; active uptake causes sharp concentration gradients of nutrient elements, moisture, and acidity. These modifications spread out from the root surface to form a cylinder of differentiated, rhizosphere soil that may be millimeters thick. Mycorrhizal hyphae extend from the root surface into the soil over distances from centimeters to meters; they bind soil particles together into aggregates and contribute to the long-term accumulation of soil organic matter. The quantity and quality of such millimeter-scale effects



**FIGURE 1** Concept of three-dimensional structure of plant-soil interactions. Processes occurring across dimensions of time (minutes to millennia) and space (microns to hundreds of kilometers) occur with respect to the species pool—both the numbers of species and the kinds of species (see text).

vary among plant species, and so patches of soil filled by roots of different species may become chemically and physically different at scales of centimeters to tens of meters, depending on plant size. These species-level effects may translate into stand-level effects. At these larger spatial scales, however, local and regional differentiation of soils due to differences in topography, geological substrate, and geological history may override the effects of plant species and constrain the nature of the plant-soil interactions. But at these scales, plant communities also exert control over the long-term accumulation of organic matter.

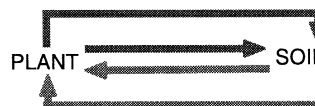
Plant-soil interactions occur over an equally great range of temporal scales. At the level of the individual plants, uptake of nutrients and water occurs more or less continuously over hours, but rates vary over seconds or minutes, depending on environmental conditions. The aggregate effects of this uptake over periods of months (seasons) or years reflects the continuity of plant community composition. Similarly, the exudation of substances used by the microbiota, and the response in microbial growth and metabolic activity, occurs over timescales of minutes to hours, but the aggregate effect on soil properties and on plant growth extend over seasons and years. Over the long term (decades to centuries or millennia), different plant communities cause different types of soil to develop.

Finally, these patterns of interaction themselves depend on the size and diversity of the available pool of plant species. In highly diverse communities, the effects of individual species may be masked by the variability among those effects, or by the overlapping nature of root and shoot systems. In communities that are species-poor, either because of human manipulation (e.g., forest plantations of a single species) or because of biogeographic and climatic constraints (e.g., boreal forests with few species), the individual species-specific effects seen at small scales may translate into effects seen over hectares or square kilometers.

Within each of the temporal and spatial scales of interaction, the mechanisms creating the interaction are subject to feedback, both positive and negative (Fig. 2). Indeed, the presence of feedback in the soil-plant system is one of its salient properties: the results of a particular mechanistic effect alters the environment so as to modify (increase or decrease) the continued activity of that mechanism. Feedbacks have been recognized as crucial in mediating such disparate processes as the supply of mineral nutrients through decomposition, and the activities of soil pathogens and mycorrhizae.

The complex interplay of soil and plants has commanded attention for a long time. Gilbert White, in "The Natural History of Selbourne" in 1778, observed

rhizosphere habitat for micro- and meso-biota  
 root activity effects on mineral weathering  
 litter amounts, timing and chemical composition - source of organic matter  
 microclimate effects: temperature and moisture  
 competitive withdrawals of water and nutrients  
 root growth creates macropores



micro- and meso-biota decompose organic matter, releasing mineral nutrients  
 physical properties affect root proliferation  
 physical properties control mobility of nutrients and water  
 minerology determines intrinsic quantities of nutrient elements present

FIGURE 2 Feedback processes in plant-soil interactions. Changes in plant growth, physiology, and so forth and/or changes in plant community composition can alter soil properties, which in turn will drive further changes to the plants. Conversely, soil properties that constrain the patterns of plant growth, physiology, and community composition will be affected by the plants that grow under those conditions. Positive feedbacks will cause the soil-plant system to evolve in a particular direction, unless deflected by a disturbance external to the system. Conversely, negative feedbacks will maintain the system in a constant state.

that "In a district so diversified with such a variety of hill and dale, aspects, and soils, it is no wonder that great choice of plants should be found." Darwin pointed out that plant growth is modified by the presence of soil fauna, especially earthworms; he noted that Gilbert White had also observed this a century earlier. The German scientist Justus Liebig (1876) founded the scientific basis for agriculture by emphasizing the role of soil in controlling vegetation and crops through the supply of mineral nutrients. Indeed, the perception of soil as the ultimate control on the development of plant communities has been a central paradigm of plant ecology for at least a century.

In a parallel development, the role of plants in shaping soils was recognized with the beginnings of soil science (Dokuchaev, 1879; cited in van Breeman 1995). Hans Jenny, in his classic work "Factors of Soil Formation" in 1941, identified plants (and animals) as one of the primary forces in the genesis of soil from bedrock, and acknowledged that soil scientists extending back to the nineteenth century had perceived plants and soils to be a "coupled system." Conversely, the critically important role of plants in causing soils to form from weathered and unweathered rock was elaborated in detailed studies of primary succession on newly exposed substrate by Crocker and Major in 1955. In the past several decades, the multiple forms of interaction between plants and soil have become a major focus for research.

Plant-soil interactions play a critical part in human interactions with the biosphere. Humans create plant

communities ranging in complexity from single species of crop plants in agriculture to attempts at the restoration of complex communities; the choice of plants is conditioned by soil qualities and alters those qualities. Human-caused changes in atmospheric chemistry—the addition of nutrients such as nitrogen and sulfur as pollutants, the increase in CO<sub>2</sub> concentration—directly and indirectly affect plants, soils, and the interplay between them. Changes in climate, driven by these atmospheric changes, have similar effects. Human-caused extinction of species, the introduction of species into novel habitats, and the spread of weedy species change plant community composition, thus driving changes in soil. Conversely, the widespread degradation of soils from erosion and overgrazing constrains the kinds of plant communities that can develop, often to the detriment of human society. These are but a few examples of the profound importance of plant-soil interactions in the management, exploitation, and conservation of biodiversity.

## II. SMALL-SCALE PROCESSES: MICRON TO MILLIMETER SCALES

### A. The Rhizosphere: Structure of the Root/Soil Interface

Plants contact soil most intimately along the surfaces of their roots. The soil in direct contact with the surface of nonwoody roots, the “rhizoplane,” differs in biological, chemical, and physical characteristics from the “bulk soil” (Fig. 3a). The chemical and physical influence of the root surface on the soils decreases rapidly with distance from the root surface; thus, there is a cylinder of “rhizosphere” soil that is influenced by the root but not directly in contact with it. Its width varies with the texture and mineralogy of the soil and the diameter and physiological activity of the roots, being from <1 to several millimeters. The rhizosphere changes longitudinally along the root with distance from the root tip, root age, physiological stress to the plant, and degree and type of mycorrhizal infection (Fig. 3a).

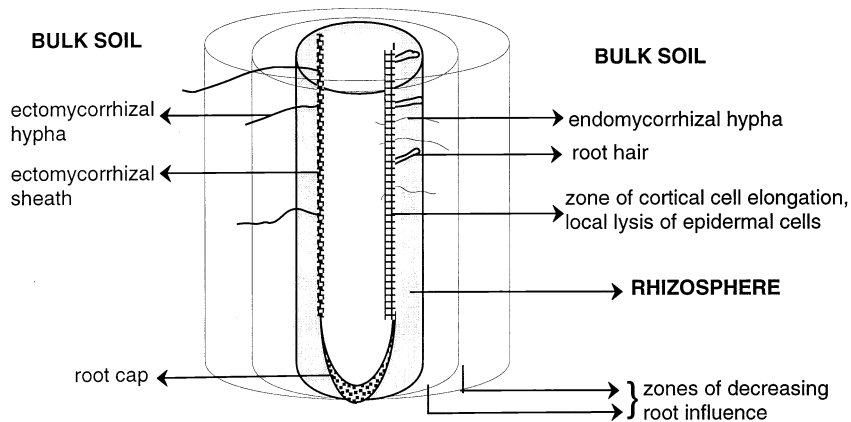
The rhizosphere is most strongly affected by rhizodeposition, the transfer of organic substances from the root into the rhizosphere. This material includes whole cells sloughed from the root cap and the epidermis, secretions of mucilage from these cells, and soluble materials leaked from intact cells and cracks caused by emerging new lateral roots (“exudate”). These materials

mix with polysaccharide secretions of the adhering bacteria. The mucilage, or “mucigel,” lubricates the root surface as it pushes past soil particles and helps to maintain the physical continuity between the root and the bulk soil. In addition, the mucigel may adsorb nutrient ions, especially cations, and prevent their removal from the root surface by leaching. Table I illustrates the diversity of compounds released from various species. The relative proportions of each class of compound may vary with position along the root; for example, amino acids are released in higher amounts near the root tips than farther back.

Studies in a variety of plants have shown that approximately 10 to 20% of the carbon fixed during photosynthesis is released to the soil through these mechanisms. This rhizodeposition accounts for 40 to 90% of the total amount of carbon transferred to the roots from the shoots; the remainder is respired by the root tissues or is used to construct new root biomass. Commonly cited values of 10 to 100 mg g<sup>-1</sup> root dry weight for soluble substances and 100 to 250 mg g<sup>-1</sup> root for insoluble materials give some indication of the amounts released. The amounts of carbon transferred from the shoot to the root system and the fraction of that carbon that is released to the soil vary greatly with plant species and soil conditions (Table II), mycorrhizal infection, soil microbial abundance, and soil physical conditions. The total amounts of carbon involved in the transfer can be quite large (an estimated 1.2 t C ha<sup>-1</sup> yr<sup>-1</sup> to 7.5 t C ha<sup>-1</sup> yr<sup>-1</sup>).

The effects of this input of carbon and nitrogen to the rhizosphere is a large increase in the density, diversity and biomass of the microbiota in the rhizosphere, compared to the bulk soil. Cell counts of microbes in the rhizosphere are usually in the range of 10<sup>10</sup> to 10<sup>12</sup> cells cm<sup>-3</sup>, compared to 10<sup>9</sup> cells cm<sup>-3</sup> in bulk soil. Figure 3b illustrates a typical distribution of microbes with respect to the root surface. Rhizosphere microbial cell densities : bulk soil cell densities range from <1 for microalgae to > 24 for many bacteria; for most microbial groups, ratios are >5. Successions of different microbial species occur (a) over time, as newly produced root tips extend into the soil and are colonized, (b) over the length of the root (from root cap to zones of suberization and secondary thickening), and (c) with distance out from the root surface. Not only is the population of microbes in the rhizosphere larger than in the bulk soil, but it differs in composition. Bacteria are more abundant than fungi, while the reverse is often true in the bulk soil, and more species of bacteria may be found in the rhizosphere. Several genera of bacteria capable of fixing atmospheric N<sub>2</sub>, including *Enterobac-*

### a structure of the rhizosphere



### b distribution of microorganisms relative to the rhizosphere

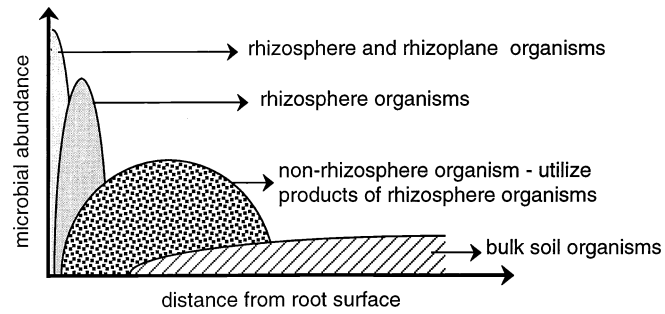


FIGURE 3 (a) A schematic diagram of the rhizosphere, illustrating different zones of the root tip (root cap, zones of cell elongation, and epidermal cell lysis), different types of mycorrhizal fungal growth (endomycorrhizal and ectomycorrhizal), and rhizosphere soil. The diagram also illustrates the zone of decreasing root influence with distance from the soil. (b) Distribution of microorganisms with distance from the roots, based on Bazin *et al.* in Lynch (1990); different species of microorganisms utilize substrates directly derived from roots or secondary materials produced by other microbes in adjacent zones.

*ter*, *Klebsiella*, *Azotobacter*, *Azospirillum*, *Bacillus*, *Pseudomonas*, and *Clostridium*, are only found in the rhizosphere; these nonsymbiotic bacteria may (but not always) contribute to greater plant growth. Species of *Pseudomonas* are particularly important, as many have been found to benefit plant growth. The composition of the rhizosphere biota changes rapidly (within months) in response to changes in the species composition of the plant community, demonstrating the close coupling of the rhizosphere community to the species—specific morphology and physiology of root systems. Variation in the abundance and diversity of plant pathogens also reflects the differences in rhizodeposition among plant species.

Structural complexity in the zone of root/soil contact is provided by extensions of the root surface into the soil matrix. Most plants produce extensive root hairs

from cells just distal to the root cap and zone of elongating cells. The zone of root hair development extends only a short distance along the root (1–4 cm), as the epidermal cells producing the hairs are sloughed off in more basal areas (Fig. 3a). Production of root hairs is itself affected by soil chemical conditions (e.g., pH, ion concentrations), texture, and water content. This zone provides increased structural complexity (e.g., a greatly expanded root surface area that is in direct contact with soil particles). Structural complexity also results from mycorrhizal infection, particularly by ectomycorrhizae, as fungal hyphae spread out from the root surface into the soil (Fig. 3a). Ectomycorrhizal fungi, in particular, produce large quantities of extra-radical mycelium; hyphae may extend for several meters from the host plant. These hyphae greatly expand the contact zone between the root and the soil.

TABLE I

Compounds Identified in Root Exudates, Classified by General Type

Type	Compounds
Simple sugars	Hexoses (glucose, fructose, galactose), pentoses (arabinose, xylose, rhamnose), di- and oligosaccharides (wide variety), uronic acids, amino sugars
Amino acids	$\alpha$ -Alanine, aspartic acid, phenylalanine, glutamic acid, leucine, serine, threonine $\alpha$ - and $\gamma$ -aminobutyric acids, $\alpha$ -aminoadipic acid, many others
Aliphatic acids	Formic, acetic, oxalic, citric, malonic, tartaric, malic, succinic, lactic, palmitic, stearic, oleic, linoleic, linolenic, others
Aromatic acids	p-Hydroxybenzoic, ferulic, o-cumaric, gallic, vanillic, sinapic, shikimic, trans-cinnamic
Complex carbohydrates	Polygalacturonic acid, other hydroxy-, keto-, di- and tri-carboxylic acids
Enzymes	Several phosphatases, peroxidase, invertase, urease, $\beta$ -glucosidase
Vitamins	B group, biotin, pantothenic, nicotinic acids, thiamin, pyridoxine, riboflavin
Proteins and peptides	Unspecified
Plant hormones	Auxins, gibberellins, kinetins, ethylene
Other	Isoflavonoids, thiophenes, benzofurans, terpenes, lectins, phytoalexins, alcohols

<sup>a</sup> Commonly occurring and/or representative specific compounds are listed; in all cases, a very large number of specific compounds have been identified in each chemical class. From Vancura and Kunc (1988).

## B. Nutrient and Water Uptake: The Function of the Root/Soil Interface

Nutrient supply to plants, at the scale of the root surface and the immediately adjacent soil, is a complex function of the chemistry of individual nutrient ions. The availability of nutrients other than nitrogen is ultimately determined by the chemistry of the soil minerals present, which in turn reflects the geology of the parent material. Weathering releases these nutrients from the crystalline lattice of the minerals into solution, making them available for uptake; however, both physical and chemical weathering rates are themselves strongly affected by vegetation. Roots

TABLE II

The Range of Values Observed for the Fate of Photosynthetically Fixed Carbon in the Root System

Plant type	Percentage of total C fixed transferred to root	Percentage of root C in root transferred to soil	Percentage of root C respired
Wheat	27–59	4–29	4–79
Barley	27–54	21–31	6–60
Maize	28–36	18–52	33–50
Pea	44–65	29–43	55–74
Yellow poplar trees	40	–	–
Douglas fir trees	73		

From Lynch (1990).

penetrate cracks and fractures in bedrock, widening them, allowing water to penetrate, and promoting chemical weathering. Chemical weathering depends on the supply of protons for the dissolution and transformation of soil minerals. Plant-derived sources supply most of acidity involved in weathering; these include CO<sub>2</sub> from root and microbial respiration, which forms H<sub>2</sub>CO<sub>3</sub> in the soil solution, organic acids in root exudates and the products of microbial decomposition of plant tissues, and hydrogen ions released from roots in exchange for nutrient cations. Soil genesis also results from the chelation and mobilization of iron and aluminum by the organic acids. As different species of plant generate different types and amounts of organic acid formation, the composition of the plant community strongly affects the nature of the weathering and pedogenic processes that take place. Thus, processes at the scale of the root surface result in broad-scale patterns of correlated vegetation and soil development.

Plant-mediated weathering and pedogenic reactions generate a supply of nutrient cations; their availability for uptake depends on a different set of intersecting plant-soil processes. This includes the diffusion rates, exchange, adsorption, and precipitation chemistry of each element, patterns of water movement through the soil, the morphology of the root system, the amount and diversity of mycorrhizal fungi, and the physiology of uptake by the roots. Nutrients are brought to the root surface in the mass flow of water and through diffusion; roots encounter nutrients as they grow through the soil. The relative importance of each path-

way varies with the different elements (Fig. 4). The availability of sparingly soluble elements, such as iron and manganese, may be enhanced by siderophores in the exudate. These are non-protein-forming amino acids, which complex with these metals, facilitating their mobility and uptake. Organic acids help to mobilize metal nutrients by lowering pH, and help to solubilize P by preferentially binding to the metals. For diffusion-limited nutrients such as P, zones of depletion develop within the rhizosphere (Fig. 5a); this does not occur with highly mobile ions such as  $\text{NO}_3^-$  (Fig. 5b). The size of the depletion zone reflects the balance of supply (diffusion and mass flow) and demand (uptake rate).

Unlike the nutrient cations derived from soil minerals, the availability of nitrogen depends on the microbially mediated decomposition of plant-derived organic matter. It is often assumed that the supply of readily metabolizable carbon from root exudate will promote increased nitrogen mineralization because it relieves carbon limitation, this is not always the case. Theoretical studies have suggested that the net effect of exudates on nitrogen dynamics will depend on the C:N ratio of the exudates relative to that of the indigenous organic matter of the soil, so that balance between the release of N through mineralization versus immobilization in new microbial biomass will vary with the quality of the exudates. Several studies with herbaceous plants have found that the presence of rhizosphere microbes can increase availability of N to the plants, but the effect varies with the availability of N from the soil and the

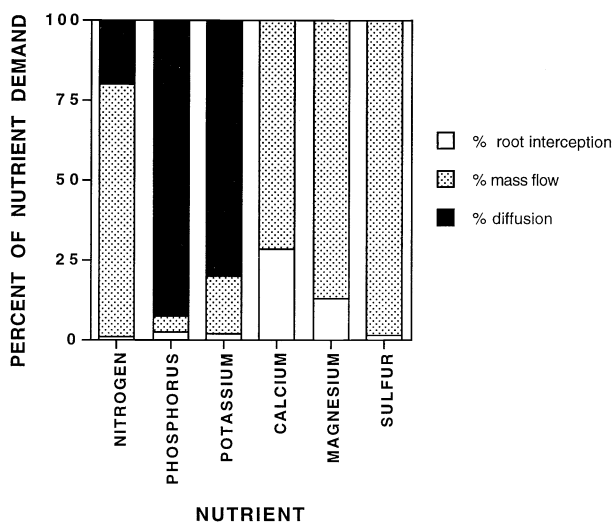


FIGURE 4 The relative importance of different aspects of root activity in the uptake of several nutrients. From Jungk, in Waisel *et al.* (1991), "Plant Roots: The Hidden Half," by courtesy of Marcel Dekker Inc.

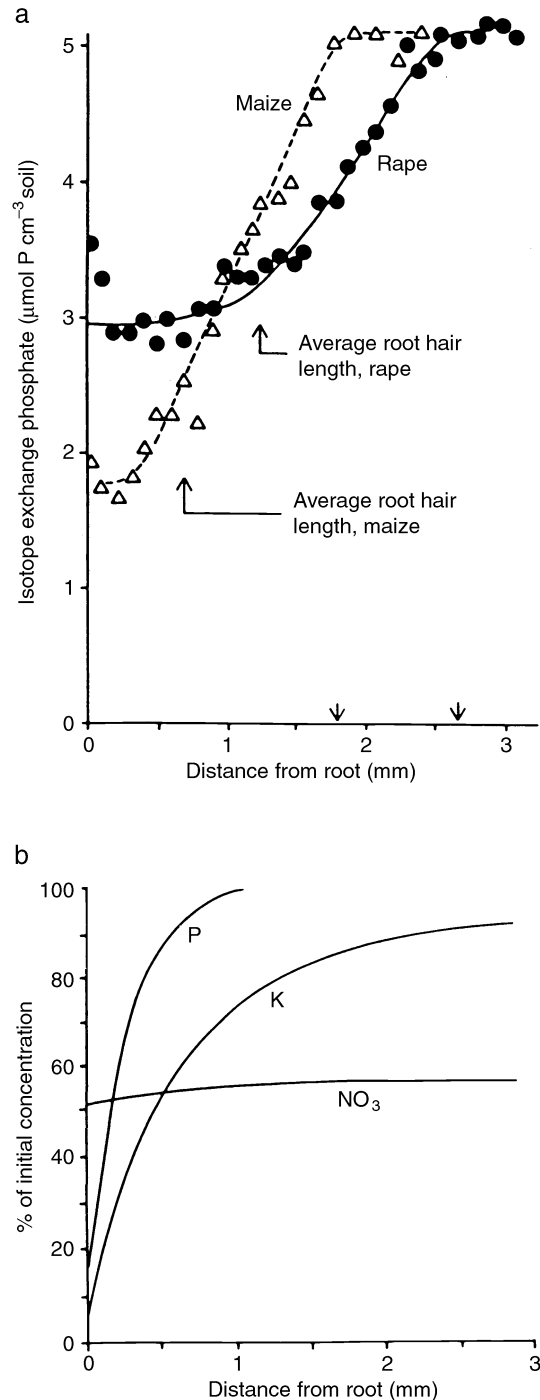


FIGURE 5 (a) Observed patterns of P depletion with distance from the surface of roots of maize and rape; the zone of low P concentrations adjacent to the root is an indicator of the distance that root hairs grow outward from the root surface. (b) calculated depletion curves (as percent of bulk soil concentration) of three nutrients; the depletion rate out from the root surface depends strongly on the chemical mobility (combination of diffusivity and likelihood of adsorption, precipitation, or other immobilizing chemical processes). From Jungk, in Waisel *et al.* (1991), "Plant Roots: The Hidden Half," by courtesy of Marcel Dekker Inc.



kind of plant grown. Alternatively, N may be immobilized within the microbial communities in response to root exudation. Thus, the effects of root exudation on the supply of mineral N to plants is a complex function of the kind of plant, the nature of the organic matter in the soil, and the amounts and composition of the exudate. Root exudate may decrease or increase N availability, even in different horizons of the same soil.

Not least among the effects of roots on the rhizosphere soil is the alteration of soil pH. In addition to the excretion of organic acids and CO<sub>2</sub>, roots exchange protons and hydroxyl ions for nutrient ions, particularly ammonium (NH<sub>4</sub><sup>+</sup>) and nitrate (NO<sub>3</sub><sup>-</sup>). Because nitrate is rapidly taken up when it is present, a net imbalance of anion and cation uptake occurs. Maintenance of electroneutrality during this requires that an anion be extruded in exchange for the nitrate ion; this anion is usually hydroxyl (OH<sup>-</sup>). Conversely, hydrogen ions (H<sup>+</sup>) are excreted during uptake of ammonium. Soil pH can rise as much as two units within the rhizosphere when nitrate uptake is rapid, whereas it may decrease by up to two units when all uptake is in the form of ammonium.

### C. Physical Structure of Soil

Plants modify the physical arrangement of soil particles in a variety of ways. In most soils, the grains of mineral material and the soil organic matter are bound together in aggregates. In the rhizosphere, the mucigel produced jointly by the plant and the resident microbes bind clay minerals together into microparticles of about 50 μm diameter. Fungal hyphae can bind together these microparticles into larger aggregates of 1 to 2 mm diameter. The products of root growth (and root-supported mycorrhizal fungal growth) are especially important in stabilizing the aggregates over time. Moreover, different plant communities create different quantities and sizes of soil aggregates, depending in part on the morphology of the root system and the extension and mass of mycorrhizal hyphae.

Pore size distributions are also affected by roots through several other mechanisms. Alternate wetting and drying of the soil induced by transpiration can affect cracking patterns, especially in clay-rich soils. Roots also preferentially grow into larger pores, or zones of lower penetration resistance, and may thereby enlarge them. The death and decay of roots also creates macropores or channels of highly porous organic matter; these large voids provide preferential pathways for the flow of water. Finally, the degree of contact between the root surface and the soil varies with root age,

TABLE III  
Effects of Roots on Reduction-Oxidation Potential (Eh, mv) of Wetland Soils, Mediated by Root Oxygen Loss

Plant species	Growth form	No plant	With plant
<sup>1</sup> <i>Acer rubrum</i>	Tree	+280	+320
<sup>1</sup> <i>Rosa palustris</i>	Shrub	+220	+240
<sup>2</sup> <i>Salix viminalis</i>	Shrub	+200	+330
<sup>1</sup> <i>Saururus cernuus</i>	Herb	+295	+310
<sup>3</sup> <i>Myriophyllum verticillatum</i>	Herb	+170 to +200	+270 to +280
<sup>4</sup> <i>Isoetes lacustris</i>	Herb	+600	+50 to +200
<sup>5</sup> <i>Ranunculus circinatus</i>	Herb	-200	+400
<sup>6</sup> <i>Myriophyllum tenellum</i> + <i>Isoetes braunii</i>	Herbs	-100	+190
<sup>7</sup> <i>Spartina alterniflora</i>	Herb	-50	+150

For each study, the maximum observed difference between planted and unplanted soils are reported.

<sup>1</sup> Havens (1997). *Wetlands* 17, 237-242.

<sup>2</sup> Grosse, W. (1997). In *Trees* (Rennenberg *et al.*, Eds.). Backhuys Publ, The Netherlands.

<sup>3</sup> Carpenter, S., Elser, J., and Olson, K. (1983). *Aquatic Bot.* 17, 243-249.

<sup>4</sup> Wiim-Andersen, S., and Andersen, J. M. (1972). *Limnol. Oceanog.* 17, 948-952.

<sup>5</sup> Flessa, H. (1994). *Aquatic Bot.* 47, 119-129.

<sup>6</sup> Jaynes, M. L., and Carpenter, S. R. (1986). *Ecology* 67, 875-882.

<sup>7</sup> Howes, B. L., Howarth, R. W., Teal, J. M., and Valiela, I. (1981). *Limnol. Oceanog.* 26, 350-360.

amount of exudation, and root physiological status (e.g., hydration).

Roots also lose molecular oxygen to the soil, as air diffuses from leaves and stems through lacunae in the cortex tissues and through cracks in the surface of the root epidermis. This "radial oxygen loss" is particularly large in wetland plants, in which the root cortex develops extensive air spaces ("aerenchyma") through the degradation of cell walls. Root oxygen loss can be sufficiently high to alter the reduction-oxidation potential of the surrounding bulk soil (Table III). As a result, a zone of 2 to 3 mm of oxidized soil (often apparent as a coating of orange-red iron oxide grains) may develop around the root surface under anoxic, waterlogged conditions. This zone of oxidized iron minerals acts as a trap for other metal ions in the soil, such as cadmium, copper, lead, and zinc, which become enriched in the surface concretions as much as 10-fold over the bulk soil concentrations. Moreover, by maintaining a higher redox potential in the soil, denitrification and methano-

genesis may be inhibited; this both decreases the presence of toxins (methane) adjacent to the root surface and also alters the net carbon and nitrogen balance of the soil.

Several species of wetland plants, including the yellow water lily (*Nuphar lutea*) and common reed (*Phragmites australis*) take advantage of a thermal diffusion mechanism to develop positive air pressures in young leaves and drive a mass flow of air into the roots. This air can effectively oxygenate long rhizomes (>2 m long in *Phragmites*) plus the surrounding soil. Only a few wetland plants employ mass flow to move air through the roots and back to the leaves; the rest rely on diffusion of oxygen through the aerenchyma. In most plants, however, the aerenchyma also provides a pathway for venting methane produced in anoxic soil to the surface; plant stems account for >90% of the methane release to the atmosphere from waterlogged soils.

Soil physical properties, including texture, bulk density, temperature, and moisture-holding ability, affect the number, size, diameter, and branching patterns of roots in a great diversity of ways. In many plants, clayey soils inhibit root growth compared to sandy soils. This reflects the ability of the roots to grow into pores between soil particles and to penetrate aggregates of particles. Mechanical impedance, reflected in the soil's bulk density, affects all aspects of root growth, root system morphology, and uptake functions. Temperature is another factor affecting root growth; optimal temperatures vary among species and are usually lower than the optimal temperatures for shoot growth. Aeration and moisture status, which are inversely related to each other, are jointly controlled by climate and by the transpiration rate of the plant community. Temperature, aeration, and moisture content also affects the morphology of the root system (root diameters, density of root hairs and frequency of branching). Not surprisingly, texture, mechanical impedance, temperature, moisture, and aeration are interrelated (as they all are affected by particle size distribution), and their effects on root growth are interactive.

### III. MESOSCALE PROCESSES: CENTIMETERS TO METERS

The microscale interactions between root surfaces and the adjacent soil are part of a larger set of interactions that occur over scales of centimeters to meters. Strategies of resource allocation by whole plants, including changes in growth rate, differential growth of above-

ground and below-ground tissues, changes in root longevity and turnover, and physiology (i.e., in changing metabolic capacity for nutrient uptake) affect the vertical and horizontal extent of root penetration through the soil, thereby affecting the spatial extent over which microscale plant-soil interactions occur. At this larger scale, the deposition of litter from the above-ground tissues becomes as important as the patterns of root growth within the soil; indeed, litter input may dominates the biological processes within the soil. However, spatial variation in soil properties, driven by topography, bedrock geology, and other exogenous factors, also affects root system morphology and growth, and condition the fate of the litter.

#### A. Nutrient Acquisition and Root Growth: Individual Plants

The acquisition of nutrients and water is strongly affected by the architecture and size of the root system, a function of the allocation of photosynthate below ground. Conversely, the availability of nutrients and water molds the growth of the root system, thus setting up a feedback interaction. The metabolic capacity to absorb nutrients not only varies among plant species, but also varies within an individual as the availability of nutrient elements changes. Nutrient uptake capacity is often inversely related to availability, because transport systems in the root cell membranes increase in activity as nutrients become more limited in supply. For example, barley plants growing in nitrogen-limited cultures have uptake rates of both ammonium and nitrate more than 200% greater than plants growing with excess N; similarly, phosphate uptake rates increase by 400% in phosphorus-limited cultures, and sulfate uptake rates increase by almost 900% in sulfur-limited culture. Similar inverse relationships between availability, as measured in soil solutions or extracts, and short-term uptake rates have been found for a wide range of wild plants, including both herbs and trees. However, while relative uptake capacity varies within a species inversely with the availability of nutrient elements, absolute rates of uptake of plant species from low-nutrient environments may be less than, equal to, or greater than plant species found in high-nutrient environments. Thus, the effects of nutrient availability on root physiology depend on both the species of plant and the variation present in the soil.

Nutrient acquisition is also affected by the abundance and distribution of the roots. Different kinds of plants are characterized by differences in the relative allocation of biomass to roots and shoots (Fig. 6). In

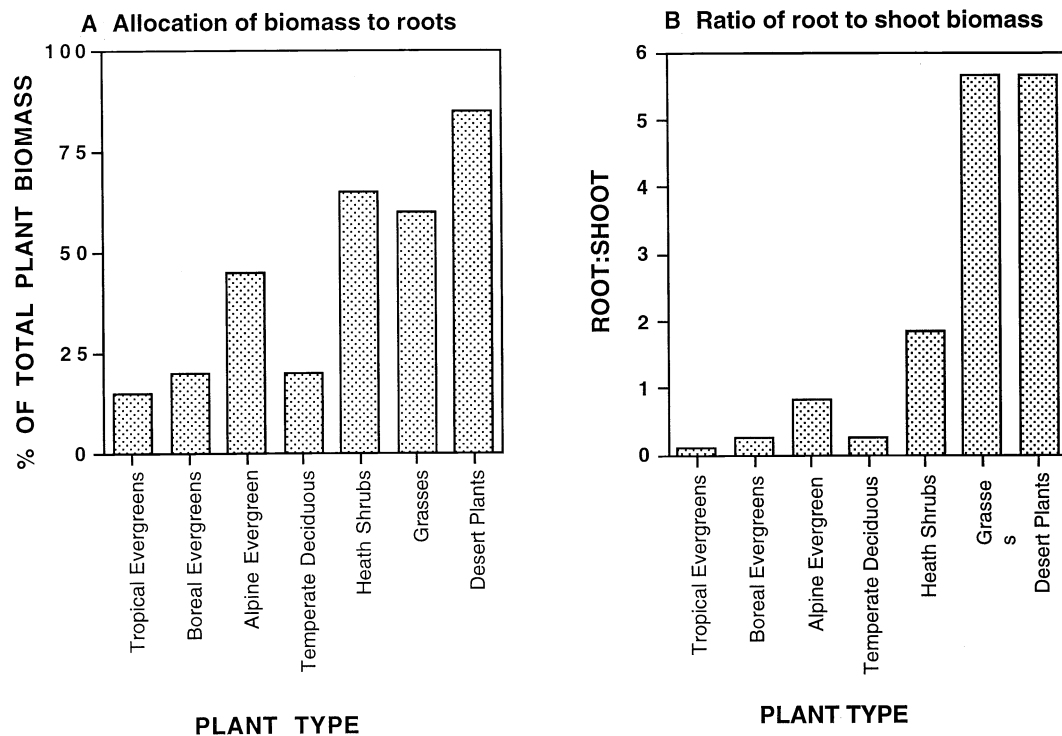


FIGURE 6 The relative allocation of biomass to roots, as (a) the percentage of total plant biomass and (b) with respect to shoot biomass, for plants from various habitats. From Klepper, in Waisel *et al.* (1991).

general, grasses tend to have higher relative allocation to roots than do woody plants, although trees and shrubs in highly nutrient-limited environments, such as alpine tundra or bogs, have higher root:shoot ratios than do woody plants from more nutrient-rich environments such as tropical or temperate deciduous forests.

Root systems also adjust to varying nutrient availability by changes in morphology (length, diameter, branch density, etc.). Root length density (the length of roots per unit volume of soil) commonly increases in zones within the soil that have higher concentrations of limiting nutrients, as has been frequently demonstrated in experiments (Fig. 7). However, such responses are not seen in all species, and the response to variations in nutrient availability depends on a plant's tendency to produce thick versus thin and fast-growing versus slow-growing roots. Increases in root length in one area, in response to a patch of high nutrient availability, are often compensated for by decreases in other parts of the root system. Furthermore, nutrient concentrations in plant tissues do not necessarily increase, and plant growth also often does not show an overall increase, suggesting that plants compensate for increased growth in one part of the root system in re-

sponse to increased nutrient availability by decreasing growth elsewhere.

The response also depends on the plant's tendency to produce herringbone versus dichotomous patterns (Fig. 8). The optimal response depends in part on the plant's ability to allocate carbon to the root system (versus to shoots, leaves, or reproductive structures), the size of an area of increased nutrients, and the chemistry (i.e., the diffusivity) of the nutrient at stake. Research by A. Fitter has shown that herringbone patterns are most efficient for mobile nutrient ions, but herringbone and dichotomous patterns are equivalent for immobile ions. Others have shown that herringbone patterns are more common in roots growing in infertile soils; the lower amount of branching allows the root system to explore a larger volume of soil. In fertile soils and in nutrient-rich microsites, branching patterns become more dichotomous; this pattern presumably permits more intensive and complete exploitation of the soil resource.

The heterogeneity of root systems in response to soil conditions reflects the high degree of heterogeneity in soil conditions at scales of centimeters to tens of meters. In agricultural fields, for example, chemical and physi-

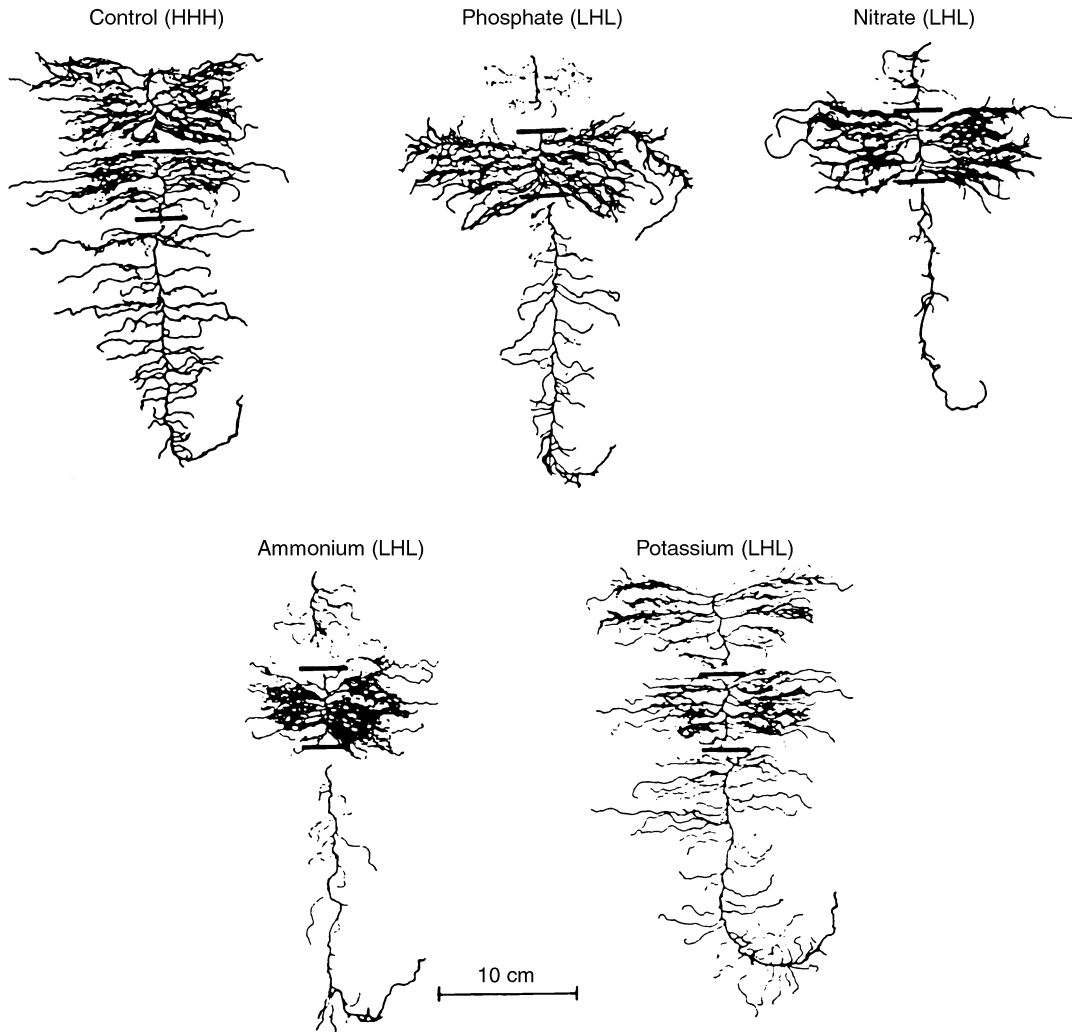


FIGURE 7 The results of a classic experiment by Drew (1975, Comparison of the effects of a localized supply of phosphate, nitrate, ammonium and potassium on the growth of the seminal root system, and the shoot, in barley. *New Phytol.* 75, 479–490) in which the roots of barley plants grew through zones of nutrient-enriched soil. Control plants received all nutrients throughout the root system; the other plants received high concentrations of the target nutrient only in the middle of the root system.

cal soil properties vary significantly between points that are only tens of centimeters apart, despite the apparent uniformity that might be expected from the nearly flat topography, plowing and planting of a single species of crop plant (Fig. 9). Extreme patchiness in one soil resource may not be paralleled by similar patchiness in other resources in the same site; the distribution of mobile, limiting elements like nitrate can vary greatly among points only 3 cm apart while immobile, nonlimiting elements (P, K) do not vary, even over large distances.

In natural communities, significant variation in nutrient concentrations and nutrient supply rates (e.g., nitrogen mineralization rates) have been documented

in soils beneath different co-occurring species of grasses and trees. Spatial variation in soil properties also is generated by topographic variation. Even small changes in elevation or geomorphic position (e.g., hill crest, upper and lower slope positions, valley bottoms) are associated with differences in soil profile development, accumulation of organic matter, soil texture, the mineralization rates of nutrient elements, and gaseous losses through microbial processes such as denitrification. At scales of tens to hundreds of meters, differences in plant community composition are clearly correlated with differences in soil properties related to topography.

Spatial heterogeneity of soils is particularly strongly developed in plant communities in which either plants

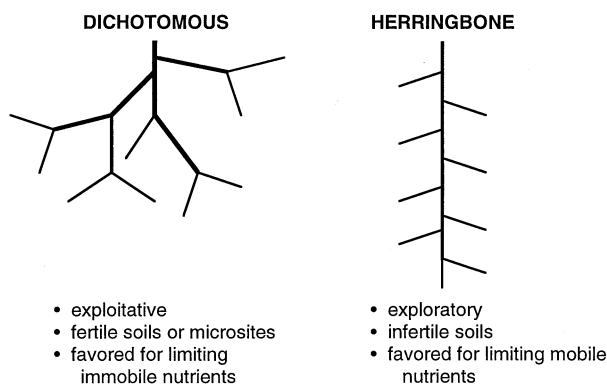


FIGURE 8 Alternative patterns of root growth. Dichotomous branching systems have more internal segments of the root system, compared to herringbone branching systems. The internal segments, or links, tend to be larger in diameter and therefore are more “expensive” to construct and maintain (in units of carbon fixed during photosynthesis). Plastic responses of roots to heterogeneous soil conditions involve changes in branching pattern, branch number and length, branching angle, root diameter and tissue density (observed as changes in specific root length and area, or length and area per unit root mass), and density of root hairs. Based on A. Fitter, in Waisel *et al.* (1991) and other papers.

are patchily distributed over otherwise barren ground (typical in many arid ecosystems) or in which differing life-forms (e.g., trees and grasses) are interspersed (e.g., savanna ecosystems). In deserts, the scattered plants form “islands of fertility” because of their ability to trap both plant litter and wind-blown fine soil particles. The

soils within and beneath the scattered shrubs or tussock grasses in these systems have higher amounts of nitrogen, phosphorus, organic matter, silt or clay-sized particles, and moisture than the intervening bare areas. Similarly, patches of woody vegetation within grasslands accumulate higher amounts of nutrients and organic matter, again leading to patches of higher fertility within the grassland landscape. Plasticity of root system morphology and physiological capacity allows plants and communities to respond to this spatial variation in soil properties.

## B. Nutrient Acquisition and Root Growth: Stand Scale

The response of individual plants to variation in soil properties manifests itself at the level of the community as changes in root biomass and the production and mortality rates of the roots. Many studies have shown that root biomass tends to be lower in more nutrient-rich sites, and experimental studies in which nutrients (particularly nitrogen) are added often confirm these observations. Similarly, it has been noted that plants typical of infertile sites have a higher ratio of root biomass to above-ground biomass than do plants from fertile sites. In a comprehensive review of available literature on nutrient cycling and root production, Vogt *et al.* (1986) found that the mass of fine roots in forest ecosystems is higher in forests with larger amounts of nitrogen in the forest floor, implying that the biomass

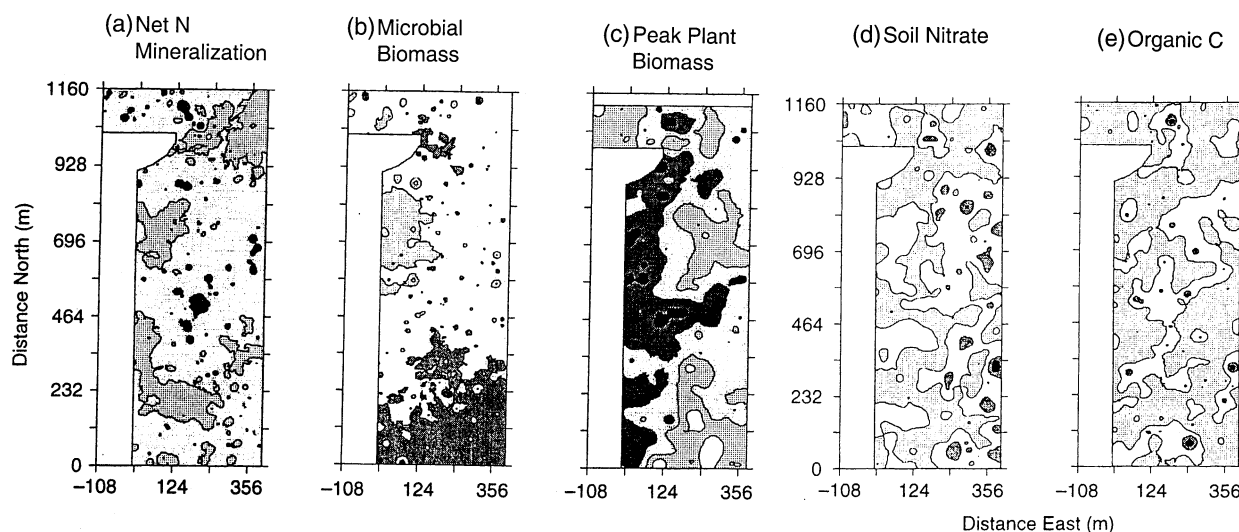


FIGURE 9 Spatial patterning of physical and biological soil properties in a soybean field in Michigan (United States). Units: (a) net N mineralization,  $\text{mg N g}^{-1} \text{d}^{-1}$ ; (b) microbial biomass,  $\text{mg C g}^{-1} \text{soil}$ ; (c) peak plant biomass,  $\text{g m}^{-2}$ ; (d) nitrate,  $\text{mg N g}^{-1} \text{soil}$ ; (e) elevation, m. From Robertson *et al.* (1997), Soil resources, microbial activity, and primary production across an agricultural ecosystem. *Ecol. Applic.* 7, 158–171.

of roots increases when nitrogen is sequestered in compounds that decompose slowly and accumulate in thick forest floors. This supports the idea that root biomass is greater in infertile than in fertile soils.

Rates of production and mortality of fine roots also respond to nutrient status, although experimental studies have documented both increases and decreases in response to nutrient additions. For example, Vogt *et al.* (1986) suggested that the rate of root turnover (input of new roots per year) is higher in forests with smaller amounts of nitrogen input in litterfall (an index of nitrogen limitation), but experimental addition to N to northern hardwood forests caused a decrease in mortality rates of fine roots. The contradictory data may reflect differences between the "normal" growth patterns of roots relative to the inherent fertility of the soil and the response of existing root systems to large, sudden changes in fertility induced by experimental nutrient additions.

### C. Competitive Interactions between Plants and Microbes

As the plant community and the soil microbiota require the same resources (nutrients and water), there may be competition between the microbiota and plants. Microbial demand for nitrogen is closely tied to the availability of degradable carbon, as carbon is often considered to be their most limiting resource. In this context, the likelihood of competitive interactions with plants depends on the C:N ratio of the available carbon resources: root exudates tend to have low C:N ratios (about 15:1), litter normally has high C:N ratios ( $>>30:1$ , and  $>100:1$  for woody plant tissues), and soil organic matter ratios are in the range of 10 to 20. Theoretical analyses have suggested that the relative mineralization versus immobilization of N derived from root exudates varies in relation to the relative availability and quality of the soil organic matter, where "quality" refers to the metabolic capacity of the microflora to degrade the components of the organic matter. When N is limiting (i.e., when C:N ratios are high), microbes may take up mineral N from the same sources (soil solution, exchange surfaces) that plants use, resulting in competition.

Experimental additions of mineral N ( $\text{NH}_4^+$ ,  $\text{NO}_3^-$ ), to forest soils have shown that the microbiota captures a substantial proportion, often a majority, of the added N, but that nitrification rates are decreased when plants are present. These results suggest that while the microbiota competes successfully for mineral N in solution, plants can limit the availability of  $\text{NH}_4^+$  to the ammonia

oxidizers (nitrifiers). However, long-term studies suggest that although microbes may outcompete plants for mineral N in the short term, the long-term activity of root systems, compared to the shorter, more localized bursts of activity of the microbes in response to changing temperature and moisture regimes, may allow plants to compete successfully. In this regard, the influence of plants on the soil moisture regime, through uptake of water and loss in transpiration, may indirectly affect competitive interactions for N by causing limitation of microbial activity due to soil drying. Furthermore, competition for N may be modified by differential use of the inorganic and organic forms of N in the soil.

### D. Litter Inputs to the Soil: Effects of Plant Tissue Chemistry

Perhaps the most widely studied and best-known interaction of plants and soil is the supply of organic matter to the soil through the deposition of litter. Litter deposition includes "fine litter," the leaves, herbaceous stems, and smallest woody twigs, "coarse woody debris," the large twigs, branches, and boles of shrubs and trees, "greenfall," of leaves and leafy twigs clipped by herbivores such as squirrels, and root litter, the dead roots of all size classes. However, most measurements of "litter" refer only to fine litter, as this is most easily measured. Litter is the largest source of organic matter supplied to the microbiota of the soil, and thus it fuels all of the microbial processes responsible for decomposition, nutrient cycling, and reduction-oxidation reactions of soil minerals and also is the ultimate source of the soil organic matter.

Litter inputs vary widely in quantity and timing. Tropical rain forests may have  $>2,000 \text{ g m}^{-2} \text{ yr}^{-1}$  litterfall, whereas tundra ecosystems may have  $<200 \text{ g m}^{-2} \text{ yr}^{-1}$ . Deposition of fine litter is seasonal in temperate zones and continuous in the tropics, but the formation of coarse woody debris may be sporadic (i.e., large inputs following major storms, or infrequent inputs at any time of year). Amounts of coarse woody debris are comparable (about 120 to 3000  $\text{Mg m}^{-2} \text{ yr}^{-1}$ ) but much more patchily distributed. Greenfall production is also sporadic, as it reflects the population dynamics of the herbivorous insects or mammals that cause it. The timing of root death is quite variable among ecosystems; in some forests, root growth and death are pulsed, but in other sites, production and mortality may occur simultaneously and not in synchrony with seasons. Assuming that root populations are at steady state, so that root production rates can be used as an estimate of root litter deposition rates, the amounts of root litter

production similarly range from roughly  $100 \text{ g m}^{-2} \text{ yr}^{-2}$  to  $1,700 \text{ g m}^{-2} \text{ yr}^{-2}$ .

The ratio of above-ground to below-ground litter inputs is also quite variable among ecosystems (Table IV). Root litter inputs can be very high in grassland ecosystems as well, with ratios of below-ground : above-ground litter mass much greater than 1. The variability of the relationship between above-ground and below-ground litter production was clearly demonstrated by Raich and Nadelhoffer (1992), who found no statistically significant relationships between the amount of fine root production and the amount of above-ground litter production over a wide geographic range of locations and forest types.

As important as the quantity and timing of input of litter is, the chemical composition of the plant tissues is perhaps more important as the ultimate control on the activity of the soil microbiota. The ability of microbes to decompose the litter material depends on (a) the ratio of carbon to nitrogen and (b) the chemical forms in which the carbon occurs. Carbon occurs in plant tissues in several forms, including "nonstructural carbohydrates," which include soluble and membrane-bound contents of the cytoplasm and vacuoles. Cell walls contain cellulose (an unbranched chain of glucose residues), hemicelluloses (branched chains of glucose residues), and pectins; these materials are variably bound to lignin, a complex phenolic material that provides strength to the tissues. Because of its molecular structure, lignin is extremely resistant to decomposition;

TABLE IV

Above- and Below-Ground Inputs of Organic Matter to the Soil in Forest Ecosystems

Ecosystem	Root litter input $\text{kg ha}^{-1} \text{ yr}^{-1}$	Above-ground input $\text{kg ha}^{-1} \text{ yr}^{-1}$	Ratio
Tropical broadleaf evergreen	n.d.	$9,438 \pm 1,104$	
Warm temperate deciduous	$5,732 \pm 1,970$	$9,369 \pm 790$	0.61
Warm temperate evergreen	$9,053 \pm 453$	$4,432 \pm 234$	2.04
Cold temperate deciduous	$2,280 \pm 920$	$3,854 \pm 213$	0.59
Cold temperate evergreen	$6,152 \pm 1,007$	$3,144 \pm 194$	2.91
Boreal evergreen	$998 \pm 208$	$2,428 \pm 204$	0.41

From Vogt *et al.* 1986; averages of values reported for several studies for each community type.

TABLE V

An Example of the Role of Plant Tissue Chemistry in Regulating Decomposition Rates

Plant	%N	% lignin	%C	C:N	<i>k</i>
White spruce	0.52	13.1	47.0	89.4	-0.757
Douglas fir	0.61	20.5	46.7	76.6	-0.850
Balsam poplar	0.58	13.6	44.6	76.9	-0.907
Aspen	0.64	14.1	45.5	71.1	-1.351
Grass	0.81	15.2	45.8	56.5	-1.391
Dogwood	0.78	7.5	41.5	53.2	-1.706
Rose	1.15	3.4	44.5	38.7	-1.841
Cow-parsnip	1.31	7.0	39.6	30.3	-2.550

From Taylor *et al.* (1989). Nitrogen and lignin content as predictors of litter decay rates: a microcosm test. *Ecology* 70, 97-104. Tissues were allowed to decompose in laboratory microcosms for 4 months. Regression analyses of the data showed that the best predictors of mass loss were the C:N ratio, %N and lignin:N, in that order. Nitrogen content was more important early in decomposition; lignin became more important as decomposition proceeded.

relatively few soil microorganisms have the requisite enzymes to degrade it. Other important carbon-containing compounds include polyphenols and tannins; these compounds act both as deterrents to herbivores and as inhibitors of decomposition. The nitrogen in plant tissues is largely found in protein; almost 50% of the total leaf nitrogen may be in enzymes and structures associated with photosynthesis.

A large number of studies over many years have shown that the ability of the soil microbiota to decompose plant material depends on both the C:N ratio and the lignin:N ratio. The C:N ratio describes the quality of the substrate (e.g., litter material) relative to the demand by the microbiota for both energy and nitrogen. As the C:N ratio of the microbiota is lower (about 3-5 for bacteria and 10-12 for fungi) than any plant material (range from about 12 for leguminous herbs such as vetch and alfalfa to 400-600 for hardwood and coniferous sawdust, respectively), decomposition will be slowed in proportion to the energy/nutrient imbalance between the decomposers in a given soil and the particular substrate available. However, decomposition rates are also affected by both the absolute amount of lignin in the plant tissues and by the lignin:N ratio. Table V illustrates typical results of a comparative study of decomposition rates (*k*) of different plant materials, showing the close relationship between tissue chemistry and decomposition. Although studies differ in identifying C:N, lignin:N, nitrogen concentration or lignin concentration as the primary factor regulating decom-

position rates, and in describing the relationship between decay rate and lignin:N as linear or nonlinear, almost all identify some subset of these characteristics as critical.

Decomposition not only produces CO<sub>2</sub>, but also results in (a) the production of microbial metabolic products that are stable in the soil, and form the passive or long-term component of soil organic matter, and (b) the mineralization of nutrients such as nitrogen, sulfur, and phosphorus, which are bonded to carbon. Release of these nutrients into the soil solution is, again, a function of the availability of the nutrients in the plant litter substrate relative to the demand by the microbiota, as modified by the efficiency of assimilation of the microbes. Thus, the chemical composition of the plants that compose a community will affect the rates of accumulation of organic matter in the soil, as well as the rates of mineralization of limiting nutrients for plant growth. Alterations of plant community structure, through disturbance, succession, or management, can precipitate changes in nutrient cycling rates and carbon sequestration rates that create feedbacks to the plant community. For example, extensive browsing by moose on birch trees in boreal forests allows white spruce to grow up in their place; the switch from readily decomposable birch leaves to recalcitrant spruce needles causes decreases in the availability of nitrogen, which in turn favors the slow-growing spruce over nitrogen-demanding birch. This positive feedback maintains spruce-dominated forests until stand-destroying fires permit birch to become reestablished.

#### IV. LARGE-SCALE INTERACTIONS: STANDS AND BIOMES

The small and medium-scale interactions described earlier contribute to larger-scale patterns that reflect, on the one hand, the structure and composition of the whole plant community and, on the other hand, geographical effects on soil structure. These patterns may change over time, in response to successional changes in the plant community and in response to the geomorphological evolution of the landscape.

##### A. Geographic Patterns of Soils in Relation to Vegetation

In the classic formulation of Hans Jenny, the formation of soils reflects not only climate, bedrock geology, time, and local topography, but also vegetation. Over

the broad geographic scale of biomes, regular patterns of soil characteristics are associated with patterns of plant community composition (Table VI). In polar regions, extreme climatic conditions constrain the development of vegetation; without organic matter inputs, microbial activity is low, weathering processes are extremely slow, and soils are barely more than bedrock material. In boreal and temperate forests, differences in the kinds of plant species result in differences in the rates of decomposition and accumulation of soil organic matter and the production of organic acids; these processes in turn affect the kinds of chemical weathering and pedogenic processes that occur. In grasslands, the high production rate of fine, nonwoody roots introduces both organic matter and the root-associated processes described earlier to depths of several meters; rapid decomposition and incorporation of surface litter combined with the pattern of root production create soils with very different properties than those of forest soils. Indeed, many studies have examined soils beneath forests that developed following long-term management of an area as grassland or cropland; rapid changes in the amounts, spatial distribution, and chemical quality of the soil organic matter cause changes in cation and metal chemistry associated with the changed chemistry of organic acid production and correlated changes in nitrogen availability and microbial activity.

##### B. Variations in Soils within Biomes: Effects of Species Composition

Within biomes, soil properties vary geographically with the species composition of the vegetation. The differential effects of particular tree species on soil properties result in parallel differences among stands when different species are dominant. These effects are clearly shown when single-species plantations are compared (Fig. 10). The effects of different plant species on soil acidity is particularly important, as changes in pH and acidity affect the mobility of cationic nutrients and can result in large losses through leaching. The differences are due both to differences in the amount of cations stored in the vegetation and also to differences in the loss of cations in water leaching through the profile. Cations lost through leaching may cause a permanent change in the quality of the soil, unless weathering rates are equal to loss rates.

Successional changes in plant communities may in part be driven by plant-caused changes in soil structure and chemistry. In many communities, high rates



TABLE VI  
Regional Characteristics of Soils and Vegetation

Vegetation	Climate	Soil order	Speed of Pedogenesis	Weathering processes	Horizons present
Cold desert	Extremely dry, cold	Entisols	Extremely slow	Precipitation of carbonates	Only C
Tundra	Dry, cold	Inceptisols	Slow	Organic acid formation clay weathering Fe-hydroxides	Thin O, A, B
Boreal forest	Very moist, cold	Spodosols	Slow to moderate	Organic acid leaching Fe-hydroxides clay weathering	Thick O, A, leached E differentiated B
Temperate forest	Moist, mesic	Inceptisols alfisols	Moderate	Organic acid formation clay weathering Fe-hydroxides	Thin O, A, variable E, deep, differentiated B
Grassland	Moist to dry mesic	Mollisols	Moderate	Nonmobile organic acids; CaCO <sub>3</sub> precipitation, Fe-hydroxides	O absent, deep A, thin B, little differentiation
Savanna	Moist to dry warm	Vertisols, Udisols	Fast	Nonmobile organic acids, clay weathering, variable CaCO <sub>3</sub> and Fe-OH formation	No O, deep A, differenti- ated, very deep B
Desert	Dry warm	Aridisols	Slow	CaCO <sub>3</sub> precipitation Fe-OH formation	No O, A, shallow B
Tropical rainforest	Wet, very warm	Oxisols	Very fast	Mobile organic acids Fe, Al-OH formation strong clay weathering	Thin O, A thin E, very deep, differentiated B

Based on Ugolini and Spaltenstein (1992). Pedosphere. In *Global Biogeochemical Cycles*. (Butcher, S. S., Charlson, R. J., Orions, G. H. and Wolfe, G. V. Eds.), pp. 123–145. Academic Press, San Diego.

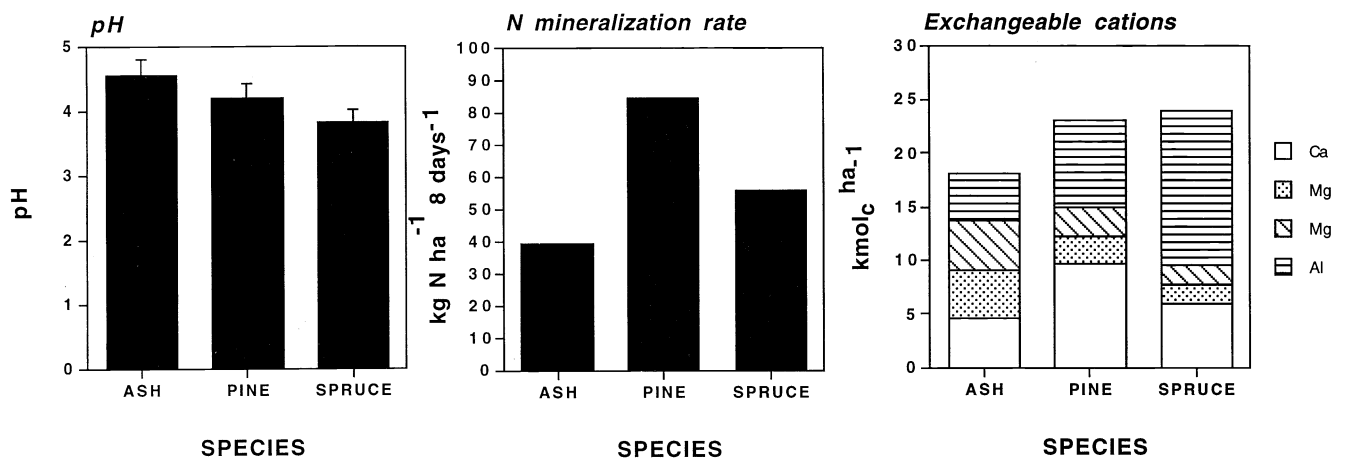


FIGURE 10 Variations in soil properties among different tree plantations. Data from Binkley and Valentine (1991), Fifty-year biogeochemical effects of green ash, white pine, and Norway spruce in a replicated experiment. *For. Ecol. Manage.* 40, 13–25.

of nitrogen cycling (mineralization and nitrification) are observed early in succession, in association with herb- or shrub-dominated communities, but these rates decrease markedly as trees come to dominate the community. These decreases in nitrogen cycling rates are associated with changes in litter chemistry (more recalcitrant material, lower nutrient concentrations) and in root quantities and distributions. A good example of this process is demonstrated by boreal ecosystems. Early in succession, N-fixing alders (*Alnus spp.*) in shrubby communities support high rates of nitrogen mineralization and nitrification; these rates decline rapidly when the alders are replaced by poplars (*Populus balsamifera*) and white spruce (*Picea glauca*). The decline in nitrogen cycling rates that accompanies the successional shift from alders to poplars is driven not only by differences in the N content of the litter, but also by the presence of high concentrations of tannins and phenolics in the spruce needles. The tannins inhibit microbial activity, while the phenolic compounds are used as a carbon source by the microbiota; however, both compounds result in N immobilization.

Invasions of trees into grasslands are associated with a wide range of changes in soil properties. For example, water content of the soils changes due to (a) differences in inputs (stem flow and drip from canopy leaves accentuate inputs, thick litter layers promote infiltration, but high rates of interception by the canopy can reduce total precipitation reaching the surface, and interception within the litter layer can prevent moisture from reaching mineral soil), (b) differences in soil porosity (large pores and channels associated with the growth of woody roots, and the growth of such roots to greater depths than the grasses), and (c) differences in losses (higher transpiration rates by trees, but shading decreases soil temperature and therefore evaporation rates, as does the accumulation of thick litter layers). Large withdrawals of cationic nutrients occur, which are sequestered in the larger amounts of perennial biomass of woody plants. Nitrogen becomes increasingly unavailable, as recalcitrant, lignin-rich litter accumulates. Finally, greater acidity of leaf litter tissue, greater production of organic acids from litter, and protons released through cation and ammonium uptake all drive acidification of the soil profile, with concomitant changes in both leaching loss of cations and weathering rates.

Just as changes in species composition may drive changes in soil properties, soil properties can affect successional patterns. Fertile soils are thought to support faster rates of succession (i.e., rates of species re-

placement), but with lower plant diversity than infertile soils. However, a wide variety of patterns have been observed, and many of the studies involve applications of nutrient fertilizer to experimental plots. Additions of nutrients have promoted the growth of early successional annual species, or the persistence of perennial herbs at the expense of woody pioneers, but opposite results have also been observed.

Successional processes have also been related to the mechanism of competition among plant species for soil resources. Tilman (1988) has shown that competitive ability is related to the capacity of a plant to obtain nutrients from the soil solution and thereby reduce concentrations below the minimum needed for competitors; this ability is balanced by the need to compete for light, so that the outcome of competition (and thereby the composition of a community) reflects the tradeoffs in competing for different resources. Plant growing on infertile soils tend to have lower nutrient requirements (low uptake rates but compensating low growth rates) and higher rates of nutrient conservation (higher efficiency of nutrient use) compared to plants typically found on fertile soils. The tradeoffs among characteristics promoting competitive success on infertile soils (strong competition for nutrients) versus success on fertile soils (competition for light being more important) ultimately determine the successional process on a given site.

Clearly, as soil properties themselves are dynamic, and change with changes in species composition, interplant interactions such as competition will change over successional time as the soil-plant system itself evolves. Such a situation has been observed in the Netherlands, where high rates of atmospheric nitrogen deposition has driven a successional change from heath-dominated (*Erica tetralix*) communities to grass-dominated communities (*Molinia caerulea*). *E. tetralix* is characterized by low growth rates, low rates of biomass loss to shed tissues, herbivory, and so on, and poorly decomposable litter, which supports low rates of mineralization but high rates of sequestration in thick organic horizons. In contrast, *Molinia* is characterized by high growth rates, high rates of tissue loss to shedding and herbivory, and rapidly decomposable litter. As expected, in experiments specifically testing competitive ability, *Erica* is a superior competitor when nutrient supplies are low, whereas *Molinia* is the superior competitor when nutrient supplies are high. Thus, the feedback relationships between plants and soils can drive successional changes in community composition.

## V. IMPLICATIONS FOR THE MANAGEMENT AND CONSERVATION OF ECOSYSTEMS

Clearly, the complex, multiscale interactions of plants and soil affect all aspects of both the biology of plants and the properties of soils. This implies that both purposeful and inadvertent human effects on plant-soil systems will have diverse, perhaps unexpected ramifications, as changes in soils affect plants and vice versa. Following are brief discussions of four major areas of current concern, in which plant-soil interactions are likely to play a major role.

### A. Forestry, Agriculture, and Ecosystem Restoration: Human Creation of Ecosystems

Humans create plant communities for a wide variety of purposes, from the production of food, fiber, and fuels to the restoration of “natural” ecosystems and the creation of beauty for its own sake (in gardens and parks). The choice of plant species is often conditioned by soil properties; certain crops cannot be grown in certain soils, or conversely particular soils may be ideally suited to particular crops. Conversely, the choice of crop plant or tree can alter soil properties, in some cases perhaps irreversibly. As described earlier, the substitution of coniferous trees for deciduous trees, a widespread practice because of the rapid growth and high economic value of many conifers, results in acidification, the loss of nutrient cations through leaching, and the creation of thick litter and humus layers. In some cases, foresters use mixtures of tree species that are specifically designed to take advantage of differences in plant-soil interactions among the species. For example, larch (*Larix* spp.) and lodgepole pine (*Pinus contorta*) are often interplanted with Sitka spruce in plantations in the British Isles, because larch and pine promote higher rates of nitrogen cycling and therefore faster growth of the spruce. The increased nitrogen availability results, in part, from more widespread roots of the interplanted pines or larches, which reduce waterlogging in the organic horizon and thereby stimulate mineralization. Nitrogen-fixing plants are often introduced into both agricultural and forest plantations in order to alter the cycling of N through the soil.

Ecosystem restoration similarly involves the deliberate establishment of particular plant species and species mixtures. The choice of species is most commonly made

on the basis of replicating the “natural” community, rather than explicitly with respect to soil properties. However, plant communities are sometimes specifically designed to alter soil properties: the introduction of fast-growing grasses to introduce organic matter in barren soils, and the parallel introduction of N-fixing species into such soils, the use of nutrient-demanding, fast-growing species to remove nutrients (through repeated harvest) from overfertilized fields being restored to native grasslands or from wetlands being used for wastewater treatment are examples. Conversely, soil properties may unexpectedly affect the course and success of restorations. The potential for such effects is clearly illustrated in efforts to restore coastal wetlands in San Diego, California. Coarse-textured sediments were used to construct the wetland, which were unlike the fine-textured sediments found in naturally occurring marshes. Although the introduced salt marsh plants survived and spread, they only attained about half the height of plants in undisturbed marshes; the cause of the problem was the low storage rate of organic matter in the coarse sediments and the concomitant low intrinsic supply rate of nitrogen from the sediments.

### B. Exotic Species Invasions—Inadvertent Human-Caused Changes in Species Composition

The spread of exotic plant species around the world is recognized as one of the major threats to biodiversity. Plant communities are being altered through invasions of species that in some cases become part of the community, but in other cases exclude and eliminate the native species. These invasions often cause economic problems, as well as challenges for conservation of native communities, as the invading species degrade grazing lands, choke stream channels, or promote wildfires.

These changes in plant community composition are likely to alter soil properties, especially when the exotic species is of a different growth form than the natives or when its litter chemistry is different. Recent studies have shown that, for example, invasions of cheat grass (*Bromus tectorum*) into desert grasslands alters the abundance and species composition of fungi, protozoa, nematodes, and soil invertebrate communities, decreases the abundance of N-fixing lichens, and thereby decreases the amount of nitrogen in the soil. Invasions of exotic shrubs (*Berberis thunbergii*) and grasses (*Microstegium vimineum*) into deciduous forests stimulates increases in nitrification and soil pH. Invasions of the hawkweeds (*Hieracium* spp.) into pastures in New

Zealand causes the accumulation of nitrogen, but decreases soil pH. Invasions of grasses (*Melinis minutiflora*) into Hawaiian shrub lands stimulates increases in net nitrogen mineralization rates. It is likely that such changes are widespread and may affect the ability of resource managers to remove the exotics and restore the native plant communities.

### C. Nutrient Pollution—Inadvertent Human-Caused Changes in Soil Properties

The profound and far-reaching changes in atmospheric chemistry that are accompanying the spread of industrial society around the world include greatly increased inputs of several nutrients to soil-plant systems, through the deposition of particles and the dissolution of soluble ions in precipitation. The most important changes include greatly elevated deposition of nitrogen and sulfur. As these are both important plant nutrients and as their chemistry is intimately linked with other chemical and biological processes within the soil, a cascade of interactive changes to the soil-plant system are taking place in many places around the world.

Deposition of sulfates at rates below the level that directly injures plants still causes problems through soil acidification. Leaching of sulfate in soil water carries with it nutrient cations, and in soils with low base saturation and a parent material that supplies few cations through weathering, cations can become seriously depleted. Calcium, an essential element for the integrity of root function, is particularly strongly affected. Hardwood forests in New Hampshire (United States) that have received large inputs of sulfate for several decades have prematurely stopped growing, and their tissues show abnormally low concentrations of calcium. Depletion of nutrient cations is accompanied by increases in the concentration of aluminum ions in the soil solution and on exchange sites, and the aluminum can be directly toxic to plants. Depletion of calcium also leads to increased sensitivity to frost damage of needles, due to change in the structure of cell membranes, as has been demonstrated for declining red spruce in the eastern United States.

Nitrogen deposition to ecosystems is now occurring at double the estimated rates prior to the industrial revolution. The nitrogen originates from agriculture (applications of fertilizers and manures, and the planting of N-fixing crops), combustion (automobiles, power plants, and also forest burning), and waste disposal (e.g., sewage). Although nitrogen is considered the most important limiting resource in most terrestrial ecosystems, an excess does not simply alleviate stress

on plants. As was discussed earlier, plants growing on nutrient-poor soils have a variety of characteristics, including patterns of both root and above-ground morphology and physiology, characteristics of life history, and intrinsic growth rates, which adapt them to such habitats and which contrast strongly with the characteristics of plants found on nutrient-rich soils. Added nutrients often cannot be absorbed or assimilated by such plants, nor can they respond with increased growth. Rather, species adapted to nutrient-rich soils become capable of competitively displacing these species. Thus, many ecosystems that had been subject to limiting supplies of nitrogen have undergone profound changes in species composition. The substitution of grass for heather in European heathlands, as discussed earlier, is a case in point. Similar changes in native plant communities are occurring elsewhere in the industrialized world. And because communities found on infertile soils tend to have high diversity of plants, including many rare species, the increasing dominance of nitrophilous species may threaten many of them.

Nitrogen additions to soil not only affect the plant community through competitive displacements and substitutions, but also affect the internal nitrogen dynamics of the soil. Evidence from a variety of ecosystems has shown that N additions stimulates intrinsic mineralization processes, thus accentuating the effects of added N. Moreover, changes in the microbiota under conditions of high N inputs can result in the development of acid-tolerant nitrifiers; the increased levels of  $\text{NO}_3^-$  not only cause changes in plant community composition, but cause leaching of cations and thus changes in soil chemistry. As plant community composition changes, the chemistry, amount, and timing of litter inputs will change, as will all other components of the plant-soil system. The long-term effects of this chain of linked changes can only be guessed.

### D. Changing $\text{CO}_2$ in the Atmosphere—Inadvertent Human-Caused Changes to the Soil-Plant System

The rising concentration of  $\text{CO}_2$  in the atmosphere is expected to affect virtually all aspects of the plant-soil system. The topic can be touched on only briefly here, simply to note the major processes and interactions that are anticipated. Plants are expected to respond to increasing  $\text{CO}_2$  by increasing their overall rate of production (particularly  $\text{C}_3$  plants), increasing the allocation of carbon to below ground (including root growth, root turnover rates, root respiration rates, and

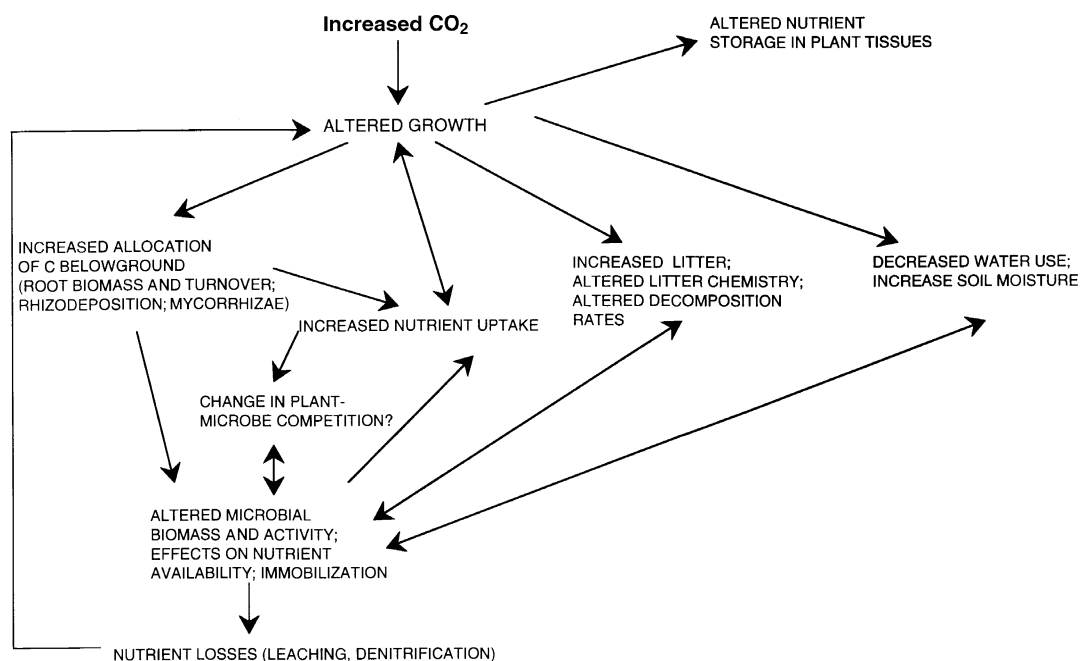


FIGURE 11 A schematic outline of the potential changes in the soil-plant system that may be induced by changes in atmospheric CO<sub>2</sub> concentrations.

rhizodeposition rates). Water use efficiency and nutrient uptake rates also often increase in experimentally enriched plants. Changes in litter chemistry (increased C:N ratios, increases in lignin, phenolics, or tannins) are anticipated but have yet to be clearly demonstrated. However, individual species responses have proven to be highly variable; studies of the growth response of a wide range of species has shown that changes in growth rate can range from a 25% decrease to a 200% increase. However, it is clear that virtually all aspects of plant morphology and chemistry that have been shown to affect soil properties are likely to change, precipitating changes in the soil in accordance with the processes discussed in this chapter. Figure 11 illustrates the complexity of the possible responses. Although a large amount of research is currently being conducted on these linkages, the complexity of the problem and the variability in response among different plants and in different soils make the discovery of general patterns and predictability difficult and yet to be achieved.

### E. Biodiversity and Ecosystem Function

The effects of changing plant community diversity on the structure and function of soil is among the most important current topics of research. Several studies

utilizing experimental grasslands have shown that such ecosystem processes as nitrogen mineralization, carbon accumulation and mineralization, nitrogen concentrations, soil respiration, and litter decomposition rates are altered (often decreased) by decreases in plant diversity, but other studies, including studies in forested ecosystems, have presented contradictory results. In many cases, particular species or particular classes of species (for example, nitrogen-fixers) have large effects. In addition, responses of soil processes to changes in plant diversity often are largest in comparing small numbers of species; once the plant community reaches a moderate number of species (three or four), there is no further change as plant species diversity increases. Reciprocal effects of diversity within the soil microbiota and soil fauna on the plant community are largely unknown.

The effects of changes in diversity within the plant community to the biology, chemistry and physics of soils, and vice versa, will be among the most important questions to be addressed in the future.

### See Also the Following Articles

FOREST ECOLOGY • GREENHOUSE EFFECT • PLANT INVASIONS • RESTORATION OF BIODIVERSITY • SOIL BIOTA, SYSTEMS AND PROCESSES • SOIL CONSERVATION

## Bibliography

- Atkinson, D. (Ed.) (1991). *Plant Root Growth. An Ecological Perspective*. Blackwell Science Publications, Oxford.
- Chapin, F. S., III, Sala, O. E., Burke, I. C., and Grime, J. P. (1998). Ecosystem consequences of changing biodiversity. *BioScience* **48**, 45–52.
- Eissenstat, D. M., and Yanai, R. D. (1997). The ecology of root lifespan. *Advances in Ecological Research* **27**, 2–60.
- Fitter, A. H., Atkinson, D., Read, D. J., and Usher, M. B. (Eds.) (1985). *Ecological Interactions in Soil. Plants, Microbes and Animals*. Blackwell Scientific Publishers, Oxford.
- Foster, R. C. (1988). Microenvironments of soil microorganisms. *Biol Fert. Soils*. **6**, 189–203.
- Glinski, J., and Lipiec, J. (1990). *Soil Physical Conditions and Plant Roots*. CRC Press, Boca Raton, FL.
- Harley, J. L., and Russell, R. S. (Eds.) (1979). *The Soil-Root Interface*. Academic Press, London.
- Hobbie, S. (1992). Effects of plant species on nutrient cycling. *Trends in Ecology and Evolution* **7**, 336–339.
- Hu, S., Firestone, M. K., and Chapin, F. S., III. (1999). Soil microbial feedbacks to atmospheric CO<sub>2</sub> enrichment. *Trends Ecol. Evol.* **14**, 433–437.
- Jenny, H. (1980). *The Soil Resource*. Springer-Verlag, New York.
- Lynch, J. M. (Ed.) (1990). *The Rhizosphere*. John Wiley and Sons, New York.
- Kaye, J. P., and Hart, S. C. (1997). Competition for nitrogen between plants and soil microorganisms. *Trends Ecol. Evol.* **12**, 139–143.
- Metting, F. B., Jr. (Ed.) (1993). *Soil Microbial Ecology*. Marcel Dekker, New York.
- Robinson, D. (1994). The responses of plants to non-uniform supplies of nutrients. *New Phytologist* **127**, 635–674.
- Tilman, D. (1988). *Plant Strategies and the Structure and Dynamics of Plant Communities*. Princeton University Press, Princeton, NJ.
- van Breeman, N. (1995). Nutrient cycling strategies. *Plant and Soil* **168–169**, 321–326.
- Vancura, V., and Kunc, F. (Eds.) (1988). *Soil Microbial Associations—Control of Structures and Functions*. Elsevier, New York.
- Vogt, K., Grier, C. C., and Vogt, D. J. (1986). Production, turnover and nutrient dynamics of above- and belowground detritus of world forests. *Adv. Ecol. Res.* **15**, 303–377.
- Waisel, Y., Eshel, A., and Kafkafi, U. (Eds.) (1991). *Plant Roots—The Hidden Half*. Marcel Dekker, New York.





# PLANT SOURCES OF DRUGS AND CHEMICALS

William H. Gerwick, Brian Marquez, Ken Milligan, Lik Tong Tan,  
and Thomas Williamson  
*Oregon State University*

---

- I. Introduction
  - II. Examples from Marine Plants Illustrating Plant Sources of Drugs and Chemicals
  - III. Examples from Terrestrial Plants Illustrating Plant Sources of Drugs and Chemicals
  - IV. Conclusions
- 

## GLOSSARY

- anticancer agent** A chemical substance that is capable of bringing about a remission or cure of the family of diseases known as cancer.
- biosynthesis** The biochemical process by which a plant produces molecules of utility or adaptation.
- cytotoxicity** The ability of a chemical compound to kill cells in an experimental system, such as in a petri dish.
- secondary metabolite** A chemical substance produced for a reason other than basic life requirements and with some adaptive or defensive value.
- 

**NATURAL PRODUCTS** from plants have formed the basis for many of our useful pharmaceuticals and agriculturals. This chapter presents examples of useful plant-derived pharmaceutical agents, which illustrate the complexity of the drug discovery process and the

importance of maintaining biological diversity in order to preserve this inherent chemical diversity and its genetic origins.

## I. INTRODUCTION

The intent of this chapter is to illustrate by example the enormous diversity of plant-derived natural products and their tremendous importance to human society. It is precisely this diversity of structure that has been, and continues to be, of such incredible value to the pharmaceutical and agricultural industries. Ultimately, this chemical diversity is rooted in the inherent biodiversity of our planet. Sadly, space limitations for this chapter relegate this treatment to only a sampling of interesting examples. These have been chosen to exemplify several concepts; various chemical adaptations typical of particular plant groups, classes of compounds of utility to society, older examples deriving from ethnobotanical information, and modern examples resulting from methodical broad-based screening programs.

It cannot be debated that the premier importance of plants to human existence is as a source of food and oxygen. Related to their serving a food function is the use of plant-derived materials as food additives to impart desirable taste or textural properties. Examples include the use of a class of biopolymer from red marine algae, the carrageenans, to stabilize emulsions in such



diverse products as ice cream and beer, and the use of the sugar alcohol sorbitol, a product of the mountain ash *Sorbus aucuparia*, as a noncariogenic sweetener with humectant (wetting) properties in toothpaste. However, this chapter will focus on another property of plants; their ability to produce unique molecular entities with potent pharmacological effects in mammalian systems. Why do plants make such compounds? Given their complicated, often exotic, structures, they certainly expend considerable biochemical energy in their production. While some hold that such compounds are vestigial in nature (previously served a function that is no longer in evidence), there is excellent experimental evidence to support the idea that most, if not all, are produced with powerful adaptive functions, such as defense against potential predators and inhibition of the growth of competing species.

Primitive peoples throughout the world make use of their indigenous flora as a source of medicines. Through the process of trial and error, these cultures have examined and discovered many plants that produce unique molecular entities with valuable biological properties. The field of "ethnopharmacology" or "ethnomedicine" seeks to obtain new pharmaceutical leads from a study of the native medicines of these primitive peoples, an approach that has been highly successful and resulted in such compounds as aspirin from willow bark (*Salix*), the anticancer alkaloid vincristine from the Madagascar periwinkle (*Catharanthus*), and the antimalarial sesquiterpene artemisinin (qinhaosu) from the Chinese herb *Artemisia annua*.

Beginning in the 1950s, comprehensive evaluations of the unique constituents of plants were undertaken. In the United States, these earliest efforts were coordinated largely by the National Cancer Institute, and so necessarily had a focus on anticancer properties. Assays for potential anticancer activity at that time were run in intact animals (*in vivo* evaluations). While even the most modern drug developments require evaluation of a drug for efficacy and toxicity in animal systems, this usually occurs only after a candidate pharmaceutical has shown the requisite property in isolated biochemical and cell-based screens. Consequently, current *in vivo* evaluations only occur on a very small fraction of the most promising candidate pharmaceuticals. In contrast, the initial screen of plant extracts in early efforts used *in vivo* techniques; consequently, they were very slow, expensive, and, perhaps worse, used test animals in large quantities.

In the late 1970s, and gaining broad acceptance in the 1980s, a major shift away from whole animal primary screening was achieved with the introduction of

cell-based screening. In the cancer effort, this largely had the endpoint of cytotoxicity to cancer cells grown in small petri dishes or, later on, 96-well trays. Ultimate expression of this idea has been realized at the National Cancer Institute wherein extracts and compounds are evaluated for their level and profile of toxicity to 60 different cancer cell lines. The cell lines have been chosen so as to represent cancers affecting nine different organ types and have been quite useful in detecting new cytotoxins as well as giving information about their molecular mechanism of action.

The past 10 years has seen a growing departure from even cell-based assays. Over the past 30 years, there have been tremendous advances in understanding at a molecular level the causes of many diseases, including cancer (Shu, 1998). This has translated into assays that screen for compounds or extracts that interfere with a specific enzymatic reaction or protein-protein interaction that has been shown to underlie a particular human disease. With the reduction in screening format from whole animal to cell to isolated protein, it is now possible to screen hundreds of thousands of compounds or extracts in a few weeks time using high throughput screening (HTS) technologies. Here, a target protein is raised in large quantities by molecular biological means in a surrogate organism, placed in a 96-well or even larger formats, and compounds or extracts evaluated for inhibitory effects using robotics that work around the clock. It is hoped that screens of this design and scope will uncover a new generation of pharmaceuticals that will have highly specific actions, targeting the underlying causes of disease, with very high potency, and leaving normal or undiseased cells untouched.

## II. EXAMPLES FROM MARINE PLANTS ILLUSTRATING PLANT SOURCES OF DRUGS AND CHEMICALS

### A. Anticancer Agents from Marine Cyanophyta (Cyanobacteria)

The marine "plants" generally include a number of groups, including microscopic as well as macroscopic forms. The major subdivisions of the microscopic forms include dinoflagellates, cryptophytes, chrysophytes, cyanophyta (= cyanobacteria), and prochlorophyta. The major groups of macroscopic marine plants include the Rhodophyta (= red algae), Chlorophyta (= green algae), Phaeophyta (= brown algae), and a number of marine angiosperms (sea grasses); the latter represent

terrestrial species that have “readapted” to life in the ocean. While interesting and significant molecules have been isolated from all of these groups, secondary metabolites of special note have been obtained from marine cyanobacteria (Jaspars and Lawton, 1998), dinoflagellates, and red and green macrophytes. Examples from these four groups follow.

### 1. Curacin A and “Chemotype” Concept

Cyanobacteria take on many different physical forms in the oceans, from gelatinous encrustations and tufts of 20 cm hairlike filaments, to microscopic unicells free in the water column. One species of pan-tropical distribution, *Lyngbya majuscula* (family Oscillatoriaceae), has been particularly plentiful in its production of structurally unique and biologically important secondary metabolites. Such an example is given by a collection made in Curaçao, in the southern Caribbean, which was reported to possess an extract with fish, snail, brine shrimp, and cancer cell toxicity (Orjala *et al.*, 1996). Bioassay guided isolation efforts led to the isolation of several structurally unrelated substances, which were each responsible for only some of the observed biological properties. The cancer cell and brine shrimp toxicity was principally due to a unique thiazoline-containing lipid, named curacin A (Fig. 1). Its mechanism of cytotoxicity was shown to involve inhibition of microtubule formation, crippling the ability of cells to properly segregate chromosomes at mitosis (Gerwick *et al.*, 1994). Several very valuable anticancer agents that are in common use in the clinic today work by this same essential mechanism (e.g., see discussion for taxol, presented later). Curacin A has been shown to interact with microtubules at the same site as the antigout drug, colchicine. Unfortunately, curacin A was found difficult to work with *in vivo* because of solubility and stability problems. Recent efforts have overcome at least some of these problems, and work continues to develop an anticancer drug patterned on this structural concept.

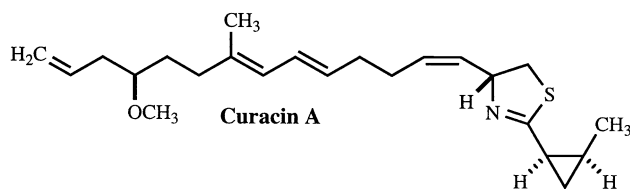


FIGURE 1 Structure of the antimetabolic lipid “curacin A,” obtained from a Caribbean collection of the blue-green alga (cyanobacterium) *Lyngbya majuscula*.

Sources of curacin A for continued drug testing have largely been from materials freshly collected in Curaçao, although the living alga does produce curacin A in laboratory culture. Interestingly, when collections of *Lyngbya majuscula* were made from approximately 20 sites along the leeward coast of Curaçao, only two chemotypes, both quite limited in extent, were found to make curacin A. If continued investigation of this phenomenon substantiates these findings, then a very intriguing conclusion must be drawn, which is of enormous significance to “biodiversity preservation.” It is widely held that in order to preserve the valuable genomic potential of a particular “species,” it is sufficient to preserve a few individuals or a limited population. This work with Curaçao collections of *L. majuscula* suggests that in order to accomplish the larger goal, it is actually necessary to consider preservation of the many, almost innumerable, “chemotypes” of a given species.

### 2. Cryptophycin

Another very promising anticancer treatment derives from the freshwater cyanobacterium *Nostoc* sp. While one of the active compounds, cryptophycin A, was first isolated as an antifungal substance, it was only several years later that, upon re-isolation, its anticancer potential was realized (Fig. 2). In common with curacin A, cryptophycin also blocks microtubule assembly processes, although at a different drug-binding site (the vinca alkaloid site). In animal testing, cryptophycin has effected actual “cures” of some tumor types. An enormous effort has been devoted to the total chemical synthesis of cryptophycin and several hundred analogs with the intent of ensuring that (a) a ready supply of the drug is available for clinical testing and (b) that advanced cancer treatment evaluations utilize the most effective molecule in this drug series. Testing of the drug in humans has recently advanced to phase II trials, and there is great hope that this agent will provide an important new tool in cancer treatment.

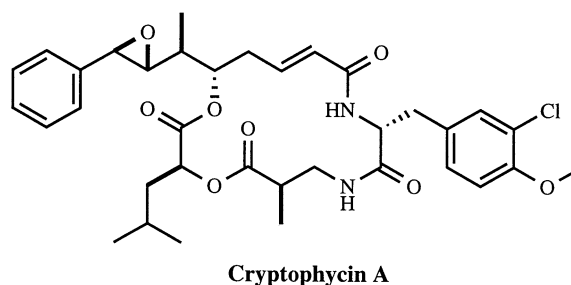


FIGURE 2 Cyclic peptide structure of the anticancer metabolite “cryptophycin.”

## B. Toxins of Impact and Utility from Dinoflagellates

Dinoflagellates represent a very diverse group of organisms, made up of both freshwater and marine varieties. Dinoflagellates are protists, which are not strictly considered plants or animals although they share characteristics of both. Some dinoflagellates are photosynthetic whereas others are parasitic to fish or to other protists. They can be very small or as big as 2 mm in diameter, such as in the case of *Noctiluca* sp. A curious feature of some dinoflagellates is that they are bioluminescent.

Dinoflagellate blooms in the world's oceans have had tremendous impact on human society; despite this importance, many aspects of these organism's chemistry and biology are poorly understood. Hundreds of people suffer from the effects of seafood poisoning and millions of fish and other marine life are killed each year as a direct consequence of large blooms of dinoflagellates. Some of the coastal species can "bloom" during the warmer months and can even make the water appear red or golden colored; hence, the term "red tide." For many years seafood poisonings and fish kills have been associated with these dinoflagellate blooms but it is only within the past decade that the chemistry behind these effects has been unraveled.

### 1. Ciguatoxin and Maitotoxin

A prime example of a dinoflagellate toxin with huge impact on society is given by Ciguatera Seafood Poisoning (CSP), which is quite prevalent in tropical areas around the globe (Yasumoto and Murata, 1993). CSP is more common than any other illness associated with consumption of tainted seafood. Generally, the poisoning results from ingestion of coral reef fish that have accumulated the toxins through their diet. Symptoms include, but are not limited to, memory loss, joint pain, miosis, erethism, cyanosis, and prostration. One of the more interesting symptoms associated with ciguatera is a neurological disturbance leading to reversal of hot and cold sensations. Most of the toxic effects of CSP are attributed to ciguatoxin, which is a metabolically oxidized derivative of a substance produced by the dinoflagellate, *Gambierdiscus toxicus*. The structure of ciguatoxin was deduced after a monumental 15-year effort by the laboratories of Dr. Paul Scheuer at the University of Hawaii and Drs. Takeshi Yasumoto and Michio Murata at Tohoku University (Fig. 3). The structures of ciguatoxin and the probable precursor produced by the dinoflagellate are shown later. These compounds illustrate the common polyether motif found in many dinoflagellate toxins.

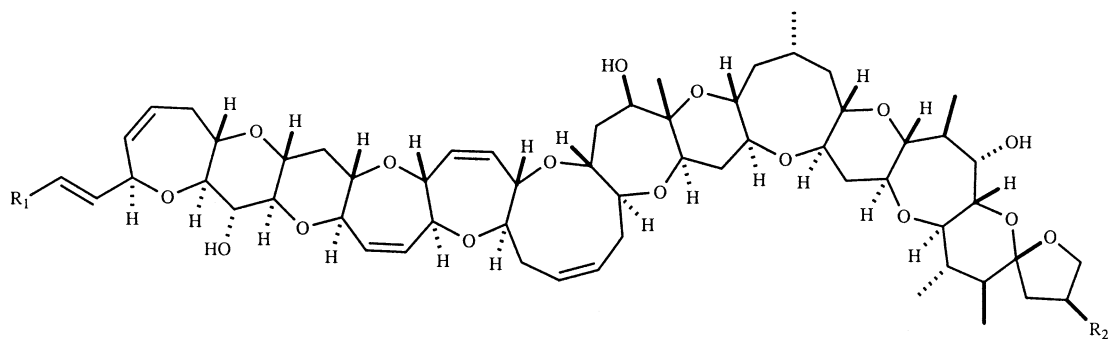
Another toxin that should be noted here is maitotoxin (Fig. 4). Probably produced by dinoflagellates related to those that produce the ciguatoxin precursor, this compound is even more toxic than ciguatoxin. It is by far the largest compound ever isolated that is not a biopolymer (e.g., protein or carbohydrate). Its structure was also elucidated by the laboratories of Yasumoto and Murata.

### 2. *Gymnodinium breve*

For many years it was thought that seafood poisonings caused by dinoflagellates were limited to tropical regions of the world. However, in 1987, 14 dead hump-backed whales were discovered in Cape Cod Bay, Massachusetts. Examination of the animals showed that they had been well until just before their deaths, appearing to be well nourished and showing no signs of ill health. A few months later, fishermen and beach-goers along the North Carolina coast began to complain of respiratory problems and eye irritation. Soon after, seafood consumers began to complain of diarrhea and dizziness. In later years and on a periodic basis, these same scenarios have been reported all along the east coast and the Gulf of Mexico. Most of these outbreaks have been connected with the dinoflagellate *Gymnodinium breve*. This dinoflagellate frequently forms large blooms, or "red tides," off the Florida coast, which result in many tons of dead finfish washing up on the beaches. This, in turn, repels many tourists and has forced many fisheries to be closed. In the early 1980s the toxic constituents produced by this organism were found to be brevetoxin A and B, along with other structurally related compounds (Fig. 5).

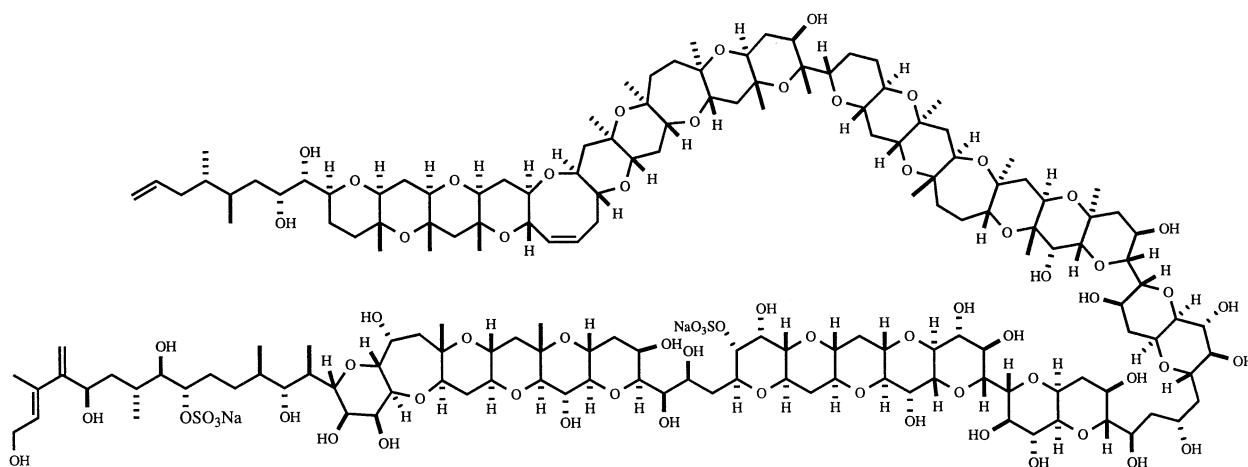
### 3. *Pfisteria piscicida*

A recent development in the economic and environmental impact of dinoflagellate blooms is the newly discovered dinoflagellate, *Pfisteria piscicida*. Outbreaks of this organism have been reported in coastal North Carolina and Maryland. Because these blooms are possibly linked to pollution from swine and chicken farms, plus the interest generated by the publication of a popular book (Barker, 1997), there is much controversy surrounding research on this topic. This particular organism is different from the other known species of dinoflagellates in that it may have a stage in its life cycle where it becomes dormant but, in the presence of large numbers of finfish, can become active and in effect ambush the fish as a predator. While research into this behavior is still in the early stages, all evidence gathered thus far seems to support this hypothesis. Less is known about the toxins produced by this organism except that



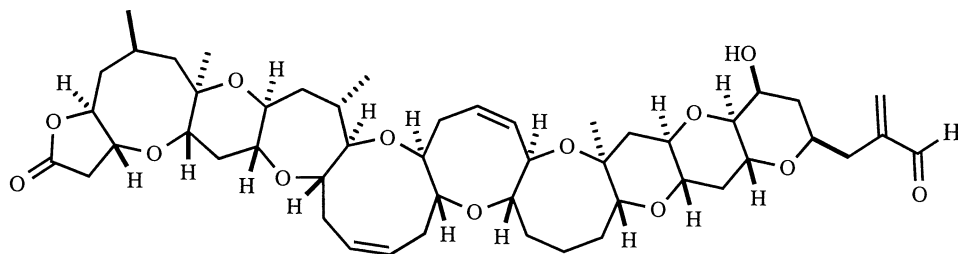
**Ciguatoxin**  $R_1 = -\text{CH}(\text{OH})-\text{CH}_2\text{OH}$ ;  $R_2 = \text{OH}$   
**Pre-ciguatoxin (CTX-4B)**  $R_1 = -\text{CH}=\text{CH}_2$ ;  $R_2 = \text{H}$

FIGURE 3 Polyether structure of ciguatoxin, the active poison in Ciguatera Shellfish Poisoning.



**Maitotoxin**

FIGURE 4 Structure of maitotoxin, the most complex nonpeptide structure ever determined.



**Brevetoxin A**

FIGURE 5 Brevetoxin A, the causative fish poison produced by the red tide organism *Gymnodinium breve*.

there appears to be two: one water soluble and one soluble only in organic solvents. The symptoms resulting from contact with this dinoflagellate are quite severe, ranging from serious memory loss, loss of motor function, to death.

### C. Kainic Acid and Domoic Acid, Neurotoxic Anthelminthics from Red Algae

Two Japanese red algae, *Digenea simplex* and *Chondria armata*, have been employed for more than 1000 years in Japan for their potent anthelmintic properties; that is eliminating intestinal worms, such as parasitic roundworms (*Ascaris lumbricoides*), whip worms (*Trichuris trichura*), and tape worms (*Taenia* spp.). Two closely related compounds, domoic acid and kainic acid (Fig. 6), have been isolated from these red algae and are responsible for these healing effects. These compounds were traditionally ingested by consumption of the producing algae, their names deriving from the Japanese names for these seaweeds, *domoi* and *kaininso*. Kainic acid has become an important tool in neurobiological research because of its potent neurotoxic effects on neurons (Brown and Nijjar, 1995). On the other hand, domoic acid has been identified as the active poison in amnesiac shellfish poisoning (ASP), which has recently struck both east and west shores of North America. The levels of the drug that are ingested for anthelmintic properties are relatively small when compared to the amount of compound observed in the recent outbreaks of ASP. However, the macrophytic red algae were determined to not be the culprit in the recent ASP outbreaks; rather, a planktonic pennate diatom, *Nitzschia pungens* f. *multiseries*, now renamed *Pseudonitzschia australis*, is the producer in this latter case.

The mechanism of action of these compounds is very similar, as is suggested by their closely related chemical structures. Both are analogs of glutamic acid, a well-known neurotransmitter found in the brain of mammals. The algae-derived substances are known as excitotoxic amino acids as they can stimulate neurons through the release of endogenous glutamate. The release of high doses of *L*-glutamate causes damage or death to various neuronal cell types. Domoic acid is two to three times more potent than kainic acid and 100 times more potent than *L*-glutamate. Although these compounds (domoic and kainic acid) are environmental neurotoxins to be avoided, they have shown considerable utility in biomedical research. For example, the resulting extensive neuronal loss seen with the appropriate dose of kainic acid is very similar to that observed for Huntington's disease and hence provides an excellent model for the study of this disease in humans. Kainic acid has also been used in the study of epilepsy.

A problem of epidemic proportion that occasionally arises from these excitotoxic amino acids is amnesiac shellfish poisoning (ASP) (Todd, 1993). It has been shown that blooms of the planktonic diatom *P. australis* result in a buildup of domoic acid in shellfish. Three highly publicized ASP outbreaks have occurred in the past 10 years in North America. The first was caused by the ingestion of mussels containing high levels of domoic acid from Prince Edward Island, Canada. The second was a large pelican kill in Monterey Bay, California. The domoic acid was found to accumulate through the food chain from diatoms to small fish to anchovies to pelicans. The last reported outbreak was in razor clams and Dungeness crabs in Oregon and Washington. The discovery of high concentrations of domoic acid led to the shutdown of these seafood harvests for several years, causing substantial economic hardship for af-

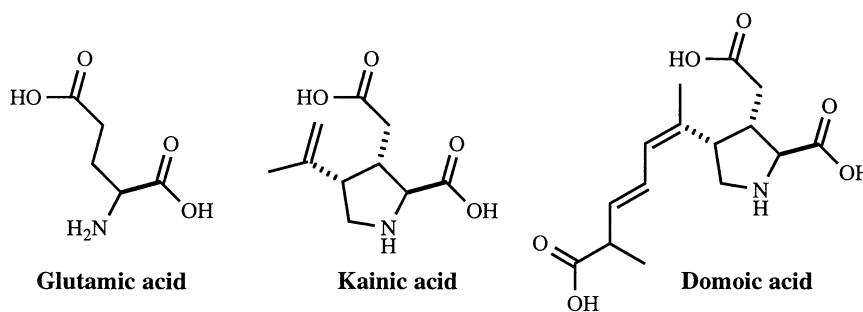


FIGURE 6 Structures of the neurotoxic amino acid, glutamic acid, present in mammalian cells and two neurotoxic congeners from red algae, kainic acid, and domoic acid.

flicted communities. As a result of these ASP outbreaks, strict guidelines have been developed for the maximum allowed domoic acid content in harvested shellfish (20 ppm).

### D. Green Algal Metabolites

Coral reef environments are habitats of high speciation and biodiversity ranging from microorganisms to top predators. This biodiversity inherently creates intense competition between species, including predator-prey relationships such as between herbivores and their algal diets. Methods that algae use to protect against predation include calcification, production of spiny appendages, and the biosynthesis of chemically noxious deterrent molecules that are either toxic or deter feeding.

The green alga *Halimeda* sp. employs both calcification and chemical defense strategies. As a result, *Halimeda* species constitute the largest macroalgal biomass in many tropical reef systems. In feeding preference studies, and analyses of reef fish stomach contents, *Halimeda* spp. have been shown to not be a food source for generalist herbivores. While calcification can provide algae with a strong structural protection against potential predators, some herbivores have coevolved specialized feeding structures that allow them to macerate even very tough materials. However, the dominance of *Halimeda* spp. in reef systems is enhanced through its production of a potent defensive and antifouling substance, a diterpenoid natural product named halimedatrial (Fig. 7). Halimedatrial is a trialdehydic secondary metabolite that is structurally similar to the iridoid aldehydes and the insect antifeedant warburganal. Halimedatrial displays a broad spectrum of activity against marine bacteria, a marine fungus, the division of fertilized sea urchin eggs, and the motility of sea urchin sperm (Paul and Fenical, 1988). Halimedatrial was also toxic to common reef fish as well as being inhibitory of reef fish feeding at ecologically relevant doses.

In the alga, halimedatrial exists in a “pro-toxin” form,

a protected tetraacetate named halimedatetraacetate. Field experiments with *Halimeda* sp. have shown that halimedatetraacetate is converted to halimedatrial upon wounding of the alga. This conversion is rapid and specific to the area of herbivory or wounding and does not take place simply upon introduction to air but is an enzymatic conversion of a “protoxin” secondary metabolite to an highly potent antipredatory substance. This activated defense of *Halimeda* represents the first to be described in the marine environment. Interestingly, a similar activated defense system that produces feeding deterrents has previously been described in Russulacean mushrooms.

## III. EXAMPLES FROM TERRESTRIAL PLANTS ILLUSTRATING PLANT SOURCES OF DRUGS AND CHEMICALS

### A. Anticancer Agents from Higher Plants

#### 1. Taxol

The story of the discovery and development of taxol as an anticancer agent is a superb example of the opportunities and problems inherent in plant-derived medicines (Wall and Wani, 1994). A National Cancer Institute-sponsored plant collection program in the Pacific Northwest region in 1960 gathered 650 plant species for screening through various cell-based and *in vivo* anticancer evaluations. The twigs and bark of the Pacific Yew, *Taxus brevifolia*, showed very good activity in a cell assay and was subsequently subjected to bioassay-directed isolation of the active compound by Wall and Wani at the Research Triangle Institute. *Taxus* species have a rich history of use by native peoples for treatment of various illnesses, including cancer. Painstaking efforts were involved in the isolation of the active component, taxol, which was shown to be a new taxane deriva-

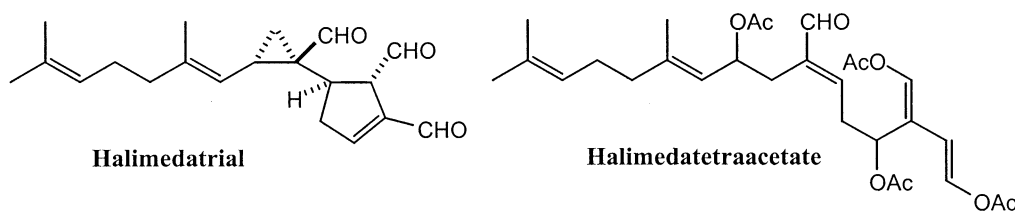


FIGURE 7 Structure of the green algal predator defense compound halimedatrial and the “pro-toxin” halimedatetraacetate.

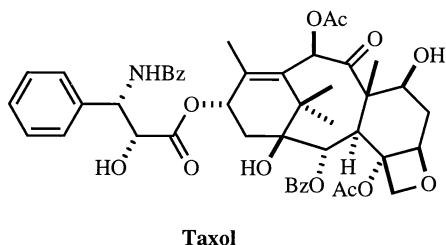


FIGURE 8 Structure of the anticancer natural product "taxol" from the tree *Taxus brevifolia*.

tive by X-ray crystallography (Fig. 8). Despite its intriguing structure and impressive activity in cell assays (but not in antileukemic *in vivo* assays!), research on taxol languished for nearly a decade due to a combination of short supply of the drug, difficulty in working with this lipid soluble substance, and lack of research funds.

Interest in the drug was rekindled when Susan Horwitz at the Albert Einstein Institute showed that taxol had the opposite effect on microtubules compared to all other known classes of antimetabolic anticancer agents. By binding to a unique drug-binding site on the  $\beta$ -subunit of tubulin, taxol promotes the stabilization of microtubules whereas other classes of antitubulin agents promote their destabilization. This mechanistic difference fueled additional evaluations of taxol, and once formulation problems were overcome, taxol emerged as the most significant new anticancer drug of the past 20 years. Currently, taxol has shown good efficacy to ovarian, breast, and lung cancers, as well as several others. However, its development was greatly impaired from severe supply shortages, a common issue for natural product derived drugs. It has been estimated that it takes the recoverable taxol content of four full grown 70 year-old trees to treat 1 patient for 1 year. At its peak, natural collections of 60,000 pounds of Pacific Yew bark were made in 1988 and 1989, which threatened the survival of this relatively common tree. Subsequently, supply of the drug has been made plentiful by harvest of the needles from cultured trees (*Taxus baccata*), extraction of the essential core of the taxol molecule, and its conversion by chemical synthesis into taxol. The 25-year development period from isolation to clinical use is typical of modern pharmaceutical development wherein many basic scientific, regulatory, and safety issues must be answered before drugs reach the market. The taxol supply problem has subsequently initiated a national program to explore how to resupply a natural product drug in large scale during its development as an anticancer agent.

## 2. Camptothecin

The same group that discovered taxol made a second major contribution in the fight against cancer. This latter compound, called camptothecin, comes from the wood and bark of the tree *Camptotheca acuminata* (Pantozis *et al.*, 1996). Monroe Wall, who was working at the eastern regional Research laboratory of the USDA in Philadelphia, Pennsylvania, discovered it in the 1950s. Previously, this group had been screening plant extracts for steroids that could be used as cortisone precursors. In 1958, samples from this entire collection of extracts were sent to the National Cancer Institute (NCI) for testing against certain cancer cell lines. Only one extract, that from *Camptotheca acuminata*, showed activity in this particular biological assay.

Camptothecin was eventually isolated through the use of the Craig Countercurrent Distribution method, which takes advantage of different compounds having varying solubilities in dissimilar solvents. In 1966 the structure of camptothecin, and its corresponding sodium salt, were published on the basis of a large amount of spectroscopic and chemical degradation data (Fig. 9). In early animal trials, camptothecin showed significant antileukemic activity to the L1210 cell line at concentrations as low as 0.2 mg/kg body weight of mice. By 1970, camptothecin had progressed to phase I clinical trials at NCI and soon progressed to phase II clinical trials. Encouraged by these results, the NCI decided to introduce the water soluble sodium salt of camptothecin into phase II clinical trials. The results of these trials were not very promising and camptothecin lost much of the interest it had attracted in its earlier phase I trials. It was later discovered that the sodium salt of camptothecin was only one-tenth as active as the parent compound against the P388 cancer line.

In 1985, it was found that camptothecin acts by a novel mechanism and interest was rekindled. These later studies showed that camptothecin and various water-soluble analogs worked by inhibiting the action of mammalian topoisomerase I (T-I, an enzyme in mam-

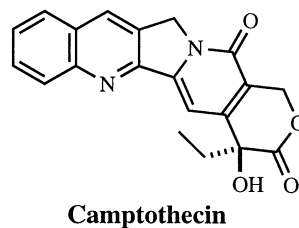
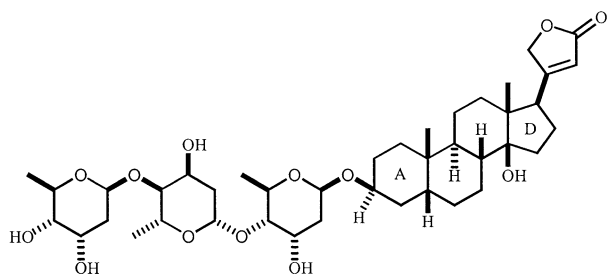


FIGURE 9 Structure of the anticancer natural product "camptothecin" from the tree *Camptotheca acuminata*.







**Digitoxin**

FIGURE 11 Structure of the cardiac glycoside digitoxin obtained from the foxglove *Digitalis* sp.

digitoxin, the favored treatment for rapid atrial fibrillation (Fig. 11).

Cardiac glycosides are naturally occurring steroids that are functionalized by sugar and lactone moieties (Trease and Evans, 1983). The most active of these possess an unsaturated lactone ring connected to C-17 of the steroid nucleus. Optimal cardiac activity also requires *cis* stereochemistry of the A/B and C/D ring junctures. Structure-activity studies on the cardiac glycosides have demonstrated that the sugar moiety aids in bioavailability and solubility, while functionalization of the steroid nucleus with hydroxyl, methyl, and lactone ring groups is essential for pharmacological activity. The unsubstituted aglycone, or steroidal portion, is less active than the glycoside (sugar attached).

In treating heart failure, the cardiac glycosides of *Digitalis* function by increasing the strength and efficiency of ventricular contraction, thus shortening the length of contraction and allowing the heart muscle a longer relaxation time between contractions. This results in recovery of the myocardium, decreased heart rate, and improved renal function through enhanced circulation.

Nearly 30 different cardiac glycosides have been identified in *Digitalis purpurea* and more than 70 from *Digitalis lanata*. Cardiac glycosides have been observed in nine other *Digitalis* species. Interestingly, none of the major glycosides of *D. lanata* are identical to those isolated from *D. purpurea*, with the most common alteration in the molecule being acetylation of hydroxyl groups in the *D. lanata* representatives. This fact indicates great chemical, biochemical, and metabolic diversity among *Digitalis* species.

Historically, a drawback of *Digitalis* treatment is its toxicity. This toxicity led to its being eschewed by most

medical practitioners until its reappearance in relatively modern treatments. The fact that no fewer than nine prescription products from *Digitalis* species are used to treat congestive heart failure, atrial fibrillation, and atrial and ventricular tachycardia demonstrates the human need and utilization of not only the innocuous plants, but also the noxious varieties, such as *Digitalis*.

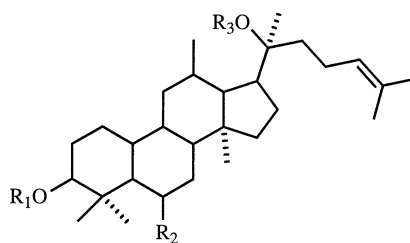
## D. Herbal Remedies

### 1. Ginseng

Perhaps the most venerable of traditional Chinese herbs is “ginseng” or “Korean ginseng,” also known by its Latin name *Panax ginseng*. Usage of *Panax ginseng* has a long and rich history in China dating back to the Han dynasty some 2000 years ago. The earliest description of its application appeared in the oldest Chinese pharmacopoeia, believed to have been written in the first century during the late Han dynasty. It is the root of *Panax ginseng* that is used extensively in Chinese medicine as an effective tonic for enhancing stamina as well as the capacity to endure fatigue and physical stress.

*Panax ginseng* was originally found in northeastern China, Korea, and eastern Siberia. Other congeners of ginseng, such as the *P. quinquefolius* (American ginseng) are also used widely as medicines. *Panax quinquefolius* was initially reported growing wild in the northeastern United States. Another common member is the Siberian ginseng (*Eleutherococcus senticosus*), which belongs to the Araliaceae family and contains distinctly different glycosides. The main source of *Panax ginseng* used in the United States and Europe comes from Korea where it is cultivated extensively. Two main forms of ginseng are currently being used; white ginseng, which is processed from the air-dried roots of five to seven year old plants, and red ginseng, which is white ginseng that has been steam treated for 2 to 4 hr.

Research on the chemistry of *Panax ginseng* suggests that the major active components of the roots are the ginsenosides (glycosides), which are derivatives of the triterpene “dammarane.” To date, more than 20 ginsenosides have been reported from the roots, leaves, and flower buds of ginseng (Fig. 12). Common sugars found attached to the protopanaxadiol core of these ginsenosides include glucose, maltose, fructose, and saccharose. In spite of the discoveries of these bioactive glycosides in ginseng, the precise pharmacological mechanisms of ginseng’s actions are unclear. However, there are extensive reports of the effects of ginseng on the function of the neuroendocrine system, the central nervous system (memory, learning, and behavior), carbohydrate and lipid metabolism, the immune system, and the car-



	<u>R<sub>1</sub></u>	<u>R<sub>2</sub></u>	<u>R<sub>3</sub></u>
<b>20(s)-protopanaxadiol</b>	<b>Glc<sup>2</sup>-Glc</b>	<b>H</b>	<b>Glc<sup>6</sup>-Glc</b>
<b>20(s)-protopanaxatriol</b>	<b>H</b>	<b>O-Glc<sup>2</sup>-Rha</b>	<b>Glc</b>

FIGURE 12 Two of a number of ginsenosides from *Panax ginseng*, a traditional herbal remedy from China.

diovascular system. For instance, ginsenoside Rb1 was demonstrated to have CNS-sedative, tranquilizing, and hypotensive actions while ginsenoside Rg1 was shown to be CNS-stimulating, hypertensive, and possess anti-fatigue properties. As this example demonstrates, many of the reports on the biological activities of ginseng's metabolites are contradictory; this may be due to the difference of ginsenoside content in ginseng root or root extracts.

## 2. Echinacea

The application of *Echinacea* as a medicinal plant is derived from its use by Native North Americans. Traditional use of *Echinacea* involved the treatment of numerous conditions, such as external applications to wounds, burns, inflammation of lymph nodes (mumps), and insect bites. Internal uses of the plant included stomach cramps, headache, coughs, chills, measles, and gonorrhoea. The root of *Echinacea* was the most important part of the plant used in these native treatments. There are also reports of the use of the juice or a paste of macerated fresh plant by Native North Americans. In view of the multitude of beneficial uses of *Echinacea* by these people, it was likely one of their most important medicinal plants.

The first pharmacological studies on *Echinacea* extracts began in the early 1950s and were mainly concerned with its nonimmunological actions. However, we now know that *Echinacea* acts mainly via stimulation of the nonspecific immune system. The main immunostimulatory components of *Echinacea* extracts were found to be lipophilic alkylamides as well as cichoric acid, which is a derivative of caffeic acid. Cichoric acid (2,3-O-dicaffeoyltartaric acid) was first isolated from

*Echinacea purpurea* and found to cause significant stimulation of phagocytotic activity in an *in vitro* granulocyte bioassay (Fig. 13). In addition, polysaccharides have also been suggested to be active components in *Echinacea* juice and aqueous extracts. A purified mixture of polysaccharides from the roots of *E. purpurea* were found to increase *in vitro* phagocytosis by macrophages 23 to 32% at concentrations of 0.01 and 0.001 mg/mL. It appears that the complete immunostimulatory properties of *Echinacea* extracts are due to the combination of activities from several different classes of compounds.

## IV. CONCLUSIONS

The examples given here clearly demonstrate the pivotal role that plants have played in the development of much of our current pharmacopeia (Shu, 1998). Indeed, it has been estimated that as much as 37% of all pharmaceutical sales are for compounds that derive, either

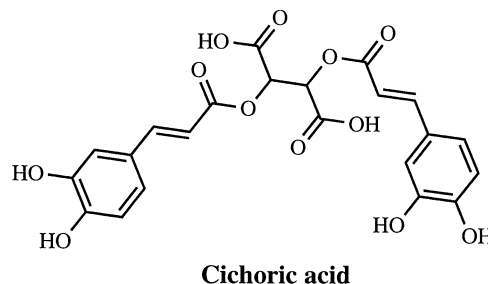


FIGURE 13 Structure of cichoric acid, one of the biologically active components of *Echinacea purpurea*.

wholly or in part, from natural products. For anticancer drugs, the percentage deriving from natural products is even higher (ca. 60%). Nevertheless, the enormous task that we require of pharmaceuticals, to have potent and selective action for a particular disease state, to be orally active and distribute with high efficiency to the target organ, and to be nontoxic to normal cells and tissues of the body, makes the development of any new pharmaceutical a long-term affair. On average, it is estimated that it requires nearly 15 years from the time of discovery to the time of clinical application of a new drug entity, a process estimated to cost approximately 650 million U.S. dollars in 1998.

Recognition of the enormous long-range value of plant-derived natural products, and the genetic sequences encoding for them, is one of the most compelling pragmatic arguments for biodiversity preservation. Indeed, the International Cooperative Biodiversity Group program in the United States has the simultaneous objectives of natural product drug discovery, biodiversity preservation, and developing short- and long-term economic benefit from ecologically preserved habitats, mainly in the tropics. One program funded through this mechanism has the innovative strategy of fostering the transmission of the folklore uses of plants and animals within a primitive society by artificially enhancing the "status" associated with the village "shaman." This is being accomplished through paying modest "salaries" to shaman and their assistants, which in turn gives these individuals a unique and desirable status with their peers.

As our understanding of the molecular basis for various diseases increases and defines new drug targets, we will need an ever-increasing source of molecular diversity to act as starting points for developing new generations of more potent and more selective drug agents. Nature's storehouse is still the largest and most chemically diverse source of unique molecular architectures. Preservation of this resource, and of the genes that encode for it, should be an overarching societal priority.

## See Also the Following Articles

BIODIVERSITY AS A COMMODITY • BIOPROSPECTING • ECONOMIC VALUE OF BIODIVERSITY, OVERVIEW • ECOSYSTEM SERVICES, CONCEPT OF • PHARMACOLOGY, BIODIVERSITY AND

## Bibliography

- Barker, R. (1997). *And the waters turned to blood: The ultimate biological threat*. Simon and Schuster, New York.
- Brown, J. A., and Nijjar, M. S. (1995). The release of glutamate and aspartate from rat brain synaptosomes in response to domoic acid (amnesic shellfish toxin) and kainic acid. *Mol. Cell. Biochem.* 151, 49.
- Gerwick, W. H., Proteau, P. J., Nagle, D. G., Hamel, E., Blokhin, A., and Slate, D. (1994). Structure of curacin A, a novel antimetabolic, antiproliferative, and brine shrimp toxic natural product from the marine cyanobacterium *Lyngbya majuscula*. *J. Org. Chem.* 59, 1243.
- Jaspars, M., and Lawton, L. A. (1998). Cyanobacteria—A novel source of pharmaceuticals. *Current Opinion in Drug Discovery and Development* 1, 77.
- Orjala, J. O., and Gerwick, W. H. (1996). Barbamide, a chlorinated metabolite with molluscicidal activity from the Caribbean cyanobacterium *Lyngbya majuscula*. *J. Nat. Prod.* 59, 427.
- Pantoazis, P., Giovanella, B. C., and Rothenberg, M. L. (Eds.) (1996). The camptothecins: From Discovery to the patient. *Ann. New York Acad. Sci.* Volume 803. New York Academy of Sciences, New York.
- Paul, V. J., and Fenical, W. (1988). Isolation of halimedatrial: Chemical defense adaptation in the calcareous reef-building alga *Halimeda*. *Science* 221, 747.
- Shu, Y.-Z. (1998). Recent natural products based drug development: A pharmaceutical industry perspective. *J. Nat. Prod.* 61, 1053.
- Todd, E. C. D. (1993). Domoic acid and amnesic shellfish poisoning—A review. *J. Food Protection* 56, 69.
- Trease, G. E., and Evans, C. W. (1983). *Pharmacognosy*, 12th ed. Bailliere Tindall, London.
- Tyler, V. E., Brady, L. R., and Robbers, J. E. (1988). *Pharmacognosy*, 9th ed., pp. 195–203. Lea and Febiger, Philadelphia.
- Wall, M. E., and Wani, M. C. (1994). Taxol: Discovery to clinic. In *Economic and Medicinal Plant Research*, Volume 6 (H. Wagner and N. R. Farnsworth, Eds.), pp. 299–322. Academic Press, London.
- Yasumoto, T., and Murata, M. (1993). Marine toxins. *Chem. Rev.* 93, 1897.



# POLLINATORS, ROLE OF

David W. Inouye  
*University of Maryland*

---

- I. What Is Pollination?
  - II. The Diversity of Pollinators
  - III. Coevolutionary History of Plants and Pollinators
  - IV. Pollinators in Natural Systems
  - V. Pollinators in Agriculture and Their Economic Value
  - VI. Conservation Biology and Pollination
- 

## GLOSSARY

- anemophily** Pollination by wind.
- anthophile** A flower-visiting animal, which may or may not be a pollinator (transfer pollen).
- cantharophily** Pollination by beetles.
- chiropterophily** Pollination by bats.
- melittophily** Pollination by bees.
- myiophily** Pollination by flies.
- nectar robber** An animal that bites holes in flowers to obtain nectar, typically from a flower with a long corolla.
- ornithophily** Pollination by birds.
- pollination** The transfer of a pollen grain (male gametophyte) to a flower's stigma.
- pollinium** A packet of pollen, characteristic of flowers in the Orchidaceae and Asclepiadaceae.
- proboscis** (plural proboscides) Tongue of an insect.
- pseudocopulation** Pollination by a male insect at-

tempting to copulate with a flower mimicking a female insect.

**sapromyophily** Pollination by carrion and dung flies.

---

**POLLINATION IS THE TRANSFER** of a pollen grain (male gametophyte) to a flower's stigma (receptive surface of the female reproductive organ), where it may germinate, grow through the style, and fertilize an ovule to produce a seed. This transfer can be accomplished by abiotic agents such as wind and water, but the majority of pollination is effected by animal pollinators seeking nutritional rewards, such as pollen and nectar, or sometimes other resources, such as floral perfumes, oils, or resins.

## I. WHAT IS POLLINATION?

The relationship between the plants and pollinators is commonly assumed to be mutualistic, with the plants benefiting from the transfer of pollen and the pollinators receiving a nutritional or other reward. However, there are also instances in which the plants appear to provide no reward and cases in which pollinators may also be seed predators. This rich diversity of relationships has been fertile ground for investigation by pollination biologists for hundreds of years. The importance of pollination for agriculture has also been a powerful stimulus for the study of plants and pollinators.

There is an important distinction to be made between “flower visitor” and “pollinator.” Visitors to flowers (anthophilous animals) cannot be assumed to be pollinators because in reality they may be nectar or pollen thieves which, owing to a mismatch in morphology or an unusual behavior, do not pollinate. For example, “base workers,” or insects that remove nectar from between the petals of a flower with an unfused corolla, or insects too small to contact the reproductive parts of a flower would not be pollinators despite the fact that they may spend much time harvesting nectar. Similarly, a small insect that collects pollen from anthers but never contacts stigmas is an example of a pollen thief that does not pollinate. Simple experiments, such as collecting stigmas of flowers to confirm pollen deposition, or the observation of transfer of a pollinium (a packet of pollen, characteristic of flowers in the Orchidaceae and Asclepiadaceae), can help to distinguish between visitors and pollinators. Unfortunately, many animals observed on flowers are probably best categorized as flower visitors because this kind of confirmation of pollination has not been conducted.

## II. THE DIVERSITY OF POLLINATORS

The list of species of animals that serve as pollinators is long and diverse. The largest group of pollinators is insects, but both flying and non-flying mammals, birds, and at least one reptile have been recorded as pollinators. Table I provides a list of pollinator classes for the world's wild flowering plants (the angiosperms; approximately 240,000 species) and estimates of the number of species in each class. Much work remains to be done to confirm the activities of these taxa as pollinators, but these numbers are the first estimates available.

### A. Invertebrate Pollinators

Although flower visitation has been observed in species from at least 16 orders of insects, only 4 include many species that regularly pollinate flowers and that seem to have been involved in coevolutionary interactions with plants: the beetles (Coleoptera), flies (Diptera), butterflies and moths (Lepidoptera), and ants, bees, and wasps (Hymenoptera). In addition, the thrips (Thysanoptera) include some pollen-eating species that may be pollinators, some stoneflies (Plecoptera) and true bugs (Hemiptera) will visit flowers and eat pollen and nectar, and some lacewing (Neuroptera), scorpion-fly (Mecoptera) and caddisfly (Trichoptera) species eat

TABLE I  
Pollinator Classes for the World's Wild Flowering Plants  
(Approximately 240,000 Angiosperm Species)<sup>a</sup>

Pollination categories	Estimated pollinator taxa
Wind (abiotic)	20,000
Water	150
All insects	289,166
Bees	40,000
Hymenoptera (bees and wasps)	43,295
Lepidoptera (butterflies and moths)	19,310
Diptera (flies)	14,126
Coleoptera (beetles)	211,935
Thrips	500
All vertebrates	1,221
Birds	923
Bats	165
Mammals other than bats	133

<sup>a</sup> The numbers are “the number of species comprising the invertebrate and vertebrate genera, families, and orders in which there are more known effective pollinators than there are pollen or nectar cheaters, robbers, or avoiders” (p. 274) (reproduced with permission from Nabhan and Buchmann, 1997).

nectar. Additional work is needed to confirm whether most of these lesser known flower visitors are indeed pollinators.

#### 1. Beetles

Beetles have been abundant since at least the Mesozoic, and it is likely that some of them have been flower visitors since the origin of the earliest angiosperms. For example, the current association of beetle pollination (cantharophily) with primitive woody angiosperms (e.g., *Magnolia*) probably dates back to the evolutionary origins of both groups. Beetle pollination is considered to be the most primitive type of pollination by animals and is not very important in cool temperate regions. It is most common in the moist tropics and to a lesser degree in arid areas. Beetles constitute the largest order of insects, and some of this diversity is thought to have arisen through the same evolutionary radiation of flowers and insects during the Tertiary that led to the origin of the other major orders of flower-visiting insects.

#### 2. Flies

Flies are probably the second most common order of flower visitors (after Hymenoptera). Fly visitors from

at least 71 families of Diptera to flowers in 137 plant families have been recorded in the literature (compilation by Inouye). The families Syrphidae (hoverflies), Bombyliidae (bee flies), and Muscidae are especially common as flower visitors. In tropical areas, the diversity of Diptera can rival or exceed that of the Hymenoptera. For example, 4856 species of Diptera have been recorded from Australasia from flower-visiting families compared to approximately 2570 bees (superfamily Apoidea), whereas for the neotropics the estimates are 2940 species for Diptera and 5630 species for bees (Roubik, 1995). In some parts of the world, including the Faroe Islands, New Zealand, and Australia, where bees are nonexistent or relatively rare, flies have filled that ecological niche of pollinators.

Pollination by flies (myiophily) is economically important; in tropical areas flies are the primary pollinators of cacao and they also pollinate mango, cashew, and tea. Roubik (1995) lists pollinators of 785 species of cultivated plants in the tropics, and 26–31 of these plants are apparently pollinated only by flies, 32 or 33 by flies as the primary pollinators, and 87–101 by flies as secondary pollinators. In Europe, flies are used commercially to pollinate protected crops of onion, chive, carrot, strawberry, and blackberry. In both tropical and temperate areas, flies appear to be more important as pollinators than has been generally recognized, and there is much need for additional research on them.

In addition to the more specialized flower-visiting flies, carrion and dung flies are pollinators of some plants in a category of pollination called sapromyiophily. The basis for attraction of the flies is floral fragrances that mimic rotting flesh or dung, and typically the flies do not receive any nutritional reward for their visits. Blood-feeding Diptera such as mosquitoes and biting flies (Tabanidae; horseflies and deerflies) are also sometimes pollinators, attracted to flowers by the nectar they offer.

### 3. Butterflies and Moths

The Lepidoptera (butterflies and moths), as a consequence of their two distinct lifestyles as primarily herbivorous larvae and commonly flower-visiting adults, have dichotomous relationships with plants. In at least one case, the pollination of yucca, the act of pollination is accompanied by oviposition and the larval yucca moths are seed predators. Although not all Lepidoptera feed as adults, most if not all that do depend on nectar as a source of both sugars and amino acids. Some of the longest-lived of all butterflies, the Neo-tropical *Hel-*

*conius* species, also feed regularly on pollen by collecting it on their proboscis and regurgitating nectar on the pollen mass to leach out amino acids. The floral rewards sought by Lepidoptera are not always nutritional; some species are attracted to flowers by the pyrrolizidine alkaloids that they produce, which the male butterflies require for the production of pheromones, for mating success, and for the establishment of multi-species leks. Lepidoptera, as agents of natural selection, may be responsible for much of the diversity of defensive chemicals in plants as well as the diversity in flowers and the rewards they offer.

Two families of moths stand out as pollinators: the hawkmoths (Sphingidae; also called sphinx moths) and the noctuid moths (Noctuidae; the largest family in the order Lepidoptera). The hawkmoths are strong and agile fliers that maintain very high body temperatures (e.g., as high as 46°C) while in flight. Some diurnal species are easily mistaken for hummingbirds (e.g., the European hummingbird hawkmoth, *Macroglossum stellatarum*). Some hawkmoths are migratory and (in successive generations) may move hundreds of kilometers during a summer. Hawkmoths include the pollinators with the longest known tongues; most of these extraordinary species are found in Madagascar. Darwin measured nectar spurs in some orchids from Madagascar that were 29 cm long, and concluded that there must be moths there that have proboscides of the same length. Forty years after his prediction, a hawk moth with a proboscis that reaches more than 24 cm in some individuals was found and given the name *Xanthopan morgani praedicta*. There is actually a large guild of long-tongued hawkmoths and flowers with matching corolla lengths in Madagascar.

Some noctuid moths are long-distance migrants, flying distances as far as 1600 km. One species is reported to travel as much as 1000 km in a day when migrating from the Baltic states to Britain. This presents the opportunity for some very long-distance gene flow on the part of their food plants. Some species migrate vertically, moving up in altitude during some times of year to avoid unfavorable seasonal weather. Before these migrations the moths may store substantial quantities of fat, and at least one such species was an important part of human diets; the bogong moth (*Agrostis infusa*) was collected while aestivating in the Snowy Mountains by aboriginal Australians.

### 4. Ants, Sawflies, Wasps, and Bees

The Hymenoptera (ants, sawflies, wasps, and bees) are the largest and most diverse group of pollinators. Ants

are only rarely recorded as pollinators, although they are not uncommon as flower visitors collecting nectar. One suggestion for why there are not more species of plants adapted for pollination by ants is that the antibiotic and other secretions on their exoskeletons are detrimental to pollen survivorship and growth. Sawflies are more common as pollinators. They have herbivorous larvae, and in some cases the adults restrict their flower visits to the larval host plant. The adults may also eat flower parts in addition to nectar and pollen. Flower visitation by wasps has been studied in part because of interest in attracting and sustaining populations of parasitoid wasps (which prey on pest insects) in agricultural situations. However, predatory and social wasps have also been recorded as flower visitors and a few species of plants are specialized for pollination by pseudocopulation by male wasps (see Section IV,B2). Some of the most specialized relationships between plants and pollinators are those between species of figs and the wasps that pollinate them; many of these relationships are reported to involve single species of wasps pollinating each fig species.

There are fewer species of bees than beetles, but a much higher proportion of bees are pollinators than are beetles. Almost all species of bees are dependent on flowers for pollen and nectar as nutritional resources as both adults and larvae, and of all the insects bees are the mostly highly adapted for flower visitation and pollination. Their behavior, morphology, and senses of vision and smell all appear to be adapted for finding and collecting floral resources. For example, the trichromatic color vision of bumblebees and honeybees includes ultraviolet, which allows them to see the ultraviolet reflectance patterns that many flowers use as nectar guides or a petal color. Humans cannot see these patterns without the help of cameras.

Although pollen (primarily for feeding larval bees) and nectar (both a larval and adult resource) are the most common resources collected by bees, some flowers offer oil as a reward. The oil is collected by bees in the family Anthophoridae; females of these species have absorbent brushes for holding the oil and sharp edges on their legs that are used to squeeze the oil out so it can be mixed with pollen as food for larvae. Some bees collect resin as a floral reward, which they then use in nest construction.

Pollination by bees (melittophily) is very important for agriculture (see Section V). The social bees are the best known crop pollinators and are the species most commonly managed for pollination or honey production. These include honeybees (*Apis*), bumblebees

(*Bombus*), and stingless bees (*Trigona*). The presence of large quantities of honey (concentrated nectar), pollen, and larvae makes social bee nests an attractive target for insects and vertebrates that prey on nests and adult bees. Bee-eaters and honey guides are examples of Old World birds that specialize on such resources.

## B. Vertebrate Pollinators

### 1. Mammals

Although invertebrates are the most common pollinators, mammals, birds, and lizards have also been documented as pollinators. Mammalian pollinators include both bats, some of which eat primarily nectar and pollen, and nonflying species such as some Australian marsupials. Most bat pollination occurs in the tropics, but some migratory bats are important pollinators in temperate North America to about 30° north latitude. In the Old World, species of the family Pteropidae are important pollinators, whereas in the New World flower-visiting bats are confined to the Phyllostomidae. Although pollination by bats (chiropterophily) is geographically widespread, the other mammalian pollinators are more restricted and less common as pollinators, including lemurs in Madagascar and several species of Australian marsupials such as sugar gliders, honey possums, and some marsupial mice. There are some reports of other mammals, including mice and giraffes, that may serve as pollinators in unusual cases. Vertebrate taxa other than mammals or birds are only rarely reported as pollinators; there is one report of a lizard that appears to be a pollinator.

### 2. Birds

Bird pollination (ornithophily) is common in many parts of the world. Several families of birds are primarily nectarivorous and undoubtedly include important pollinators. These include the hummingbirds (Trochilidae; New World), honeyeaters (Meliphagidae; Australia, New Zealand, and parts of Asia), sunbirds (Nectariniidae; Africa and southwest Asia to the Philippines), sugarbirds (Promeropidae; South Africa), flowerpeckers (Dicaeidae; Asia and Australasia), and Hawaiian honeycreepers (Drepanididae; Hawaii). Other families that include some nectarivores or for which there are a few records of flower visitation include the Thraupidae (honeycreepers and some tanagers; New World), Icteridae (orioles; New World), white-eyes (Zosteropidae; Africa, Asia, and Australia), and Psittacidae (lorikeets and hanging-parrots; Southeast Asia and Australia). Many flower-visiting birds are recently extinct in the

south Pacific islands, but in at least one case an introduced species has taken over the role of pollinator for some of the endemic plants that were pollinated by the extinct birds.

### III. COEVOLUTIONARY HISTORY OF PLANTS AND POLLINATORS

The “abominable mystery” to which Darwin referred, the origin and relationships among flowering plants, is still being revealed. Beetle pollination is thought to be the primitive condition in flowering plants, but by 100 million years ago butterflies and moths had joined beetles as important pollinators. Recently discovered fossil evidence indicates that the origin of the angiosperms (flowering plants) may date back as far as the late Jurassic, suggesting that the origin of the coevolutionary relationship between flowers and pollinators may also be older than previously thought. Much of the tremendous biodiversity found in the flowering plants is thought to have arisen through evolutionary interactions with pollinators, and certainly much of it is maintained through their flower-visiting activities that result in production of seeds. The remarkable diversity of flower size, shape, odor, rewards, color, and pollination mechanisms is the result of these coevolutionary relationships.

### IV. POLLINATORS IN NATURAL SYSTEMS

#### A. Obligate and Facultative Mutualisms

The relationships between plants and pollinators range from obligate to facultative in nature. Although one study in Brazil found that 43% of the plants studied were visited by only a single species of pollinator, other studies have found tens of visitor species to a single plant species. In only a few cases have researchers quantified the relative importance of different species in cases with multiple pollinators of a single plant species. A common theme of multiyear studies is variation, both temporal and spatial, in the abundance of these different groups of pollinators, which can include diverse taxonomic groups including both vertebrate and invertebrate pollinators (e.g., hummingbirds and bumblebees) visiting a single plant species. By having several groups of potential pollinators, it is more likely that these plants will not suffer from pollen limitation (a shortage of

pollination); if one group of pollinators is unusually low in abundance at a particular site or during a particular year, it is likely that another group will not be low in abundance. Thus, it can be important for there to be a diversity of pollinators available.

The obligate nature of some plant–pollinator relationships extends to pollinators because some insects appear to visit flowers of only a single species of plant. In such cases there must be a good match in the phenology of life cycles of both plant and insect to ensure success for both partners, and a concern about global climate change is the potential for such phenological relationships to become mismatched. There are similar cases of plant species visited by only a single species of pollinator. Although a plant or pollinator may thus have an obligate relationship with a single partner, the partner may be involved in relationships with multiple species.

Although we assume that most plant–pollinator relationships are mutualistic in nature, some examples appear to involve a mixture of positive and negative interactions from at least the plant’s perspective. Several studies have demonstrated seed predation or other forms of herbivory by insects that may also serve as pollinators. These examples include beetles and palms, moths and yuccas, wasps and figs, as well as flies and moths and some flowers that they pollinate. Some of these relationships (e.g., yucca flowers and their moths) can be further complicated by the presence of closely related species of insects that are herbivores but not pollinators. Much detailed study is needed to dissect the complicated interactions in such relationships.

There is no doubt that long-lived or highly mobile pollinators must interact with a large number of plants because their life spans or travels extend beyond the temporal or spatial availability of single plant species. Thus, a group of plant species, even if they are not sympatric, could be considered mutualists if in concert they sustain and share a particular pollinator. This kind of multispecies relationship has significant conservation implications because it may be necessary to conserve plant species that flower widely separated in time and space to conserve a (e.g., migratory) pollinator.

#### B. Cheaters in Pollination

##### 1. Pollen and Nectar Robbers

The nutritional rewards that pollinators find in flowers—pollen and nectar—are sometimes harvested by flower visitors that may pierce or bite holes in the flowers to obtain the resources “illegitimately.” Some species of birds (*Diglossa*, the flower-piercers) and bees (*Xylo-*



*copa*, carpenter bees; some short-tongued *Bombus*, or bumblebees; and some stingless bees) may obtain most of their nectar in this fashion. Once these primary nectar robbers have created holes, other species may learn to use them (secondary robbers). Typically, the flower species that are robbed have long corollas, and the nectar robbers do not have mouthparts long enough to obtain the nectar without robbing. Although biologists at least as far back as Darwin have assumed that nectar robbing would have a significant negative influence on the robbed plants, recent evidence suggests that in fact most insect nectar robbers that have been studied have a beneficial or neutral effect on the plants. In the process of moving around on the flowers, or perhaps while collecting pollen from the same flowers, robbing insects may effect pollination. Even in the absence of such a direct effect, robbers may indirectly increase the fitness of plants they rob by influencing the behavior or legitimate pollinators. The reduced quantity of nectar in a robbed flower may induce a legitimate pollinator to visit more flowers or to fly greater distances between flowers, thereby increasing gene flow.

## 2. Pollination by Deception

Cheating in pollination biology is not restricted to the pollinators. Plants also provide fascinating examples of deception, attracting some pollinators by falsely advertising opportunities for feeding or sex. The orchid family is perhaps best known in this regard because many species of orchids have flowers with no nectar and pollen that is not accessible or attractive to pollinators. In such species, it appears that the pollinators gain nothing from visiting (and pollinating), and quickly learn to avoid the flowers. Other orchids have flowers that in both morphology and odor (production of insect sex pheromones) sufficiently resemble female insects that males attempt to copulate with them and in the process pollinate (pseudocopulation).

## V. POLLINATORS IN AGRICULTURE AND THEIR ECONOMIC VALUE

Our knowledge of the pollination biology of crop species is surprisingly sparse. For example, the pollination requirements of about one-third of the crop species grown in the European Union are unknown. These plants include at least 264 species that are cultivated as crops, or gathered from the wild, for human and/or livestock food or for the essential oils they contain. The

best known pollinators of cultivated food plants are honeybees, which have been widely introduced throughout the world (see Section VI). However, honeybee populations in the United States have been decimated in recent years by the introduction of two species of parasitic mites and are also being threatened by the introduction of a hive beetle. Most feral colonies of honeybees in the United States have disappeared in the past decade and the number of managed colonies has declined to less than half of its peak. This decline has resulted in a growing appreciation of the roles of other, native, species in the United States and elsewhere as pollinators of crops.

These problems have also prompted some attempts to quantify the economic value of pollination services provided by both managed and wild pollinators. A recent model estimates that U.S. consumers realize \$1.6–5.7 billion in annual social gains that would be lost if honeybee services for 62 crops were reduced. The potential value of non-honeybee pollinators in the U.S. agricultural economy is estimated at \$4.1–6.7 billion each year. The global value of pollination services has been estimated to be \$117 billion per year.

### A. Introduced Species of Pollinators

Undoubtedly the largest-scale introduction of pollinators has been the global traffic in the European honeybee (*Apis mellifera*). Such continental-scale introductions have a long history. For example, European colonists in North America imported honeybees as early as 1641. This species went on to establish feral colonies throughout much of North America and probably through subsequent introduction throughout much of the neotropics. They were also introduced to Australia. Probably the second largest introduction of pollinators has been the spread of bumblebees (*Bombus*). The first introduction of bees anywhere specifically for pollination was that of European bumblebees to New Zealand in 1885, and subsequently they were also introduced to Australia. The second successful introduction of a pollinator was a fig-pollinating wasp (*Blastophaga psenes*) into California in 1899.

In these cases, the incentive for introducing bees or wasps was for pollination of introduced species of plants that were not attractive to or could not be pollinated by native pollinators (e.g., clover in New Zealand and figs in California). Recently, international transport of bumblebees has been the result of their use in pollination of greenhouse crops such as tomatoes. In the 1980s, techniques were developed for the year-round commercial-scale propagation of bumblebee colonies (which

normally have an annual life cycle of a few months). Commercially produced colonies of bumblebees have now largely replaced the tedious hand pollination of tomato flowers with electronic vibrators (tomato flowers produce no nectar but are visited by bees that “buzz” the flowers to release pollen from poricidal anthers so that they can collect it for feeding their larvae, and during this process they pollinate the flowers) and the use of chemical sprays to induce fruit set. In the early 1990s a European species of bumblebee (*Bombus terrestris*) was introduced to Japan for use in greenhouse pollination; it has subsequently escaped and is now established in the wild. Although an effort is being made to exterminate the species in the wild (and to develop native species for use in greenhouse pollination), it is likely to be unsuccessful.

Other species of pollinators have also been introduced. The wasp pollinators of figs have been introduced to New Zealand and the United States, where they pollinate naturalized figs. The Japanese white-eye, an introduced passerine in Hawaii, has taken over the role of pollinator of two native plants that lost their original pollinator to extinction. An Asian bee is now the primary pollinator of alfalfa plants grown for seed production in the United States, and a Japanese bee is used in the United States and Canada for pollination of apples.

## B. Pollinators as Vectors and Victims of Engineered Genes

As genetic engineering of crop plants has become more common, and as these crops are approved for cultivation in the field, concerns have arisen about the potential for gene flow through pollen transport between the engineered genomes and wild plants. One management tool to prevent, or at least minimize, this potential is to plant a buffer zone of unengineered plants around a field of engineered plants in hopes of intercepting pollinators carrying pollen before they can visit related native plants.

There is also the potential for pollinators to become victims of engineered genes. For example, laboratory studies of monarch butterfly caterpillars showed that if they eat significant quantities of corn pollen from plants engineered to have the BT toxin in their leaves, this can kill them. Corn is wind pollinated, and a likely scenario is that large quantities of corn pollen could be distributed onto leaves of milkweed plants at the margins of corn fields. The potential detrimental effects of this on the monarch butterfly and other herbivores remain to be studied.

## VI. CONSERVATION BIOLOGY AND POLLINATION

### A. Threats to Plant–Pollinator Mutualisms

Pollinators face a large variety of threats of anthropogenic origin, including habitat fragmentation, a variety of effects of agriculture, pesticides and herbicides, and the introduction of both pollinators and plants. Fragmentation of habitats is likely to affect nonflying pollinators most strongly because it may be difficult for them to locate and visit patches of flowers that are isolated by areas of different, perhaps nonpreferred, habitat. However, even winged pollinators may be reluctant to leave a preferred habitat, such as undisturbed forest, to fly across a newly created pasture to another patch of forest. If the fragments are small enough to affect the size of pollinator populations, there may be negative genetic consequences for the smaller populations.

Modern agriculture can have a variety of negative impacts on pollinators. The creation of large fields that are disturbed regularly by plowing can prevent ground-nesting bees from establishing populations, and if only a single plant species is available there may not be enough pollen and nectar resources to support the life cycle of a pollinator species. The use of pesticides to control agricultural pests can have a negative impact on pollinators, and herbicides may remove species that could provide resources for the pollinators. Pesticide use in nonagricultural areas, such as spraying of large tracts of forest to control lepidopteran herbivores, has also been shown to have strong negative impacts on nontarget pollinator populations.

The introduction of exotic plants, such as weedy plants, could have potential implications for the pollination of native plants. If the exotics are close relatives of native species, there is potential for hybridization. Also, if the exotics are prolific producers of nectar or pollen, they may draw pollinators away from native species that may then suffer a deficit in seed production. The potential for these kinds of consequences is not well understood, but this area is beginning to attract attention from researchers.

The ecological consequences of widespread introductions of *A. mellifera* and *Bombus* for native plants and pollinators are not well-known because of a lack of data from before the introductions. We can speculate that the introduction of honeybees to North America may have had significant consequences for species of bumblebees that have the same proboscis length be-

cause the perennial honeybee colonies have many more workers than the annual bumblebee colonies. The spread of *A. mellifera scutellata*, the African subspecies ("Africanized honeybees"), from its point of release in Brazil to as far north as the United States was much more recent and has provided opportunity for some studies of their impact on native Neotropical pollinators (relatively minor, it appears).

Species from every major group of pollinators, invertebrate and vertebrate, have been classified as endangered, and recent extinctions have been documented for others. Thus, pollinators are just as susceptible to the current human-induced mass extinction as any other group of organisms.

## B. Potential Management Solutions

A first step in solving problems in the conservation of plant–pollinator relationships is gathering information about both the nature of the relationships and the problems plants and pollinators face. Probably the best way to conserve pollinators is to preserve habitat that includes their food plants and nest sites. Given that the growing human population is resulting in rapid habitat destruction, the establishment of protected areas is an important conservation tool. In agricultural situations, management techniques as simple as changing the timing of pesticide application, leaving buffer zones around fields where bees can nest and food plants can grow, or providing suitable artificial nesting sites can make a significant difference. Domestication of wild bees may also play a role in their maintenance in agricultural situations.

If pollinators are locally extinct, it may be possible to reintroduce them once the factors that caused the original extinction are addressed. If the species is globally extinct, plants that were dependent on that pollinator could be maintained through hand pollination or the introduction of a suitable replacement. Although the bird was not introduced for this purpose, the presence of the Japanese white-eye in Hawaii has had the

effect of replacing an extinct endemic avian pollinator. Removal of introduced pollinators, such as the honeybee, may also help to preserve populations of native pollinators.

We have taken some important steps toward conservation of pollinators and thus the relationships they have with plants. However, much basic knowledge about pollination and pollinators remains to be learned and it remains to be seen how many of these species can be conserved.

## See Also the Following Articles

BETLES • BUTTERFLIES • COEVOLUTION • FLIES, GNATS, AND MOSQUITOES • HYMENOPTERA • MOTHS

## Bibliography

- Buchmann, S. L., and Nabhan, G. P. (1996). *The Forgotten Pollinators*. Island Press, Washington, DC.
- Dafni, A. (1992). *Pollination Ecology: A Practical Approach*. IRL/OUP, Oxford.
- Faegri, K., and van der Pijl, L. (1979). *The Principles of Pollination Ecology*, 3rd ed. Pergamon, Oxford.
- Grant, V., and Grant, K. A. (1965). *Flower Pollination in the Phlox Family*. Columbia Univ. Press, New York.
- Jones, C. E., and Little, R. J. (eds.) (1983). *Handbook of Experimental Pollination Biology*. Scientific and Academic Editions, New York.
- Kearns, C. A., and Inouye, D. W. (1993). *Techniques for Pollination Biologists*. Univ. Press of Colorado, Niwot.
- Kearns, C. A., Inouye, D. W., and Waser, N. M. (1998). Endangered mutualisms: The conservation biology of plant–pollinator interactions. *Annu. Rev. Ecol. Syst.* **29**, 83–112.
- Matheson, A., Buchmann, S. L., O'Toole, C., Westrich, P., and Williams, I. H. (eds.) (1996). *The Conservation of Bees*. Academic Press, New York.
- Nabhan, G. P., and Buchmann, S. L. (1997). Services provided by pollinators. In *Nature's Services. Societal Dependence on Natural Ecosystems* (G. C. Daily, ed.), pp. 133–150. Island Press, Washington, DC.
- Proctor, M., Yeo, P., and Lack, A. (1996). *The Natural History of Pollination*. Timber Press, Portland, OR.
- Real, L. (1983). *Pollination Biology*. Academic Press, Orlando, FL.
- Roubik, D. W. (ed.) (1995). Pollination of cultivated plants in the tropics. Agricultural Services Bulletin No. 118. Food and Agriculture Organization, Rome.



# POLLUTION, OVERVIEW

William H. Smith  
Yale University

---

- I. Introduction
  - II. Perturbation of the Carbon Biogeochemical Cycle
  - III. Perturbation of the Nitrogen Biogeochemical Cycle
  - IV. Perturbation of the Sulfur Biogeochemical Cycle
  - V. Trace Pollutants of Global Importance
  - VI. Conclusion
- 

## GLOSSARY

- acid deposition** Precipitation or dry fall-out with a pH of less than 5.0.
- anthropogenic** Related to the activities of human beings.
- biodiversity** Total inventory of genes, organisms, species, populations, communities, and their habitats.
- biogeochemical cycles** Pathways and storage of chemical elements through the biota and geologic resources, including the atmosphere, hydrosphere, and lithosphere.
- biomagnification** The increase in body burden of persistent pollutants in organisms at higher trophic levels in food webs.
- climate change** Variability of weather over time.
- ecotoxicology** The study of chemical stressors on ecological resources.
- food web** The pathways of energy and nutrient movement between organisms in an ecosystem.

**hydrologic cycle** The movement of water between the surface of the earth and the atmosphere.

**nitrogen saturation** Nitrogen levels in excess of the biological nitrogen needs of an ecosystem potentially causing adverse ecological effects.

**ozone** Tri-atomic oxygen naturally present (and beneficial) in the stratosphere but present in the troposphere (due to human activities) in amounts potentially toxic to the biota.

**pollution** Materials or chemicals in excess of natural levels caused by the activities of humans.

---

**BIODIVERSITY INCLUDES THE** total inventory of genes, organisms, species, populations, communities, and their environments, and it integrates all of their complex interactions. This chapter focuses on the approximately 1.4 million species of living organisms known to presently exist on earth while recognizing that this total may represent fewer than 15 percent of the actual number (Raven and Wilson, 1992).

## I. INTRODUCTION

Over geologic timescales, species numbers have been very dynamic. Massive species disappearances have occurred as follows: dinosaurs and marine reptiles in the Cretaceous (66 million years ago), reptiles and cono-

donts in the Triassic (210 million years ago), trilobites and marine animals in the Permian (250 million years ago), marine reef plants and animals in the Devonian (365 million years ago), and nautiloids and small ocean animals in the Ordovician (440 million years ago). Their mass extinctions are presumed to have resulted from earth impacts (asteroid collision?), climatic change, or large-scale alteration of biogeochemical cycles.

Changes in major element biogeochemical cycles over geologic time have been responsible for the appearance and disappearance of life forms. Changes in the oxygen cycle, for example, which allowed increased concentrations of oxygen gas in the troposphere and ozone in the stratosphere, allowed the evolution of life on earth to proceed. Major episodes of volcanic activity, resulting in marine eruptions and the release of sulfur containing and particulate pollutants, has periodically destroyed life.

Contemporary losses of species are recognized to result from physical stressors (e.g., habitat loss, harvesting), biological stressors (e.g., exotic organism introductions), and chemical stressors (changes in biogeochemical cycles, pollution). This chapter focuses on the latter group of stressors. Presently humans are recognized to have the ability to alter global and regional biogeochemical cycles via air, water, and soil pollution. As a result, the impact of pollution on biological diversity is a contemporary concern of major dimension (Barker and Tingey, 1992; Peters and Lovejoy, 1992).

## II. PERTURBATION OF THE CARBON BIOGEOCHEMICAL CYCLE

Mineral carbon is located in a limestone ( $\text{CaCO}_3$ ) reservoir from which it may be leached into a mineral solution as dissolved hydrogen carbonate ( $\text{HCO}_3^-$ ) formed when dissolved  $\text{CO}_2$  reacts with  $\text{CaCO}_3$ . In the atmosphere carbon exists as carbon dioxide ( $\text{CO}_2$ ). Carbon dioxide in the atmosphere is transferred to the biosphere via photosynthesis. Organic (fixed) carbon in the biosphere is ultimately returned to the atmosphere as  $\text{CO}_2$  via microbial decomposition of organic matter. Another fraction of carbon is ultimately fixed as coal, petroleum, and natural gas. This fraction may return to the atmosphere as  $\text{CO}_2$  via combustion.

In recent decades, human activities have accelerated the return of  $\text{CO}_2$  to the atmosphere via increased deforestation, biomass burning (Nepstad *et al.*, 1999), and fossil fuel combustion (IPCC, 1996.)

The accelerating input of  $\text{CO}_2$  to the atmosphere

from human deforestation and fossil fuel combustion (very approximately 8 billion metric tons annually) (Fig. 1) is a significant driver of global climate change. In addition, the incomplete combustion of fossil fuels generates a variety of toxic and precursor compounds that contaminate the atmosphere and ultimately may cause adverse effects on plant and animal populations. Both of these stressors (i.e., climate change and air pollutants) have significant potential to reduce biological diversity.

### A. Climate Change

Climate is defined as the time-averaged value of meteorological quantities. Over time, climate, like an ecosystem, is characterized by change not constancy. In geological terms, the climate of the earth is most typically characterized by extended "moderate periods" with equable weather the year round, lack of ice caps, and generally warm seas. Humans evolved after the last "moderate period" and our development has been in a period of climatic revolution. This period of revolutionary climatic alteration has been characterized by a complex of "cycles within cycles." Over the past 3000 years, for example, the general evidence suggests that the northeastern portion of North America has become cooler and more moist. Over the past several hundred years, however, particularly during the first half of the present century, there is evidence for a moderating trend. There is general agreement that there has been a systematic fluctuation in recent global climate characterized by a net worldwide warming of approximately  $0.5^\circ\text{C}$  between the 1880s and the early 1940s. Warming has generally continued over the past three decades.

The regulation of global climate is complex and incompletely appreciated. Numerous hypotheses have been proposed to explain the forces responsible for the variability of climate. The most plausible of these include variations of the solar constant, changes in solar activity, passage of the solar system through an interstellar gas-dust cloud, variation in the velocity of the earth's rotation, gigantic surges of the Antarctic ice sheet, changes in the earth's orbital parameters, and alterations in the interactions between glaciers and oceans. Climate may respond rapidly and dramatically to small changes in these independent variables.

Added to this complexity and uncertainty is the suggestion that the activities of human beings, particularly land use activities and atmospheric contamination, are currently influencing global and regional climates. Numerous trace gases of the atmosphere, including water vapor, carbon dioxide, methane, nitrous oxide, halocar-

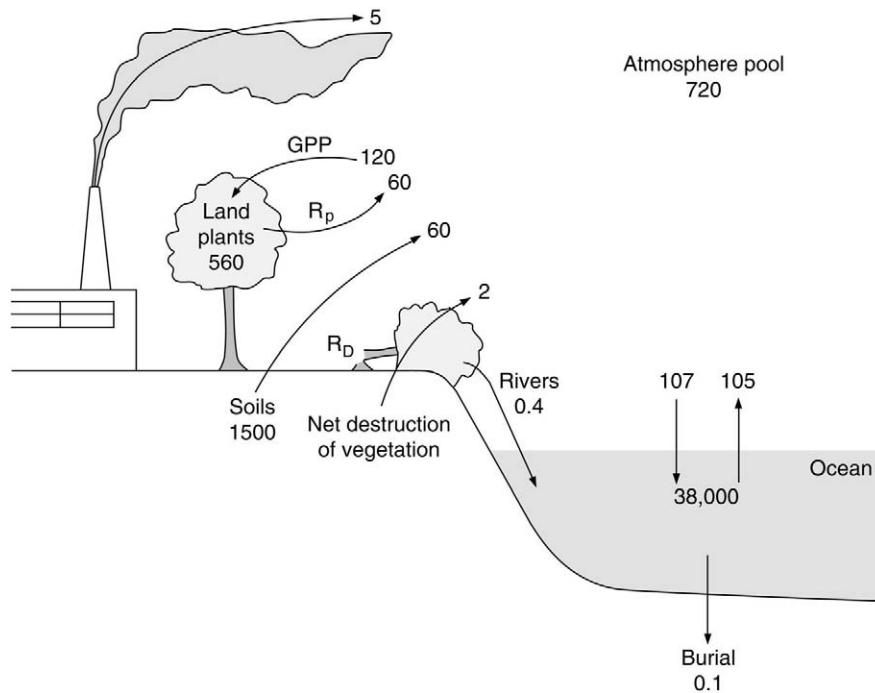


FIGURE 1 The global carbon cycle. The storage (numbers without arrows) and the annual flows (numbers with arrows) are shown for the global system. The units are pentagrams of carbon (=  $10^{15}$ g C = one giga-ton = 1 billion metric tons). From Schelesinger (1992).

bons, and ozone (greenhouse gases), have strong infrared absorption bands. As a result, these gases can have a significant effect on the thermal structure of the atmosphere because they absorb within the 7- to 14- $\mu$ m atmospheric window, which transmits most of the thermal radiation from the surface of the earth and troposphere to space. Presumably a primary result of more carbon dioxide and other “greenhouse” gases in the atmosphere will be warming. While incoming solar radiation is not absorbed by carbon dioxide and these trace gases, portions of infrared radiation from the earth to space are. Over time, the earth will become warmer. While the forces controlling global temperature are varied and complex, as suggested, the increase of 0.5°C since the mid-1800s is generally agreed to be at least partially caused by increased carbon dioxide.

General circulation models, widely used to predict future climates, are numerical models of the earth-atmosphere system that solve the basic equations for atmospheric motion and provide boundary conditions of the earth and ocean. General circulation models divide the landscape into a series of regions called *grid boxes*. Each grid box consists of a series of layers, which represent land, ocean, and layers of the atmosphere. Global weather is simulated using difference equations,

which detail the physics and dynamics of energy and material movement among the boxes. To evaluate the climatic change hypotheses, models are initially run with current atmospheric conditions. They are then run with altered atmospheric trace gas concentrations (for example, carbon dioxide) to simulate a new climate. The difference between these two equilibrium climate simulations provides an estimate of how the climate system may change. Most general circulation models predict warming and overall intensification of evaporation and precipitation in the warmer climate regimes of doubled atmospheric carbon dioxide.

Using a mid-range scenario of driving variables, an increase in global mean surface air temperature relative to 1990 of about 2°C by 2100 has been estimated. Under any scenario, the average rate of warming will probably be greater than any seen in the past 10,000 years (IPCC, 1996). General circulation model predictions of the change in the hydrologic cycle with doubling of atmospheric carbon dioxide are not clear. While precipitation is generally estimated to increase, warming will intensify evaporation. As a result, soil moisture may not increase with warming; it may even decrease. The movement of water into and out of the soil-vegetation system has a large influence on the hydrologic cycle

and is very variable regionally. Models have predicted that in the United States, portions of the Northwest could become wetter, portions of the Southern Great Plains could become somewhat drier, and portions of the Northeast could become considerably drier.

## B. Ecosystems and Climate Change

The consequences of a warmer global climate, with even very modest changes in regional temperature and hydrologic patterns, can have profound effects on ecosystems. Natural ecosystems, as opposed to artificial ecosystems designed and intensively managed by humans, may be the most vulnerable to climate change. We are least able to intervene and facilitate adaptations of natural ecosystems because of their expansiveness, our limited knowledge about fundamental ecological processes, and the absence of effective global institutions to assume responsibility for adaptations.

A shift in climate zones may exceed the ability of vegetation, for example grassland, savanna, or forest systems, to adapt through migration. If climate zones shift hundreds of miles over the next century, rates of vegetative dispersal and colonization, which are on the order of tens of miles in a century, may be inadequate to maintain present-day eco-region patterns. With regard to terrestrial ecosystems, climate change may also adversely influence habitats by changing patterns of major stressor groups including fire regime, drought regime, and insect and disease relationships.

Alteration of ecosystems could lead to a change in species diversity, including a loss of species. The potential for migration of plants and animals to new suitable habitats is not known, but barriers such as water bodies, roads, or development could impede migration. Isolated ("island") species may find themselves in climate zones that are no longer suitable for their survival.

Montane species, unless they can migrate to higher elevations, may face extirpation or extinction. A study of a 3°C warming in the Great Basin National Park in eastern Nevada concluded that it would cause 20 to 50 percent of species in individual mountain ranges to go extinct (Murphy and Weiss, 1992).

### 1. Forests

Tropical montane cloud forests are unique systems due to their dependence on regular cycles of cloud formation for their existence. Still *et al.* (1999) have simulated environmental conditions for such forests at CO<sub>2</sub> doubling and found that a combination of reduced cloud contact and increased evapo-transpiration could threaten the persistence of these systems. Pounds *et al.*

1999 have, in fact, provided evidence that alterations in bird, reptile, and amphibian communities in the highland forests at Monteverde, Costa Rico, have occurred and may be related to recent increases in air temperature.

The vulnerability of forest ecosystems to climate change is a function of the forest location, biology, and human management activity. Forests at greatest risk will generally be those located in the northern portions of temperate latitudes. Altitude, proximity to continental margins, and distance from large water bodies will be important influencing factors. Plant species that are able to use increased CO<sub>2</sub> efficiently will have an advantage over other species. Similarly, species that distribute seeds widely may fare better under climate change. A mixed-species forest may tolerate a wider variety of changes than would a single-species forest. Individual plants with a low tolerance for climate fluctuations would be least adaptable to ongoing climate change. Individual populations with little genetic diversity among plants might prove to be at greatest risk of long-term decline. The most intensively managed industry and private forest land may be least at risk of catastrophic loss or long-term decline because mitigation efforts to reduce such effects will be undertaken. Many private forest managers have both the financial incentive (and resources) and the latitude to protect against extensive loss from climate-related threats. They can use several available techniques: short rotations to reduce the length of time a tree is exposed to an unsuitable climate; planting better-adapted varieties developed through selection, breeding, or genetic engineering programs to reduce vulnerability; and thinning, weeding, pest management, irrigating, improving drainage, or fertilizing to improve general health (Table I).

The first observable effects of climate change on ecosystems will not be so much climate related as weather related. The near-term effects of climate change will be driven by changes in weather extremes and mediated through those stressors that have always been the primary controllers of ecosystem structure and health: insects, disease, wind, and fire. Even in regions where productivity may be ultimately improved, the transition period could be extended and punctuated by sudden dieback and decline. Forest ecosystems are complex, long-lived systems that can only slowly adjust to climate but that can suddenly be threatened by weather-related stresses.

The near-term response of forest systems to climate change will involve complex reactions to new averages, new patterns, and new extremes in weather variables. Some forest species that are specialized to current cli-

TABLE I  
Forest Ecosystem Vulnerability to Climate Change Caused by Alteration of the Global Carbon Cycle

Forest location	Forest biology	Forest management
Higher latitude	Small, fragmented range	Fragmented forests
Higher elevation	No or few migration corridors	Less-diverse forests
Continental interior	Low genetic variance	High stand density
Maritime sites	Low species diversity	Inappropriate species
Forest-range boundaries	Genetically specialized to site	
Low-productivity sites	History of widespread dieback	
High-fire-risk sites	Heavy seed	
Drier sites		

From Smith *et al.*, 1993.

mate conditions may not thrive. Altered patterns of exposure to high and low temperatures could mean that winter chilling requirements will not be met. Flowering, seed-formation, and seed-dispersal processes could be disrupted, especially if pollinators do not adjust to changing conditions. With longer growing seasons, trees might add more “light” earlywood relative to the “dense” latewood that forms at the end of the growing season. This would mean a lower-quality wood for structural lumber and higher costs for pulp mills. Changes in early growing season weather conditions, particularly moisture and frosts, may affect the establishment of seedlings. Warmer and moister weather might favor the spread and boost the significance of certain fungal diseases and insect pests. Elsewhere, the drying effects of higher temperatures could be especially damaging, especially where the frequency of drought is increased. Associated with droughts would be higher risks of secondary threats from forest fires and insects. Insect damage may increase, for example, if insect pests produce more generations or persist longer during the tree-growing season.

The initial effects of climate change will not at first be easily recognized as distinct from the effects of the normal regulators of forest growth and development. The potential initial effects of climate change can be illustrated by current weather-related stresses on selected highly valued tree species in several regions of the United States. These potential vulnerabilities, by region, are presented in Table II.

## 2. Coastal-Marine

Coastal zones are very productive ecosystems, very rich in species, and very perturbed by human activities. In addition, they may be at substantial risk from global

climate change. These risks stem from changes in water column temperature or other water quality parameters, change in ocean current dynamics, and alterations of sea level.

Widespread extinctions are not likely, but expansive alteration of community distributions and composition are possible. Most marine organisms, in contrast to long lag times discussed for forest ecosystems, have high mobility, large ranges, high fecundity, and rapid growth rates. As a result, they may be able to adapt to climate change more efficiently than terrestrial systems (Ray *et al.*, 1992).

Coral reefs are the most species rich of marine ecosystems. These reefs are largely restricted to tropical waters, but generally do not occur where temperatures exceed 30°C for extended periods. As a result, even small increases of 1 to 2°C for the surface waters of tropical oceans could have important implications for reef ecosystems.

Temperate latitude estuarine and tidal wetland and marsh ecosystems will be at substantial risk if sea level increases over the next several decades. For each cm of rise in sea level, beaches may erode 1 m landward. Storm surges, a major eroding force, will increase, especially in areas of modest shore slopes. For every 10 cm rise, saltwater wedges in estuaries and tidal rivers may advance 1 km. Sea-level rise will also increase salinity intrusion into coastal freshwater aquifers (NAS, 1987).

## 3. Fresh Water

Freshwater ecosystems will present challenges for native biota if global climate continues to warm. Changes in precipitation and evapotranspiration can combine to offset surface-water and groundwater quantity and quality, low and high flow conditions, and drought and



TABLE II  
Forest Vulnerabilities of Selected United States Forest Systems to Changes Resulting from Climate Alteration.

Region	Tree	Potential stressors/key climate factor and vulnerability
Northeast/North Central	Maple	Insect defoliators/warm, dry weather <i>Armillaria</i> (root decay)/drought Deep root freezing/lack of winter snow cover <b>Vulnerability:</b> High potential for damage with warmer temperatures; drier conditions
Southeast	Loblolly pine	Southern pine beetle/prolonged hot, dry weather Fungus (fusiform rust)/warm, moist weather Fire (favors longleaf or shrub)/warm, dry weather Storm damage/increase in coastal storms <b>Vulnerability:</b> Potential for much warmer weather (with increase or decrease in precipitation) to reduce productivity.
Rocky Mountain/Pacific Southwest	Ponderosa pine	Borers, bark beetles/drought Fire/drought or lightning <b>Vulnerability:</b> Resistant to near-term climate change, though productivity may decrease
Pacific Northwest (Coastal)	Douglas fir	Most stressors not strongly weather related <b>Vulnerability:</b> Resistant to near-term climate change, though productivity may decrease
Alaska	Spruce	Spruce beetle/warm weather (speeds insect development), moisture problems, erratic freezes <b>Vulnerability:</b> High potential for rapid effects because climate plays pervasive role.

From Smith, 1993.

flood frequency. In addition, water column temperature is a key determinant of habitat quality. Stream temperature, for example, is a fundamental habitat criterion for all salmonids. For example, the habitat of brook trout in the southern Appalachian Mountains is already fragmented by natural and anthropogenic forces. Climate warming would directly and indirectly fragment and limit brook trout range even more via temperature change and hydrologic cycle change. Flebbe (1997) used geographic information system techniques to evaluate warming effects on brook trout distributions. Both suitable area and stream length for these fish were seen to decrease as suitable habitat is increasingly restricted to mountaintops.

### III. PERTURBATION OF THE NITROGEN BIOGEOCHEMICAL CYCLE

The atmosphere is 78 percent by volume elemental nitrogen ( $N_2$ ). The  $N_2$  molecule is extremely stable and only very specialized pathways allow this nitrogen to be made available to the biota. Elemental nitrogen is

incorporated into chemically bound forms (fixed) by biochemical processes mediated by microbes. High-energy lightning discharges can produce nitrogen oxides. The nitrogen incorporated in the biota is mineralized to the inorganic form during the decomposition of biomass. The production of gaseous  $N_2$  and  $N_2O$  by microorganisms and the evolution of these gases to the atmosphere completes the nitrogen cycle through a process called denitrification (Fig. 2). While elemental nitrogen and elemental oxygen ( $O_2$ ) do not react at ambient temperature and pressure conditions in the troposphere, they do react at the elevated temperatures of combustion processes. As a result, all anthropogenic activities involving combustion, from industrial and power generation combustion to automobile and truck engine combustion, function to inject large amounts of nitrogen oxides ( $NO$  and  $NO_2$ ) to the lower atmosphere. These oxides are extremely important pollutants and have substantial potential significance for biological diversity. Nitrogen oxides present risks to the biota directly in the form of acid rain and nitrogen saturation and indirectly as they act as precursors for the formation, along with hydrocarbons, of ozone in the troposphere.

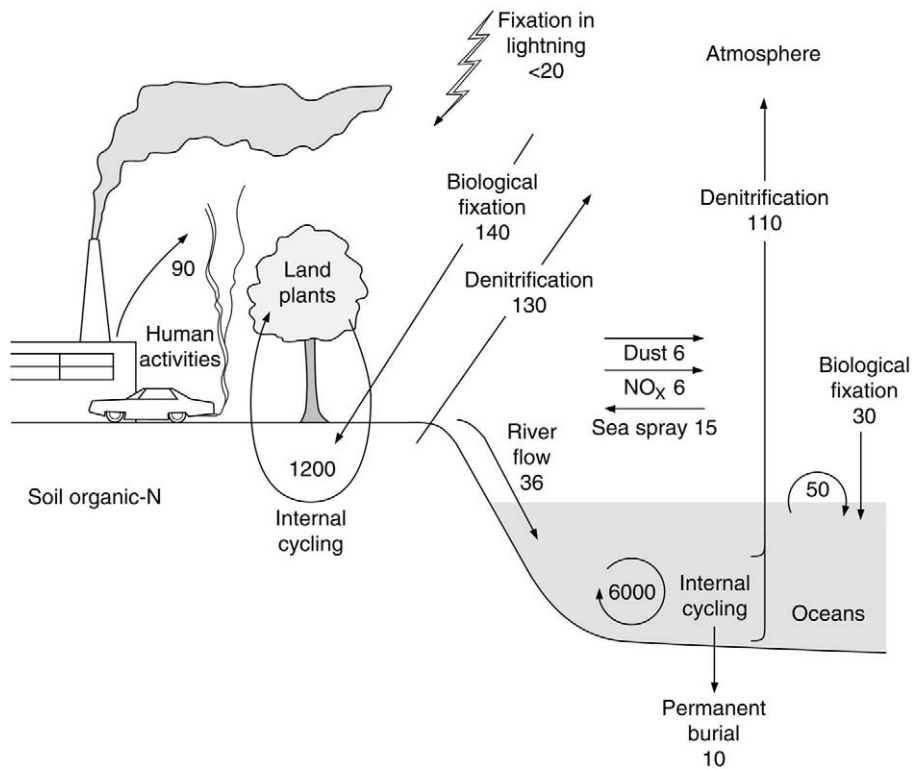


FIGURE 2 The global nitrogen cycle. The storage (numbers without arrows) and the annual flows (numbers with arrows) are shown for the global system. The units are  $10^{12}$ g nitrogen. From Schelesinger (1992).

### A. Acid Rain

Acid rain is appropriately described as an old environmental problem with a new image. Acid rain, more than any other environmental contaminant, has focused societal concern on ecosystem toxicology. Natural rain, including precipitation in relatively clean or unpolluted regions, is naturally acidic, with a pH in the range of 5.0 to 6.0. This natural acidity results from the oxidation of carbon oxides and the subsequent formation of carbonic acid. Formic and acetic acids, originating primarily from natural sources, may also contribute minor amounts of acidity to precipitation.

In regions downwind from electric-generating power stations employing fossil fuels, industrial regions, or major urban centers, precipitation can be acidified below pH 5.0. Precipitation with a pH less than 5.0 is designated acid rain. This human-caused acidification of precipitation results primarily from the release of sulfur dioxide ( $\text{SO}_2$ ) and nitrogen oxides ( $\text{NO}$  and  $\text{NO}_2$ ) from smokestacks and tailpipes. The sulfur and nitrogen oxides are subsequently oxidized to sulfate ( $\text{SO}_4^{2-}$ ) and nitrate ( $\text{NO}_3^-$ ), hydrolyzed, and returned to earth

as sulfuric ( $\text{H}_2\text{SO}_4$ ) and nitric ( $\text{HNO}_3$ ) acids. This additional acidification of precipitation by human activities can readily reduce the pH of precipitation in downwind regions to between 4.0 and 5.0 on an annual average basis. Individual precipitation events that have a pH in the 3.0 to 4.0 range are not uncommon.

The atmosphere deposits acidity onto the landscape both during and in between precipitation events. In the latter case, termed dry deposition in contrast to wet deposition, the acids are delivered in the gas phase or in association with fine particles (aerosols). Acid deposition is a term that includes acid delivery in the form of precipitation (rain, snow, fog, and cloud moisture) plus dry deposition. In view of the importance of both wet and dry deposition in acid transfer from the atmosphere to the biosphere, acid deposition is a much more appropriate descriptor than acid rain.

### B. Aquatic Ecosystems

The pH range for undisturbed lakes and streams with full compliments of native biota, especially selected finfish species important to humans, typically is within

the range of 6 to 8. Lakes and streams with a pH <6 do occur naturally. Land use practices, for example, conversion from agriculture to forest in the watershed of a lake can, over time, also cause lake acidity to increase. Aquatic resources acidified from the input of acid rain are also well documented, particularly in the northern United States, eastern Canada, and Scandinavia.

Fortunately, most surface waters in the United States have sufficient buffering (acid-neutralizing) capacity to be resistant to acidification via atmospheric deposition. In the instance of high acid deposition inputs and poorly buffered aquatic systems, however, declines in water column pH can be caused by acid rain. In the mid-1980s a National Surface Water Survey was conducted in areas of the United States known to contain lakes and streams with little capacity to neutralize acids. This survey identified the Adirondaks and mid-Atlantic Highlands region as having both sensitive aquatic resources and high levels of acid deposition.

The adverse response of aquatic biota to acidification is very well documented (NAPAP, 1991). Sensitive species may be stressed or lost at small increases in acidity.

Acid-sensitive mayfly and stonefly species can be impacted at pH levels near 6.0 and sensitive finfish species, for example, fathead minnows, can be lost at pH 5.6 to 5.9. Acid sensitive phytoplankton, zooplankton, and benthic macroinvertebrate species may also be lost. Acid-tolerant species may increase significantly with acidification. Survival of fish in acidic waters is primarily related to pH level, inorganic chemistry of aluminum, and the concentration of calcium. Other relevant considerations include food web effects, spatial and temporal variation in exposure to acidity and aluminum at various life stages, and behavioral avoidance. Critical pH levels for selected aquatic organisms are presented in Fig. 3.

### C. Forest Ecosystems

The adverse impact of acid deposition on forest systems is focused on the montane forests of eastern North America. Prevailing wind patterns and the importance of cloud moisture delivery of acidic materials places these high-elevation forests at special risk. Both morbidity and mortality of large numbers of forest trees are

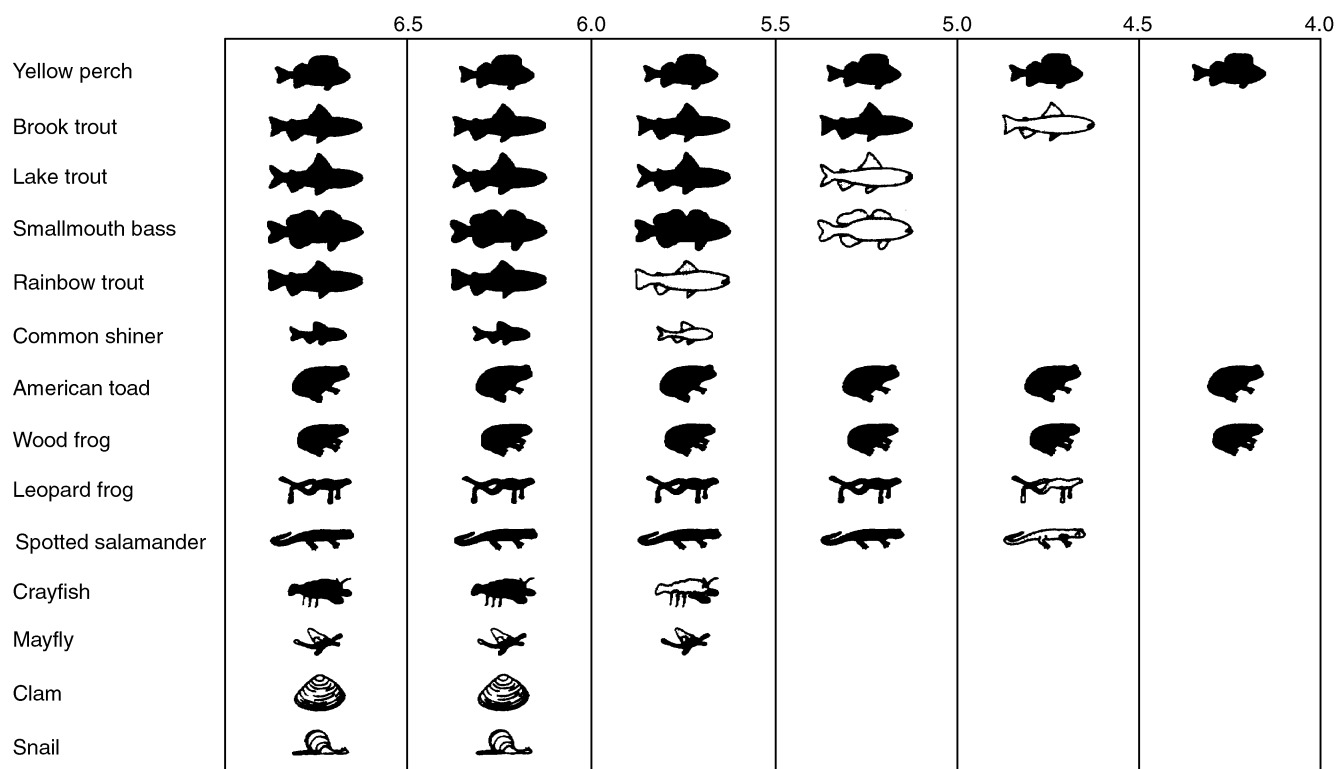


FIGURE 3 Critical pH levels for selected aquatic organisms where the American toad, wood frog, leopard frog, and spotted salamander are shown in the embryonic stage, and the crayfish, mayfly, clam and snail are selected species. From NAPAP (1991).

easily documented in montane forests. Acid deposition may be directly or indirectly involved in this symptomatology. The primary mechanisms of proposed effects are foliar weathering, foliar and soil nutrient leaching, aluminum and heavy metal soil toxicity, and predisposition to greater microbial disease and insect activity (Smith 1989, 1990).

#### D. Nitrogen Saturation

For forests in temperate latitudes, the input of nitrogen via atmospheric deposition may approximate 0.5 kg N ha/yr in clean atmospheric conditions. This low input combined with acidic organic soils in cool climates with high lignin content and slow rates of decomposition and, therefore, slow internal nitrogen turnover cause most forest systems to be nitrogen deficient. Numerous forests provided with supplemental nitrogen exhibit enhanced growth rates indicating that they were nitrogen limited prior to fertilization. In forest locations with less than clean atmospheric conditions (acid rain), forests may receive up to 27 kg N ha/yr if both dry and wet deposition sources are inventoried. At elevated levels of atmospheric input and normal dynamics of internal nitrogen cycling, forests can be fertilized by nitrogen in acid rain and exhibit increased growth and productivity.

If high levels of nitrogen deposition persist, however, nitrogen availability may reach a level that exceeds the vegetative and microbial nitrogen demand and cause the system to enter a condition designated nitrogen saturated. If sustained, evidence has been presented to indicate that nitrogen saturation can be detrimental to forest systems and lead to decreases in foliar biomass, imbalances in foliar nutrient concentrations, enhanced nitrification, and possibly tree morbidity and mortality. Nitrogen saturation of forest soils may lead to high levels of nitrate in stream waters of forest catchments. Excess nitrogen loading of surface waters can impose stresses on aquatic species.

#### E. Nitrogen Oxides as Tropospheric Ozone Precursor

In the presence of sunlight nitrogen dioxide is dissociated and forms equal numbers of nitric oxide molecules and oxygen atoms. The oxygen atoms rapidly combine with molecular oxygen to form ozone. This ozone then reacts with the nitric oxide, on a one-to-one basis, to reform nitrogen dioxide. The steady-state concentration of ozone that is produced by this cycle is very small.

When hydrocarbons, aldehydes, or other reactive atmospheric constituents are present, however, they can form peroxy radicals that oxidize the nitric oxide back to nitrogen dioxide. With reduced nitric oxide available to react with the ozone, the latter may accumulate to relatively high concentrations.

High concentrations of ozone are toxic to vegetation. Morbidity and mortality of agricultural and forest plants caused by ozone are well documented. California provides numerous examples of ozone stress on wildland ecosystems. In southern California the predominant native shrub land vegetation consists of chaparral and coastal sage scrub. The former occupies upper elevations of the coastal mountains, extending into the North Coast ranges, east to central Arizona, and south to Baja California; the latter occupies lower elevations on the coastal and interior sides of the coast ranges from San Francisco to Baja California. Westman (1979) applied standard plant ordination techniques to these shrub communities to examine the influence of air pollution. The reduced cover of native species of coastal sage scrub documented on some sites was statistically indicated to be caused by elevated atmospheric ozone. Sites of high ambient ozone were also characterized by declining species richness.

Ponderosa pine is one of five major species of the "mixed conifer type" that covers wide areas of the western Sierra Nevada and the mountain ranges, including the San Bernardino Mountains, in southern California from 1000 to 2000 m elevation. Other species represented include sugar pine, white fir, incense cedar, and California black oak. The response of these five major tree species to ozone in the San Bernardino National Forest is variable as shown by field surveys and seedling exposures. Ponderosa pine exhibits the most severe foliar response to elevated ambient ozone. An aerial survey conducted by the USDA Forest Service indicated 1.3 million ponderosa (or Jeffrey) pines on more than 405 km<sup>2</sup> (100,000 acres) were stressed to some degree. Mortality of ponderosa pine has been extensive. White fir has suffered slight damage, but scattered trees have exhibited severe symptoms. Sugar pine, incense cedar, and black oak have exhibited only slight foliar damage from oxidant exposure. A 233 ha study block was delineated in the northwest section of the San Bernardino National Forest in order to conduct an intensive inventory of vegetation present in various size classes and to evaluate the healthfulness of the forest. Ponderosa pines in the 30 cm diameter class or larger were more numerous than other species of comparable size in the study area. These pines were most abundant on the more exposed ridge crest sites of the sample area. Mortality

of ponderosa pine ranged from 8 to 10 percent. The loss of a dominant species in a forest ecosystem clearly exerts profound change in that system. Miller (1989) concluded from his investigations that the lower two-thirds of the study area will probably shift to a greater proportion of white fir. It was judged that incense cedar will probably remain secondary to white fir. Sugar pine was presumed to be restricted by lesser competitive ability and dwarf mistletoe infection. The rate of composition change was deemed dependent on the rate of ponderosa pine mortality. The upper one-third of the study area, characterized as more environmentally severe due to climatic and edaphic stress, supports less vigorous white fir growth. Following the loss of ponderosa pine in this area, sugar and incense cedar may assume greater importance. Miller judged, however, that the natural regeneration of the latter species may be restricted in the more barren, dry sites characteristic of the upper ridge area. California black oak and shrub species may become more abundant in these disturbed areas. Generally, the data support the hypothesis that forest succession toward more tolerant species such as white fir and incense cedar occurs in the absence of fire. In the presence of fire, pine may be favored by

seedbed preparation and elimination of competing species.

The changes in forest composition caused by ozone in this southern California forest have created a management concern, as well as ecological change, because the forest is intensively used as a recreational resource and the loss of ponderosa pine is judged to reduce the aesthetic qualities of the forest.

#### IV. PERTURBATION OF THE SULFUR BIOGEOCHEMICAL CYCLE

The natural sulfur cycle is complex in that it involves several gaseous species, poorly soluble minerals, and several species in solution (Fig. 4). Sulfur enters the atmosphere from natural sources as hydrogen sulfide ( $\text{H}_2\text{S}$ ) from active volcanoes and the decay of organic matter in anaerobic environments (swamps, tidal flats), sulfur dioxide ( $\text{SO}_2$ ) from active volcanoes, and particles of sulfate salts (e.g. ammonium sulfate) from sea spray. Approximately one-third of all sulfur compounds and 99 percent of the  $\text{SO}_2$  reaching the troposphere comes

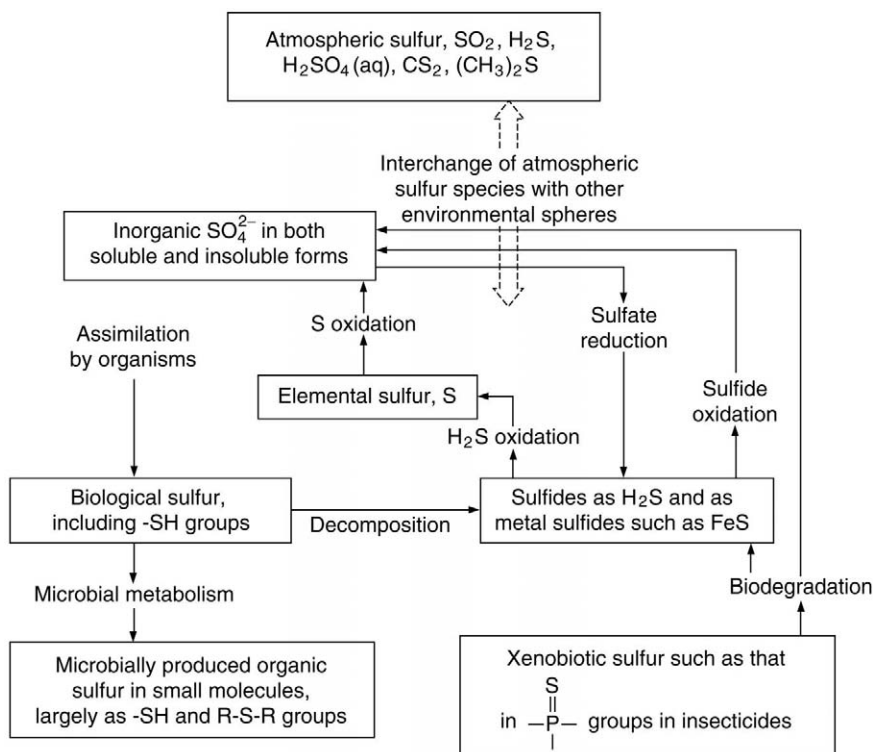


FIGURE 4 The global sulfur cycle. From Manahan (1994).

from human combustion activities. The combustion of coal and oil for the production of electricity and the smelting of metal-bearing ores have historically been major sources of SO<sub>2</sub> to the atmosphere.

The industrial and energy sector release of SO<sub>2</sub> over multiple decade timescales has reduced biological diversity in local environments surrounding point source facilities in many parts of the world.

In North America, for example, we can cite the cases of Ducktown, Tennessee, and Sudbury and Wawa, Ontario. The Ducktown, Tennessee, area consists of approximately 243 km<sup>2</sup> (60,000 acres) and was originally covered with southern deciduous forest. Mining operations were initiated in the basin in 1850 and smelting operations were most active between 1890 and 1895. By 1910 gross forest simplification resulting from excessive sulfur dioxide had created three new vegetative zones surrounding Ducktown. In a 27 km<sup>2</sup> (10.5 mile<sup>2</sup>) area closest to the source vegetation was devastated and largely eliminated. All trees and shrubs were destroyed, and only a few, isolated islands of sedge grass occurred in the outer portions of this zone. A belt of grassland ecosystem, 68 km<sup>2</sup> (17,000 acres) in size, surrounded the barren zone. The principal grassland species was broomsedge. A transition zone of somewhat indefinite boundary and consisting of approximately 120 km<sup>2</sup> (30,000 acres) was located beyond the grassland. Few trees were located along the inner edge of the transition zone. Sassafras, red maple, sourwood, and post oak were common in the middle of the transition forest. The uninfluenced forest beyond the impact of the smelter consisted principally of mixed oaks, hickory, dogwood, sourwood, black tupelo, and some eastern white pine. The distance of vegetative impact extended 19 to 24 km (12 to 15 miles) to the north and approximately 16 km (10 miles) to the west of the smelter. Eastern white pine damage was recorded 32 km (20 miles) from the industry.

Sheet and gully soil erosion has been excessive in the acutely damaged inner zone. Micrometeorological changes in the inner zone relative to the surrounding forest have been substantial; summer air temperature averages are 1 to 2°C higher, while the winter air temperature averages 0.3 1°C lower, the soil temperature is 11°C higher in the summer; the wind velocities are 5 to 15 times higher and rainfall is consistently lower.

Reforestation of Copper Basin has not been rapid due to severe erosion, low soil nutrients and moisture, and high winds. Recent reforestation efforts have employed both pine and hardwood species (Smith, 1990).

In the Sudbury area, a century of sulfur dioxide fumigation, copper and nickel particulate deposition, fire, soil erosion, and increased frost action have interacted to create 10,000 ha (25,000 acres) of barren land and 36,000 ha (89,000 acres) of stunted, open birch-maple woodland. Three large nickel and copper smelters have historically discharged several thousand tons of sulfur dioxide daily into the surrounding atmosphere. Sulfur dioxide emissions from this area have historically approximated 10% of the North American sulfur dioxide total and 25% of the smelter total. Extensive simplification of the mixed boreal forest ecosystem surrounding this region has occurred primarily via the mortality of eastern white pine throughout a 1865 km<sup>2</sup> (720 square mile) area to the northeast of the Falconbridge, Copper Cliff, and Coniston smelters. Acute impact on balsam has been recorded in excess of 40 km (25 miles) from the source of sulfur dioxide. Red oak, red maple, and red-berried elder are more tolerant and may exist in disturbed forests as close as 1.6 km (1 mile) to the smelters. Morbidity and mortality of forest trees in the Sudbury region continued to spread at least until the construction of the world's tallest smokestack of 403 m (1250 feet) at Copper Cliff in 1972 and the closing of the Coniston smelter also in 1972.

Soil erosion has followed the destruction of surrounding forest ecosystems. Rainfall has been made highly acidic, commonly less than pH 3.0 in the early 1970s. Elevated nickel and copper concentrations in soils have been recorded to distances of 50 km (31 miles). Eroded sediment, contaminated with trace metals, has contaminated area lakes and water resources. This contamination, along with acidification of surface waters, has resulted in numerous aquatic species extirpations.

Forest destruction in the vicinity of an iron smelter in Wawa, northern Ontario, exhibits a more discrete pattern, one smelter relative to three, and a shorter history of impact resulting in less equilibration of vegetative response than in Sudbury. The Wawa smelter initiated operations in 1939 (significantly expanded in 1949) and has released as much as 100,000 tons of sulfur dioxide annually to the surrounding atmosphere. Negative impact is primarily confined to a strip northeast from the plant in the direction of the prevailing wind. Symptoms of sulfur dioxide damage may be observed for at least 32 km (20 miles) to the northeast. The mixed boreal forest in the Wawa area consists mainly of white spruce, black spruce, balsam fir, jack pine, white cedar, larch, and white pine in the dominant layer. Mountain maple and *Pyrus decora* are common in the understory.

## V. TRACE POLLUTANTS OF GLOBAL IMPORTANCE

Persistent organochlorine compounds, including the polychlorinated biphenyls (PCBs) and organochlorine pesticides, are among the most important anthropogenic contaminants with global status. These pollutants biomagnify in food webs and have resulted in acute and subtle ecotoxicological effects in aquatic and terrestrial organisms. Despite increased regulation and reduced use, urban and industrial areas remain a significant source of PCB pollutants. The insecticide DDT has seen reduced or terminated use in Canada, the United States, and western Europe, but increased use in Mexico, Central America, India, and eastern Europe. Both PCBs and DDT (including its breakdown products DDD and DDE) can enter the atmosphere via volatilization, or in combination with aerosols, and be transferred from the atmosphere to the biosphere via both dry and wet deposition. Long-distance atmospheric transport and subsequent deposition have been observed to contaminate numerous ecosystems remote from primary sources including estuarine and lacustrine systems. Less research has been directed to documenting contamination of terrestrial systems, especially forests.

Northern and boreal forests develop a thick organic forest floor and in recognition of the tendency of trace organic pollutants to bind to the surface (adsorb) or penetrate (absorb) into soil particles, it is especially important to examine these organic soils for trace pollutants. Such an examination of forest floors of remote northern New England hardwood and montane forest soils has revealed the presence of numerous PCBs and DDT and its breakdown products (Smith *et al.*, 1993).

While the soil and water column concentrations of these pollutants are relatively low, their low water solubility but high lipid solubility allow the processes of bioaccumulation and bioconcentration to elevate these potential toxics to levels sufficient for adverse biological effect (Allen-Gil *et al.*, 1997a, 1997b). These elevated concentrations can cause change in community composition (Ferraro and Cole, 1997).

## VI. CONCLUSION

Human beings have—through their efforts to generate electricity, manufacture things, and transport themselves and their goods—managed to alter major biogeochemical cycles of the earth. In addition, the manufacture and use of a variety of environmentally persistent

chemicals have resulted in the global circulation of materials with potentially biologically toxic properties.

Perturbation of the carbon cycle via increased combustion and deforestation and the resulting global warming presents a high risk to biological diversity. Continued global warming will result in species extirpations and possible extinctions as we progress to the middle of the twenty-first century. Changes in the nitrogen cycle caused by human activities are less impressive than the alteration of the carbon cycle but still have the extended-term potential to change both terrestrial and aquatic ecosystems and to result in species losses. Changes in the sulfur cycle, while substantially less widespread than alterations in the carbon and nitrogen cycles, have proven historically to have been sufficient to cause numerous local extirpations in the vicinity of point sources.

The atmosphere, while highly structured (layered) horizontally, has no vertical partitions. As a result, the injection of chemically persistent toxins anywhere in the world means they may begin a journey of global circulation and eventual deposition. Accumulation of these toxins via environmental media or food-web exposure can place certain species at risk of loss from receiving environments.

Pollutants of global importance and the ability of humans to alter fundamental biogeochemical cycles must be recognized as important mechanisms for the loss of biological diversity. Efforts to monitor the trends of both pollutants and major chemical cycles along with species at risk must be made a high priority for contemporary environmental management. The loss of species and communities must be recognized as a primary reason for increased international efforts to reduce the production and release of pollutants.

### See Also the Following Articles

ACID RAIN AND DEPOSITION • CARBON CYCLE • CLIMATE CHANGE AND ECOLOGY, SYNERGISM OF • GREENHOUSE EFFECT • NITROGEN AND NITROGEN CYCLE

### Bibliography

- Allen-Gil, S. M., Gubala, G. P., Wilson, R., Landers, D. H., Wade, T. L., Sericano, J. L., and Curtis, L. R. (1997a). Organochlorine pesticides and polychlorinated biphenyls (DCBS) in sediments and biota from four U.S. Arctic lakes. *Archives of Environmental Contamination and Toxicology* 33, 378–387.
- Allen-Gil, S. M., Landers, D. H., Wade, T. L., Sericano, J. L., Lasorsa, B. K., Crecelius, E. A., and Curtis, L. R. (1997b). Heavy metal, organochlorine pesticide and polychlorinated biphenyl contamination in Arctic ground squirrels (*Spermophilus parryi*) in northern Alaska. *Arctic* 50, 323–333.

- Barker, J. R., and Tingey, D. T. (Eds.) (1992). *Air Pollution Effects on Biodiversity*. Van Nostrand Reinhold, New York.
- Ferraro, S. P., and Cole, F. A. (1997). Effects of DDT sediment-contamination on macrofaunal community structure and composition in San Francisco Bay. *Marine Biology* 130, 323–334.
- Flebbe, P. A. (1997). Global climate change and fragmentation of native brook trout distribution in the southern Appalachian Mountains. In *Wild Trout VI: Putting the Native Back in Wild Trout* (R. E. Gresswell, P. Dwyer, and R. H. Hamre, Eds.). Proceedings of the 6th Wild Trout Conference, August 17–20, 1997, Bozeman, MT.
- Intergovernmental Panel on Climate Change (IPCC). (1996). *Climate Change 1995. Technical Summary of the Working Group I Report*. Cambridge University Press, Cambridge.
- Manahan, S. E. (1994). *Environmental Chemistry*, 6<sup>th</sup> ed. Lewis Publishers, Boca Raton, FL.
- Miller, P. R. (1989). Concept of forest decline in relation to western U.S. forests. In *Air Pollution's Toll on Forests and Crops* (J. J. MacKenzie and M. T. El-Asbury, Eds.). Yale University Press, New Haven, CT.
- Murphy, D. D., and Weiss, S. B. (1992). Effects of climate change on biological diversity in western North America: Species losses and mechanisms. In *Global Warming and Biological Diversity* (R. L. Peters and T. E. Lovejoy, Eds.), pp. 355–368. Yale University Press, New Haven, CT.
- National Academy of Sciences (NAS). (1987). *Responding to Changes in Sea Levels: Engineering Implications*. Marine Board, National Research Council, National Academy Press, Washington, DC.
- National Acid Precipitation Assessment Program (NAPAP). (1991). *1990 Integrated Assessment Report*. Washington, D.C.
- Nepstad, D. C., Verissimo, A., Alencar, A., Nobre, C., Lima, E., Lefebvre, P., Schlesinger, P., Potter, C., Moutinho, P., Mendoza, E., Cochrane, M., and Brooks, V. (1999). Large-scale impoverishment of Amazonian forests by logging and fire. *Nature* 398, 505–508.
- Peters, R. L., and Lovejoy, T. E., Eds. (1992). *Global Warming and Biological Diversity*. Yale University Press, New Haven, CT.
- Pounds, J. A., Fogden, M. P. L., and Campbell, J. H. (1999). Biological response to climate change on a tropical mountain. *Nature* 398, 611–615.
- Raven, P. H., and Wilson, E. O. (1992). A fifty-year plan for biodiversity surveys. *Science* 258, 1099–1100.
- Ray, G. C., Hayden, B. P., Bulger, A. J., Jr., and McCormick-Ray, M. G. (1992). Effects of global warming on the biodiversity of coastal-marine zones. In *Global Warming and Biological Diversity* (R. L. Peters and T. E. Lovejoy, Eds.), pp. 91–104. Yale University Press, New Haven, CT.
- Schelesinger, W. H. (1992). *Biogeochemistry – An Analysis of Global Change*. Academic Press, New York.
- Smith, W. H. (1989). Effects of acidic precipitation on forest ecosystems in North America. In *Acidic Precipitation*, Vol. II (D. C. Adriano and A. H. Johnson, Eds.), pp. 165–188. Springer-Verlag, New York.
- Smith, W. H. (1990). *Air Pollution and Forests*, 2<sup>nd</sup> ed. Springer-Verlag, New York.
- Smith, W. H. (1993). Forests. In *Preparing for an Uncertain Climate*, Vol. II, Chapter 6. Publication No. OTA-0–567. U. S. Congress, Office of Technology Assessment. U.S. Government Printing Office, Washington, D.C.
- Smith, W. H., Hale, R. C., Greaves, J., and Huggett, R. J. (1993). Trace organochlorine contamination of the forest floor of the White Mountain National Forest, New Hampshire. *Environmental Science Technology* 27, 2244–2246.
- Still, C. J., Foster, P. N., and Schneider, S. H. (1999). Simulating the effects of climate change on tropical montane cloud forests. *Nature* 398, 608–610.
- Westman, W. E. (1979). Oxidant effects on Californian coastal sage scrub. *Science* 205, 1001–03.
- Wuebbles, D. J., and Edwards, J. (1991). *Primer on Greenhouse Gases*. Lewis Publishers, Chelsea, MI.







# POPULATION DENSITY

Brian H. McArdle  
*University of Auckland*

---

- I. Measuring Population Density
  - II. Spatial Patterns
  - III. Determinants of Spatial Density Variation
  - IV. Consequences of Spatial Density Variation
  - V. Temporal Patterns in Density
  - VI. Determinants of Temporal Density Variation
  - VII. Consequences of Temporal Density Variation
  - VIII. Transspecific Patterns
  - IX. Conclusions
- 

## GLOSSARY

**autocorrelation** Densities from adjacent samples are likely to be more similar than are ones far apart.

**compound Poisson distribution** A family of probability distributions of numbers per area where some areas have a higher expected density than others.

**Poisson distribution** The probability distribution of numbers per area if the organisms are distributed completely at random in space.

---

**POPULATION DENSITY** can be simply defined as the number of organisms per unit area/volume. Its importance to ecologists is that it relates directly to population size (multiply by the total area/volume and one obtains total numbers). Its importance to organisms is that it defines (to some extent) the number of individuals with

whom an individual can (or has to) interact. However, despite such a simple definition, it remains a difficult concept. This is clearly evident when we consider why people measure it.

## I. MEASURING POPULATION DENSITY

Generally (although inevitably not exclusively), population density of a species of organism is measured either as a feature of a natural (or manipulated) system that responds to the environment or as a predictor (natural or manipulated) or causal agent of some response of the ecosystem.

### A. As a Response Variable

Population density is often used as a simple relative measure of how an organism responds to local conditions. If conditions are not good for the species, the density will be low (organisms will have died or moved out of the sampled area), whereas if conditions are good the density will be high (organisms will have reproduced and/or immigrated into the area). In this way, changes in density can provide insight into the natural history of the preferences and tolerances of individuals of the species. Of course, if the species is regulated by density-dependent processes (e.g., mortality or emigration) then the relationship of density with the attractiveness of the environment can be obscured. Even though the environment changes in a positive way, there may be no increase in density.

Sometimes, density can be used as an explicit proxy for population size, which of course is what many ecologists want to know about. This is particularly true in applied ecology (e.g., conservation and fisheries science). Unfortunately, the link between population density and population size is not always direct. Therefore, definitions of rarity that use either population density or species range are likely to be misleading compared with a definition that uses the product of the two.

The main problem lies in defining the area to be sampled. If it includes the entire population of interest, then the density multiplied by the area gives total population size. However, if the area does not include the whole population then this simple calculation does not work. However, it will perhaps give a relative measure of the population so that changes in the population size will be reflected in changes in the density. Unfortunately, density-dependent processes can weaken this link. If, as the population increased in size, the population was unable to expand the area it inhabited, then the density would increase in proportion to the size. However, if the population was able (or driven by density-dependent migration) to expand its range, then the density could remain constant while the population grew. Since range expansion appears to be common, density should only be used as a proxy for population size when the range is constrained, as on islands. In most studies, therefore, density simply gives the number of organisms present in some defined study area. This will seldom correspond to a biological population.

### B. As a Causal Agent

In many studies, changes in density are seen as causing changes in the population of the same or other species. For example, many experiments are performed each year in which the density of organisms is manipulated (or simply observed) to investigate the response of individuals in the same species (e.g., intraspecific competition) or other species (e.g., predation or interspecific competition). In such studies, the biological effect of density is generally determined by the number of direct or indirect interactions between individuals (of the same or different species). Density in this case is often acting as a proxy for the number or probability of interactions. There are problems associated with this use of density. In particular, the area has to be defined carefully. If the area for which the density is being estimated includes habitat unsuitable to the study species, then the zeros measured from such sites are not due to sampling error (sampling zeros) but reflect a genuine inability of the species to live there (structural zeros). Includ-

ing structural zeros in the density estimate would lead to an underestimation of the density that individuals of the same species actually experience. Of course, if it is the response of other species that is of interest, then these areas of habitat may be relevant (other areas may not). Again, the purpose for which the density is to be used determines how one defines the area over which it is to be measured.

There are additional problems. It is well-known that even if one defines the extent of the study area, estimating the number of organisms can be difficult. Perhaps less well appreciated is that natural surfaces can be generally considered as fractal (or at least approximately so). The effective area of a given area of the earth's surface experienced by an elephant is considerably less than that experienced by a mite. The area used to calculate the density of the organism is nearly always anthropocentrically biased. The more similar the organism is to humans with regard to size and habit, the more accurate the estimate of effective area is likely to be, and therefore the more relevant the estimate of density.

Of course, a simple estimate of density can still be used as a relative measure of the effective density. However, it must be remembered that organisms are seldom (if ever) uniformly distributed over a study site. Typically, some parts of the site have a high density and others a low density. If most of the organisms are located in one small part of the site, the average density for the site will reveal little about the density most experienced. The marked nonlinearity of most ecological dynamics means that inferences based on the average density are unlikely to be particularly relevant. The situation is similar though more complex when one is interested in the density of one species experienced by another (e.g., in predation or interspecific competition studies).

Population density is therefore nearly always used as an approximate proxy for the real abundance measure of interest. The degree to which it can be regarded as relevant in a particular study depends to a large extent on issues of scale.

## II. SPATIAL PATTERNS

Within a species' range, we can expect the density to vary with the habitat from high (optimal) to low (suboptimal). The effect of this on the ground depends on the scale at which one is sampling. For example, at the largest scale it has often been noted that the density of a species tends to be highest in the core of its range and lowest at the periphery. However, much local variation can still exist within the species range at a number

of scales down to that of the individual organism (spatial pattern often depends on spatial scale). Similarly, the range and regions of optimal conditions may be temporally variable. For small animals such as insects, the vagaries of wind and weather can lead to short-lived increases in different geographic regions at different times (Fig. 1). Other species (e.g., habitat specialists) tend to be far more predictable. Of course, temporal stability can be dependent on the scale: The spatial pattern might be relatively stable at one scale (e.g., geographic region) but not at another (e.g., locally within habitat patches).

It is worth noting that although we may often refer to the density as patchy, the pattern of variation is often more continuous. The pattern is better shown with contour maps rather than with a tiling or mosaic. The edges of so-called patches may not be real discontinuities but rather areas of lower density. The density within the patches may not be constant. Therefore, although we may often talk in terms of patches, we must think in terms of contours. Of course, thinking in terms of patches is an implicit acknowledgment that the presence/absence data are often more reliable and easily interpreted than abundance or density data. This is because the organism counts produced by most sampling programs are subject to considerable statistical sampling error.

### A. Statistical Patterns in Spatial Variation

One of the most common ways of investigating population density is to count organisms in small representative portions (e.g., quadrats) of a larger study area. Then, one estimates the mean density or total numbers in the study site. One can also deduce many characteristics by examining the frequency distribution of the counts per sampling unit.

The most obvious feature of real data is that the frequency distribution of these counts is long tailed and heavily skewed. (Fig. 2): Low densities are common, but very high values sometimes occur. As a general rule, a relatively small number of sampling units contain most of the organisms, and many of the sample units are empty. This indicates a fundamental difference between total number of organisms and the typical density experienced by members of another species (e.g., a predator or competitor) or the typical density experienced by members of the population.

Thus, the total number of animals in the whole area is estimated by the arithmetic mean  $\times$  total area. However, the typical number of animals found in a quadrat, i.e., the density of the species most commonly experi-

enced by other species (e.g., predators) searching at the same scale as the quadrats under study, is given by the median number per quadrat. This is approximated by the geometric mean,  $\text{antilog}[\text{mean}(\log[\text{count}])]$ , or more commonly by the pseudogeometric mean,  $\text{antilog}[\text{mean}(\log[\text{count} + 1])]$ . The number 1 is added to each count to overcome problems with zeros. However, the density most commonly experienced by members of the target species is neither of these. Since most of the animals are in a few quadrats, most experience very high densities. Thus, the best estimate of the effective conspecific density, the average density experienced by a member of the population at the spatial scale of the quadrat, is

$$\frac{\sum N_i^2}{\sum N_i}$$

which gives a density that is very different from the other two average densities.

Statistically, there are many distributional models that could fit these kinds of data. The most commonly used are members of the compound Poisson family (e.g., negative binomial and Poisson log-normal). That these long-tailed distributions fit in no way suggests that the mathematical processes that generate them have anything to do with the biological processes that generated the numbers. However, there are a few interesting ideas that can be derived from these distributions that may have biological relevance in some situations, and which when combined provide some understanding of one of the few ecological laws discovered to date (discussed later).

The simplest way of viewing these models suggests that the expected density of organisms at any point in space derives from some long-tailed continuous distribution (a gamma for the negative binomial, and a log-normal for the Poisson log-normal), but that the number actually recorded is from a Poisson with that expected value. There are many ways of parameterizing these distributions, but perhaps the most useful is with the mean ( $m$ ) and shape parameter ( $k$ ). The shape parameter defines the relationship between the variance and the mean. The relationship is often displayed as:

$$\sigma^2 = \mu + (1/k)\mu^2,$$

indicating that the variance of the counts increases approximately as the square of the mean, with the rate determined by the shape parameter. A small  $1/k$  leads to small variances for a given mean, whereas a large  $1/k$  leads to large variances. Thus, large  $1/k$  values suggest high aggregation (some quadrats containing high

1969

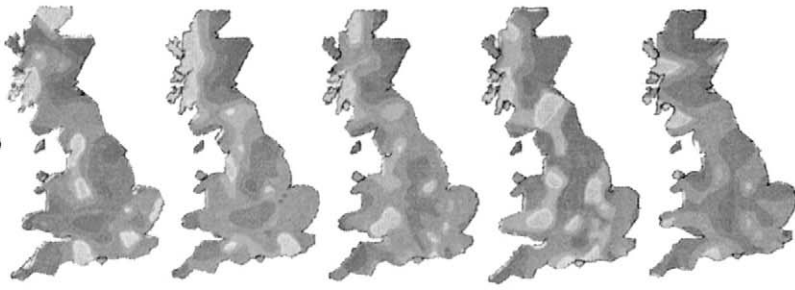
1970

1971

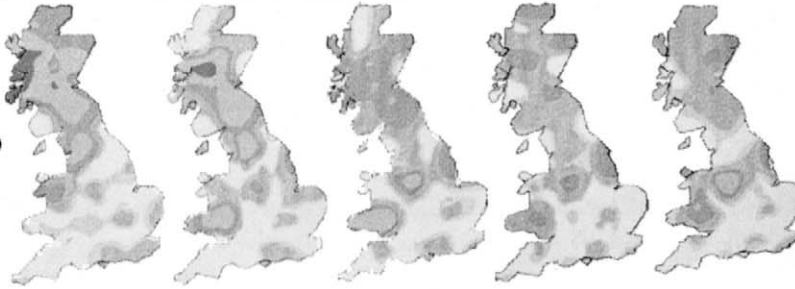
1972

1973

(a)



(b)



(c)



(d)



(e)



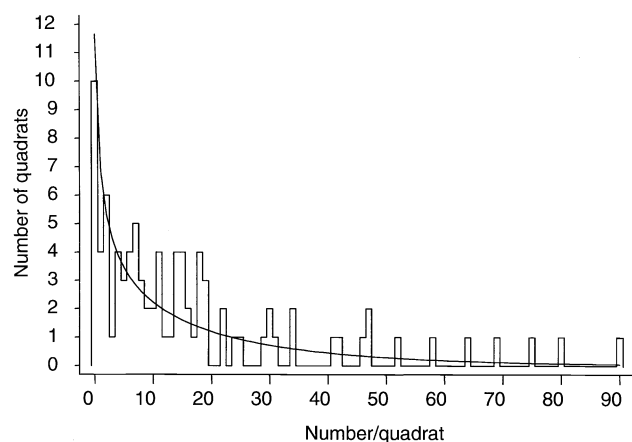


FIGURE 2 Frequency distribution of density (per 0.25 m<sup>2</sup>) of the New Zealand cockle (*Austrovenus stutchburyi*). A compound Poisson (negative binomial) distribution is fitted. Note the extremely crowded quadrats; more than 50% of the animals are in quadrats with densities of more than 30 per 0.25 m<sup>2</sup>, although the mean density is only 16.9 (median = 11). The estimated shape parameter  $k$  of the negative binomial is 0.61.

numbers of organisms and some quadrats containing low numbers), and low  $1/k$  implies low aggregation. ( $1/k = 0$  suggests a Poisson distribution for organisms whose dispersion is indistinguishable from random.) Therefore,  $1/k$  and related statistics are commonly used as aggregation indices. If the statistical basis of the compound Poisson was in fact correct (i.e., expected density at a point varying spatially, and the number actually collected a Poisson variable), then  $1/k$  separates the Poisson sampling error from the underlying distribution of expected density. Indeed, it is the coefficient of variation squared of the underlying distribution of expected density, i.e., the ratio of the variance of the gamma or log-normal with the squared mean:

$$1/k = ((\sigma^2 - \mu)/\mu^2).$$

Compound Poisson distributions have a biologically relevant property. If a population of organisms is dis-

tributed in an aggregated fashion and if some of them randomly die, the mean density changes but the shape or aggregation parameter of a compound Poisson distribution does not. This is of course as it should be: Random mortality leaves the degree of aggregation unchanged. Therefore, if the compound Poisson family of distributions is appropriate, the basic relationship between mean and variance of animal densities (counts) should be

$$\sigma^2 = \mu + \frac{1}{k}\mu^2$$

if mortality is operating randomly in space. Of course, there is no need for this to be true (indeed, it might be considered unlikely). Real data might be expected to show a general relationship similar to this, but the degree of aggregation will change with the mean density (possibly due to density-dependent processes). This is the case for virtually all sets of field data. Therefore, the compound Poisson models are not necessarily always appropriate, but they often fit the available data.

## B. Taylor's Power Law

One of the few ecological laws that have some universality was discovered in 1961 by L. R. Taylor. He worked with populations that had been sampled at many sites on many occasions. He showed that the variance of the densities over the sites on an occasion when plotted against the mean density over the sites on that occasion resulted in a very tight power curve. A plot of  $\log(\text{variance})$  against  $\log(\text{mean})$  gave a generally straight line [Taylor's power plot (TPP)]. This implies that the formula is  $\sigma^2 = \alpha\mu^\beta$  (Taylor's power law). As might be expected from the compound Poisson model, the slope of this line is often close to 2.

This plot and the formula are not just applicable to organism abundances. They are commonly used by statisticians searching for transformations to remove

FIGURE 1 Maps showing the variability of regional relative density in five British moth species with very different dynamics. (a) *Xanthorhoë fluctuata* (golden carpet moth) is a widespread species, numerically fairly stable but with mobile holes in its distribution. (b) *Cerapterix graminis* (antler moth) has a distribution limited by its host plants (fine upland grasses), but within these limits the distribution of density is constantly changing. (c) *Euxoa nigricans* (golden dart moth) has a spotted distribution that is spatially and temporally chaotic. (Consider how difficult it is to define the range of species such as this.) (d) *Spilosoma lutea* (buff ermine moth) has a numerically very stable population with an unstable northern limit. (e) *Callimorpha jacobea* (cinnabar moth) is highly concentrated in suitable areas but has a highly mobile distribution (reproduced with permission from Taylor and Taylor, 1979).

the variance–mean relationship and therefore variance heterogeneity, a problem for many traditional methods of statistical inference. They try to find a transformation that gives a slope of zero to the log(variance)—log(mean) relationship.

Real data are seldom quite so obligingly simple as Taylor's original model implies. Indeed, the dominant pattern is linear in the log–log plot but curved at low densities. Alternative formulae have been suggested to accommodate this. The modified compound Poisson (or Nelder) model  $\sigma^2 = \mu + \alpha\mu^\beta$  generally performs well, as does a split straight line, although the threshold that the latter implies seems unlikely to some biologists. Since some populations are demonstrably not Poisson even at the lowest densities, alternatives such as  $\sigma^2 = \gamma\mu + \alpha\mu^\beta$  and  $\sigma^2 = \gamma + \mu + \alpha\mu^\beta$  have also been used.

Although there may be a variety of models, the basic pattern in the data is clear, but the interpretation is not. In all these models, the parameters  $\alpha$  and  $\beta$  have a simple statistical interpretation, but what they imply about the population dynamics is less clear. The parameter  $\alpha$  clearly represents the variance (and therefore the level of aggregation) when the mean density is 1. In the modified compound Poisson (or Nelder) model it represents the value of  $1/k$  when the mean is 1. If  $\beta = 2$ , then this suggests that the level of aggregation is constant at  $\alpha$  at all densities. Thus, contrary to much that has been written about the use of the TPP, it is  $\alpha$  that tells us about aggregation and not  $\beta$ . The slope of the relationship ( $\beta$ ) indicates how rapidly aggregation changes with the mean and not the level of aggregation. To make this clear, we can rewrite the modified compound Poisson model as

$$\frac{1}{k} = \frac{\sigma^2 - \mu}{\mu^2} = \alpha\mu^{\beta-2}.$$

Aggregation is therefore related to density by a power relationship (i.e., a linear relationship in a log–log plot). Thus,  $\beta - 2$  indicates how aggregation varies with the mean. Of course, if  $\beta \approx 2$  the level of aggregation arguably does not change with density. If  $\beta < 2$  then aggregation decreases as density increases. This is consistent with density-dependent migration out of higher concentrations into lower ones, evening out density variation as the overall mean density increases. If  $\beta > 2$  then the organisms are becoming more concentrated—for example, if population growth were restricted to areas of higher density with no diffusion. Of course, it would be very unwise to link any value of  $\alpha$  or  $\beta$  to a particular dynamic process. One reason is that

$\alpha$  and  $\beta$  depend on the particular context (habitat, sampling method and scale, and even the age structure of the population) of the study and not, as suggested by Taylor, on a characteristic slope  $\beta$  for a species and an intercept  $\alpha$  depending on the sampling situation. This makes it difficult to interpret TPPs with any confidence. However, with the spatial variance  $\times$  multiple times data sets originally studied by Taylor, the patterns of variation implied by  $\beta$  can help us understand the biology of the species at least for the site and scale studied.

Interpretation is much more difficult if the TPP is generated by a spatial variance  $\times$  multiple sites sampling scheme. Here, the variance is over samples (e.g., quadrats) taken within a site, but each point on the graph is given by a different site. Since habitat is likely to vary over sites, we are likely to be looking mainly at patterns of habitat variation and utilization. A  $\beta > 2$  might imply that sites with higher means had greater local concentrations of resources than sites with lower means. A  $\beta < 2$  might suggest that where the resources were abundant, rather than being locally concentrated they were distributed more uniformly. Of course, the concept of resources is a broad one that includes predator- or competitor-free space and other dynamic features of the environment.

### C. Density–Biomass Thinning Laws

In this article, population density has been defined as the number of organisms per unit area. However, there is no reason why density should not be measured using other properties of the population. Probably the most common alternative to abundance is biomass. It should therefore not be surprising that there is a relationship between the two measures of density. However, nonbiologists might be surprised by the relationship that seems to exist in plants.

#### 1. Plants: $-3/2$ Thinning Rule (Yoda's Law)

In 1963, K. Yoda and coworkers noted that if a cohort of plants were monitored over the course of a growth season, there was a clear negative relationship between the biomass and numerical density as the plants “self-thinned.” This relationship seems to be well modeled by a power relationship:

$$B = cN^{-b}$$

where  $B$  is the biomass density,  $N$  is the density, and  $c$  and  $b$  are constants. The value of  $-b$  has generally

been thought to be between  $-0.3$  and  $-0.8$  with an ideal value at  $-0.5$ . It is generally assumed to be the result of density-dependent mortality and allometric patterns of growth interacting with the way in which trees pack together in the canopy. It is called the  $-3/2$  thinning rule because it is commonly formulated in terms of  $w$ , the biomass per individual, as

$$w = cN^{-a}$$

where  $a = b - 1$ . The exponent is now between  $-1.3$  and  $-1.8$  with an ideal value of  $-1.5$ .

In recent years, its generality has come into question. There does seem to be a generally negative relationship between density and biomass in most investigated species, but there is little evidence that there is any ideal value of  $a$ , let alone one of  $-1.5$ .

## 2. Animals: $-4/3$ Thinning Law

In contrast to the empirical origin of the plants' thinning law, zoologists derived an analogous rule for animals from theory. The postulated relationship is again a power function  $w = cN^{-a}$ , where  $-a = -4/3$  rather than  $-3/2$  as in plants. To date, it has been most convincingly shown in salmonid fish but it seems unlikely that all animal species have the same slope.

## III. DETERMINANTS OF SPATIAL DENSITY VARIATION

Although the existence of spatial variation in the population density of a species is almost trivially self-evident, the processes that can produce it are many and varied. This reinforces the problems of inferring process from pattern. The mere existence of spatial variation at any given scale is seldom enough to suggest the processes that generated it. There are three main classes of processes that generate spatial variation in density.

### A. External Forcing

The most obvious (and at many scales probably the most common) cause of spatial variability in population density is variation in the quality of the available habitat. If the environment varies spatially at any scale, so will the density of many species. A patchy environment usually produces a patchy distribution of organisms. Of course, some species will be affected more than

others. There is enormous variation in the tolerance of organisms to variation in environmental factors.

Another external influence (which might be viewed as a special case of environmental variation) might be the distribution or spatial response of predators or competitors. Predator- or competitor-free space can be as significant a resource as food or space with suitable environmental conditions. Other things being equal, higher density can be expected in such areas.

### B. Lifestyle Characteristics

Of course, interactions among conspecifics can also lead to spatial variation in density. Social aggregations of many kinds are common in animals and at some scales can lead to variation in density that is independent of the environmental conditions. Conversely, random variation in density at local scales can be reduced by spacing behaviors (such as territoriality).

Some of the possible conspecific interactions are not strictly social. Many species modify the local environment, which makes it more attractive to recruits (e.g., larval settlement). This can also result in the formation of aggregations.

Other biological processes result in strong aggregations. Some species have no choice regarding where they will live (e.g., most plants). Those that start life close to their parent (e.g., many forms of vegetative reproduction, nondispersing seeds, and colonial animals) will stay there. Aggregations can hardly be avoided.

### C. Population Dynamics

Perhaps one of the most topical areas of research concerns the possibility that spatial patterns of density variation could be the result of intrinsic dynamic processes—an emergent property of the dynamic system of which the population is a part. Arguments are often taken from complexity or chaos theory and often refer to the behavior of cellular automata models (such as the Game of Life referred to in most popular texts on complexity and chaos theory).

In fact, as one might expect, the tendency of spatially distributed dynamic systems to produce a spatial pattern is related to a very simple property of the system which in ecology has a simple biological interpretation. In 1992, M. J. Phipps postulated that spatial pattern will only be produced in cellular automata models if areas neighboring high-density sites are more likely to achieve high density than other areas. He called this



the neighborhood coherence principle. In ecology, this predicts that only species that attract conspecifics either by recruitment (e.g., vegetative growth or limited dispersal from place of birth) or by movement (e.g., flocking) or that somehow reduce the death rate in areas near high densities are likely to show intrinsically generated spatial patterning that is independent of environmental forces. All the sophisticated theory results in a trivial prediction that social and other aggregative behaviors are the main reason why spatial pattern emerges in populations for which the environmental factors have little influence. Generally, the more important question is the following: Given that the biology of the animal predisposes it to aggregation, why is it not more obvious? Density-dependent forces such as predation and emigration that tend to reduce high densities will tend to prevent major concentrations. It is the balance of these ecological forces that produces the observed patterns. This is not to say that there may not be other system dynamics that can produce spatial patterning, but they will almost certainly be similarly spatially non-random in their action or somehow be related to the forcing processes described previously and, above all, they will be biologically obvious. There is nothing mysterious about the action of nonlinear dynamics in producing spatial patterns. It depends on the properties of the system and in a biological context seems generally to agree with common sense.

#### IV. CONSEQUENCES OF SPATIAL DENSITY VARIATION

##### A. Population Dynamic Consequences

The existence of spatial variation in population density implies consequences at both the individual and the population levels (and other scales in between). The most obvious is that the state of the individual is frequently influenced directly and indirectly by the surrounding density of conspecifics (intraspecific density-dependent processes). The direct effects may include behavioral changes (e.g., increased propensity for emigration), physiological changes (e.g., increased stress hormones and reduced or enhanced sexual activity), or even morphological changes (e.g., the social phase of locusts). It must be appreciated, however, that the density perceived by an individual will seldom be that measured by researchers. Again, the issue of scale is crucial. Measuring density averages counts over areas and reduces variation at smaller scales. These may, however, be the scales at which the individual's perception

operates. Local high densities determine the perceived density per individual and the resulting consequences.

Similarly, the indirect consequences of local high density can be very different than expected if one simply examines the average density at a larger scale. The impact of predators, parasites, and pathogens can be very different if the organisms are concentrated rather than uniformly distributed. Spatial variation in density can be as important as differences in overall average density.

These considerations have led to an increasing appreciation that population dynamic models should be spatially explicit. A move in this direction started with optimal foraging models, but recently entire populations have been modeled in this way. One class of such models is those that describe metapopulations. A metapopulation is a population of subpopulations. Although this is usually an oversimplified spatial pattern—all organisms are in discrete, fixed clusters or in transit between them—the behaviors are quite different from those in a simple uniform population model. Rate constants (e.g., birth or death rates) averaged over a spatially distributed population will seldom adequately reproduce the behavior of a population. It is worth noting that models that use discrete partitions of space do not behave in the same way as those based on continuous space (i.e., on differential calculus). Therefore, although ecologists are currently attempting to produce more realistic models by incorporating space, the way it is incorporated may influence the conclusions.

##### B. Genetic/Evolutionary Consequences

Although the population dynamic effects of density variation may be important at a variety of scales, the evolutionary consequences have been considered only at larger scales: A subpopulation of organisms becomes spatially isolated from the rest and allopatric speciation may then occur. Work by D. S. Wilson on multilevel selection theory has incorporated the spatial organization of organisms into groups at different scales to expand the range of evolutionary forces.

One genetic consequence of low density and high aggregation in less mobile organisms is the possibility of genetic drift and inbreeding. Such populations are possibly more vulnerable to disease and can show depression of reproduction rates if the aggregations are stable and isolated.

#### V. TEMPORAL PATTERNS IN DENSITY

When population density is plotted against time, a logarithmic scale is often used because populations can

fluctuate so widely, over many orders of magnitude, that the arithmetic scale is inadequate. Locusts are an obvious example. However, for many species the magnitude of the fluctuations depends on the temporal and spatial scale over which the density was determined. For example, insect species with a single synchronized generation per year will fluctuate enormously within the year—huge numbers of eggs and first-instar larvae decline throughout the year to much smaller numbers of adults—but might not fluctuate much between years. If one averaged or totalled density over the year, much of the variation would be smoothed out. Similarly, sampling at the same time each year would also not provide information about within-year variability. The spatial scale over which the density is averaged is also crucial. Variation in the density of starlings through time will appear very different when measured at the scale of the bird table, the garden, or the suburb.

### A. Shape of Distribution

Although the nature of the distribution of organism counts sampled in space is well-known, far less effort has gone into characterizing the distribution of counts in time. Part of the reason, of course, is the difficulty in obtaining long time series of suitable data. Also, often the data are not independent; rather, they are autocorrelated: The value at one time will be more similar to recent values than those from later times. This makes characterizing the distribution difficult. At this stage all that can be sensibly stated is that, like the spatial distributions, they are discrete and long tailed, and the members of the compound Poisson family could be used as an initial approximation. However, the autocorrelation often leads to multimodality in real, short time series.

### B. Red-Shifted Variance

One observed characteristic of such time series of density data is that, generally, the variability (measured, for example, as  $SD[\log(N)]$ ) increases with the length of the series. There are at least three phenomena which can produce this effect, and they are not mutually exclusive.

1. Long-term trends: If density is consistently increasing or decreasing over time (as commonly occurs due to human impact and climate change), then its variance will naturally increase as long as the trend continues.

2. Internally generated autocorrelation in the series:

Autocorrelation (or serial correlation) is present when the density at the next sample occasion is similar to the current value, perhaps because some of the same organisms are still living (interval between samples  $<$  generation time) or because the number living at the next time period depends on the current breeding population. In this situation, short time series cannot vary widely, but as the series becomes longer the variance gradually increases to some asymptotic level. This is of course particularly common with long-lived species that, naturally, we sample at a frequency convenient to us but not necessarily relevant to the species.

3. Externally forced autocorrelation: If the physical environment fluctuated in an autocorrelated fashion then the population's density might vary in response. Such autocorrelated time series are characteristic of "red-shifted" spectra of environmental variables (characteristic of climatic factors).

### C. Taylor's Power Law

If time series of densities are produced for many sites, then a plot of the variance over time against the mean over time for the sites will provide a close relationship, as occurs for spatial variances and means. Essentially the same models describe this relationship as those used for the spatial TPP. Temporal variance increases with the mean as does spatial variance. Again, the slope is usually approximately 2. The intercept  $\alpha$  measures the variability when the average density at a site is 1, and therefore it describes a sort of baseline temporal variability at this spatial and temporal scale. The slope  $\beta$  describes how this variance changes for sites that have higher or lower densities. A slope of 2 suggests that there is essentially no change in variability at higher densities. Values higher than 2 indicate that the species is more variable at denser sites. This is a behavior characteristic of species subject to local explosions in numbers at certain sites. A good site is sometimes good and sometimes bad. A bad site is always bad. A slope of less than 2 means that sites with generally high densities tend to be less variable than poor sites. This is consistent with density dependence, although other habitat-driven explanations are possible.

## VI. DETERMINANTS OF TEMPORAL DENSITY VARIATION

Although the processes that produce temporal variability in density are similar to those that produce spatial

variability, they are not identical. The effects of choices of scale tend to be greater. However, the same three classes of factors are discussed.

### A. External Forces

Fluctuations in the environment will drive temporal variability of density in many species. Invertebrates in particular are at the mercy of the elements, although vertebrates are demonstrably not immune. The extreme winter of 1962 and 1963 in the British Isles resulted in a marked decrease in the densities of many bird species that took years to recover. The densities of many marine organisms are driven by El Niño-like phenomena (as are many terrestrial organisms). The threshold-like effects of many environmental variables can lead to large fluctuations in organism densities. For example, a species may not breed until the temperature is above some value (which may only happen every few years) or unless rain falls (especially in desert species).

Of course, the consequences for a species of such fluctuations in the physical environment may be indirect. Food, competitors, disease, and predation may also be driven variables, in many cases amplifying or extending the direct effects. Clearly, an organism's lifestyle (food/nutrient requirements, vulnerability to disease or predation, etc.) can also influence its susceptibility to such forces.

### B. Life History Characteristics

Life history characteristics, such as synchronized generations or life stages (perhaps dormant) that are not available for sampling (e.g., the seed bank or the nymphs of the 21-year cicada), can lead to apparently major changes in density. Arguably, the variation is the result of a particular definition of density. Whether such a definition is useful will depend on the question under investigation.

Some species have reproductive strategies that amplify fluctuations in density. Physiological and/or morphological changes (usually triggered by habitat characteristics) can lead to dramatic increases in population growth, e.g., the shift from sexual to asexual reproduction when conditions are good.

An organism's mobility (or its ability to disperse propagules) can also influence its temporal variability. If it has an aggregated, patchy, spatial distribution (e.g., metapopulation structure) the degree of connectedness between the parts of the population and the degree to which they fluctuate independently can influence the extent of the variation of the whole population.

### C. Population Dynamics

By definition, density-dependent processes should be major determinants of the degree of temporal variation. In fact, the extent to which such processes influence density is largely unknown. The problem is twofold:

1. It has proved easier to study the apparent consequences rather than the process. That is, it is easier to show that changes in density over a period depend on densities prior to that period (the statistical detection of density dependence) than it is to show the ecological process that produced the relationship. However, there is little doubt that internal density-dependent processes cause many species to fluctuate less than they might otherwise.

2. The second process is more subtle. Complexity and chaos theory predict that certain combinations of high growth rate and strong density-dependent dynamics can lead to chaotic behavior in the resulting population densities. Since this can lead to apparently random fluctuating densities, it is difficult to separate the effects of density dependence from the results of external forces. Indeed, since these driving variables may be the result of chaotic processes (e.g., the weather or market forces), it is difficult to determine how such density-dependent effects can be distinguished without much more data about environmental factors than is generally available.

## VII. CONSEQUENCES OF TEMPORAL DENSITY VARIATION

### A. Increased Probability of Extinction

The most widely discussed consequence of low density is the increased probability of extinction. Although it is difficult to demonstrate empirically, it seems obvious that a population whose size (not necessarily density) varies widely is more likely to go extinct (all things being equal) than a population whose size varies less. Once a population reaches low densities within a restricted range, chance events can determine whether it has a future. If the numbers are sufficiently low a single death or failure to breed can lead to eventual extinction. It is worth reiterating that such stochastic extinctions are the result of low population sizes and not densities.

The probability of extinction can (occasionally) be influenced by low density per se. For example, regardless of the population size, at very low densities Allee effects may increase the chances of extinction. That is,

if the density of a sexually reproducing species becomes so low as to affect the probability of locating a mate, then the population can start a slide to extinction from which it cannot escape. Of course, if the population responds by aggregating, then the distinction between local and regional density becomes crucial. Ultimately, it will be the fate of the aggregations, and therefore the total number of organisms, that determines the destiny of the population.

### B. Cascade Effects

Of course, fluctuating densities of one species can have knock-on effects on other species. Patterns of density variation affect man's exploitation of the natural world. Outbreaks of pests or periods of low fish density have economic consequences, and assessing the risk of such events is a lucrative job for well-paid consultants.

Such an anthropocentric view ignores the cascade of effects that are likely to propagate up and down the food chain, for example, the changes consequent on fluctuations in the density of kelp on the west coast of North America. Such effects are usually difficult to demonstrate unequivocally.

### C. Evolutionary Consequences

Temporal variation in density can produce genetic bottlenecks with consequent loss of genetic variability. However, if the population rebounds rapidly the reduction in heterozygosity may be minimal.

## VIII. TRANSSPECIFIC PATTERNS

### A. Comparing Densities across Species

It should be apparent that both the measure and the scale of a density must be appropriate for the processes or patterns of interest. In particular, the appropriate area might not be the same for different species. If perceived density of conspecifics is required, the appropriate area would probably depend on the movement patterns of individuals of the different species and would likely differ among species. If perceived density by a predator species is required, then the appropriate area would be defined for the predator and would likely be the same for most prey species. Since the biological relevance of a density depends on the processes being investigated, it is likely to vary between species. As a result, it is extremely difficult to interpret patterns of simple arithmetic density measured across species at

some arbitrary scale. This is not to say that such patterns are not interesting (indeed they are sometimes tantalizingly so); however, interpreting the causes and consequences of these patterns will be more difficult because the measure and scale chosen will be to a greater or lesser extent irrelevant to the dynamics of most of the species.

In addition, many other factors complicate such transspecific comparisons, e.g., phylogenetic relatedness which leads to statistical nonindependence and possible artifactual detection of pattern. Transspecific patterns of population density must therefore not be taken at face value. Their meaning is seldom as simple as it appears.

### B. Distribution of Density over Species

It has long been observed that organism abundances are approximately lognormally distributed over species, regardless of the size of the sample area. Preston (1962) suggested that this distribution, which he named the canonical lognormal, had a remarkably constant property,  $(2(\text{SDlog}_2)^{-0.5}) \approx 0.2$  (where  $\text{SDlog}_2$  is the standard deviation of the abundances after transforming using log base 2). This translates into a  $\text{SDlog}_{10} \approx 1$ . In fact, subsequent work has demonstrated that nature is not so kind, but the  $\text{SDlog}_{10}$  does appear quite consistent. An explanation based on the law of large numbers was given by R. May in 1975, who suggested that this phenomenon is inevitable when large numbers of species are involved. Alternative models have been suggested, notably MacArthur's broken-stick model and the geometric log series. Although the lognormal model does not imply any ecological generating process, the other models usually suggest such a process. These (and other more complex niche partitioning models) often fit as well as a lognormal, but there is a danger that workers will infer a process from an adequate fit. The amount of information in a single sample frequency distribution is clearly not sufficient to distinguish between generating processes, especially given the difficulties of defining and comparing densities across species.

### C. Transspecies $-3/2$ Thinning Law

The  $-3/2$  thinning law has a transspecies extension (although the word thinning seems inappropriate). If we plot the maximum recorded density of species against their biomass density we should get a classic allometric power curve  $w = cN^a$  with an exponent  $a$  of close to  $-0.5$ . Enquist *et al.* (1998) suggested that for the curve  $a$  is closer to  $-0.333$  (a  $-4/3$  thinning law

like that suggested for the intraspecific relationship in animals). They suggest a theoretical support for this value.

The existence of a negative relationship between body size and density in animals has long been accepted. However, research suggests that there are two main shapes to the relationship on a log–log plot: linear negative (Fig. 3a) and polygonal (Fig. 3b). Other forms, usually nonlinear with a generally negative slope, have occasionally been suggested for other groups, but currently it is difficult to assess their generality.

The way in which the densities are estimated clearly influences the form of the relationship; in the more than 500 studies investigated by Blackburn and Gaston (1997), the scale of the area considered (“local” versus

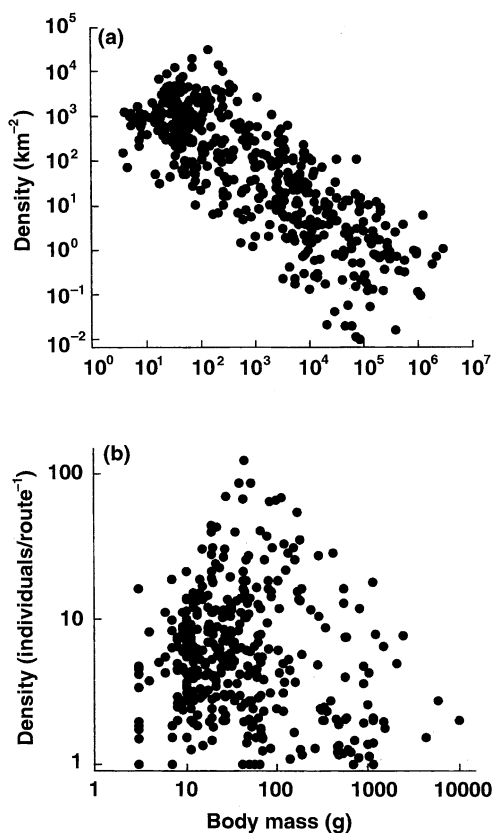


FIGURE 3 Density–body size relationships showing the two main types. Note that despite having generally negative slopes, both have peak densities at intermediate body masses. This is a common phenomenon of unknown significance. (a) Selection of mammals of the world [reproduced with permission from Damuth (1987) as cited in Blackburn and Gaston, 1997] and (b) density (individuals per BBS route,  $n = 380$ ) for North American land birds (reproduced with permission from Brown and Maurer (1987) as cited in Blackburn and Gaston, 1997).

“regional”) was the most important factor. This reinforces the idea that until we know and understand the nature of the density that is being measured we cannot interpret any relationships observed with which it is involved.

The linear negative relationships had highly variable slopes and there was no tendency for them to cluster around  $-0.333$ . Indeed, all the slopes (including manifestly polygonal ones) had a mean of  $-0.51$  and a mode near  $-0.75$ .

## D. Relationships with Other Traits

### 1. Breeding Systems

Characteristically low-density species (especially plants) rely less on outcrossing and sexual reproduction in general than do more common species.

### 2. Reproductive Investment

It has been suggested that rare species (restricted range or low density) produce smaller numbers of propagules per unit time, although they often have longer life spans.

### 3. Dispersal Ability

In general, studies have shown that rare species (with restricted ranges and lower regional density) have poorer dispersal abilities than more common ones.

### 4. Homozygosity

As expected from theory, there is ample evidence that less dense species have lower genetic diversity than more common species.

### 5. Resource Usage

Since density, as previously noted, is often driven by environmental factors, it is not surprising that low-density species are frequently associated with restricted or uncommon conditions and resources.

### 6. Trophic Status

One of the oldest ecological concepts, the ecological or Eltonian pyramid (Elton, 1927), predicts that the higher a species is in the food chain, the lower its density. There is ample evidence that this is usually true; but there are enough exceptions to remind us that few ecological “rules” have universal generality.

### 7. Latitude

For many years it has been suggested that density for a species tends to be higher in higher latitudes. (This surely has to be true for blood-sucking insects.) In

part, this is a prediction of the density compensation hypothesis, which predicts that in communities of low diversity (e.g., islands) density will be higher. Although there is evidence to support this prediction for islands, the evidence for latitude is more equivocal.

## 8. Geographic Range

The relationship between abundance and range was first noted (publicly) by Darwin and has occupied many workers since. That it exists is not in doubt (Fig. 4), but whether it implies anything of biological interest is debatable. The following is the central question: Does such a relationship imply anything about differences in life history between rare and common species? It is easy to show that in an environment that varies spatially in its attractiveness or ability to support species (as in the compound Poisson model described earlier) but in which all species have similar properties besides mean density (or at least whose properties are independent of their average density), there must be a positive relationship between density and the proportion of sites at which they are found. Such a relationship therefore does not imply anything about the life history or lifestyle of common compared to rare species.

In order to demonstrate that such differences do exist, it is necessary to compare the observed distribution of organisms with an appropriate null such as that previously discussed. This model is sometimes called the “sampling” or “artifact” model and has been shown to fit some data very well. However, it is not simply a sampling phenomenon, disappearing if a sufficient number of organisms have been collected or a sufficient

number of areas are sampled. It is predicted by many population dynamic processes in which rate parameters (e.g., birth rate and death rate per individual) are independent of local density. Efforts to investigate the density range relationship in the past have been made more difficult by the use of density at sites at which the species is actually present—local density (treating all zeros as structural and therefore ignoring them).

In fact, there is an easier approach to the problem. Since the probability of being found at a site depends on the degree of aggregation as well as on the mean, the null (the artifact model) assumes that aggregation at the chosen scale is constant (or at least that common species have the same distribution, and therefore mean, of aggregation as rare species). A trend showing that common species were less aggregated (lower  $1/k$ ) than rare ones would show that common species were found at more sites than the null suggests, i.e., they have wider ranges than would be expected if they had the same life history characteristics as the rare ones. Such a trend might be considered interesting (although there might be many explanations—some trivial—for such a pattern). A trend showing that they were more aggregated would be even more interesting. It is worth noting that most of the studies performed to date use either snapshots of the spatial pattern of species or maps accumulated over some extended period of time. Both of these ignore the temporal variation of range (and mean density) that is typical of many groups of organisms. If such temporal information is available, examining each species separately and demonstrating life history phenomena for each and then searching for a trend for those over mean density would be more productive. I discuss this in more detail in the next section.

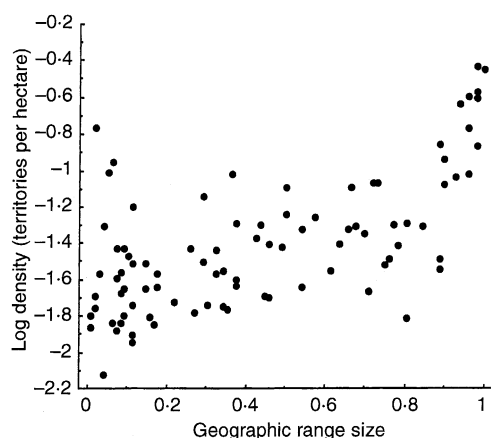


FIGURE 4 The relationship between local density ( $\log_{10}$  territories  $\text{ha}^{-1}$ ) and the proportion of sites occupied on farmland Common Bird Census plots in Britain in 1975 (reproduced with permission from Gaston *et al.*, 1997).

## 9. Transspecific TPPs

Although Taylor's power plots were originally defined for single species, it is possible to plot variance mean relationships across species, with each species contributing a single point. Such plots are clearly difficult to justify if the area over which the data are collected is not equally relevant to all species and the means and variances used are not truly representative of the species. However, if these problems are ignored, then the transspecific TPP can address the density–range problem by plotting the level of aggregation against mean density. The null described in the previous section now becomes equivalent to a TPP slope ( $\beta$ ) of 2 or a horizontal line (slope =  $\beta - 2$ ) on a  $\log(1/k)$  vs  $\log(\text{mean})$  plot. If  $\beta < 2$  then aggregation appears to decline as mean density increases; the range of common species is greater than the null would suggest. However, before

accepting an alternative of differences in the life history characteristics of rare and common species, another null must be considered and rejected. If all the species have individual slopes  $\beta < 2$  (a very unlikely event in most groups), then a transspecific TPP will tend to have a slope of less than 2 as well, even if the rare and common species have the same life history characteristics. In most situations, it will be impossible to check this since if there were enough information to produce spatial variance TPPs for each species, it would be unnecessary to produce a transspecific plot in which each species was represented by a single point. It would be more informative to fit separate TPPs and search for trends in the parameters.

## IX. CONCLUSIONS

The concept of density is an amorphous one. It has to be specifically determined for a particular purpose. The area used for both the sample unit and the study area is crucial and can only be determined after considering what aspect of density is relevant to the current study, and it may even then be difficult to determine since it depends on information about the species' biology that may not be available (e.g., degree of aggregation and movement patterns). This makes comparisons across species especially difficult to interpret (particularly from published studies). That there is significance in all these patterns seems obvious, but interpreting them

rigorously is a considerable (possibly insuperable) challenge.

## See Also the Following Articles

MEASUREMENT AND ANALYSIS OF  
BIODIVERSITY • POPULATION DYNAMICS • POPULATIONS,  
SPECIES, AND CONSERVATION GENETICS

## Bibliography

- Blackburn, T. M., and Gaston, K. J. (1997). A critical assessment of the form of the interspecific relationship between abundance and body size in animals. *J. Anim. Ecol.* **66**, 233–249.
- Gaston, K. J., and McArdle, B. H. (1994). The temporal variability of animal abundances: Measures, methods, and patterns. *Philos. Trans. R. Soc. London* **345**, 335–358.
- Gaston, K. J., Blackburn, T. M., and Lawton, J. H. (1997). Interspecific abundance–range size relationships: An appraisal of mechanisms. *J. Anim. Ecol.* **66**, 579–601.
- Kunin, W. E., and Gaston, K. J. (1997). *The Biology of Rarity. Causes and Consequences of Rare–Common Differences*. Chapman & Hall, London.
- McArdle, B. H., Gaston, K. J., and Lawton, J. H. (1990). Variation in the size of animal populations: Patterns, problems and artefacts. *J. Anim. Ecol.* **59**, 439–454.
- Phipps, M. J. (1992). From local to global: The lesson from cellular automata. In *Individual Based Models and Approaches in Ecology: Populations, Communities and Ecosystems* (D. L. DeAngelis and L. J. Gross, Eds.), pp. 165–187. Chapman & Hall, New York.
- Taylor, R. A. J., and Taylor, L. R. (1979). Population dynamics. In *The 20th Symposium of the British Ecological Society* (R. M. Anderson, B. D. Turner, and L. R. Taylor, Eds.), pp. 1–27. Blackwell, Oxford.



# POPULATION DIVERSITY, OVERVIEW

Jennifer B. Hughes  
Stanford University

---

- I. Origins and Detection
  - II. Importance
  - III. Extent
  - IV. Human-Induced Changes
- 

## GLOSSARY

**allele** One of several forms of a gene.  
**chromosome** A long, threadlike structure in the nucleus of a cell composed of DNA and protein.  
**conspecific** Of the same species.  
**gene flow** The incorporation of genes from one group of individuals into the gene pool of another group.  
**gene pool** The total of all genes in a group of individuals.

---

**EVERY SPECIES IS COMPOSED OF ONE OR MORE POPULATIONS**—localized groups of individuals of the same species. Population diversity is the variety of these groups. This article surveys the origin and detection of genetic differences between populations, discusses the importance of populations to humanity, and reviews estimates of the extent of global population diversity and human-driven modifications of this diversity.

## I. ORIGINS AND DETECTION

Populations are usually defined ecologically or genetically. The ecological entity is called a demographic unit, a group of individuals whose population dynamics are not significantly influenced by migration from nearby conspecific groups. In other words, the fluctuations in population size of one group are largely independent of those of other groups. The genetic entity is called a Mendelian population, a genetically distinguishable group of individuals evolving independently of other groups. Demographic units may be Mendelian populations and vice versa, but the two are not necessarily the same. Moreover, populations are not always clear, discrete units (as is the case when defining any taxonomic boundaries). Although the majority of individuals in a population breed with members of the same population, some individuals may move and interbreed with individuals from other populations, making a clear cutoff between most populations impossible. Here, a genetic definition of a population is used—that is, the groups must be significantly genetically different from one another.

Just as populations can be defined in two general ways, so can population diversity. Population diversity can refer to the amount of genetic divergence between two populations. Thus, two populations that are situated nearby and exchange genetic material frequently are not as diverse as two populations that are completely isolated from one another, perhaps because of a geographic barrier. Population diversity can also be defined



as the number of populations in an area. Thus, one can consider population diversity at many scales, for instance, the number of populations in a habitat, on a continent, or on the entire planet. This definition of population diversity—the total number of populations in an area—is used here.

The following sections review the processes by which conspecific groups diverge into genetically distinct populations and how genetic differences are detected.

## A. Variability within Species

Much biodiversity exists within a species. This variation occurs at three different levels. First, there is genetic variation within individuals of sexually reproducing species. Individuals carry multiple copies of most genes (two copies in diploid species such as humans), and these copies (alleles) may be different. Second, there is genetic variation between individuals. Except in the case of identical twins, no two individuals will have the same alleles for every gene. Finally, there are genetic differences among populations. A population may contain alleles that other populations do not have and vice versa. In addition, two populations may differ in the proportion of individuals that carry alleles shared by both groups.

## B. Processes of Divergence

Individuals are not usually spread out evenly within a species' range but are clustered into localized groups. When these groups have limited genetic exchange, they may diverge into genetically distinguishable populations. (Methods to detect genetic divergence between populations are discussed in the next section.) There are many mechanisms by which populations may diverge, including mutation, genetic drift, and natural selection. These agents of microevolution can result in genetic differentiation by acting on populations at different rates or in different directions.

### 1. Mutation

Mutations change the gene pool of a population by substituting one allele for another. Generally, mutation is not a strong force in genetic divergence between populations because it occurs so rarely at any particular gene. In other words, frequencies of alleles shared between groups are not likely to change dramatically through mutation. If a new allele is produced by mutation and increases in frequency in a group, however, then genetic divergence between groups results from

the presence of the unique allele in one group and not in another.

### 2. Genetic Drift

Gene frequencies may change from one generation to the next simply because of chance. Particularly in a small population, random sampling error in the reproduction of alleles may contribute to changes in allele frequencies from one generation to the next. For instance, an allele that is only represented in a few individuals may be completely lost in the next generation if those few individuals by chance fail to reproduce or pass down that particular allele to their offspring.

### 3. Natural Selection

Natural selection is the differential reproductive success of genetically different individuals resulting from the interaction of the individuals' inherited traits and their environment. Thus, different environmental conditions or biotic communities in a habitat can result in different selection pressures on populations. Some alleles may be beneficial to the reproduction of the individual carrying them in one population, whereas in another population the same allele may be relatively neutral or harmful to the carrier. In other words, selection in these two groups operates at different rates or directions, contributing to their genetic divergence.

### 4. Factors Influencing Divergence

Two factors, population size and dispersal, greatly influence the rate and importance of the previously mentioned agents of divergence. The smaller the population size, or number of individuals in a group, the greater the effect of genetic drift. The probability of gene frequencies in the next generation being representative of the previous generation decreases as the number of reproducing individuals decreases.

Dispersal, or the migration of individuals from one group to another, affects the amount of gene flow between groups. When an individual successfully reproduces in another group, the genetic composition of the new group becomes more similar to the group from which the immigrant originated. Thus, gene flow has the effect of homogenizing genetic composition, thereby decreasing the rate of divergence resulting from the three agents discussed previously.

## C. Detection of Divergence

In general, the degree of genetic divergence between groups of organisms is measured by comparing the amount of variability within a group to the variability

between groups. Thus, the genetic diversity embodied by the individuals in each group must be assessed in order to detect genetic divergence. Genetic diversity may be estimated indirectly by measuring morphological characteristics and migration rates or directly by examining the genetic material. A survey of some of these methods is provided in the following sections.

### 1. Morphological Variation

The variation that is observed among individuals is the result of their genetic composition, their environment, and the interaction between these factors. Thus, morphological variation will in some part represent underlying genetic variability, but it will also be intertwined with variation caused by environmental factors. Nevertheless, in the absence of direct genetic methods, morphological variation of a suite of traits is often used to resolve the population structure of a species. There have been attempts to tease apart genetic and environmental factors by using breeding experiments. In these studies, the environment is kept constant and morphological traits of parents and offspring are measured. One can then estimate the heritability, or the variation resulting from the genetic background, of a trait in the experimental conditions.

### 2. Migration Rates and Population Size

The migration rate of reproductive individuals into a group gives an estimate of the amount of gene flow, or exchange of genetic material, with outside groups. The lower the gene flow into a group, the greater the impact of genetic drift. The other critical factor affecting the degree of genetic drift in a population is its size: The smaller the population, the greater the expected drift. Thus, a combination of migration rate and population size estimates gives an indirect estimate of the amount of genetic divergence that occurs resulting from genetic drift.

### 3. Protein Variation

A relatively direct method of detecting genetic divergence is to examine a sample of proteins (allozymes) from individuals between groups. The composition and structure of a protein are determined by the sequence of nucleotides (the subunits that make up the DNA molecule) comprising that gene. Alleles of the same gene have differences in their sequences, and each allele gives instructions for making a particular protein. The proteins can be distinguished by their rate of movement through a gel medium that is exposed to an electrical field. The movement rates of proteins differ because of different sizes and net charges of the molecules. By

sampling many individuals in a group, one can assess the number of unique alleles and their frequencies. Thus, the genetic divergence between groups can be estimated from the variation of a random sample of proteins. The greater the differences in the proteins and their frequencies among groups, the greater the genetic divergence.

### 4. Chromosomal Variation

A direct method of surveying the genetic material of an individual is to examine the chromosomes. Under a microscope, differences in the number, size, or shape of the chromosomes of an individual can be observed. Differences in the morphological features of chromosomes reveal differences in genetic composition. Below the species level, however, most variation in DNA will not be apparent from chromosome morphology; in other words, the chromosomal variation among populations of the same species is relatively small. Thus, genetic divergence between populations often cannot be detected with this method.

### 5. DNA Variation

Specific sections of an organism's DNA can be isolated and reproduced using the polymerase chain reaction method. This section of DNA can then be examined by a variety of methods that vary in their resolution. For instance, restriction fragment length polymorphism analysis offers a broad examination of DNA differences. The isolated DNA is exposed to proteins that break the DNA wherever a particular sequence of nucleotides occurs. The length of these fragments can be distinguished by their rate of movement through a gel medium that is exposed to an electrical field (similar to the method used to detect protein variation). The greater the difference in the pattern of fragment lengths, the greater the difference in the piece of DNA.

Determining the exact nucleotide sequence of an isolated piece of DNA offers the greatest resolution of genetic composition. The genetic divergence between groups can then be estimated by comparing the similarity of the sequences.

## D. Indices of Divergence

A variety of indices are used to evaluate the population structure or the genetic divergence within a species. One of the most common indices reported is Sewall Wright's  $F_{ST}$ . This statistic can be calculated from the allele frequencies of a gene or genes and describes the variation among groups relative to the variation within the groups. The greater the variation that can be attrib-

uted to differences between groups rather than within groups, the greater the  $F_{ST}$  value. Other measures of divergence describe the genetic distance between groups. These measures are indices of similarity or dissimilarity and are calculated using any of the types of data described previously, including protein frequencies and DNA sequences. Groups that are more isolated and for longer periods of time are expected to be more genetically distant.

## II. IMPORTANCE

Many of the benefits that biodiversity confers on humanity are delivered through population diversity. These benefits include aesthetic enjoyment, species conservation, discovery and improvement of pharmaceuticals and agricultural crops, replenishment of stocks of economically valued species, and delivery of ecosystem services.

Again, population diversity is defined here as the number of populations in an area. Certainly, characteristics of a population, such as the area it occupies and the number of individuals it includes, will affect the benefits provided by a single population. In general, however, as the number of populations of a species increases, its area occupied and number of individuals increase. Thus, the following discussion focuses on numbers of populations and ignores population characteristics.

### A. Aesthetic Value

Populations of organisms, their physical environments, and the interactions between them make up natural ecosystems. People value the aesthetic benefits of the populations in these habitats. For instance, bird watchers enjoy bird populations, hikers appreciate shade trees, and divers seek out reef fishes. Others simply enjoy the scenery created by the sum total of populations in an area.

### B. Species Conservation Value

By definition, populations are essential to the conservation of species. Specifically, the number and size of populations influence the probability of species extinction; a species with many large populations is less susceptible to extinction than a species with a few small populations. Migrants from other populations can prevent the extinction of a population by contributing individuals when numbers are low or supplying genetic

variation needed to adapt to changing environmental conditions. If local extinction does occur, individuals from other populations can found a new population in the area. The threat of rapid global climate change makes population diversity within a species especially critical; a species with many populations is more likely to include the variation necessary to adapt to new conditions.

### C. Genetic Value

Populations of the same species may produce different types or quantities of defensive chemicals, compounds that may be of medicinal or agricultural value to humans. The story of the development of penicillin provides a clear example of the importance of population diversity to pharmaceuticals. The widespread use of penicillin as a therapeutic drug did not occur until 15 years after Alexander Fleming's discovery of the compound in common bread mold. One reason for this delay was a search to find a population of the mold that produced greater quantities of penicillin than did the original population.

Population diversity among wild crop relatives supplies genetic material to agricultural strains. The world's three main crops (rice, wheat, and maize) are widely planted in genetically uniform strains. The emergence of a new disease or pest can therefore threaten large fractions of the harvest at one time. Wild crop relatives contain genetic variation that does not exist in the agricultural populations, including genetic resistance to a variety of diseases and pests. Thus, when an agricultural strain is susceptible to a disease or pest, researchers search for genetic resistance in wild crop relatives. Thousands of populations may have to be tested until one is found that carries the genetic resistance that can be used to protect the crop. For example, when the grassy stunt virus emerged as a serious threat to the rice crop in Southeast Asia in the late 1960s and 1970s, a search for resistant rice varieties was conducted. Only one strain of a wild rice at the gene bank of the International Rice Research Institute was found to resist the virus. Genetic variation in wild populations will also be crucial in providing genetic material to sustain crop yields in the face of human-induced changes in growing conditions.

### D. Direct Economic Value

The reduction of economically important species often has direct consequences for local peoples. For instance, the decline of fish harvests leads to the loss of income

for fishermen and the loss of an important source of protein for much of the human population. Populations of economically valued species not only act as a safety net against species extinction but also increase the species' harvest level. If more populations exist, then more populations can be harvested. Populations may also increase harvests indirectly. A collection of populations that are connected by dispersal is called a metapopulation. In a metapopulation, dispersal of individuals from other populations can rescue a population from extinction when its numbers are critically low. Over the long term, a population will be larger if it is a component of a larger metapopulation than if it is part of a smaller metapopulation or completely isolated.

## E. Ecosystem Service Value

Ecosystem services—services that natural ecosystems provide to human civilization for free—are perhaps the most important benefits that populations confer on humanity. They include natural processes such as purification of air and water, detoxification and decomposition of waste, generation and maintenance of soil fertility, pollination of crops and natural vegetation, and pest control. Populations deliver these services, and therefore population diversity at global, regional, and local scales affects the quality and quantity of the services provided. (This article focuses on how the number of populations affects ecosystem services, although the size and density of populations are also important factors.)

### 1. Global Population Diversity

Higher global population diversity probably enhances the delivery of global ecosystem services such as regulation of biogeochemical cycles and stabilization of climate. For example, the larger the area remaining under natural tree cover in the Canadian taiga, the greater the amount of carbon stored there. Although deforestation in this region may not result in the extinction of any tree species, a large-scale loss of tree populations would affect the global balance of greenhouse gases in the atmosphere.

### 2. Regional Population Diversity

For many ecosystem services, global population numbers will not be as important as the population diversity in the region of interest. For these services, it is necessary not only that there be many populations somewhere in the world but also that they exist in the region where the services are to be provided. These services include water purification by forests

and wetlands and mitigation of floods and droughts by forests. Loss of these services occurs when the populations in forest and wetland habitats are destroyed in one area, regardless of the continued existence of these populations elsewhere.

New York City provides an excellent example of the value of regional population diversity for water purification. The city was renowned for its clean water, which came from the Catskill Mountains 100 miles to the north. Natural purification processes, performed by populations of plants and soil organisms, were sufficient to cleanse the water for most of the city's history, but in recent years land development and related human activities reduced the effectiveness of these processes. In 1996, city water officials floated an environmental bond issue to purchase land, freeze development on other lands, and subsidize the improvement of septic tanks. It is expected that these actions will restore and safeguard the local populations of the organisms that filter and purify the water. If so, an investment of \$1 billion in natural purification services will have saved city taxpayers \$6–8 billion, the additional avoided cost (over 10 years) of building a water treatment plant.

Regional population diversity is also essential for pest control. This function of populations is easy to take for granted, but it is dramatically illustrated when one kind of organism is transplanted to a new environment that lacks populations of predators that usually keep its numbers in check. A classic case is the importation of prickly pear cacti (*Opuntia*) into Australia by early settlers. In the absence of their normal predators, the cacti spread over approximately 25 million ha in New South Wales and Queensland. About half of the area was so densely covered with the cactus that the land could not be used for farming or ranching, and the costs of poisoning or removing the *Opuntia* were more than the value of the land. The problem was solved by importing a cactus-eating moth, *Cactoblastis cactorum*, from the South American homeland of the *Opuntia*. Once regional populations of the moth were established, the *Opuntia* populations were decimated. Although the cactus can still be found in Australia, it now occurs only as scattered clumps.

The importance of regional populations of native pollinators for agriculture is made clear by their decline (chiefly a result of pesticides and habitat destruction). For example, for more than 60 crops grown in the United States, farmers have to pay honeybee keepers to bring hives to the fields or orchards to be pollinated. This service is estimated to cost farmers more than \$60 million a year and the federal government more than \$80 million in subsidies. Problems within the bee-keep-

ing industry (disease and hybridization with the Africanized honeybee) are increasing these costs.

### 3. Local Population Diversity

The number of populations at a particular location—the local diversity of species—affects local ecosystem function. In many greenhouse and field experiments, plant productivity has been found to increase with the diversity of plants. Higher diversity also seems to be associated with greater stability of plant productivity. For instance, more diverse grasslands appear to be more resistant to drought and grazing disturbance than less diverse areas. Thus, local population diversity likely influences the amount and variation of services provided by an ecosystem. Also, because regional and global services are performed by an assemblage of local ecosystems, this scale of population diversity will affect the delivery of larger scale services as well. For example, global biogeochemical cycles are probably influenced not only by the total number of tree populations on the planet but also by the diversity of trees within a habitat.

Furthermore, traits that affect local ecosystem functioning may differ among populations. Treseder (2000) studied multiple populations of two ecologically important tree species in Hawaii. She found genetically based differences among populations of *Metrosideros polymorpha* in the decomposability of leaf litter and among populations of *Acacia koa* in the potential to fix nitrogen. Thus, these populations are not interchangeable; if they were replaced with other populations of the same species, the functioning of the native ecosystem would be altered.

## III. EXTENT

Although populations are critical to humanity, little is known about the extent of population diversity. Hughes *et al.* (1997) made a rough estimate of the total number of populations on the planet. They used a Mendelian population definition and restricted the estimate to eukaryotes (although the diversity of bacteria and viruses is probably enormous, information on their diversity is scarce).

The evaluation of global population diversity involved three steps. First, Hughes *et al.* (1997) estimated the average number of populations per unit area from literature on population differentiation. Second, they calculated the average range size of a species from published range maps. The product of these two numbers yields an approximation of the average number of populations per species. Finally, they multiplied this product

by the number of species on Earth to arrive at an estimate of global population diversity.

### A. Populations per Area

To quantify the number of populations per area for an average population, Hughes *et al.* (1997) searched 15 journals for genetic studies on population differentiation. They found 81 articles that provided appropriate data to quantify the number of populations per area. Most of the species were vertebrates (35), followed by plants (23), arthropods (19), mollusks (4), and a flatworm.

For each species, Hughes *et al.* (1997) determined whether the sampling locations in the studies were for separate populations or a single population. If the genetic differentiation between localities was statistically significant, then all the localities were called separate populations. The number of populations per area was calculated as the number of sampling points divided by the area that was sampled. If the samples were not significantly differentiated, then it was assumed that all of the samples were taken from one population and that the size of a population was the size of the sampling area. For many species, an intermediate amount of differentiation was found. For example, although not all the samples were significantly different, some clusters of sampling locations were significantly different from others. In this case, the number of significantly diverged clusters was used as the number of populations.

Comparing genetic differentiation across studies is difficult because each study uses slightly different methods for detecting genetic differentiation. Thus, the estimates of populations per area from each study were rounded to the nearest order of magnitude. The average of these estimates yields an estimate of one population per 10,000 km<sup>2</sup> for an average species. Note that the arithmetic average of orders of magnitude is equivalent to the geometric mean of the untransformed estimates. Therefore, all else being equal, the estimate of populations per area is conservative.

There are several notable biases with this method. First, the species represented give a taxonomically biased sample of the earth's species diversity. For example, arthropods make up about 65% of the planet's species, whereas birds account for probably less than 0.01%. However, in this estimation arthropods account for only 20% of the species, whereas birds account for more than 11%. Second, the sampling intensity of each study limits the number of populations that can be estimated. This constraint makes the estimate conservative since in many cases additional sampling in the

study area may have revealed further differentiation. Finally, the molecular markers used in the studies may not always reveal notable differences between groups, again causing the estimate to be conservative.

### B. Average Range Size

The average range size of a species was calculated from more than 2400 species range maps for birds, mammals, fish, and butterflies (Hughes *et al.*, 1997). Weighting each of the taxonomic groups equally, the mean range size of a species is 2.6 million km<sup>2</sup>. The average range size of the most species-diverse group, the butterflies, is 2.2 million km<sup>2</sup> and this was conservatively used as the estimate of the average range size of a species.

The shaded areas of species distribution maps almost always include habitats in which populations do not occur. Therefore, estimating species' ranges from these maps will probably inflate the population diversity estimate. In addition, the majority of species included occurred in temperate regions; however, it is estimated that two-thirds of species diversity exists in the tropics. Because species' range sizes in some taxa tend to increase toward the poles, the bias toward temperate areas may inflate the estimate of average range size. In contrast, the range maps that were used were restricted to one continent; therefore, the range sizes of intercontinental species were probably underestimated.

### C. Species Diversity

The product of the average number of populations per area and the average range size of a species gives an average of 220 populations per species. Multiplying this number by three estimates of global species richness (5, 14, and 30 million) yields estimates of 1.1, 3.1, and 6.6 billion populations globally. This calculation assumes that a species' range size and its number of populations per area are independent. If these two factors are strongly correlated, however, the estimate may be quite inaccurate. However, it is not known whether a correlation exists, let alone whether the correlation is positive or negative.

## IV. HUMAN-INDUCED CHANGES

With the expansion of the human enterprise, human beings are having a significant impact on population diversity. Populations are being both created and destroyed on a large scale as a direct result of human activities. As species are introduced to nonnative habi-

tats, new populations are formed. At the same time, habitat disturbance and destruction are driving the extinction of populations. Although most public and scientific attention to biodiversity is centered on species, a few recent studies have attempted to evaluate the extent to which humans are modifying population diversity.

### A. Introductions and Extinctions in Western Australia and California

Hobbs and Mooney (1998) collected data on the deletions and additions of populations in Western Australia and California. They gathered records of species that are threatened, extinct, or introduced in these areas. The number of regional extinctions gives a very conservative estimate of population extinction since most, if not all, of the species originally had more than one population in the region. Similarly, it is reasonable to assume that the threatened species have already lost populations. These species are endangered in large part because of range contractions—that is, the extinction of populations, thereby reducing the geographic distribution of the species. The number of introduced species is also a lower bound estimate of the number of new populations. To be recorded as an introduced species, at least one population must be established, but many species may have already established multiple populations.

For California, Hobbs and Mooney (1998) gathered data for butterflies, amphibians, reptiles, birds, mammals, and plants. Of 8274 total species, 1109 are introduced (13.4%), 71 endangered, and 49 extinct. The endangered and extinct categories include subspecies, which are most certainly genetically distinct populations. For Western Australia, the study reported information for plants, mammals, birds, reptiles, amphibians, and freshwater fishes. Of more than 12,000 plants, 1032 have been introduced, 232 are threatened, and 52 have gone extinct. The percentage of all introduced species is 7.3%, which is lower than that in California. In California, no extinctions have been reported for reptiles and amphibians, and in Western Australia no extinctions have been recorded for birds, reptiles, amphibians, or freshwater fishes. As mentioned previously, however, these results mask most of the population extinctions. In the groups for which no extinctions were recorded, other studies report significant species' range contractions. For instance, the loss of native vegetation in Western Australia has resulted in range contractions and population extinctions of species that rely on this habitat.

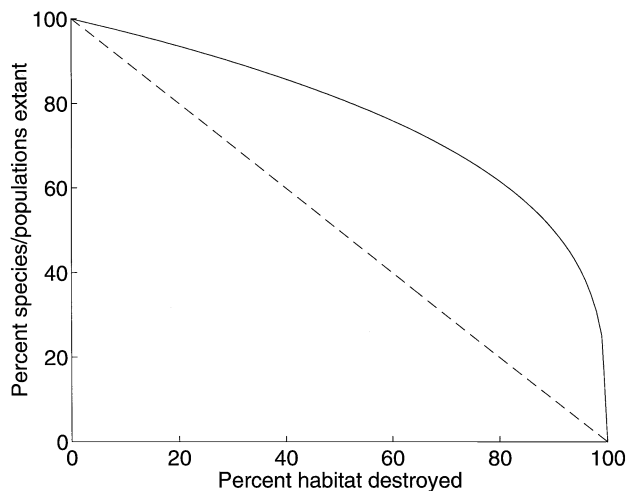


FIGURE 1 Predicted relationships between the area of habitat destroyed and the extant number of species and populations. The species curve (solid line) is  $S = cA^z$  (see text), where  $z$  is 0.30. The population curve (dashed line) is linear.

## B. Extrapolating Population Extinctions from Tropical Deforestation Rates

Most estimates of species extinction rates are based on species–area relationships and the rate of habitat loss due to deforestation. The most commonly used model of the species–area relationship is shown in Fig. 1. This model is  $S = cA^z$ , where  $S$  is the number of species,  $A$  is the area (or habitat size) in which the species are found, and  $c$  and  $z$  are constants estimated from empirical studies. This model is in agreement with a multitude of empirical studies on a variety of taxa and scales. With this formula, the rate of species extinction in an area can be approximated from the rate of habitat loss. Figure 1 also illustrates a widely cited rule of thumb: A decrease in habitat area by 90% should result in a decrease in species diversity by approximately 50%.

### 1. Population–Area Relationship

There is no comparable work describing a population–area relationship. Hughes *et al.* (1997) reason that population numbers and area probably correspond approximately in a one-to-one manner. The basis for the difference between the species– and population–area relationships is their size; a population occupies a small area relative to a species. When a substantial amount of habitat is destroyed, populations that were encompassed by the area will be lost. In contrast, many fewer species will go extinct because other populations of the species exist elsewhere. Thus, the rate of population

extinction will be faster than that of species extinction. If the one-to-one relationship does hold, then when 90% of an area is destroyed about 90% of the populations in the original area will be lost, in contrast to only 50% of all species (Fig. 1).

### 2. Extinction Rate

The rate of population extinction in tropical forest regions can be estimated with the population–area relationship in Fig. 1. Assuming that tropical deforestation is occurring at 0.8% per year, global population diversity is 3 billion, and two-thirds of all populations exist in tropical forest regions, Hughes *et al.* (1997) calculate that 16 million populations per year, or approximately 1800 per hour, are being exterminated in tropical forests alone. This rate is three times higher than conservative estimates of species extinction.

## C. Impacts of Changing Population Diversity

Impacts resulting from human-caused extinction and introduction of populations are already evident and widespread. The extinction of population diversity results in the loss of the benefits outlined previously, and the establishment of populations of invasive species often disrupts native communities, altering ecosystem functioning or causing the extinction of native species. For instance, approximately 20% of all endangered vertebrates are thought to be threatened by invasive species in some way.

Furthermore, it is unclear how often the addition of introduced populations (or “weedy” populations) will compensate for the loss of native population diversity. For services such as pest control, there is abundant evidence that such compensation will be rare. Pesticides destroy populations of natural enemies as well as the intended pest. As a result, obscure organisms, released of predation pressure, often are promoted to pest status when no new weedy species fill the role of the natural predators. Similarly, weedy species are not compensating for the loss of pollinator populations in the United States.

For other services such as flood control and soil retention, the potential for substitution by weedy populations, at least in the short term, may sometimes be high. Over the long term, however, the ability of weeds to maintain services is unknown. For example, weedy ground cover on bare slopes may prevent soil erosion, but it may be poor at maintaining soil fertility. Moreover, the capacity for large-scale technological substitu-

tion of ecosystem services by human-made communities of populations appears limited. The Biosphere 2 project is a case in point. Despite hundreds of millions of dollars invested in development and operating costs, scientists failed to engineer a system that could support eight people with food, air, and water for 2 years. The failure dramatically illustrates humanity's dependency on the life-support services that populations in natural ecosystems provide for free.

## D. Ameliorating Changes in Population Diversity

There are many ways in which human modification of population diversity can be slowed and its impacts alleviated.

### 1. Controlling Invasions

In general, the impact of newly introduced populations on an ecosystem is highly unpredictable. In fact, most invasions do not produce noticeable effects. The invasive species that do have significant impacts, however, can cause widespread ecosystem disruption and economic loss. In theory, removing invasive populations can restore the native ecosystem (unless species have gone extinct). In practice, however, complete removal or even partial control of invasives is rarely successful. One of the few examples of a successful program to control invasive populations is the *Opuntia* cactus case described previously.

There are three broad methods to control invasive populations: biological, chemical, and manual. One reason that introduced species can be very harmful to an ecosystem is that the species are free from their natural enemies and therefore overabundant. Biological control involves reducing the invasive species' numbers by introducing a predator or parasite from its native habitat. Chemical controls are poisons (usually synthetic) such as pesticides and herbicides. Manual control entails the physical removal or killing of individuals. All three of these fixes are usually expensive and in the case of biological and chemical controls risk doing more harm than good to the ecosystem. An introduced predator might attack native species, and chemicals may harm native species or human beings. Thus, the best control is to prevent invasive populations from establishing in the first place. Unfortunately, this is an extremely difficult challenge—most introductions are accidental. Furthermore, with the accelerating movement of products and people around the world, the rate of introductions will certainly increase in the future.

### 2. Preservation

Protecting intact habitat from destruction or degradation is the most straightforward method to prevent the extinction of populations. If population diversity and habitat area are indeed related in a linear manner, then the more land that is preserved, the more population diversity that is protected. In most cases, however, limiting almost all human access to a piece of land is usually not feasible because of competing uses, such as agriculture and urbanization. Nevertheless, these competing uses do not eliminate all prospects of conserving population diversity and the benefits they provide. In fact, in some parts of the world, many populations appear to be surviving in the "countryside"—that is, in human-dominated areas (Daily, 2000). Thus, a promising approach involves managing the countryside to prevent further degradation (and perhaps some restoration) of population diversity in these areas.

### 3. Restoration

Unlike species, populations can be restored and, along with them, at least some of the benefits that they once provided. Although genetic variation unique to the extinct populations will be lost, groups of individuals can be reintroduced to an area and in a relatively short time may evolve significant differences from other populations. Reestablishing populations of a native species is not a simple task, however. Often, the native ecosystem is degraded and little is known about the biological requirements of the species.

## See Also the Following Articles

DIVERSITY, COMMUNITY/REGIONAL LEVEL • ECONOMIC VALUE OF BIODIVERSITY, OVERVIEW • ECOSYSTEM SERVICES, CONCEPT OF • EXTINCTIONS, MODERN EXAMPLES OF • HUMAN IMPACT ON BIODIVERSITY, OVERVIEW

## Bibliography

- Daily, G. C. (1999). Developing a scientific basis for managing Earth's life support systems. *Conserv. Ecol.* 3(2), 14.
- Hobbs, R. J., and Mooney, H. A. (1998). Broadening the extinction debate: Population deletions and additions in California and Western Australia. *Conserv. Biol.* 12, 271–283.
- Hughes, J. B., Daily, G. C., and Ehrlich, P. R. (1997). Population diversity: Its extent and extinction. *Science* 278, 689–692.
- Treseder, K. K., and Vitousek, P. M. (2000). Potential ecosystem-level effects of genetic variation among populations of *Metrosideros polymorpha* from a soil fertility gradient in Hawaii. *Oecologia* (in press).







# POPULATION DYNAMICS

Alan Hastings  
University of California, Davis

---

- I. Introduction
  - II. Density-Independent Dynamics of a Single Species without Age Structure
  - III. Density-Independent Age Structure
  - IV. Density Dependence
  - V. Role of Stochasticity
  - VI. The Simple Two-Species Interactions
  - VII. Conclusions
- 

*POPULATION DYNAMICS* is the changes in the numbers of a species at a single location. Understanding the dynamics of a population requires incorporating both within species and between species interactions. This chapter presents a conceptual approach to the study of population dynamics with brief illustrative examples, beginning with single species approaches and finishing with interactions between two species.

## GLOSSARY

- Allee effect** Reduction in growth rate of a population at low densities.
- competition** The effect of the members one species reducing the growth rate of the population of another species through interference or exploitation of common resources.
- density dependence** Effects on the growth rate of a population due to interactions among the members of that population.
- endogenous forces** Forces within a population that affect the dynamics of the population.
- exogenous forces** Forces outside a population that affect the dynamics of the population.
- population** Members of a single species that can potentially interact with each other.
- predation** An interaction between two species in which one species consumes the other.
- 

## I. INTRODUCTION

The term *population* refers to the members of a single species that can interact with each other. Thus, the fish in a lake, or the moose on an island, are clear examples of a population. In other cases, such as trees in a forest, it may not be nearly so clear what a population is, but the concept of population is still very useful.

Population dynamics is essentially the study of the changes in the numbers through time of a single species. This is clearly a case where a quantitative description is essential, since the numbers of individuals in the population will be counted. One could begin by looking at a series of measurements of the numbers of particular species through time. However, it would still be necessary to decide which changes in numbers through time are significant, and how to determine what causes the changes in numbers. Thus, it is more sensible to begin with models that relate changes in population numbers through time to underlying assumptions. The models

will provide indications of what features of changes in numbers are important and what measurements are critical to make, and they will help determine what the cause of changes in population levels might be.

To understand the dynamics of biological populations we start with the simplest possibility and determine what the dynamics of the population would be in that case. Then deviations in observed populations from the predictions of that simplest case would provide information about the kinds of forces shaping the dynamics of populations. Therefore, in describing the dynamics in this simplest case it is essential to be explicit and clear about the assumptions made. We would not be arguing that the idealized population we describe would ever be found, but that focusing on the idealized population would provide insight into real populations, just as the study of Newtonian mechanics provides understanding of more realistic situations in physics.

## II. DENSITY-INDEPENDENT DYNAMICS OF A SINGLE SPECIES WITHOUT AGE STRUCTURE

The very simplest description of population dynamics looks only at a single population of a single species. Thus, complications due to the effect of one species on the growth or dynamics of another are initially ignored. Any differences among individuals within the species are assumed insignificant. Thus, differences in age, sex, genetics, spatial location, and other individual characteristics are all ignored. Moreover, we focus on a single population, so any effects of immigration or emigration are ignored as well. We also ignore any effect of randomness, which might arise either from the effects of small population sizes or environmental variability.

We still need to make one more fundamental assumption when describing the dynamics of a population in order to consider what effect members of the population have on each other. The simplest case is density independence—in which the demography of the population does not depend on the size (in numbers) of the population. We assume that the per capita rates or numbers of births and of deaths do not depend on the size of the population. This is not the same as assuming that the overall birth and death rates do not depend on the population size, but it is equivalent to assuming that the overall birth and death rates are proportional to the population size.

To describe these dynamics under these assumptions in detail, we need to split our discussion into two cases, depending on the way reproduction occurs. For some

species, such as humans or many microorganisms, births and deaths occur essentially continuously through time. For other species, such as many butterflies or annual plants, organisms reproduce once each year and then die. We will treat these two possibilities separately and then turn to more complex cases.

For the case of overlapping generations and continuous reproduction, we can write our description of population dynamics using the verbal equation

$$\begin{aligned} \text{Rate of change of population} \\ = \text{Rate of births} - \text{Rate of deaths.} \end{aligned}$$

With our assumption of density independence, we can write the rate of births as  $bN$ , where  $b$  is the per capita birthrate and  $N$  is the population size. Similarly, the rate of deaths is  $mN$ , where  $m$  is the per capita death (mortality) rate. The verbal equation can then be written as the mathematical equation

$$dN/dt = bN - mN$$

or, letting  $b - m = r$ , the intrinsic rate of increase (or “little  $r$ ”) as

$$dN/dt = rN. \quad (1)$$

We can determine an explicit solution to this model expressing the population size at any future time  $t$  in terms of population size at time 0 and the intrinsic rate of increase,  $r$ . The solution of the model (1) is exponential growth, namely

$$N(t) = N(0)e^{rt}, \quad (2)$$

where  $e$  is the base of the natural logarithms,  $N(0)$  is the population at time 0, and  $N(t)$  is the population at time  $t$ . We will turn to ecological the implications of this formula after we discuss the alternate case of growth in discrete time.

For the case of nonoverlapping generations, appropriate for annual plants, we can write down a verbal model as follows:

$$\begin{aligned} \text{Numbers next year} \\ = \text{Per capita reproduction times numbers this year.} \end{aligned}$$

In mathematical terms, this equation becomes

$$N(t + 1) = RN(t), \quad (3)$$

where  $N(t)$  is the population size in year  $t$ , and  $R$  is the number of individuals in the population next year

produced by each individual in the population this year. This model, too, can be solved explicitly,

$$N(t) = R^t N(0), \quad (4)$$

which is geometric growth. In fact, if the growth rates in the two cases were matched so that  $R = e^r$ , or  $\ln R = r$ , the two population descriptions (2) and (4) would provide exactly the same numerical values (at the discrete times  $t = 1, 2, 3$ , etc.).

The implications of the solutions (2) and (4) are not that we would expect any real population to grow at these rates, but, as emphasized perhaps first by Malthus several centuries ago, that exponential or geometric growth is very rapid. If  $r$  is positive or  $R$  is greater than 1, after relatively short times, the predictions of exponential or geometric growth are that populations will become very large. Thus, exponential growth cannot last, and the fundamental question of population dynamics becomes what prevents exponential growth and what are the consequences for population dynamics of these forces that prevent exponential growth.

Long time density independent growth in natural systems obviously does not occur, with the closest examples following introduction of nonnative species. For example, the number of collared dove in Great Britain increased exponentially from less than 10 to more than  $10^4$  from 1955 to 1963, but the population essentially stopped growing shortly thereafter. Other introduced species have behaved similarly.

### III. DENSITY-INDEPENDENT AGE STRUCTURE

There are many ways to extend our understanding of the dynamics of a single population from the simplest cases of density independent growth we have just covered, and to explore what prevents exponential growth. There are two basic ways to extend our model for a single species (and remain within the context of looking at a single species). We could look at the effects of members of the population on each other and how they affect the dynamics. We could also look at the effect of differences among members of the population on the dynamics of the population. We will first look at the effect of differences among individuals, focusing on differences due to age. Other differences are also important and we focus on age because it is a critical factor, has been well studied, and is easy to understand and measure. Does including the effect of age prevent exponential growth?

It is clear that within many populations some individuals are too young to reproduce, while others are too old. Thus, just looking at the numbers of individuals in the population is not enough to predict the future numbers. When looking at the effects of age structure, we also face the issue that the proportion of females and males may affect the growth of the population. Since we will still ignore the role of density dependence, we will take the approach of only looking at the female members of the population, and therefore only looking at births of females to females. We thus implicitly assume that males do not affect the dynamics of the population, which may be valid for some natural populations, such as moose, but not all, such as birds with parental care.

The information we want to include is the probability of births (or rate of births in continuous time) to a female of a given age, and the probability (or rate) of death at each age. From this information, and from the current distribution of females of different ages in the population, we should be able to predict the future population. Once again, there are different formulations for this model, depending on whether the model is phrased in continuous or discrete time.

In discrete time, one way to write this model is known as the Leslie matrix model. We first decide on a census interval, for example five years in humans. The population is described by a vector (which is just a vertical list of numbers) in which each entry is the number of females in each age class, 0 to 5, 6 to 10, 11 to 15, 16 to 20, and so on. The length of the age classes and the census times have to be the same. Then from data we determine the average number of births to, and probability of survival of, a female in each age class, over a 5-year interval. These data are presented in a life table, which summarizes the relevant demographic information. Life tables have been constructed from numerous species, ranging from insects, to small mammals, to long lived reptiles, to plants, to humans. The data in a life table can be used as the entries in a matrix, known as the Leslie matrix, in which the first row are the birthrates and the entries under the diagonal are the survival probabilities.

Analysis of this model reveals information about both the rate of growth of the population and the distribution of individuals of different ages. The key concept is the stable age distribution—the unique proportions of individuals of different ages that is preserved through time in the model. Thus, if the population starts at the stable age distribution, it will remain in the stable age distribution. Also, for a population in the stable age distribution, the proportional change in the numbers in each age class, and therefore the population as a

whole, is the same at each time step. This is simply geometric growth once again.

What happens if the population is not in the stable age distribution? It is clear that then the growth rate may change through time and will depend on the ages of the individuals—for example, a human population of all 0–5-year-olds would have no births, while one composed only of 20–25-year-olds could have many births. Yet, under some simple assumptions (essentially that reproduction is not confined to a single age class), if the population does not start in the stable age distribution, it will approach the stable age distribution through time. The result is that geometric or exponential growth occurs again, but only in the limit.

The analogous model in continuous time provides similar results, though there is a difference in the conceptual development. In this case, which can be traced back several centuries to the work of the Swiss mathematician Euler, the focus is typically on the rate of births through time rather than the individuals of all ages at a single time, through the use of a renewal equation. In words, the dynamics are described as

Rate of births at time  $t$

= Sum over all ages  $j$  of the birthrate at time  $t - j$   
 × the probability of surviving to age  $j$   
 × the birthrate at age  $j$ .

Using the assumption of exponential growth this equation can be reduced to an equation for the growth rate of the population, the Euler-Lotka equation. This equation (or its discrete time approximation) is the basis for most estimates of the growth rate of populations.

The demography of populations and the ways in which population growth rates are calculated are worth emphasizing because they play a central role in conservation biology. A calculation of the growth rate of a population is often used to determine if the population is endangered. By looking at the effect of changes in different demographic parameters on the growth rate, one can determine which conservation efforts may have the largest effect on the growth rate and therefore the survival of the species in question.

## IV. DENSITY DEPENDENCE

### A. Continuous Time

We now need to continue our search for processes that will prevent populations from growing exponentially

and what the consequences of these limiting processes are for population dynamics. The simplest answer is that as the density of a species increases, either the birthrate goes down or the death rate goes up, because the members of the species compete with each other (due to the fact that the food supply or some other resource is limited). This kind of intraspecific competition is one of the fundamental reasons why populations may not grow exponentially, and we will explore its consequences both for species with overlapping generations and those that reproduce once per year. In both of these cases and all that follows we will be ignoring age structure, for simplicity.

One of the earliest formal descriptions of these dynamics was given by Verhulst, who produced what is now known as the logistic equation. In this equation, the growth rate of the population is described by the product of the exponential growth term as in model (1) and the term representing reduction in per capita reproduction due to interactions with other individuals. In modern terms, the model is written as follows:

$$dN/dt = rN(1 - N/K),$$

where the term  $K$  is called the carrying capacity of the population.

The logistic equation can be solved explicitly (Fig. 1), but the most important conclusion is that any population reaches the carrying capacity,  $K$ . This qualitative conclusion also holds for similar assumptions. The search for the mechanisms leading to density dependence, and even direct evidence for density dependence in natural populations, is difficult.

Another aspect of density dependence is to reduce population growth rate at low population levels, perhaps through difficulties in finding a mate. This concept, the Allee effect, provides a lower limit to popula-

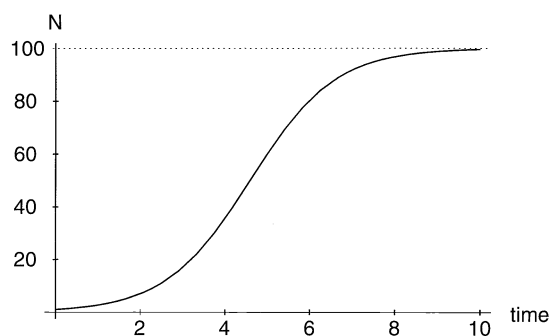


FIGURE 1 Explicit solution of the logistic equation with  $r = 1$ ,  $K = 100$ , and  $N(0) = 1$ .

tion size. If the population is too small, it may actually decline and thus can go extinct.

## B. Discrete Time

An alternate formulation based on discrete time leads to very different conclusions. Here the idea is that the number of individuals next year will simply be a function of the number of individuals this year, with density dependence. The history of models like this began with examples from fisheries, such as the Ricker model:

$$N_{t+1} = N_t e^{r(1-N_t/K)}.$$

The dynamics and behavior of this model are captured in the behavior of a much simpler model, the quadratic model, usually written as follows:

$$x_{t+1} = rx_t(1 - x_t),$$

where  $x_t$  is a scaled measure of population size and  $r$  is a measure of the rate of population growth. The density-dependent or carrying capacity parameter is absorbed in the scaling of  $x_t$ . Beginning with work of Robert May in the early 1970s, much of the interesting behavior of this model and its implications for population dynamics were elucidated.

The dynamics of density dependence in discrete time differ greatly from those in continuous time because of the presence of the implicit time delay imposed by the discrete time framework. As the population tends toward equilibrium, it can overshoot the equilibrium, with striking consequences. The overall dynamics of the quadratic model as a function of the parameters is as follows. If the growth rate is too low,  $r < 1$ , the population will die out. If  $1 < r < 2$ , the population monotonically approaches an equilibrium.

For larger values of the growth rate, the dynamics differ from the continuous time model. If  $2 < r < 3$ , the population approaches an equilibrium, but in an oscillatory fashion. For  $r > 3$ , different dynamics emerge. At first the population exhibits stable cyclic behavior of alternating low and high years. For still larger values of  $r$ , the population has cycles of longer period, until above about  $r = 3.57$ . For these large values of  $r$ , the population is typically chaotic, meaning that the population fluctuates and exhibits very sensitive dependence on initial conditions.

The search for this kind of behavior in natural populations has proven very difficult, but the attempts have greatly increased our knowledge of population dynamics. Laboratory populations including flour beetles have

been shown to behave chaotically. Further study of this phenomenon is discussed later in the context of disease dynamics.

## V. ROLE OF STOCHASTICITY

The discussion so far has assumed that no chance, or stochastic, factors play a role in the dynamics of the population. In many cases, this is far from the truth. For very small populations, demographic stochasticity, resulting from the chance that many births or deaths in a row might occur, has a large effect on the dynamics. In small populations, extinction may occur by chance even if the expected outcome would be for the population to grow.

A second important source of variability is the influence of the environment. Random fluctuations including year-to-year variability in weather, can cause large changes in the growth rate of a population. These effects need to be taken into account to understand the dynamics of natural populations. Historically, ecologists phrased this as a dichotomy, asking whether density independent (e.g., weather) or density-dependent factors were responsible for controlling populations. The more modern version of this question is discussed later in the context of disease dynamics.

## VI. THE SIMPLE TWO-SPECIES INTERACTIONS

### A. Overview and Definitions

Thus far, we have focused only on effects of the environment or within a single species, intraspecific interactions, on population dynamics. A fundamental question is, what is the role of interspecific interactions on population dynamics? The consequences of the interactions among many species fall within the realm of community ecology, but we can understand many of the simple effects of species interactions by taking a population dynamic approach, starting with the interactions between two species. This study will begin to highlight the important issue of which interspecific interactions act to prevent exponential growth of natural populations. The study of the interactions between two species has a long history in ecology, beginning with the seminal theoretical work of the American Lotka and the Italian Volterra, and the experimental work of the Russian microbiologist Gause.

The first step in a study of two species interactions

is to classify the interactions, primarily according to the effects each species involved has on the other. We will then outline the dynamical effects of each interaction. An interaction where one species eats another, such as wolves preying upon moose, is called a predator-prey interaction. The predator gains a benefit and the prey is harmed. Interactions between herbivores and plants also are similar, so these interactions can more generally be called exploiter-victim or consumer-resource. Interactions in which one species is harmed and one gains we will treat separately because their dynamics are different or special. One of these is the interaction between disease-causing organisms and their hosts, and the other is the interaction involving insects (parasitoids) that lay their eggs inside developing stage of other insects (hosts).

A second important two-species interaction is between two species that share a resource, such as two species that both eat the same food items. This is a competitive interaction, where each species reduces the growth rate of the other. The resource that is the object of competition can also be nonbiological, such as light or water for plants or nutrients. Other competitive interactions might be more active and involve the production of chemicals, which would harm the competitor.

A third kind of two-species interaction is one in which both species benefit, as might occur between a pollinator and the species it pollinates. The pollinator gains, typically, a food resource, and the species being pollinated is aided in its reproduction. This is known as mutualism.

This classification does not exhaust possible two-species interactions, but these are the major interactions that have been the subject of most study. Many of the dynamical consequences of other interactions can be deduced by considering the cases we have outlined.

## B. Predation

We now begin with a discussion of the dynamical effects of each interaction, beginning with the predator prey interaction. The description of these dynamics began with the work of Lotka and Volterra, and we start here with the simplest case they considered. We assume that in the absence of the predator, the prey population grows exponentially, in accordance with our examination of the single species case earlier. In the absence of the prey, we assume that the predator population declines exponentially. To complete the description of the dynamics, we need only describe the interaction between the predator and the prey and the consequences for the dynamics of both populations. The

simplest case is to assume that encounters between predator and prey are random, so they occur at a rate proportional to the size of each population. We also assume that there is a fixed probability that the prey is killed in such an interaction, and that each prey individual that is killed contributes a fixed amount to the rate of growth of the predator population. These assumptions lead to the classic equations,

$$dH/dt = rH - bHP$$

and

$$dP/dt = cbHP - kP,$$

where  $H$  is the size of the prey population,  $P$  is the size of the predator population,  $r$  is the intrinsic rate of increase of the prey in the absence of the predator,  $k$  is the death rate of the predator population in the absence of the prey, the encounter rate is proportional to  $b$ , and  $c$  measures the conversion of prey deaths to increases in the predator population.

The analysis of the model shows that there is an equilibrium, but that the equilibrium is neutrally stable. Thus, all other solutions cycle. From this analysis we draw the important conclusion that predator prey interactions lead to cycles, but the model is unsatisfying because the solution depends completely on the initial conditions. The cycles essentially result from delayed feedbacks.

Modifying the model changes the conclusions from the unsatisfying one of neutrally stable cycles. If the equations are modified so that there is density dependence in the prey population, the system approaches an equilibrium in an oscillatory fashion. Changing the description of the consumption of prey by predators can then produce stable cycles. This presence of cyclic behavior is the most important conclusion from the study of predator-prey interactions.

There are numerous examples of interactions between predator and prey in natural systems, which at least exhibit a tendency toward cyclic behavior, ranging from the classic example of hare and lynx in Canada to wolves and moose on Isle Royale. The observations in many of these cases are somewhat equivocal because it is often difficult to determine what is really causing the cycles, partly because it is difficult to isolate a single predator prey pair within a larger community.

## C. Host-Parasitoid Dynamics

By contrast, there is another kind of exploiter victim relationship that does have more tightly coupled species interacting. The overwhelming majority of animal spe-

cies are insects, and approximately one-fifth to one-sixth of all insect species are parasitoids: they have a life cycle where they lay their eggs in the developing stages of other insects, the hosts, and consume them from within so the host is killed and one or more adult parasitoids emerge. Many insect pests of crops have important parasitoids. The interaction between a host and a parasitoid is essentially another version of the interaction between a prey and a predator, but the approach and the questions are somewhat different, so we treat this case separately. Here, the typical approach is discrete in time, which is appropriate for host insect species with single generations per year. The host species is assumed to grow exponentially (or geometrically) in the absence of the parasitoid. Unlike the predator-prey interaction, the parasitoid life cycle is tightly coordinated with the host because the parasitoid lays its eggs in the developing host. Thus the dynamics of the interaction are described by the probability that an individual host is parasitized, which could depend on the population density of the host and of the parasitoid.

The equations describing this interaction were first developed by Nicholson and Bailey, but these initial equations predicted that the populations would be unstable with cycles of ever-increasing amplitude. Thus, efforts since then have been focused on determining what would allow host and parasitoid to persist in a stable fashion. Suggestions like aggregation of host or parasitoid, or density dependence have been studied in models. Yet, as recent careful work by Murdoch and his colleagues on red scale and its parasitoids in California shows, the search for the stabilizing feature in natural systems is quite difficult. This is a system where a host and its parasitoids coexist stably, yet the search for the mechanisms that stabilize this interaction has been able to reject many of the conventional explanations.

#### D. Disease Dynamics

Another very important interaction that is similar to the predator-prey system is the interaction between diseases caused by bacteria or viruses and their hosts. However, in this case the focus is very different, since we do not count the number of the disease organisms and instead count only the number of diseased hosts. These systems have become very important subjects of study in population biology partly because of the very good records of the number of diseased individuals, in particular for human diseases.

The classic models in this area, which go back to the early work of Kermack and McKendrick in the 1920s, break up a population into three categories: sus-

ceptible, those individuals that can become infected; infectives, those individuals which are infected and infect others; and removed, those individuals which no longer can infect others. The simplest case is to look over relatively short times, so that overall population sizes are constant, and look at epidemics. The models, called SIR models, are framed in terms of differential equations where the key parameters are the contact rate, the recovery rate, and the total population size. The rate of infection is assumed to be equal to the product of the contact rate, the number of susceptibles, and the number of recessives. The recovery rate for individuals is assumed to be just a constant. Then the dynamics depend on a single nondimensional parameter, the reproductive number,  $R_0$ , which is the number of new infectives produced by each infective, when the population is essentially composed of all susceptible individuals. If  $R_0$  is less than one, the epidemic dies out, and if  $R_0$  is greater than one, the epidemic will grow. This simple model fits many epidemics very well, as shown by the original work of Kermack and McKendrick looking at the 1905–1906 Bombay plague epidemic.

Disease dynamics have become central to some of the most important current questions in population ecology—namely, the relative importance of endogenous (within the system) versus exogenous (outside the system, often called stochastic) forces in determining the dynamics of population. This question, which is central to understanding the importance of any of the interactions outlined in this chapter, has been dealt with most clearly within the realm of childhood disease (chicken pox, measles, whooping cough, etc.) dynamics because the quality of the data is so good and because the models are more mechanistic.

The data, such as that for measles, have several important features, exhibiting both seasonal behavior driven in part by the school cycle and more complex multiyear cyclic behavior. There have been many different attempts to use extensions of the SIR model to understand the dynamics in a quantitative fashion. The approach taken is essentially to explicitly include the stochastic forces as an additional term in the model and to find the best fit (in a statistical sense) of the model to the data. This procedure illustrates the importance of using these parameter estimation approaches since the standard SIR model with constant infectivity does not provide a very good fit to the data, implying that the model is inadequate in some way. Further investigation shows that making the infectivity seasonally variable, corresponding to the role of school vacations in reducing the likelihood that children infect each other, greatly



improves the fit of the model. Moreover, the rather complex dynamical patterns in the disease dynamics are shown to be generated by a combination of exogenous and endogenous forces: both the stochasticity and the nonlinearity (and seasonality) in the SIR model are critical to producing the observed dynamics.

### E. Competition

The simplest models of competition are essentially phenomenologically based and start with the logistic model for a single species as their basis. The effect of each species on the other is to reduce the per capita growth rate. From this simple model of competition, several important ecological paradigms have emerged. The basic question to ask is what allows two species to coexist, rather than having a superior competitor eliminate an inferior one. The outcome of competitive interactions ends up being determined, in part, by the relative importance of interspecific effects as compared to intraspecific effects. Coexistence requires that the interspecific effects be weaker than the intraspecific effects.

In contrast to the predator-prey interaction, the dynamics of competition view is a monotonic approach to equilibrium without any cycles. However, if interspecific effects are strong enough, there can be a kind of priority effect. The species that wins in competition may depend on the initial densities in that case.

A more mechanistic approach to competition sheds some light on the issue of coexistence. With interference competition, in which the two species might, for example, affect each other by the production of allelopathic chemicals, interspecific effects may be stronger than the intraspecific ones. With exploitation competition, in which the two species use common resources, the relative strength of interspecific effects and intraspecific effects depends on the differences in resource use—for example, differences in the size of seeds used. From this view comes what was an extraordinarily important principle, the competitive exclusion principle of Gause, which stated that competing species have to differ in order to coexist.

The search for evidence of competition in natural populations is also difficult, but several large-scale surveys have demonstrated that there is relatively good evidence for competition. Classic work in this area includes careful studies of lizards in the Caribbean, which

has shown competition for food and avoidance of competition by different species perching at different heights in trees. Competition for space has also been well documented in intertidal systems, in which removal of one species allows another to expand the range of habitat it uses. Recent work has also focused on apparent competition, in which one species “competes” with another by allowing a common predator to increase its density, which then suppresses the competitor.

## VII. CONCLUSIONS

Population dynamics is one of the fundamental areas of ecology, forming both the basis for the study of more complex communities and of many applied questions. Understanding population dynamics is the key to understanding the relative importance of competition for resources and predation in structuring ecological communities, which is a central question in ecology.

Population dynamics plays a central role in many approaches to preserving biodiversity, which until now have been primarily focused on a single species approach. The calculation of the intrinsic growth rate of a species from a life table is often the central piece of conservation plans. Similarly, management of natural resources, such as fisheries, depends on a population dynamics approach to determine the largest sustainable yield.

### See Also the Following Articles

DISEASES, CONSERVATION AND • PARASITIDS •  
POPULATION DENSITY • POPULATION GENETICS •  
PREDATORS, ECOLOGICAL ROLE OF • SPECIES  
INTERACTIONS

### Bibliography

- Begon, M., Mortimer, M., and Thompson, D. J. (1996). *Population Ecology*, 3rd ed. Blackwell Science, Oxford.
- Crawley, M. J. (Ed.) (1992). *Natural Enemies*. Blackwell Scientific Publications, Oxford.
- Hanski, I., and Gilpin, M. E. (Eds.) (1997). *Metapopulation Biology*. Academic Press, San Diego, CA.
- Hastings, A. (1997). *Population Biology. Concepts and Models*. Springer-Verlag, New York.
- Quinn, T. J., III, and Deriso, R. B. (1999). *Quantitative Fish Dynamics*. Oxford University Press, Oxford.



# POPULATION GENETICS

Brian Charlesworth  
*University of Edinburgh*

---

- I. Variation within Populations
  - II. Deterministic Population Genetics
  - III. Random Genetic Drift
  - IV. The Interaction of Drift with Deterministic Forces
  - V. Conclusions
- 

## GLOSSARY

- allele frequency** The frequency of a variant form of a genetic locus within a population.
- genetic drift** Evolutionary change caused by random sampling of genotype frequencies in a finite population.
- genotype** The state of an individual with respect to a defined genetic locus or set of loci.
- heritability** The proportion of the variance in a trait that is due to additive genetic effects.
- inbreeding** Matings between close relatives.
- mutation rate** The frequency with which new mutations arise per generation.
- neutral mutations** Mutations whose effects on fitness are either nonexistent or so small that their fate is controlled by genetic drift rather than selection.
- phenotype** The state of an individual with respect to a trait of interest.
- polymorphism** The existence at intermediate frequencies of two or more variants at a locus within a population.

**selection** The differential survival or reproductive success of individuals, associated with differences in phenotype or genotype.

---

*DARWIN'S THEORY OF "DESCENT WITH MODIFICATION"* implies that all of the stupendous diversity of life on Earth is ultimately traceable to genetic diversity within populations. The study of the nature and causes of within-population variation, and of the mechanisms by which it is transformed into differences between populations over space and time, is the province of population genetics. The subject involves both theoretical modeling of evolutionary processes, based on knowledge of the mechanisms of inheritance, and the testing of these models using data on variation and evolution in natural and artificial populations.

## I. VARIATION WITHIN POPULATIONS

### A. Types of Phenotypic Variation

Since evolutionary change depends on the existence of genetic variation within populations, measurement of the extent of such variation is crucial (Lewontin, 1974). Variation at the level of externally visible phenotypes can be divided into three categories.

#### 1. Discrete Variation

This involves traits which can be divided into a small number of discrete categories, such as eye color in hu-

mans or shell color and pattern in the land snail *Cepaea* (Ford, 1975). It is often controlled by one or a few genes, and usually it involves relatively superficial traits such as color patterns. Only a relatively small proportion of the phenotypic variation of interest to evolutionists is of this kind.

## 2. Quantitative Variation

Quantitative variation is all-pervasive. This can involve either meristic traits, such as bristle number in *Drosophila*, in which there is a large number of discrete categories, or continuously varying metrical traits such as body size. Variation in typical quantitative traits is known to be under the joint control of environmental effects, accidents of development, and sets of genes whose individual effects are small relative to the total range of variation in the traits (Falconer and Mackay, 1996). Statistical methods that utilize the degree of resemblance between close relatives enable the determination of the proportion of the total phenotypic variation that is contributed by additive genetic causes—the heritability, which controls the rate of response to selection on a trait (see Section II,D). Heritabilities for quantitative traits typically are between 20 and 80%, corresponding to the fact that artificial selection is highly effective in changing the mean value of almost every trait that has been examined (Falconer and Mackay, 1996).

## 3. Concealed Variability

A more subtle form of phenotypic variation is concealed variability, i.e., variability that is only exposed when homozygous genotypes are produced by close inbreeding. This is responsible for the increased variation among inbred lines when a set of such lines is made from a random-bred base population and for inbreeding depression, which is the decline in the mean values of fitness-related traits such as viability and fecundity with inbreeding (Falconer and Mackay, 1996). Both of these phenomena reflect, at least in part, the widespread occurrence of recessive or partially recessive rare alleles in random-mating populations (see Section II,F), whose phenotypic effects are only fully exposed when they are made homozygous and are therefore not evident in randomly mating populations.

In *Drosophila*, special breeding methods involving the use of genetically marked chromosomes with inversions that suppress crossing over have shown that up to 50% of haploid genomes carry recessive lethal genes. These contribute about half the inbreeding depression manifested when fully homozygous genotypes are produced; a similar magnitude of inbreeding depression is

caused by genes of small effect (Crow, 1993). The net fitness of fully inbred *Drosophila* is only a few percent of that of outbred flies. Even more extreme effects of complete inbreeding are likely in vertebrates, which have much larger genomes. The deleterious fitness effects of inbreeding have probably played a major role in promoting the evolution of mechanisms of inbreeding avoidance, such as the self-incompatibility loci of flowering plants.

## 4. Interpreting Phenotypic Variation

The previously mentioned basic facts were established by the early 1950s and led to an active debate concerning the causes of natural variation (Lewontin, 1974). In one view, championed by H. J. Muller, variation is mostly due to rare deleterious alleles maintained by mutation pressure at a large number of loci; the coexistence of alleles at a locus at intermediate frequencies (polymorphism) is characteristic of only a small number of loci. In the other view, advocated by Dobzhansky, polymorphism is the norm, and it reflects variation that is actively maintained by selection. In the absence of any means of identifying loci without the prior existence of genetic variability, no unbiased survey of the extent of genetic variation at individual loci was possible with the methods of classical genetics, and so this question could not be answered.

# B. Molecular Variation

## 1. Protein Electrophoresis

The previously mentioned situation was transformed by the development of molecular genetics. Gel electrophoresis of soluble enzymes and proteins provides a rapid and simple method for surveying populations for variants affecting the structure of a large number of different proteins and hence genes (Lewontin, 1974; Hartl and Clark, 1997). The results of such surveys reveal that a high fraction of loci coding for soluble proteins are polymorphic in the sense of having at least one rare variant whose frequency exceeds 5%; the average individual from a randomly mating population is typically heterozygous for a significant fraction (several percent) of such loci. Despite some biases in the methodology, particularly the inability of electrophoresis to detect many types of amino acid sequence changes and the restriction of the method to soluble proteins, it is clear that protein polymorphism is not an exceptional situation.

## 2. Measurement of Variation in DNA Sequences

The introduction of recombinant DNA technology has meant that population geneticists can now study variation at the level of the nucleotide sequence. Surveys of within-species DNA sequence variation of nuclear genes have been most intensively carried out in *Drosophila*, but comparable results are emerging from other species (Li, 1997). The basic conclusion is that variants due to single nucleotide changes are the most abundant source of variation in natural populations. For silent substitutions in third coding positions, which do not change the amino acid sequence, and for changes in introns and flanking sequences, the probability that two randomly chosen alleles from a *Drosophila* population differ at a given site (the nucleotide site diversity) is typically of the order of 1% or a few percent, depending on the species. The level of this type of variability is about 10 times higher in the bacterium *Escherichia coli* and one-tenth as high in humans (Li, 1997). For most genes, diversity is much lower for replacement changes, which alter the amino acid sequence. In addition to single nucleotide polymorphisms, DNA variability is contributed by small insertions and deletions of sets of nucleotides and by insertions of transposable elements, mostly in noncoding regions. Other types of variability include variation in the sizes of tandem arrays of microsatellite and minisatellite loci, which are often highly polymorphic and provide useful genetic markers (Bruford and Wayne, 1993; Hartl and Clark, 1997). The density of such loci, however, is low in relation to the total size of the genome. The total level of variability at the level of DNA sequences is about two orders of magnitude greater than that revealed by electrophoresis because of the high degree of variability at silent and noncoding nucleotide sites relative to replacement sites.

## 3. Interpreting DNA Sequence Variation

Although variation at the level of DNA sequences must underly heritable phenotypic variation, it is difficult to relate the two, except for intensively studied human genetic diseases. The abundance of variation in both protein and DNA sequence might seem to vindicate Dobzhansky's view of the causes of natural variation. However, it is likely that much of the silent and noncoding variability is close to neutrality with respect to effects on fitness (see Section III,C). Nevertheless, there is a real possibility that selection also frequently influences variation and evolution in protein and regulatory sequences (Hartl and Clark, 1997; Li, 1997). The role of deterministic forces in variation and evolution within populations will thus be considered next.

## II. DETERMINISTIC POPULATION GENETICS

### A. Allele and Genotype Frequencies

If we focus on a given nucleotide position, the basic descriptor of the state of a population is the set of frequencies of the four alternative states, A, T, G, and C. If recombination within a gene is ignored, we can consider the set of all nucleotide sequences observed at a locus as alternative alleles, whose frequencies characterize the state of the population with respect to this locus. Mendelian inheritance implies that this state is not changed in the absence of evolutionary forces (the occurrence of intragenic crossing over and gene conversion at low frequencies means that in practice this is a good approximation rather than an exact description). This is enshrined in the Hardy–Weinberg principle, which states that the frequencies of diploid genotypes in a random mating population with a set of  $n$  alleles with frequencies  $p_1, p_2, \dots, p_n$  rapidly reach equilibrium values given by the multinomial expansion of  $(p_1 + p_2 + \dots + p_n)^n$ . The importance of this result is that existing natural variation is preserved by Mendelian inheritance. This removes Darwin's difficulties over the rapid loss of variation under blending inheritance, which led him to adopt a theory of the inheritance of acquired characteristics for which there is no empirical foundation (Fisher, 1930).

### B. Mutation

#### 1. Types of Mutations and Their Rates

The ultimate source of natural variation is known to be spontaneous mutation, defined as a heritable change in the genetic material and which occurs without reference to the adaptive utility of the phenotypic consequences of the change in question. The most abundant mutations are nucleotide substitutions, but small deletions and insertions due to slippage during DNA replication are relatively common as well (Drake *et al.*, 1998). Insertions of transposable elements, large deletions and duplications, duplications and deletions of entire chromosomes or haploid genomes (aneuploidy and polyploidy), and chromosome rearrangements such as inversions and translocations also occur and contribute to evolutionary changes in genome structure. Rates of spontaneous mutation in organisms with DNA genomes are extremely low due to the operation of complex enzymatic systems which repair lesions in DNA; the rates of nucleotide changes per site per cell generation

in DNA-based microbes are between  $1 \times 10^{-10}$  and  $5 \times 10^{-10}$ . Similar values apply to higher eukaryotes such as *Drosophila* and humans, but the rate per organism generation is between 10 and 50 times higher owing to the many cell divisions that occur during the production of germ cells. Per locus rates of mutation to alleles with major phenotypic effects are substantially higher—on the order of  $10^{-5}$  per generation in *Drosophila*, mice, and humans (Drake *et al.*, 1998). This is not surprising given the large number of nucleotides in the coding and regulatory regions of typical loci. Rates of change in copy numbers in microsatellite and minisatellite loci in mammals are much higher, however, up to  $10^{-3}$  per locus per generation (Bruford and Wayne, 1993). Mutation rates in viruses with RNA genomes are also extremely high due to their lack of repair mechanisms (Drake *et al.*, 1998).

## 2. Evolution under Mutation Pressure

Given the low rate of mutation at the nucleotide level in DNA-based organisms, the time scale of mutational change in the frequencies of the four alternative states at a nucleotide site is very large—on the order of 1 billion generations. Mutation pressure at this level is thus an extremely weak force and is easily opposed by other evolutionary factors. For most purposes, therefore, mutation can be regarded simply as a source of new variation and as unimportant as a cause of directed evolutionary change (Fisher, 1930). This statement needs to be qualified when the aggregate effects of mutations affecting a particular phenotype are considered; the numbers of loci affecting a single quantitative trait are sufficiently large that increases in variability due to mutation are detectable in stocks that are initially genetically uniform. The rate of increase in variance per generation is typically on the order of  $10^{-3}$ , relative to the nongenetic variance in the trait (Falconer and Mackay, 1996). Given that fitness-related traits are affected by many genes, and that most phenotypic changes caused by mutations are harmful to the organism, there is a tendency for the mean values of fitness components to decline under mutation pressure when selection is relaxed by experimental manipulations (Crow, 1993; Drake *et al.*, 1998).

## 3. Mutation and Selection

These considerations imply that a relatively weak force of selection, far smaller than is measurable experimentally, can prevent the spread of deleterious mutations at a locus. This explains the fact that amino acid polymorphisms are usually much less frequent than silent or noncoding polymorphisms. The same obser-

vation applies to comparisons of sequences between different species (Kimura, 1983). A very important form of selection is thus purifying selection, whereby deleterious mutations are constantly being eliminated from the population (see Section II,F). Similarly, quantitative traits are often subject to stabilizing selection, such that individuals with extreme trait values are less fit than individuals with intermediate values (Falconer and Mackay, 1996). Variability in quantitative traits is maintained, at least partially, by a balance between the input of new variation by mutation and its elimination by stabilizing selection (Falconer and Mackay, 1996).

## C. Selection at a Single Locus

### 1. The Basic Model

A large body of theory has been developed to describe the action of natural selection on Mendelian variation. In the simplest case, a single locus with a pair of alternative alleles, A and a, is postulated. A randomly mating, infinitely large population is assumed. Ignoring sex differences, the relative fitnesses of the three possible genotypes AA, Aa, and aa in a diploid can be written as 1,  $1 - hs$ , and  $1 - s$ , respectively, where  $s$  is the selection coefficient and  $h$  is the dominance coefficient.  $s$  measures the strength of selection ( $s > 0$ , when a is disfavored by selection);  $h$  measures the extent to which the fitness of the heterozygote is reduced by the presence of a. The fitnesses are assumed to be constant over time.

Fitness in this context is most easily understood in terms of viability selection in a population with discrete generations, where the relative probabilities of the three genotypes surviving from egg to adult are equivalent to the three fitnesses. In general, selection may involve many different aspects of the life history, especially female fecundity and male mating success. More elaborate models have been developed to study these cases, including extensions of the theory to populations with overlapping rather than discrete generations, but the basic conclusions are similar (Crow and Kimura, 1970; Ewens, 1979; Hartl and Clark, 1997).

If the frequencies of the two alleles in one generation are  $p$  and  $q$ , respectively, the change in frequency of A over one generation is

$$\Delta p = spq \{q + h(1 - 2q)\} / \bar{w}, \quad (1a)$$

where  $\bar{w} = 1 - 2pqhs - q^2s$  is the population's mean fitness.

An interesting equivalent form, from Wright

(1977), is

$$\Delta p = \frac{1}{2} pq \frac{d \ln \bar{w}}{dp}. \quad (1b)$$

Equation (1b) implies that gene frequency change is in the direction of the gradient of mean fitness with respect to gene frequency. A more detailed analysis of the dynamics of a single locus with an arbitrary number of alleles, and arbitrary but constant fitnesses, shows that mean fitness increases monotonically as allele frequencies change so that stable equilibria in gene frequency space correspond to local maxima in mean fitness: This is known as Wright's adaptive landscape (Crow and Kimura, 1970; Wright, 1977; Ewens, 1979; Hartl and Clark, 1997).

## 2. Directional Selection

Regarding the case of a pair of alleles, when  $0 \leq h \leq 1$  (so that the fitness of the heterozygote is bounded by the fitnesses of the homozygotes), A is favored by selection and will progress to fixation. This case is referred to as directional selection. If A is initially rare so that second-order terms in  $p$  can be neglected. Eq. (1) can be approximated by  $\Delta p = s(1-h)p$ . The initial rate of change of log gene frequency ( $\approx \Delta p/p$ ) is therefore proportional to the fitness difference between the genotypes Aa and aa. This reflects the fact that, with random mating, rare alleles are overwhelmingly carried in heterozygotes (frequency  $2pq$ ) since the frequency of homozygotes ( $p^2$ ) is negligible. If A is completely recessive so that  $h = 1$ , Eq. (1a) for rare A is approximated by  $\Delta p = sp^2$  so that the logarithmic rate of increase in frequency of A is proportional to  $sp$ , which tends to zero with decreasing  $p$ . This reflects the extreme rarity of the favored homozygotes, and it implies that rare recessive alleles are only weakly selected in randomly mating populations (Haldane, 1932).

## 3. Survival of Favorable Mutations

The previous conclusion is reinforced by calculation of the probability of survival of new mutations. Even in very large populations, a new mutation is likely to be represented initially in only one or a few individuals. Since reproduction is subject to random variation, described by the probability distribution of the numbers of surviving offspring per mated individual, there is a finite chance in each generation that all carriers of a rare mutant gene will fail to transmit it (Fisher, 1930; Haldane, 1932). The chance that a single copy of a new favorable mutation ultimately survives random loss from a large population of constant size is approximately  $2s(1-h)$ , assuming a Poisson distribution of

offspring number (Haldane, 1932; Crow and Kimura, 1970). Most favorable mutations are thus likely to be lost from the population a few generations after they arise; on average, 148 occurrences of a mutation with a heterozygous selective advantage of 1% are needed for a 95% chance that one will survive.

This implies that there is a considerable random element to adaptive evolution at the genetic level. If there are several different loci at which mutations that provide adaptations to a given pressure of selection can occur, it may be a matter of chance which locus actually responds to selection in a given population. The same pressure of selection can therefore result in the divergence of isolated populations at the level of the genotype, even if the same phenotype is evolving. The numerous different genes involved in the adaptation of different human populations to malaria parasites illustrate this principle (Hill and Weatherall, 1998).

The formula for survival probability implies that a recessive favorable mutation has a zero chance of survival in an infinite population; calculations based on diffusion theory (see Section IV,D) show that the probability for a randomly mating population of size  $N$  in this case is approximately  $0.8\sqrt{s/N}$  (Crow and Kimura, 1970). This means that recessive autosomal mutations are unlikely to become established by selection in randomly mating populations of even moderate size. As first pointed out by Haldane (1932), it is thus no accident that nonrecessive alleles have been established by selection in cases of recent adaptation involving genes of major effect, such as industrial melanism and pesticide resistance, despite the fact that most spontaneous mutations are recessive with respect to their effects on the phenotype (Haldane, 1932; Ford, 1975).

## 4. Time Course of Gene Frequency Change

Once a favorable allele rises to a sufficiently high frequency that random loss is unlikely, its progress in a large population can be calculated by integration of the differential equation which approximates Eq. (1) when selection is weak. This procedure yields expressions for the time needed to change gene frequencies by a given amount (Haldane, 1932; Crow and Kimura, 1970; Hartl and Clark, 1997). These can be used to estimate selection intensities in experimental and natural populations by comparing observed trajectories of gene frequency change with the theoretical predictions. This method is particularly useful for microbes, which have short generation times and can be grown in very large artificial populations. Selection coefficients on the order of 0.5% can be measured in microbial experiments (Hartl and Clark, 1997).

The results of these calculations show that the time

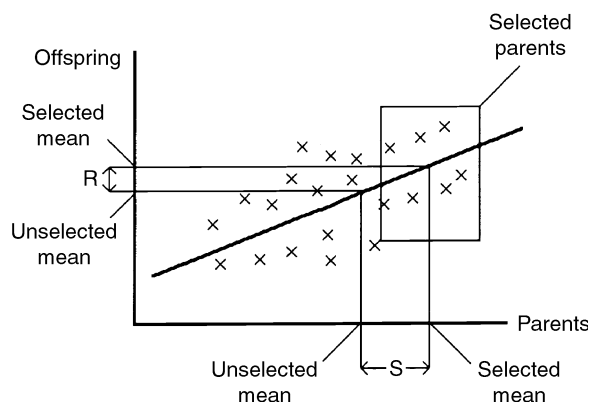


FIGURE 1 Each cross indicates the location of a point determined by the mean values of the phenotypic values of a pair of potential parents ( $x$  axis) and of their progeny ( $y$  axis). The straight line is the regression of  $y$  on  $x$ , whose slope is equal to the heritability. Families that are successful in reproducing are enclosed in the rectangular box.

required to cause a given amount of change in gene frequency is inversely proportional to  $s$  and depends only logarithmically on the initial frequency of the favorable allele, except when it is recessive. This implies that selection on nonrecessive alleles can cause evolutionary change on a timescale of the order of a few multiples of  $1/s$ , almost regardless of the initial conditions. Even a very weak force of selection can thus transform a population in a period of time that is negligible in geological terms (Fisher, 1930; Haldane, 1932). This conclusion is historically very important because it removes many of the objections to natural selection as a potent force in evolution.

## D. Selection on Quantitative Traits

### 1. Predicting the Response to Selection

This question can also be considered in relation to quantitative traits, which are more relevant to phenotypic evolution than traits controlled by single genes. The standard model of quantitative genetics (which involves several simplifying assumptions) implies that the response to selection on a single quantitative trait is governed by the equation

$$R = h^2 S, \quad (2)$$

where the selection response,  $R$ , is the difference in mean value between one generation and the next;  $h^2$  is the heritability of the trait (see Section I,A,2) (there is no connection with the dominance coefficient introduced previously); and  $S$ , the selection differential, measures the intensity of directional selection applied to the trait and is equal to the difference in mean between the

parental individuals who have survived selection and the population mean before selection (Fig. 1). Given the fact that  $h^2$  is usually substantially greater than zero, this implies that populations can respond rapidly to directional selection on quantitative traits, as is indeed the case (Falconer and Mackay, 1996).

### 2. Rates of Change under Selection

In a sexually reproducing population, the ability of recombination to put together new combinations of alleles means that selection can eventually produce genotypes with trait values that far exceed the range of variation that existed in the original population, simply by causing the fixation of alleles which increase trait value and that were segregating in the original population. In the absence of new variability created by mutation, the selection response will eventually cease when all favorable alleles have been fixed by selection. Given the evidence for a significant input of new variation by mutation for typical quantitative traits, a relatively gentle pressure of selection in a large population will never fully deplete variation and therefore sustained responses to selection are possible (Falconer and Mackay, 1996). The breeder can produce substantial changes in a single trait over less than 100 generations, comparable in magnitude to geologically rapid changes that have taken thousands of generations in nature. It is therefore unlikely that lack of genetic variation is often a serious constraint on the ability of populations to undergo phenotypic evolution, unless there is intense selection for a novel combination of traits, which cannot immediately be produced by recombination among existing genotypes (Wright, 1977).

### 3. Selection on Multiple Characters

In general, natural selection acts on suites of characters, not on a single trait in isolation. A multivariate generalization of Eq. (2) has been derived by Lande (1988), and it is useful for the interpretation of data on selection on multiple traits:

$$\Delta \bar{z} = G \nabla \ln \bar{w} \quad (3)$$

where  $\Delta \bar{z}$  is the change per generation in the vector of mean values for the set of traits,  $\nabla \ln \bar{w}$  is the gradient vector of log mean fitness with respect to the set of trait means, and  $G$  is the matrix of additive genetic variances and covariances among the set of traits. The similarity to Eq. (1b) is evident.

Equation (3) shows that, in general, we can only predict the effect of selection on a trait if we know the extent to which it is genetically correlated with other traits that are also the target of selection. Short-term evolutionary changes in a given trait may be due, at

least in part, simply to the fact that it covaries with another trait that is the target of selection. However, provided that the  $G$  matrix is nonsingular and fitnesses are constant in time, a stable equilibrium state corresponds to a maximum in the surface of mean fitness with respect to the mean trait values, indicating that selection will ultimately bring the population close to the optima with respect to each individual trait unless there are strong constraints on the structure of genetic variation. This implies that the concept of an adaptive landscape can be applied to quantitative traits (Lande, 1988).

## E. Maintenance of Variation by Selection

The role of selection in preserving variation rather than destroying it, called balancing selection, is now discussed. This form of selection was unknown to Darwin, and its discovery is one of the most important contributions of population genetics.

### 1. Heterozygote Advantage

The simplest case is when the heterozygote at a locus with two alleles has a higher fitness than the two homozygotes, which was first investigated by Fisher in 1922. Let the fitnesses of  $AA$  and  $aa$  relative to  $Aa$  be written as  $1 - s$  and  $1 - t$ ; Eq. (1b) now yields the result that there is a unique stable equilibrium at which the frequency of  $A$  is  $p^* = t/(s + t)$ , to which the population converges from any starting point other than fixation for  $A$  or  $a$  (Crow and Kimura, 1970; Ewens, 1979; Hartl and Clark, 1997). The classic example of heterozygote advantage is the polymorphism for the  $\beta$ -globin variant in humans that causes sickle-cell anemia when homozygous, a disease which is effectively lethal under natural conditions. The maintenance of this allele in human populations subject to severe malarial infections is due to the selective advantage of heterozygous carriers, conferred by their resistance to malaria, as suggested by Haldane in 1948. Several human globin gene polymorphisms, as well as some other human polymorphisms, are also maintained by resistance to malaria (Hill and Weatherall, 1998).

### 2. Frequency-Dependent Selection

Balancing selection can also be caused by negative frequency-dependent selection, in which the relative fitness of a genotype decreases with its frequency in the population. This obviously acts to inhibit the fixation of an allele which may initially have a selective advantage over other alleles when introduced into a population. Genetically controlled resistance to parasitic dis-

ease is subject to this form of selection since the abundance of a parasite decreases as the number of susceptible hosts diminishes, thereby reducing the selective advantage to a host allele that causes resistance to a particular parasite genotype (Li, 1997). A similar frequency dependence may affect alleles controlling virulence in the parasite population. Genes thought to be involved in disease resistance, such as the major histocompatibility complex (MHC) loci of vertebrates, are often highly polymorphic and segregate for large numbers of alleles; molecular analyses show clear evidence for balancing selection (see Section IV,C). The role of frequency-dependent selection in these cases remains to be established, however. Frequency dependence is inherent in the dynamics of the self-incompatibility loci of flowering plants, which also have very high allele numbers (Hartl and Clark, 1997). Similarly, Batesian mimicry, in which an edible species mimics a distasteful model, exhibits frequency dependence since a predator is more likely to mistake a rare mimetic form for the model than a common one (Fisher, 1930; Ford, 1975). It should be noted that frequency-dependent selection does not necessarily lead to polymorphism; its outcome depends on a delicate balance of the selective parameters, and dynamic complexities such as limit cycles or even chaos are possible. With frequency-dependent selection, maxima in mean fitness do not necessarily correspond to stable equilibria.

### 3. Temporal Variation in Fitnesses

Temporal fluctuations in relative fitnesses can also lead to the maintenance of polymorphisms by selection; in the case of two alleles at a locus, a sufficient condition is that the geometric mean fitness of the heterozygote over generations exceeds that of both homozygotes (Crow and Kimura, 1970; Hartl and Clark, 1997). Similarly, spatial variation in the direction of selection can maintain variation. This can happen in two ways. The first requires strong density-dependent regulation within different environmental patches so that the number of adults emerging from a patch is largely independent of the genetic composition of the eggs laid in that patch. In this case, opposing directions of selection in different patches can maintain polymorphism even if there is complete random mating among patches (Crow and Kimura, 1970). Directional selection in opposite directions on alleles in males and females is a special case. Second, restricted migration among populations subject to different directions of selection can result in polymorphism within each population accompanied by genetic differentiation among populations. Clinal variation in allele frequencies or quantitative trait values



results from geographic gradients in selection pressures, coupled with the smoothing effect of migration, as in the case of Bergmann's rule, which states that the mean body sizes of mammal populations increase with higher latitudes (Hartl and Clark, 1997).

#### 4. Meiotic Drive

Antagonism between the effects of selection at different levels may also maintain variation. The best studied cases involve the phenomenon of meiotic drive or segregation distortion, in which one allele (D) at a locus when heterozygous causes the destruction of gametes carrying the alternative allele (d). In animals, this is usually found to occur only in males. Provided that the fertility of Dd males is affected less than linearly by the destruction of the d sperm, D will gain a transmission advantage. It will spread through the population unless there is a countervailing selective disadvantage at the level of individuals (Hartl and Clark, 1997). In the best studied cases of this kind, the SD system of *Drosophila melanogaster* and the t-haplotype system of house mice, the primary disadvantage seems to come from sterility of DD males.

### F. Mutation–Selection Balance

The balance between recurrent mutations to variants that impair the functions of gene products and selection against them is probably a major factor in maintaining genetic variation, given the fact that higher organisms have tens of thousands of genes (Drake *et al.*, 1998). The large number of changes in a coding sequence that can impair the function of a gene product implies that the process of mutation to deleterious alleles at a locus can be regarded as effectively irreversible, provided that the wild-type allele predominates in the population. If the deleterious alleles at a locus are completely recessive, with selection coefficient  $s$  and a rate of origination by mutation from wild-type of  $u$ , their equilibrium frequency in a randomly mating population (assuming  $s \gg u$ ) is  $q^* = \sqrt{u/s}$ . Even lethal alleles ( $s = 1$ ) can thus reach appreciable frequencies if they are completely recessive; with  $u = 10^{-5}$ , for example,  $q^* = 3 \times 10^{-3}$ . However, experimental studies of lethal mutations in *Drosophila* show that they usually impair the viability of heterozygotes with wild-type by 2 or 3%; detrimental alleles with more minor effects ( $s$  on the order of a few percent) appear to be much less recessive, with  $h$  values of approximately 0.25 (Crow, 1993). With random mating, the much greater frequency of heterozygotes than mutant homozygotes means that selection on heterozygotes controls the frequencies of mutations; in this case,

$q^* = u/(hs)$ . Indirect experimental evidence indicates that the mean of  $hs$  for detrimental mutations is on the order of 1% in *Drosophila* so that  $q^*$  is approximately  $10^{-3}$  with  $u = 10^{-5}$ . The frequency of heterozygous carriers is approximately  $2q^*$  so that the mean number of heterozygous detrimental mutations per individual in a genome such as that of *Drosophila* with approximately 15,000 genes is  $30,000 \times 10^{-3} = 30$  with these assumptions. The total rate of mutation to lethal mutations per haploid genome in *Drosophila* is 0.01; assuming  $hs = 0.02$ , the mean number of heterozygous lethals per diploid individual is on the order of 1, in agreement with the direct estimate mentioned previously.

### G. Genetic Load

#### 1. General Considerations

The previous discussion leads to the consideration of the effect of selection on the fitness of the population as whole; if there is a large amount of variability with respect to loci under selection, it is obvious that the mean fitness of the population must be much less than that of the best genotype. If fitnesses are measured relative to a value of 1 for the optimal genotype in the system under consideration, the reduction in fitness can be conveniently measured by the genetic load, defined as  $L = 1 - \bar{w}$ . In the case of viability selection, no genotype can have a survival probability greater than 1, so  $L$  provides an upper bound to the probability that a zygote survives to maturity. More generally,  $L$  measures the proportion of the population that dies or fails to reproduce as a result of selective differences among genotypes. The effects of multiple loci on mean fitness can be calculated by assuming that different loci have independent effects, so that the fitness of a multilocus genotype is given by the product of the fitnesses of all the single-locus genotypes which contribute to it. If  $L_i$  is the load contributed by the  $i$ th locus, the mean fitness with respect to  $m$  independent loci, relative to the value for the optimum genotype, is:

$$\bar{w} = \prod_{i=1}^m (1 - L_i) \approx \exp - \sum_{i=1}^m L_i. \quad (4)$$

#### 2. Mutational Load

With nonrecessive deleterious alleles maintained by mutation, the load for a single locus at equilibrium is approximated by  $2q^*hs = 2u$ . The total load is  $1 - \exp - U$ , where  $U$  is the mean number of new deleterious mutations in a diploid individual. Lethal

mutations contribute relatively little to this since their total mutation rate is very low, but detrimental have a major effect: Assuming 15,000 genes with a mutation rate of  $10^{-5}$ , the mutational load for a *Drosophila* population would be about 0.26 (Drake *et al.*, 1998). For a genome of 80,000 genes, as in mammals, the load would be 0.80, a considerable burden of selective loss.

The existence of a large mutational load suggests that there is an adaptive advantage to a reduction in the mutation rate; this can be studied theoretically by calculating the rate of spread of a rare modifier gene that reduces  $U$  by a small amount  $\delta U$ . If the modifier recombines freely with autosomal loci subject to mutation and selection, it has a selective advantage of  $hs \delta U$ . This raises the question of why mutation rates are not closer to zero; this probably reflects the fact that there is a fitness cost to the necessary repair systems so that  $U$  is adjusted to a level at which the costs and benefits of increased repair balance (Drake *et al.*, 1998).

### 3. Segregational Load

Similar calculations can be performed for models of balancing selection, yielding estimates of the segregational load. In the case of heterozygote advantage, the load due to a single locus is  $st/(s + t)$  (Crow and Kimura, 1970). Equation (4) can be used to determine the segregational load contributed by a large number of polymorphic loci with independent effects. This can be considerable, even if selection is weak. For example, 10,000 loci each with  $s = t = 0.001$  would yield a mean fitness of only 0.0067. This is so low that only a very high fecundity species would be able to produce the two surviving offspring per adult needed to maintain itself. This implies that either most molecular variation has very slight or no effects on fitness or the assumption of multiplicative fitnesses is unrealistic.

An extreme alternative to multiplicative fitnesses is truncation selection. Genotypes at a set of loci are assumed to be ordered with respect to their fitnesses as determined by the multiplicative fitness model; a fixed proportion of the population, containing the set of genotypes with the highest fitnesses, is allowed to survive. This is equivalent to assuming that individuals compete for a limiting resource, and that only the fittest succeed. Under these conditions, a much larger number of loci can be exposed to selection for a given total  $L$  than with multiplicative fitnesses, for the same selection intensity per locus (Crow and Kimura, 1970). Less extreme forms of departure from multiplicativity can have similar but smaller effects on the total load.

### 4. Substitutional Load

Genetic loads also apply to adaptive evolution. Consider the case of a biallelic locus, where  $A$  is initially deleterious and held at a low frequency. If there is a change in the environment so that  $A$  becomes favored by selection, it will start to increase. However, in any generation before it reaches fixation, the mean fitness of the population will be reduced below its final value of 1 because of the presence of the disfavored  $a$  allele. There is thus a load associated with the substitution of  $a$  by  $A$ , reflecting the fact that natural selection cannot instantly transform a predominantly  $a$  population into one which is fixed for  $A$ . The sum of the loads for each generation over the course of a gene substitution is Haldane's cost of selection,  $C$ ; if the population size is  $N$ , the total number of individuals eliminated by selection is  $CN$ . Providing selection is not too strong,  $C$  is proportional to minus the logarithm of the initial frequency of  $A$ ; a typical value is 30 (Crow and Kimura, 1970).

The effect of changes at multiple loci can be derived as follows. Assume that there is a steady rate of change in the environment so that each generation  $K$  loci start to experience gene substitutions of this kind.  $K$  is the rate of gene substitution in the genome as a whole such that after a long period of evolutionary time,  $T$ , the population will differ from its ancestral state by  $KT$  substitutions. If a gene substitution takes  $t$  generations to complete,  $Kt$  loci will be segregating in any given generation, each associated with an average load of  $C/t$  relative to a population which is fixed for the favorable alleles. The mean fitness under multiplicative fitness, relative to a population that is fixed for favorable alleles at all currently segregating loci, is then given by  $(1 - C/t)^{Kt} \approx \exp - CK$ .

Data on rates of protein evolution suggest that  $K$  for an average amino acid site is about  $1.5 \times 10^{-9}$  per year in mammals (Kimura, 1983). With 80,000 loci coding for proteins with average size of 300 amino acids,  $K$  for the genome is 0.036. With  $C = 30$ , the mean fitness is 0.34, assuming one generation per year. A much higher load would be found if changes at silent and noncoding sequences are also taken into account. This finding of a high substitutional load associated with molecular evolution was one of the main motivations for the development of the neutral theory of molecular evolution, which asserts that most evolution at the molecular level is caused by the random sampling of alleles in finite population size, genetic drift, and not by natural selection (Kimura, 1983). The substitutional load can be considerably reduced by modifications to the assumption of multiplicative fitnesses, such as truncation

selection, so that this argument loses much of its cogency. Nevertheless, there are other good reasons to take the neutral theory seriously (see Section III,C).

## H. Multiple Loci

### 1. No Selection

So far, it has tacitly been assumed that evolutionary change involving more than one locus can be modeled by assuming that alleles at different loci are distributed independently of each other in the population and have independent effects on phenotypes and fitness. Although this may be a good approximation for many purposes, it is necessary to examine the consequences of relaxing these assumptions. Deviations from independence among loci in randomly mating populations can be described by linkage disequilibrium parameters, which measure the extent to which the frequencies of the different multilocus gamete types or haplotypes depart from the frequencies expected by randomly combining alleles at different loci. In the simplest case of a pair of loci, each with two alleles (A and a and B and b), there are four haplotypes: AB, Ab, aB, and ab. Let the frequencies of these be  $x_1$ ,  $x_2$ ,  $x_3$  and  $x_4$ . If the allele frequencies at the two loci are  $p_A$  and  $p_B$ , we can write the haplotype frequencies as  $p_A p_B + D$ ,  $p_A(1 - p_B) - D$ ,  $(1 - p_A)p_B - D$ , and  $(1 - p_A)(1 - p_B) + D$ , respectively, where  $D$  is the coefficient of linkage disequilibrium. It is easily seen that  $D = x_1 x_4 - x_2 x_3$ . If the frequency of recombination between the two loci is  $c$  ( $0 \leq c \leq 0.5$ ), the value of  $D$  in the next generation in an infinitely large, randomly mating population with no selection is  $D(1 - c)$ .

In the absence of evolutionary forces other than recombination, the extent of nonrandom association between a pair of loci thus decays exponentially at a rate that is determined by the frequency of recombination. This result can be generalized to associations between multiple loci (Crow and Kimura, 1970). Unless populations are far from equilibrium, departures from linkage equilibrium at a set of loci require the operation of forces tending to generate nonrandom associations between alleles at the different loci. Their magnitude will be at least on the order of the recombination frequencies among them. One such force is genetic drift (see Section III), which can cause randomly generated linkage disequilibrium among loci for which  $c$  is on the order of the reciprocal of the population size (Ewens, 1979).

### 2. Selection on Several Loci

Another possible force causing linkage disequilibrium is epistatic selection, in which the difference in fitness

between genotypes at one locus varies according to the genotypes at the other loci in the system. If the fitness effects of different loci combine additively, there is no epistasis, and it can be shown that polymorphic equilibria under random mating exhibit no linkage disequilibrium (Ewens, 1979). With epistasis, selection tends to preserve haplotypes which contain favorable combinations of alleles, whereas recombination breaks them down. Linkage disequilibrium is not necessarily present if linkage is sufficiently loose in relation to the strength of epistasis. Multiple alternative stable equilibria may occur in multilocus systems so that the fate of a population can be affected by the initial conditions from which evolution starts. Stable equilibria do not necessarily correspond to maxima in mean fitness in the space of haplotype frequencies, another violation of the adaptive landscape principle. However, if epistatic selection is weak in relation to the frequency of recombination, populations tend to converge to trajectories where linkage disequilibrium is nearly constant (quasi-linkage equilibrium), and mean fitness increases monotonically at a rate approximately equal to the additive genetic variance in fitness according to Fisher's fundamental theorem of natural selection (Ewens, 1979). This has provided a very useful tool for the analysis of the dynamics of multilocus systems (Barton and Turelli, 1991).

### 3. The Evolution of Close Linkage

There are two biologically important features of systems with strong epistatic selection. The first is that such selection may impose strong constraints on the degree of linkage between polymorphic loci. Suppose that the population is initially segregating for alleles A and a at one locus but is initially fixed for b at a second locus. If a mutation B arises at this locus which interacts with the alleles at the first locus, such that AB is selectively favored but aB is disfavored, B may be unable to invade the population unless  $c$  is below some threshold value. Only mutations at loci that are sufficiently closely linked to the first polymorphism in the system will be able to establish subsequent polymorphisms. This process has probably been important in the evolution of some of the classic examples of supergenes (systems of very closely linked loci held in strong linkage disequilibrium by selection), such as Batesian mimicry in butterflies (Ford, 1975) and sex chromosomes. Similarly, if ab and AB are both fitter than Ab and aB, a population fixed for ab may only evolve a two-locus polymorphism if there is a double mutation to AB and if  $c$  is sufficiently small. This probably occurred in the evolution of meiotic drive systems, which require combinations of alleles at several loci that are individually disfavored.

Second, there is a selective advantage to modifier alleles that reduce the frequency of genetic recombination between the two loci once a two-locus polymorphism has been established. If suitable genetic variation in recombination rates is available, this will eventually lead to such close linkage that the system has the appearance of a single locus (Fisher, 1930). This principle has wide generality; analysis of the conditions for spread of rare modifiers of recombination rates has shown that randomly mating populations under epistatic selection generating linkage disequilibrium at a system of loci will always tend to evolve closer linkage (Barton and Charlesworth, 1998). Since genetic recombination is a near-universal feature of living organisms, these findings have led to the search for situations that promote rather than repress recombination; these involve forces such as mutation and environmental change that perturb populations away from equilibria under selection (Barton and Charlesworth, 1998).

### III. RANDOM GENETIC DRIFT

The discovery that random sampling of allele frequencies in finite populations may be a significant factor in evolution is another major contribution of population genetics. This process has two aspects: The first is the tendency for a population of finite size to become genetically uniform, owing to the fact that there is an increasing tendency as time passes for all the copies of a gene at a locus to be descended from a single ancestral allele (Fig. 2). The second is the tendency of isolated populations to diverge in allele frequencies over time, since independent trials of a population with the same initial state will arrive at different allele frequencies by chance.

The first process is closely related to the increase in homozygosity that accompanies the inbreeding of close relatives; both are conveniently studied by means of the concept of identity by descent. Two alleles at the same locus drawn from a population are said to be identical by descent if they trace their ancestry back to a single ancestral allele. The extent to which a population has progressed toward genetic uniformity can be measured by its inbreeding coefficient, defined as the probability that a pair of randomly sampled alleles are identical by descent (Hartl and Clark, 1997). This is always measured with respect to an initial generation, in which all the alleles at a locus are arbitrarily decreed to be nonidentical (Fig. 2).

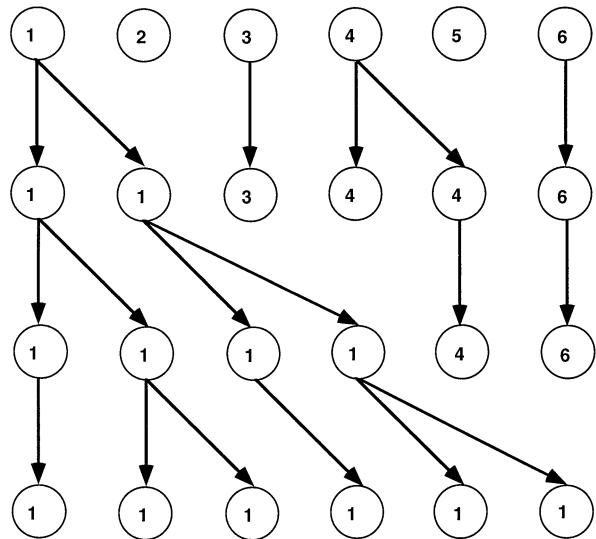


FIGURE 2 Each row of circles indicates the alleles at a locus present in a given generation. The distinct copies present in the initial generation are labelled 1–6. The arrows indicate the successful transmission of an allele to the next generation.

#### A. Increase in Homozygosity

The simplest Wright–Fisher model of genetic drift assumes a discrete-generation, randomly mating population of  $N$  hermaphroditic individuals with no selective differences among genotypes at the locus under consideration. New individuals are formed by random sampling (with replacement) of gametes produced by the parents. With diploid inheritance, such that there are  $2N$  gene copies at a locus among the breeding adults, the following recursion relation for  $f_t$ , the inbreeding coefficient in generation  $t$ , is obtained:

$$1 - f_t = \left(1 - \frac{1}{2N}\right) \left(1 - f_{t-1}\right). \quad (5)$$

This follows from the fact that the probability that a pair of randomly sampled alleles are both derived from the same allele in the previous generation is  $1/(2N)$ , in which case they are identical by descent. The probability that they come from two different alleles is  $1 - 1/(2N)$ ; their probability of identity is  $f_{t-1}$ .

Equations (5) shows that  $f_t$  tends asymptotically to 1; if  $N$  is moderately large,  $1 - f_t$  decays exponentially with a half-life of  $1.4N$  generations. This result can be generalized to more realistic types of breeding system by means of the concept of the inbreeding effective population number ( $N_e$ ) (Crow and Kimura, 1970). This utilizes the fact that genetic drift in these more general cases can be described by matrix equations or

higher order difference equations, such that a constant rate of decay of  $1 - f_i$  is reached asymptotically, yielding an expression of the same form as Eq. (5). The asymptotic decay constant can thus be equated to  $1 - 1/(2N_e)$ . For example, a population with  $N_f$  breeding females and  $N_m$  males has  $N_e = 4 N_f N_m / (N_f + N_m)$ . In general,  $N_e$  is less than the census number of breeding individuals, often considerably so (Crow and Kimura, 1970).

## B. Differentiation of Populations

The effect of drift in causing genetic differentiation among populations is described by the variance in allele frequencies among the set of populations in question. In the Wright–Fisher model, the sampling of allele frequencies at a neutral biallelic locus with alleles A and a and frequencies  $p$  and  $q$  will generate a binomial distribution of the new frequency of A, with mean  $p$  and variance  $pq/(2N)$ . Repetition of this sampling process results in a probability distribution of allele frequencies with steadily increasing variance (Crow and Kimura, 1970). The variance of allele frequency,  $\sigma_p^2$ , can be related to the inbreeding coefficient as follows. Assume that we have an infinitely large set of completely isolated populations, all founded with the same initial frequency,  $p_0$ . At some arbitrary time, the set of populations will have a mean gene frequency  $p_0$  (since drift does not change the mean) and variance  $\sigma_p^2$ . If  $N$  is reasonably large, the genotype frequencies within each given population will be in Hardy–Weinberg proportions; the mean frequencies of AA, Aa, and aa over the set of populations are equal to  $p_0^2 + \sigma_p^2$ ,  $2(p_0q_0 - \sigma_p^2)$ , and  $q_0^2 + \sigma_p^2$ , respectively.

This is the Wahlund effect: Genotype frequencies averaged over a set of populations that are individually in Hardy–Weinberg proportions show an excess of homozygotes and a deficiency of heterozygotes compared with Hardy–Weinberg expectation, whose value is determined by the variance in gene frequencies among the populations. This is a purely algebraic result, independent of the causes of the variation. It can be related as follows to the effects of genetic drift in causing increased homozygosity. Assume that a diploid individual is formed by sampling two random alleles from the same population. The probability that the alleles are identical by descent is  $f$ ; they are then both A in state with probability  $p_0$ . The probability that they are non-identical is  $1 - f$ , in which case the probability that the two alleles are both A is  $p_0^2$ . The net probability that the individual is AA is thus  $fp_0 + (1 - f)p_0^2 = p_0^2 + fp_0q_0$ . Comparison with the previous expression

shows that

$$\sigma_p^2 = fp_0q_0. \quad (6)$$

This establishes that the variance in gene frequency is proportional to the inbreeding coefficient under the Wright–Fisher model, and therefore its change over time is governed by Eq. (5). This is often but not always true under more general models of population structure, with  $N_e$  replacing  $N$  in the binomial formula for the variance conditional on the current gene frequency. In some circumstances, particularly when the population size changes in time, the conditional variance in gene frequency requires a different denominator in order to be represented by the binomial formula. In this case, a variance effective population number is computed (Crow and Kimura, 1970).

## C. Molecular Evolution and Variation

### 1. The Neutral Theory

These simple models of genetic drift can readily be applied to the study of molecular evolution and variation, assuming selective neutrality at the loci in question. Neutral theory allows for the possibility that many mutations are subject to purifying selection and are rapidly eliminated from the population (see Section IV,D,3); it is claimed that the fate of the bulk of the mutations that are not removed by purifying selection is determined by drift rather than selection (Kimura, 1983). This theory thus constitutes a useful null hypothesis, which can be tested against data on molecular evolution and variation by means of predictions of several types.

### 2. The Rate of Neutral Evolution

Consider first the rate of molecular evolution as measured by the rate of gene substitution,  $K$  (see Section II,G,4). In a population of  $N$  breeding adults, there are  $2N$  allele copies at an autosomal locus. As shown in Section III,A, Eq. (5) implies that the population tends to homozygosity with probability 1. This means that the remote descendants of the current population will all trace their ancestry back to just one of these  $2N$  alleles. Under neutrality, the probability that a given allele is the ancestor is thus  $1/(2N)$ . It follows that the probability of fixation of a new neutral mutation in a population of size  $N$  is  $1/(2N)$ ; the probability that it is lost is  $1 - 1/(2N)$ . If the rate of mutation to neutral variants is  $u$  per generation, the expected number of new mutations that enter the population is  $2Nu$ , of which only  $1/(2N)$  are destined for ultimate fixation; the expected number of mutations that ultimately become fixed is  $2Nu/(2N) = u$ .

If the process of mutation and drift has reached a stationary state, so that the expected number of new substitutions must equal the number of substitutions that go to completion each generation, we obtain the fundamental equation of neutral molecular evolution:

$$K = u \quad (7)$$

This relation holds for any level in the genetic hierarchy, from the nucleotide site through the locus to the genome as a whole, provided that  $K$  and  $u$  are defined appropriately. Since  $K$  can be determined by comparisons of DNA sequences among species with known divergence time (see Section II,G,4) and  $u$  values are known from molecular genetics, Eq. (7) can be used to test the neutral theory.

One prediction is that genomic regions whose sequences are essentially functionless, such as pseudogenes and the internal parts of introns, should evolve at the mutation rate since they are necessarily unconstrained by selection. If adaptive Darwinian evolution is a minor factor in molecular evolution, the rate for sites in these regions should be much higher than that for functionally significant regions, where selection is expected to eliminate most mutations [see Equation (10)]. These regions do indeed evolve at the rates expected from mutation rates, and other regions evolve more slowly (Kimura, 1983; Li, 1997). This does not, however, rule out a role for positive selection in fixing variants in selectively constrained regions; it could still be true, for example, that most replacement substitutions (see Section I,B,2) are deleterious, but a minority are advantageous rather than neutral, so that changes that are fixed in evolution are adaptive rather than neutral. There are also some exceptional cases of higher rates of replacement versus silent substitutions in coding regions; this is strong evidence for a positive role of selection on the amino acid sequences concerned (Li, 1997).

Another prediction of Eq. (7) is that the rate of molecular evolution should be constant over long periods of time since it depends only on the the mutation rate. In contrast, the rate of evolution under natural selection is expected to be highly variable since the theory described earlier suggests that populations will tend to adapt quickly to a new environment (or go extinct), after which change will be slow or nonexistent. This is borne out by the observations of comparative biology and paleontology, which show that evolution at the external phenotypic level is generally highly episodic and triggered by ecological opportunities such as

the occupation of vacant niches. Studies of DNA and protein sequence evolution suggest that the rate of evolution of a given molecule is much less variable among different lineages, or within the same lineage at different times, than is true for the external phenotype, especially when noncoding or silent substitutions are considered (Kimura, 1983; Li, 1997). This has generated the concept of a molecular clock, which is used to estimate the times of divergence of taxa when paleontological data are absent. However, there is evidence for more variability in rates of evolution than predicted by the simplest form of the neutral theory, especially for amino acid sequences. This suggests a role for selection, although the interpretation of rate variability is controversial (Li, 1997).

### 3. Neutral Polymorphism

The process of fixation of neutral variants is a slow one; calculations based on diffusion theory (see Section IV,D) show that the mean time to fixation of a new neutral mutation in a random mating population (conditioned on ultimate fixation) is approximately  $4N_e$  generations. While the variant is on the way to fixation, it causes a polymorphism. Similarly, variability is contributed by the large fraction of new mutations that are destined for ultimate loss; the mean time to loss is  $(N_e/N) \ln(2N)$  generations, a much shorter time than the time to fixation. In the neutral theory, polymorphism is simply a phase of molecular evolution. Although there is a constantly shifting set of variants at any one locus, the mean amount of variability can be determined assuming a statistical equilibrium between drift and mutation, using the following argument (Kimura, 1983).

The simplest version of this applies to a single locus, which is assumed to have no recombination. Thus, there are many possible sequences; new neutral mutations are assumed to occur with probability  $u$  per generation, and each mutation represents a sequence that has not been observed before. This is the infinite alleles model. Let  $h_t$  be the probability that two randomly sampled alleles are distinct in sequence in generation  $t$ . An argument similar to that leading to Eq. (5) shows that, neglecting terms in  $u/(2N)$ ,  $h_t$  for a Wright–Fisher population obeys the equation

$$h_t \approx \left(1 - \frac{1}{2N}\right)h_{t-1} + 2u(1 - h_{t-1}). \quad (8a)$$

At equilibrium, rearrangement of this equation yields

the relation

$$h \approx \frac{4N_e u}{4N_e u + 1}. \quad (8b)$$

More generally,  $N_e$  can be substituted for  $N$  in Eq. (8b). The parameter  $\theta = 4N_e u$  thus controls the equilibrium level of variability under the neutral model. A large amount of neutral variability can be maintained, provided  $\theta$  is sufficiently large, e.g., with  $N_e = 10^5$  and  $u = 2 \times 10^{-7}$ ,  $\theta = 0.08$ , and  $h = 0.074$ . This value of  $h$  is similar to the mean per locus heterozygosity for electrophoretic alleles in mammalian populations (Lewontin, 1974; Kimura, 1983).

This model can be modified to predict the equilibrium level of diversity per nucleotide site by assuming that the units of observation are the individual sites, not the entire locus. If we assume that  $\theta$  is much less than 1 so that at most one variant is segregating in the population at each site, Eq. (8) can be applied to yield the equilibrium value of  $\pi$ , the nucleotide site diversity (see Section I,B,2), such that  $\pi = \theta$  to the assumed order of approximation. This is Kimura's infinite sites model, which is widely used in the interpretation of data on molecular variation.

## D. The Coalescent Process

### 1. General Considerations

The growing body of data on DNA sequence variation within populations has stimulated interest in the development of statistical tests of the agreement of observed patterns of variation with the predictions of the neutral theory. In order to conduct such tests, it is essential to have predictions concerning the properties of statistics describing samples of alleles from populations and not merely of properties of the populations from which the samples are drawn since these cannot be observed directly. A powerful method for deriving properties of samples has recently been developed and is known as coalescent theory (Hartl and Clark, 1997; Li, 1997). It is based on the following principle. Consider a pair of alleles at an autosomal locus sampled from a Wright-Fisher population. As shown in Section III,A, there is a probability  $1/(2N)$  that they are derived from a common ancestral allele in the previous generation, i.e., that they coalesce. If they fail to coalesce, which has probability  $1 - 1/(2N)$ , they have a probability  $1/(2N)$  of coalescing in the next generation back, and so on.

There is thus a geometric distribution of the time back to the common ancestral allele, such that the probability of time  $t$  is  $(1/2N)(1 - 1/[2N])^{t-1}$ . From the well-known

properties of this distribution, the mean time to the common ancestor is  $2N$  and the variance is  $(2N)^2$ . Assume that mutations occur at rate  $u$  per site in the gene, such that each mutation that arises in the line of descent connecting the two sampled alleles is at a different site (the infinite sites assumption, see Section III,D,C,3). If there are  $m$  nucleotide sites in the sequence in question, the number of mutations conditioned on  $t$  has a mean and variance of  $2tmu$  since the total time separating the alleles is  $2t$ , and the conditional number of mutations follows a Poisson distribution. The mean and variance of the number of differences between the two alleles are thus  $m\theta$  and  $m\theta + 0.5(m\theta)^2$ , respectively. The result for the mean corresponds to that derived from Eq. (8) since the nucleotide site diversity is simply the number of differences between the pair of alleles divided by  $m$ .

This can be extended to a sample of  $n$  alleles from a population. If  $N$  is sufficiently large, the chance of more than one coalescent event per generation can be neglected. There are  $n(n-1)/2$  possible pairwise allelic combinations so that the probability of a coalescent event is  $n(n-1)/(4N)$ . The time to the first coalescent event is therefore distributed approximately exponentially, with a mean of  $4N/(n[n-1])$  and variance equal to the square of the mean. The time from this to the next coalescent is also exponentially distributed, replacing  $n$  with  $n-1$ , and so on. The process can be represented by a genealogical tree, whose nodes with  $k$  and  $k-1$  alleles are separated by a time  $t_k$  that is exponentially distributed with mean  $2/(k[k-1])$  on the coalescent timescale of  $2N$  generations (Fig. 3). As before,  $N_e$  replaces  $N$  for more general models of breeding structure. The rate of coalescence evidently diminishes as the number of nodes decreases.

This representation can be used to derive many important results, and it also provides a rapid means of simulating genetic processes, since time can be rescaled to the coalescent timescale and the properties of a sample of alleles represented by generating genealogical trees from samples drawn from the relevant exponential distributions, with mutations scattered randomly over the branches of the tree. The expected pairwise difference between all  $n(n-1)/2$  pairs of alleles on the infinite sites assumption is readily seen to be  $m\theta$ . Its variance is also known (Li, 1997). This statistic provides the obvious means of estimating the nucleotide site diversity in the population ( $\theta$ ) by equating the observed mean pairwise difference to its expectation. However, another statistic, the total number of segregating sites in a sample of  $n$  alleles ( $S_n$ ), has better statistical properties under the infinite sites model. These can be obtained as follows. The total size of a tree is the sum of  $kt_k$  over the entire tree. Application of the properties of the

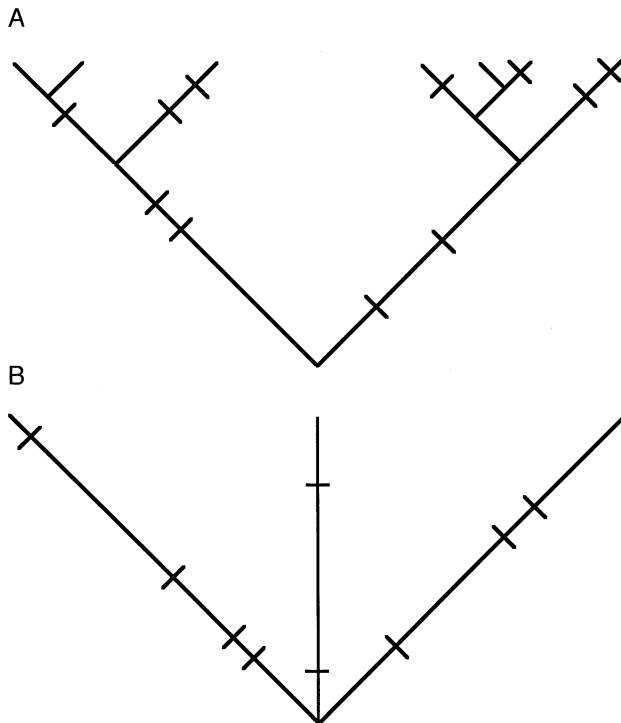


FIGURE 3 Two types of gene genealogies. (A) The genealogy of a set of alleles in a population of constant size. The tips of the tree represent six alleles sampled from the population. Each node corresponds to a coalescent event. Each cross line represents a new mutation in the DNA sequence of the gene in question, which has arisen since the common ancestor of the sample; mutations in internal branches of the tree give rise to at least two variants in the sample. (B) The genealogy of a set of three alleles in a population that experienced a recent reduction in size to one allele, followed by rapid expansion with no opportunity for coalescent events.

exponential distribution shows that the expectation of the tree size in units of coalescent time is simply  $A = 2\left(1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{(n-1)}\right)$ . The number of sites in the sample that are segregating for nucleotide variants is the number of mutations that appear on the tree. Its expectation is  $2N\mu A = m\theta A$ ; therefore,  $\theta$  can be estimated by dividing  $S_n$  by  $mA$ . The variance of this estimator is substantially lower than that based on pairwise diversity.

## 2. Tests of Neutrality

Several tests for departures from neutrality have been devised based on the properties of the coalescent process (Li, 1997). One widely used method is Tajima's  $D$  test, which uses the fact that departures from neutrality have different effects on the pairwise difference estimate of  $\theta$  than on the  $S_n$  estimate. For example, a population bottleneck of reduced size that completely removes variability at a locus will be followed by a long period of recovery, during which new mutations are all at low frequencies.

This means that the pairwise diversity, which is weighted by the frequencies of variants, will be much lower in relation to the observed number of segregating sites than in an equilibrium population, reflecting the fact that the genealogical tree is like a "star phylogeny" in this case (Fig. 3). Purifying selection on the variants in question will have a similar but much smaller effect. The opposite pattern is expected if there has been a partial bottleneck, in which rare variants have been lost from the population, or if variation is affected by balancing selection. The Tajima test computes the ratio of the difference between the two estimates of  $\theta$  to the standard deviation of their difference (calculated using the neutral equilibrium model) and compares the result with critical values obtained from coalescent simulations. Other tests have been derived from related principles (Li, 1997).

## IV. THE INTERACTION OF DRIFT WITH DETERMINISTIC FORCES

### A. Population Subdivision

#### 1. General Considerations

One important complication of the neutral theory is population subdivision; species in nature are not simple randomly mating populations but instead are usually distributed over wide geographic areas, with limited migration between localities. Random genetic drift can thus produce significant genetic differentiation among local populations if migration is sufficiently restricted. This creates two problems: How to describe data on allele or nucleotide site variant frequencies when there is differentiation among populations, and how to relate models of the underlying evolutionary processes to the data.

#### 2. Partitioning Variability

The basic method used to summarize data collected from a set of populations sampled from the same species is to partition variability into a within-population component that describes the mean level of variability within a sample and a between-population component that measures the mean difference between alleles sampled from different populations. This can be done either using data on allele frequencies, as in the case of electrophoretic loci or microsatellite loci, or using nucleotide site variant frequencies. In many cases, it is also possible to organize the populations sampled into a hierarchy of decreasing geographic scale, such as subspecies within a species, races within subspecies, and local populations within races. Variability can then be partitioned into variation between and within each level of the hierarchy (Hartl and Clark, 1997). A frequent practice is to nor-



malize each lower level measure of between-population variability relative to variability at the next higher level of the hierarchy, generating a set of Wright's  $F$  statistics or the related  $G$  statistics of Nei (Hartl and Clark, 1997).

The procedures will be illustrated here for the case of DNA sequence variation in a population divided into a set of equal-ranking subpopulations. The total nucleotide site diversity,  $\pi_T$ , is defined as the fraction of nucleotides that differ between a pair of alleles drawn randomly from the set of samples as a whole. The within-population nucleotide site diversity,  $\pi_S$ , is the mean fraction that differ between a pair of alleles drawn from the same subpopulation. The amount of between-population differentiation can be measured by the between-population component of nucleotide site diversity,  $\pi_{T-S} = \pi_T - \pi_S$ . Alternatively, a measure of nucleotide site divergence,  $\pi_B$ , can be defined as the mean fraction of nucleotide sites that differ between pairs of alleles drawn from a pair of different subpopulations, and another measure of between-population differentiation is defined as  $\pi_D = \pi_B - \pi_S$ . Two normalized measures of differentiation can be defined as  $G_{ST} = \pi_{T-S}/\pi_T$  and  $F_{ST} = \pi_D/(\pi_S + \pi_D)$ . Both measures are widely used in the literature and are often similar numerically, especially if many subpopulations are sampled.

### 3. The Island Model

Although the previously mentioned measures of population differentiation are useful as descriptive tools, they can also be used to estimate the evolutionary parameters that determine the extent of population differentiation. This requires the development of models of the joint effects of genetic drift, mutation, and migration, one of the most complex problems in theoretical population genetics. The simplest and most widely used model is the island model, which assumes that the species is divided into  $n$  distinct subpopulations or demes, which each behave according to the Wright-Fisher model with population size  $N$ . After reproduction has occurred within each deme, a fraction  $m$  of each deme's genes are replaced with genes drawn randomly from the other  $n - 1$  demes. Coalescent theory can be used to determine the mean coalescence times of pairs of alleles sampled from the same population ( $t_0$ ) or from different populations ( $t_1$ );  $t_0 = 2N$  and  $t_1 = 2Nn(1 + [n - 1]/[4Nnm])$ . In the infinite sites model, the mean fraction of nucleotides that differ between a pair of alleles is equal to the product of the mutation rate per site and twice their coalescence time (see Section III,D) so that the expected nucleotide site differences between alleles can be derived directly from the corresponding coalescence times.

An important and somewhat counterintuitive conclusion is that the within-population nucleotide site diversity,  $\pi_S$ , is equal to  $4Nnu$ , i.e., it depends on the total number of individuals in the set of populations in the same way as does the diversity in a panmictic population under the infinite sites model, and it is independent of the migration rate (with the proviso that  $m > 0$ ). As expected, the other diversity measures depend inversely on  $Nm$ , with large between-population divergence being possible only when  $Nm < 1$ ; for large  $n$ , both  $F_{ST}$  and  $G_{ST}$  are approximately equal to  $1/(1 + 4Nm)$ . Values of these statistics that are close to zero are generally taken to indicate relatively little population differentiation, whereas values close to one imply considerable differences among local populations relative to within-population variability.

### 4. Other Models of Population Structure

Although simple to analyze, the island model is not very realistic biologically since dispersal is limited in most species so that local populations are most likely to acquire immigrants from nearby. Attempts to generate useful results from more realistic models have taken two directions. The first involves maintaining the assumption of a set of discrete demes but allowing for differences in deme sizes. Migration is described by a migration matrix  $M$ , such that  $m_{ij}$  is the probability that an allele in deme  $i$  originated in deme  $j$  in the previous generation. However, it is difficult to obtain transparent general results for such a model. Under the infinite sites model, however, the result that  $\pi_S = 4Nnu$  still holds if migration is conservative, i.e., migration does not change the sizes of local populations (Maruyama, 1977; Nagylaki, 1986). Results on genetic diversity within and between populations are only available for special cases, such as the stepping-stone model, in which demes of size  $N$  in a linear or planar array receive migrants with probability  $m$  only from immediately adjacent populations (Maruyama, 1977). This model also permits the analysis of the dependence of degree of genetic differentiation between populations on the distance between them. The results show that extensive differentiation between populations is possible with a one-dimensional array of populations, even if there is a considerable amount of migration, whereas the results for a two-dimensional array are similar to those for the island model.

The second approach assumes that individuals are distributed over a one or two-dimensional spatial continuum, with density  $\rho$ . Migration is represented by the probability density that an individual moves a given distance between birth and reproduction. If migration follows a random walk, the variance of the migration

function,  $\sigma^2$ , is sufficient to describe the process. This is Wright's isolation-by-distance model. A mathematically correct formulation of this model presents formidable difficulties, but it is possible to obtain explicit results for the case of a one-dimensional continuum by treating it as the limit of the corresponding discrete population model (Nagylaki, 1986).

## B. Effects of Directional Selection at Linked Loci

Another complication in interpreting data on DNA sequence variation is that even neutral variation may be affected by selection at linked sites. The classic example is hitchhiking, whereby a new favorable mutation arises as a unique event and spreads through the population. In the absence of recombination, variants at linked loci will be dragged to fixation as the favorable allele sweeps through the population so that variability at these loci is eliminated. With recombination, the magnitude of this hitchhiking effect decreases with the ratio of the frequency of recombination between the selected and neutral locus ( $c$ ) to the selection coefficient ( $s$ ) at the selected locus, and it is negligible when this ratio is on the order of 1. The effects of hitchhiking by favorable mutations are thus only likely to be manifest at sites that are closely linked to the target of selection. There are several natural examples of such hitchhiking events, including an increased frequency of a DNA sequence variant linked to the sickle-cell anemia mutation.

However, hitchhiking effects can also be caused by deleterious mutations, a process termed background selection (Hartl and Clark, 1997; Li, 1997). A neutral variant which is tightly linked to a deleterious mutation that is in the process of elimination from the population will have a high chance of being eliminated before it can unhitch itself by recombination. As discussed previously, the genomes of higher organisms contain many loci subject to mutation to deleterious alleles; therefore, the relatively weak effects of selection against mutations at single loci may have a large cumulative effect on neutral variability.

Studies of genetic variation in *Drosophila* have shown a strong relationship between the amount of recombination in the region where a gene is located and the level of genetic variation that it exhibits (Hartl and Clark, 1997). Similarly, species or populations of hermaphroditic organisms with high rates of self-fertilization, in which the absence of heterozygotes means that genetic recombination is effectively absent, seem to also show reduced levels of molecular variation. Both hitchhiking by favorable mutations and background selection can

account for these patterns; current research is attempting to distinguish between them (and other possible explanations).

## C. Effects of Linkage to Balanced Polymorphisms

Selection at linked sites can also cause increased variability at neutral sites if selection is balancing rather than directional in nature. This can be understood in terms of a single locus with two alleles,  $A$  and  $a$ , maintained by strong balancing selection in a randomly mating population. This means that the population is effectively divided into two subpopulations represented by the two allelic classes. If the two alleles are equally frequent, a neutral site linked to this locus with recombination frequency  $c$  has a probability  $0.5c$  of crossing over from one subpopulation to the other; this is equivalent to the migration rate  $m$  in the case of an island model with just two islands. This leads to an expansion of the coalescent time at the neutral site by a factor of  $1 + 1/(4N_e c)$ , where  $N_e$  is the effective population size. This implies that the maintenance of variability by selection is accompanied by a corresponding increase in nucleotide site diversity at sites that are very closely linked to the target of selection, as seen at the *Adh* locus in *D. melanogaster* and at the MHC loci in mammals (Li, 1997). Increases in neutral variation among subpopulations of a species may also occur at linked sites when there is local selection, causing large among-population differences in allele frequencies in different populations at sites closely linked to the targets of selection. Variation in the behavior of neutral variability along a chromosomal region can thus provide valuable evidence on the action of selection.

## D. Diffusion Equations

### 1. General Considerations

A full treatment of genetic drift must deal with properties other than summary statistics of allele frequency distributions of the type considered so far, particularly if selection is to be studied. The difficulties involved in exact analytical treatments of the properties of genotype frequency distributions are great, so resort is usually made to approximations involving the use of diffusion equations, which treat genotype frequencies and time as continuous variables and assume that all evolutionary forces are sufficiently weak that second-order terms in their effects on frequencies are negligible. The standard approach is to assume a single biallelic locus. Using the

continuity assumptions, the state of the population can be described by either of the following two partial differential equations, writing  $\phi(x, p, t)$  for the probability density of the frequency,  $x$ , of allele A at time  $t$ , given initial frequency  $p$ :

$$\frac{\partial \phi}{\partial t} = \frac{1}{2} \frac{\partial^2(\phi V_{\delta x})}{\partial x^2} - \frac{\partial(\phi M_{\delta x})}{\partial x} \quad (9a)$$

and

$$\frac{\partial \phi}{\partial t} = \frac{V_{\delta p}}{2} \frac{\partial^2 \phi}{\partial p^2} + M_{\delta p} \frac{\partial \phi}{\partial p}, \quad (9b)$$

where  $M_{\delta x}$  is the expected change in gene frequency and  $V_{\delta x}$  is the variance of the change in gene frequency, both conditioned on  $x$ .  $M_{\delta x}$  can be equated to the change in gene frequency in an infinite population; under random sampling of allele frequencies,  $V_{\delta x} = x(1-x)/(2N_e)$ .

Equation (9a) is the forward Kolmogorov equation and Eq. (9b) is the backward Kolmogorov equation (Crow and Kimura, 1970; Ewens, 1979). Multidimensional versions of these equations describe systems with multiple alleles and multiple loci but are usually difficult to use.

## 2. Stationary Distributions

Equation (9a) is most useful for describing the probability density function. However, even for the simple case of a biallelic locus, a full general solution of this equation to yield the density as a function of time has been obtained only for some special cases, such as no selection. It is most useful for studying the properties of stationary distributions of gene frequency, when drift comes into statistical equilibrium with mutation, migration, and selection.

The study of such distributions has led to some important conclusions, most notably that selection is effective at countering the effects of drift in a randomly mating population when  $N_e s$  is  $\gg 1$  but is ineffective when  $N_e s$  is  $\ll 1$ . When the first condition holds, there is little scatter around the mode of the gene frequency distribution; when the second condition holds, there is a high probability that the population is close to fixation, or fixed, for the disfavored allele. If  $N_e s$  is on the order of 1, both drift and selection are significant forces.

## 3. Fixation Probabilities

The previous conclusion can also be derived by consideration of the probability of fixation of an allele,  $P(p)$ , for which a general formula was found by Kimura using the backward equation (Crow and Kimura, 1970;

Ewens, 1979). With intermediate dominance, such that  $h = 0.5$  in Eq. (1), and assuming that a single copy of the mutation is present initially ( $p = 1/[2N]$ ), this formula simplifies as follows, first discussed by Fisher (1930):

$$P\left(\frac{1}{2N}\right) = \frac{1 - \exp\left(-\frac{N}{N_e} s\right)}{1 - \exp(-2N_e s)}. \quad (10)$$

When  $s > 0$ , the fixation probability tends to  $(N/N_e)s$  as  $N$  tends to infinity. When  $N_e = N$  (so that the Wright-Fisher model applies), this is equivalent to the branching process result for the survival probability of a favorable mutation in a very large population (see Section II,C,3). The asymptote is approached when  $N_e s \gg 1$  so that a selection coefficient larger than  $1/N_e$  is sufficient to ensure that a new favorable mutation behaves as though the population is infinite. Conversely, a selection coefficient of this order ensures that a deleterious mutation ( $s < 0$ ) is almost certain to be eliminated from the population.

This led Fisher to conclude that random genetic drift is unlikely to be effective as an evolutionary force on the grounds that (i) most species have numbers of breeding individuals at least in the tens of thousands, and usually in the hundreds of thousands or even millions, and (ii) it is unlikely that a gene would have such a small effect on the phenotype that its average effect on fitness over evolutionary time would be on the order of  $10^{-4}$  or less. Although a compelling argument with regard to genetic changes that affect the phenotype, this view has been challenged by the neutral theory of molecular evolution. As already noted, the causes of protein sequence evolution remain controversial, but studies of the statistical properties of between- and within-species patterns of silent substitutions at third coding positions support the idea that this is affected by both drift and selection in bacteria, yeast, and *Drosophila*. There is a tendency for certain triplets that code for the same amino acid to be favored by selection, but there is also evidence that this selection is so weak that disfavored nucleotide changes can drift to fixation within a species. Estimates of the intensity of selection can be obtained by comparing observed distributions of frequencies of silent-site variants with those predicted by solutions of the diffusion equations (Hartl and Clark, 1997).

## E. Muller's Ratchet

Selection may become ineffective if recombination is restricted among the loci subject to selection as a

result of linkage disequilibrium generated by random drift (the Hill–Robertson effect; Barton and Charlesworth, 1998). One process that has been much studied in this context is Muller’s ratchet, which operates in a finite population subject to mutation to deleterious alleles at many loci. Consider, for example, the case of a haploid asexual population in which mutations occur exclusively from wild-type to deleterious alleles but not in the opposite direction (see Section II,F) and where  $N_e s \gg 1$ . If the selective effects of mutations at different loci are identical, a population can be characterized by the frequencies of genomes containing 0, 1, 2 . . . mutations. If the frequency of the mutational class containing the lowest number of mutations (the least-loaded class) is sufficiently small, it will be lost from the population after a finite number of generations. Given the assumed irreversible nature of mutation and the lack of opportunity for genetic recombination, the least-loaded class cannot be reconstituted and will be replaced by the class with one more mutation. This class is now vulnerable to stochastic loss in the same way. There is thus a repetitive process of loss of successive least-loaded classes, in which the loss of each class can be regarded as a turn of Muller’s ratchet. The ratchet can only operate at an evolutionarily significant rate if the equilibrium number of individuals in the least-loaded class in an infinite population ( $n_0$ ) is relatively small (e.g.,  $< 100$ ); if it does operate, there will be an approximately exponential decline in the mean fitness of the population. Given the results discussed previously, which suggest that a typical *Drosophila* individual may carry 30 or more detrimental alleles so that the frequency of mutation-free individuals is approximately  $\exp(-30) = 9 \times 10^{-14}$ , even a very large asexual population of a higher organism is likely to be vulnerable to the operation of the ratchet, leading to its eventual extinction. Asexual prokaryotes, with their much smaller genomes and enormous population sizes, are unlikely to suffer this fate. This may account for the fact that very few species of higher organisms, with their large genomes, are asexual, whereas prokaryotes and mitochondria have very low levels of recombination (Barton and Charlesworth, 1998). It may also contribute to the evolutionary degeneration of Y chromosomes, which are usually largely devoid of active genes.

### F. Group Selection

Another possibility that violates the supremacy of individual selection occurs when species are subdivided

into local populations, among which migration is so restricted in relation to the reciprocal of local population size that drift can overcome its homogenizing influence (see Section IV,A). An allele that is deleterious within its local population may drift to fixation locally in opposition to selection since  $N_e$  for the local population is much smaller than for the species as a whole. This raises the possibility of group or interdeme selection. If the allele in question causes its carriers to be altruistic in the sense of conferring increased fitness on the members of their deme at the expense of suffering a loss in fitness to themselves, demes in which the allele reaches high frequency or fixation will achieve a higher mean fitness. This may render them less susceptible to extinction and more able to contribute to the pool of migrants or to found new demes. Selection among demes can therefore result in the spread of an allele that causes a loss in individual fitness but benefits the population as a whole (Haldane, 1932; Hartl and Clark, 1997).

Although this is a theoretically viable mechanism, there are reasons to doubt that it is widely applicable to evolution in nature. First, studies of molecular genetic variation within and among populations indicate that many species lack extensive differentiation among local populations (see Section IV,A,2), suggesting that migration is so effective that the necessary conditions for group selection are often not met. Second, it does not seem capable of producing a stable evolutionary outcome: If an altruistic allele becomes fixed in a set of populations, a “selfish” opponent that is reintroduced by migration or mutation will have a good chance of spreading through the species since the outcome of the conflict between group and individual selection is probabilistic and not deterministic.

### G. Kin Selection

Altruistic behavior, such as the sterility of the worker castes in social insects, is generally believed usually to be due to kin selection. This is based on the fact that an altruistic genotype that aids relatives may experience a selective advantage, even in a large, randomly mating population, if the fitness benefit  $b$  to the recipients of the altruism is sufficiently large in relation to the cost  $c$  to the altruists (Fisher, 1930; Haldane, 1932; Hartl and Clark, 1997). This is because the genetic similarity of related individuals means that the relatives of an altruist have a greater chance than average of carrying an allele that promotes altruism; according to Hamilton’s rule, there can be an increase in the frequency of an altruistic allele

when  $br > c$ , where  $r$  is a measure of the relatedness of the altruists to the recipients. There is an extensive theoretical literature on the correct way to calculate  $r$  under different conditions; this result has played an important role in the evolutionary interpretation of social behavior in animals.

## H. The Shifting Balance Theory

The second objection to group selection is overcome by the related model of Wright, the shifting balance theory of evolution (Wright, 1977; Hartl and Clark, 1997). Wright postulated that epistatic interactions in fitness among alleles at different loci are widespread, resulting in multiple stable equilibria under selection. The simplest case is a haploid two-locus, two-allele model, with  $ab$  and  $AB$  both fitter than  $Ab$  and  $aB$ . Fixation for  $ab$  or  $AB$  is stable against introduction of  $Ab$  and  $aB$ , whereas fixation for  $Ab$  or  $aB$  is unstable to the introduction of  $ab$  and  $AB$ . With constant fitnesses and loose linkage, locally stable equilibria are approximated by the peaks in the surface of mean fitness as a function of the allele frequencies at the loci concerned (see Sections II,C,1 and II,H,2). A population will approach the peak that is the nearest attractor rather than the highest peak in the landscape. Genetic drift can cause a local population to travel down the adaptive valley separating the current equilibrium to the zone of attraction of a neighboring peak, and selection can then bring it to the new equilibrium. If this is associated with a higher mean fitness than the surrounding peaks, the process of interdeme selection can cause the species as a whole to acquire the genotype associated with this peak, and hence improve in mean fitness. In contrast to the group selection model of altruism, the new equilibrium is dynamically stable, and so there is only a low probability of reversal. This process has the attractive feature that it allows the species to acquire adaptively superior genotype or character combinations that would require multiple simultaneous mutations to be produced in a large population. Its drawback is that it requires a delicate balance between restricted migration, local population size, and the nature and intensity of selection if it is to operate with any frequency. It is difficult to distinguish the end products of the shifting balance process from ordinary individual selection, which does not have such stringent requirements, and it is unclear to what extent it has played an important role in adaptive evolution (Kimura, 1983).

## V. CONCLUSIONS

This article necessarily omitted many important topics. It focused on the basic general principles governing evolutionary change in populations; space did not permit more than a brief mention of the application of these principles to wider biological problems, including life history evolution, the evolution of genetic and sexual systems, the evolution of social behavior, speciation, and the interpretation of macroevolution. Very little useful can be said about natural variation and evolution at almost any level without taking population genetic concepts into account.

## See Also the Following Articles

DIVERSITY, MOLECULAR LEVEL • EVOLUTION, THEORY OF • INBREEDING AND OUTBREEDING • PHENOTYPE, A HISTORICAL PERSPECTIVE

## Bibliography

- Barton, N. H., and Charlesworth, B. (1998). Why sex and recombination? *Science* 281, 1986–1990.
- Barton, N. H., and Turelli, M. (1991). Natural and sexual selection on many loci. *Genetics* 127, 229–255.
- Bruford, M. W., and Wayne, R. K. (1993). Microsatellites and their application to population genetic studies. *Curr. Opin. Genet. Dev.* 3, 939–943.
- Crow, J. F. (1993). Mutation, mean fitness, and genetic load. *Oxford Surv. Evol. Biol.* 9, 3–42.
- Crow, J. F., and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper & Row, New York.
- Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics* 148, 1667–1686.
- Ewens, W. J. (1979). *Mathematical Population Genetics*. Springer-Verlag, Berlin.
- Falconer, D. S., and Mackay, T. F. C. (1996). *An Introduction to Quantitative Genetics*, 4th ed. Longman, London.
- Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford Univ. Press, Oxford. (2nd ed., 1958, Dover, New York)
- Ford, E. B. (1975). *Ecological Genetics*, 4th ed. Chapman & Hall, London.
- Haldane, J. B. S. (1932). *The Causes of Evolution*. Longmans Green, London.
- Hartl, D. L., and Clark, A. G. (1997). *Principles of Population Genetics*, 3rd ed. Sinauer, Sunderland, MA.
- Hill, A. V. S., and Weatherall, D. J. (1998). Host genetic factors in resistance to malaria. In *Malaria Parasite Biology, Pathogenesis and Protection* (I. W. Sherman, Ed.), pp. 445–455. ASM Press, Washington, D.C.
- Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge Univ. Press, Cambridge, UK.
- Lande, R. (1988). Quantitative genetics and evolutionary theory. In *Proceedings of the Second International Conference on Quantitative Genetics* (B. S. Weir, E. J. Eisen, M. M. Goodman, and G. Namkoong, Eds.), pp. 71–84. Sinauer, Sunderland, MA.

- Lewontin, R. C. (1974). *The Genetic Basis of Evolutionary Change*. Columbia Univ. Press, New York.
- Li, W.-H. (1997). *Molecular Evolution*. Sinauer, Sunderland, MA.
- Maruyama, T. (1977). *Stochastic Problems in Population Genetics*, Lecture Notes in Biomathematics No. 17. Springer-Verlag, Berlin.
- Nagylaki, T. (1986). Neutral models of geographical variation. In *Stochastic Spatial Processes* (P. Tautu Ed.), Lecture Notes in Mathematics No. 1212, pp. 216–237. Springer-Verlag, Berlin.
- Wright, S. (1977). *Evolution and the Genetics of Populations. Vol. 3. Experimental Results and Evolutionary Deductions*. Univ. of Chicago Press, Chicago.





# POPULATION STABILIZATION, HUMAN

Alene Gelbard

Population Reference Bureau

---

- I. History of Human Population Growth
  - II. Causes and Effects of Human Population Growth
  - III. Population Prospects: 2000–2050
  - IV. Responses to Human Population Growth
  - V. Conclusion
- 

## GLOSSARY

**family planning** The conscious effort of couples to regulate the number and spacing of births through artificial and natural methods of contraception. Family planning connotes conception prevention to avoid pregnancy and abortion, but it can also refer to efforts of couples to induce pregnancy.

**fertility** The actual reproductive performance of an individual, a couple, a group, or a population measured by a variety of rates, including the crude birth rate (number of births per 1000 population) and total fertility rate.

**life expectancy** The average number of additional years a person could expect to live if current mortality trends were to continue for the rest of that person's life; most commonly cited as life expectancy at birth.

**mortality** Deaths as a component of population change. Measures include the crude death rate (number of deaths per 1000 population).

**population change** An increase or decrease in the size, composition, or distribution of a population resulting from the interaction of births, deaths, and migration in a population in a given period of time.

**population composition** The distribution of a population, usually by age and sex, measured by the number and proportion of males and females in different age groups.

**population momentum** The tendency for population growth to continue beyond the time that replacement-level fertility has been achieved because of the relatively young age structure of the population.

**replacement-level fertility** The level of fertility at which a couple replaces itself in the population. In low mortality countries, a total fertility rate of 2.1 is considered replacement because some women will die before the end of their childbearing years.

**stable population** A population with an unchanging rate of growth and an unchanging age composition as a result of age-specific birth and death rates that have remained constant over a sufficient period of time. Zero population growth is a special case of population stabilization in which no growth occurs because the number of births and deaths are the same.

**total fertility rate** The average number of children that would be born to a woman during her lifetime if she were to pass through her childbearing years conforming to the age-specific fertility rates of a given year.

---

**HUMAN POPULATION CHANGE** has three components: births, deaths, and migration. On a global level,



only births and deaths determine changes in population size. In the 1990s, more than 70 million people were added to the world's total population each year and it took only 12 years for the world's population to increase from 5 to 6 billion people. This growth will continue into the twenty-first century. How much growth will occur and how quickly, and when or whether this growth will stabilize, depends on a host of factors. In 1994, the world community agreed to the need to stabilize global population growth. This article describes the history of global population growth, factors associated with this growth, and how countries view this growth. It concludes with a discussion of prospects for the stabilization of population growth in the future.

## I. HISTORY OF HUMAN POPULATION GROWTH

For much of our history, humans have struggled to survive. By 1 A.D., perhaps 300 million people lived on Earth, a paltry total after millions of years of human existence. For most of the next 2000 years, population growth was exceedingly slow. High birth rates were often offset by frightful mortality from wars, famines, and epidemics. The bubonic plague, for example, reduced the populations of China and Europe by one-third in the fourteenth century.

Despite dramatic spikes in mortality rates, the number of births exceeded the number of deaths during the seventeenth and eighteenth centuries and population growth proceeded at a slightly faster pace. As shown in Fig. 1, world population was about 790 million in 1750 and reached 1 billion in approximately 1800.

During the next century, in Europe and in a few other areas throughout the world better hygiene and public sanitation reduced the incidence of disease. Expanded commerce made food supplies more widely available and improved nutrition. The wild fluctuations in mortality of previous centuries began to recede, and life expectancy began a slow rise. Population grew more quickly and more steadily. Total world population was nearly 1.7 billion by the beginning of the twentieth century and reached 2 billion within the next 30 years.

The nineteenth-century surge of population growth occurred primarily in the more developed countries. The population of Europe more than doubled between 1800 and 1900, whereas the population of North America increased nearly 12 times—fueled by immigration from Africa and Europe. In 1800, about one-fourth of the world population lived in the now more developed regions of Europe (including Russia), Japan, and

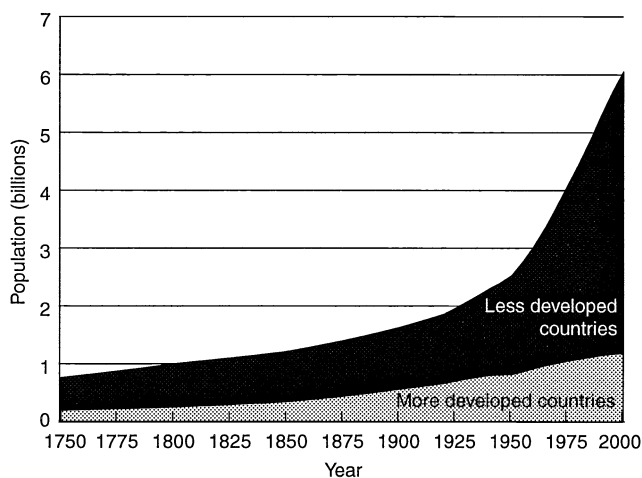


FIGURE 1 Population growth in more developed and less developed countries, 1750–2000 [sources: United Nations (1998) and estimates by the Population Reference Bureau].

North America, but this share increased to about one-third by 1900.

Less developed countries grew more slowly than more developed countries in the nineteenth century, but they already held the bulk of the world's inhabitants. Asia, dominated by China, had 62% of the world population in 1800, and Africa had 11%. Latin America and the Caribbean accounted for only about 2% of the world's population. Like North America, Latin America would see most of its population growth in the twentieth century.

### A. Demographic Transition

The improvement in human survival and the consequent explosion of population growth marked the beginning of the shift from high to low mortality and from high to low fertility that is known as the “demographic transition.” This shift occurred throughout Europe, North America, and many other areas in the nineteenth and early twentieth centuries. It gave rise to the dominant model of demographic change, which most demographers assume applies to all countries. In the classic demographic transition, the trend of high birth and death rates (and minimal population growth) is disrupted by a long-term decline in mortality. Mortality rates eventually stabilize at low levels. Birth rates also begin a long-term decline and decrease to about the same level as mortality rates. With birth and death rates at similar low levels, the equilibrium of slow population growth is regained.

The pace of change in a country will vary depending

on its culture, level of economic development, and other factors. As countries pass through the various stages of the transition, population growth from natural increase (birth rate minus death rate) accelerates or declines depending on the gap between the birth rates and death rates. Not all countries will follow the same path to low fertility and low mortality as did European countries. Also, there may be additional stages of transition that have not been identified, e.g., long-term population decline. However, the demographic transition theory provides a useful framework for assessing demographic trends and projecting future population size.

### B. Population Change: 1900–1950

When the twentieth century began, more developed countries were entering a new stage of the demographic transition. In 1900, life expectancy at birth was 47 years in the United States and between 45 and 50 years in Europe, Japan, and Australia, which was a slight increase from an average of about 40 years during the nineteenth century. A revolution in health had already begun, and life expectancy would reach unimaginably high levels by mid-century. These improvements in health reflected scientific advances of the previous century. Louis Pasteur, Robert Koch, and others had identified disease-causing “germs,” and Joseph Lister introduced antiseptic practices that were eventually adopted by hospitals. Mortality was also declining because of better personal hygiene and public sanitation projects that removed garbage and sewage from city streets and provided safer drinking water. Death rates for infectious diseases began to decrease well before vaccines and antibiotics were widely available.

Infants and young children benefited most from this health revolution. In the more developed countries, the infant mortality rate (IMR; number of deaths to infants less than 1 year of age per 1000 births) was about 200 in the 1800s—about 2 of every 10 babies died before their first birthday. In the early 1900s, the IMR decreased below 100 in the United States and many European countries. It was lower than 50 in nearly all these countries by the 1950s. United States life expectancy at birth increased to 56 years by 1920 and to 68 years by 1950. Average life expectancy was even higher in some European countries by 1950.

Although birth rates had decreased during the latter part of the nineteenth century, women were still having relatively large families in 1900. According to the United Nations, an American woman had four or five children on average; a European woman had somewhat fewer. Fertility decline quickened after 1900.

During this same period, most of Africa, Asia, and

Latin America were still in the predemographic transition stage of high mortality and high fertility.

### C. Population Change: 1950–2000

The second half of the century brought many new demographic trends and patterns. The more developed countries completed their transition to low mortality and low fertility. Population growth slowed and even turned negative in a few countries. Populations grew older. The more developed countries also experienced sometimes disruptive changes associated with baby booms and baby busts, crises in health, and waves of immigrants and refugees.

In less developed countries, the second half of the century brought decades of rapid population growth and swelling streams of migrants from rural to urban areas. Some countries appeared to be rushing through the various stages of the demographic transition, whereas others appeared to be following a new path of demographic change.

### D. Mortality, Fertility, and Natural Increase

In Europe, population growth accelerated as countries recovered from the devastating effects of World War II. The rapid decline in death rates of the early part of the century slowed considerably, in part because infant and childhood mortality had already decreased to very low levels. By 1975, the IMR was 10 in Japan, 16 in the United States, and 15 in much of Europe. According to the U.S. Census Bureau, U.S. life expectancy increased by less than 10 years in the second half of the century, from 68 years to 76 years, after increasing by more than 20 years during the first half.

After World War II, “baby booms” were commonplace in Europe, although they were more modest than the baby boom that occurred in the United States between 1946 and 1964. By the mid-1970s, however, total fertility rates (TFRs; which reflect the average number of children per woman) in many European countries had decreased below 2, the level at which a couple replaces itself in the population. A TFR must be slightly higher than 2.0 (about 2.1 in low-mortality countries) to reach replacement level because some women will die before the end of their childbearing years. When the TFR remains lower than 2 for a prolonged period, populations may experience a natural decrease because deaths will outnumber births.

European fertility had decreased during the 1930’s Great Depression, but in the mid-1980s TFRs sank to

record low levels and showed little sign of recovery. By the late 1990s, the TFR was 1.2 or less in Belarus, Bulgaria, the Czech Republic, Estonia, Italy, Latvia, and Spain.

The fertility decline began in Western Europe during a period that saw delayed marriage, more divorce, high inflation, and an increase in the percentage of women attending college and working outside the home. These same social and economic factors favored lower fertility in the United States, in which the TFR reached an all-time low in 1976 at 1.7 children per woman. Below-replacement fertility also occurred in Eastern Europe and the former Soviet Union after 1990.

Two decades of low fertility have halted population growth in nearly all of Europe and Japan. In many cases, a decline in population was avoided only by the flow of immigrants from abroad. In the late 1990s, 14 European countries experienced natural decrease, or fewer births than deaths each year.

Natural decrease will spread to other countries as low birth rates drastically reduce the number of people entering the childbearing ages. Although some countries have a net population gain from immigration, this is not expected to generate enough growth to stave off eventual population decline if fertility levels continue to remain low. It is possible that total fertility rates will increase again in some countries as recent trends toward later marriage and childbearing stabilize and couples who delayed childbearing complete their families. However, at the end of the twentieth century, not one major industrialized country had fertility above replacement level.

Among the more developed countries, only a few traditional immigration countries (Australia, Canada, New Zealand, and the United States) can expect significant long-term population growth. They have younger age structures and more immigration than do Europe and Japan, which contributes to momentum for continued growth.

Fertility and mortality patterns have been very different among less developed countries in the past 50 years. Gains in life expectancy accelerated after 1950. The average life expectancy at birth in less developed countries increased from 41 to 62 years between 1950 and 1995, according to United Nations (UN) estimates. The IMR decreased from 178 to 68 deaths per 1000 births during the same period.

Average life expectancy increased to more than 60 years in east Asia and Latin America by the early 1970s and to about 70 years by the late 1990s. The IMR decreased to about 29 in east Asia and 36 in Latin America by 1998.

Progress has been much slower in sub-Saharan Africa

and south central Asia. In the 1950s, about 180 infants died per 1000 births in these regions. By the 1990s, the IMR was still close to 100 in sub-Saharan Africa and was nearly 80 in south central Asia.

The pace of mortality decline in some areas has been slowed by the spread of HIV/AIDS, and many experts predict dramatic declines in life expectancy in some countries of sub-Saharan Africa. Worldwide, nearly 14 million people have died from HIV/AIDS since the beginning of the pandemic in the 1980s. An additional 33 million are infected with the virus. Most will die within the next decade. The UN agency that tracks the HIV/AIDS pandemic, UNAIDS, estimates that there are nearly 16,000 new infections daily, 1600 of which are to children.

## E. Population Growth

The general reduction in death rates after 1950 led to explosive population growth in many less developed countries. For the world as a whole, growth rates peaked during the 1960s and early 1970s at about 2% annually. The population total for less developed countries increased from 1.7 to 4.7 billion between 1950 and 1998. Population growth would have been even higher if fertility rates had not started to decrease in less developed countries. The pattern and pace of decline varied tremendously, depending on economic and social development, government policies, family planning use, and other factors (see Box 1).

### Box 1

#### The Reproductive Revolution

The "reproductive revolution" was one of the most remarkable events of the second half of the twentieth century. The development of family planning methods such as the pill and the IUD, simpler sterilization techniques, and contraceptives that can be injected or implanted under the skin made it easier and safer for women to avoid unintended pregnancies. Increased access to these methods and socioeconomic changes that motivated couples to limit their family size drove the fertility declines of the past few decades. Family planning use increased from less than 10% for married women of childbearing age in the 1960s to more than 50% of this group of women in the 1990s (Fig. 2).

Before 1960, women's choices of family planning methods were limited to such methods as withdrawal, rhythm, diaphragms, foams or jellies,

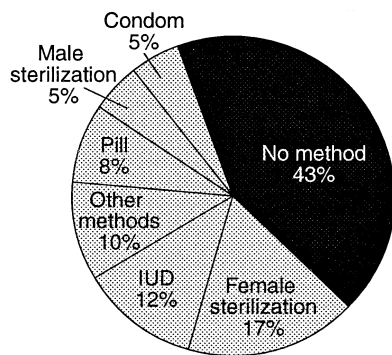


FIGURE 2 Contraceptive methods used by currently married women, ages 15–49, throughout the world, 1990s (source: United Nations, 1996).

or ineffective methods such as herbal medicines or douching. Women's options improved immensely when the pill and the modern IUD became available after 1960. In the 1990s, about 20% of women worldwide relied on one of these two methods. New contraceptives, including injectables and implants, became available in many countries in the 1980s. They have become popular methods in some African countries. Female sterilization has been widely adopted in Asia and Latin America and is the most popular method worldwide. An estimated 17% of married women ages 15–49 rely on female sterilization to prevent pregnancy.

The dramatic increase in family planning use caused fertility to decline much more rapidly in the less developed countries than it had during the fertility transition in the more developed countries. Organized family planning programs and government promotion of family planning use were important components of this phenomenon. Some experts credit family planning programs for 40–50% of the fertility decline in less developed countries since the 1960s (Bongaarts, 1995).

An estimated 120 million couples worldwide want to delay or prevent another pregnancy but are not using family planning (Miller *et al.*, 1999). If unmarried sexually active women were included, the number would be much higher, according to survey data (Shane, 1997).

Family planning use varies widely throughout the world. Less than 10% of women use family planning in Mali, for example, and less than 20% use it in Pakistan. However, more than 60% of married women use family planning in Brazil, Mexico, Thailand, and many other less developed countries.

The expansion of family planning services has been controversial in some countries. There have also been many obstacles to their use. Many women report that they fear adverse health effects from specific methods. Others want to practice family planning but are dissuaded by their husband's disapproval, their limited decision-making powers, or family pressures to have more children. Some methods are opposed for religious reasons. Difficulties in obtaining and transporting supplies and a shortage of trained medical personnel have also restricted access to family planning services.

Political and cultural barriers have limited access to family planning, especially for young people. In some countries, unmarried adolescents are denied access to family planning services on the assumption that such access would promote promiscuity. However, about 40% of girls in less developed countries give birth before age 20. The pace of fertility decline in Africa, south Asia, and other high-fertility regions will be affected by whether young couples delay their first birth until they are in their 20s. This delay lengthens the interval between generations and lowers average fertility. Health analysts estimate that women age 15–19 face twice the risk of dying from pregnancy and child-birth as do women in their 20s. In many countries, children born to mothers under age 20 are 1.5 times more likely to die before their birthdays than are children born to mothers in their 20s (Shane, 1997).

A majority of less developed countries provide family planning services. In many countries, family planning methods are also widely available from pharmacies and private health clinics. Not all women have easy access to family planning, but the expansion in the choices of methods and availability of services throughout the world during the past 40 years has been truly revolutionary.

During the 1950s and 1990s, childbearing levels declined significantly in many regions. Figure 3 illustrates that declines were greatest in Asia and Latin America and much smaller in sub-Saharan Africa. Despite these declines, all developing regions had TFRs above replacement level in 1998 and more than 100 countries had TFRs of 3.0 or higher.

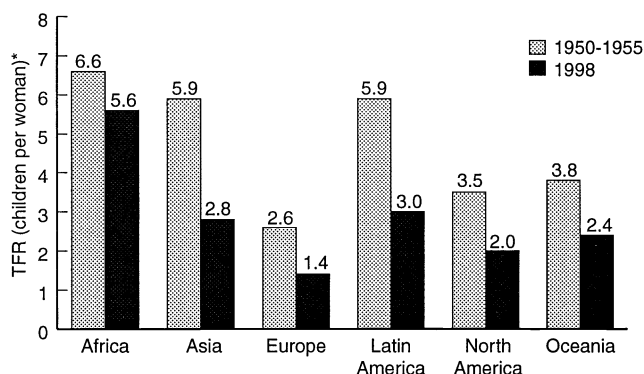


FIGURE 3 Fertility decline in world regions, 1950–1998. \*Total fertility rate (TFR) is the average number of children a woman will have under prevailing birth rates [sources: United Nations (1998), Table A.20, and Haub and Cornelius (1998)].

## F. Changing Age Profiles

Fertility, mortality, and migration trends are reflected in the age and sex profiles of the world's countries. The decades of high fertility rates in the less developed countries meant ever-increasing numbers of young people, illustrated by the broad base of the age–sex pyramid shown in Fig. 4. Improvements in infant mortality have also contributed to the expanding youth population. Children under age 15 comprised one-third of the population in the less developed countries in 1998 and even greater proportions in some regions. In sub-Saharan Africa, children comprised nearly one-half (45%) of the population. Elderly people ages 65 or older comprised only 5% of the population in all less developed countries and 3% of the population in sub-Saharan Africa.

The base of the population pyramid for less developed countries shows some narrowing—the result of declining fertility in many countries beginning in the 1980s. However, even with declining fertility rates, the young age structure creates considerable momentum for future growth because the population reaching childbearing ages continues to expand. Women currently have fewer children than women did in the past, but today there are more women having children.

## II. CAUSES AND EFFECTS OF HUMAN POPULATION GROWTH

The demographic processes of fertility and mortality are influenced by biological, cultural, economic, geographic, political, and social factors. These factors affect demographic processes directly and indirectly through

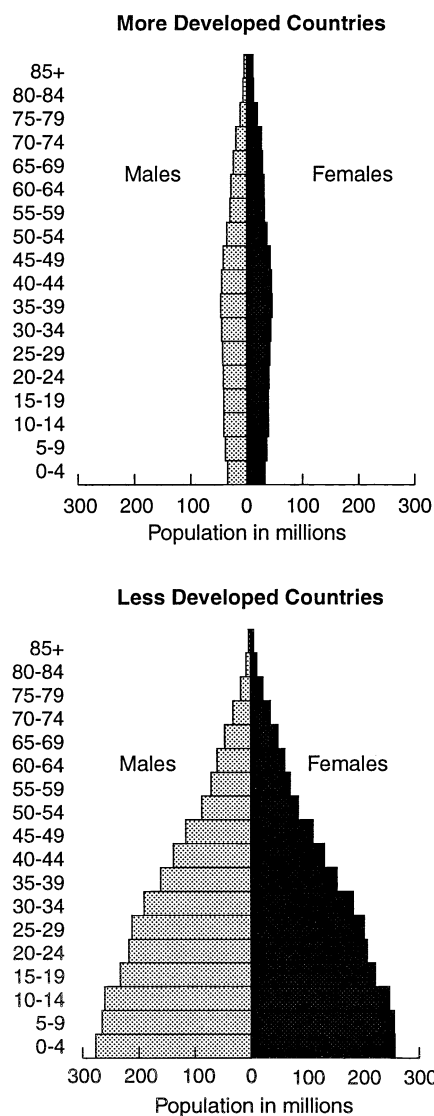


FIGURE 4 Population by age and sex in more developed and less developed countries, 1998 (source: United Nations Population Division).

a web of interdependent variables. Cultural traditions that encourage girls to marry at a young age, for example, can contribute to high fertility rates because women will spend more years exposed to the risk of becoming pregnant. Early marriage can also lead to higher mortality because health risks to the infant and mother are greater when childbearing starts in adolescence.

In the 1980s, demographer John Bongaarts identified four variables that account for most differences in fertility rates. These “proximate determinants” of fertility are (i) the proportion of women married or in a sexual union; (ii) the percentage of women using contracep-

tion; (iii) the proportion of women who cannot conceive a pregnancy, especially during the infertile period following childbirth (postpartum infecundity); and (iv) the level of abortion.

The importance of each proximate determinant depends on cultural, economic, health, and social factors within a population. The proportion of women in a sexual union is partly determined, for example, by the age at marriage, the proportion of women who never marry, and levels of divorce. Cultural mores about sexual activity and childbearing outside marriage also play a role.

In societies in which women marry young, and in which nearly all childbearing takes place within marriage, changes in the age at marriage can significantly affect fertility. In the Arab countries of the Middle East, for example, an increase in the average marriage age for women led to significant fertility declines in some countries (Rashad and Khadr, 1999).

The length of postpartum infecundity usually depends on how long women breast-feed their babies. Breast-feeding releases hormones in the nursing mother that can prevent her from becoming pregnant. Postpartum infecundity is not a significant factor in such countries as the United States, in which women usually breast-feed their babies only for a few months, but it is important in sub-Saharan Africa and other traditional societies in which women commonly breast-feed their babies for 2 years. In most populations, contraceptive use and abortion are the primary determinants of fertility levels.

Education and poverty are among the most important influences on the proximate determinants and consequently have a strong indirect effect on fertility. Low levels of education and poverty are related, and they are also related to health and to levels of economic development, urbanization, and environmental conditions.

### A. Education

Although researchers cannot determine all the reasons why, education is associated with lower fertility and mortality. A formal education may act as a catalyst for changes in values and behavior. Education may make people more receptive to new ideas, such as family planning. Social scientists note that education does not have the same effect in all cultural settings, and that many other factors (such as women's status) may explain much of the association (Riley, 1997; Jeffrey and Basu, 1996; Knodel and Jones, 1996).

More educated women have higher rates of family

planning use, smaller families, and healthier children than other women in the same society. Where educational levels are high, women are likely to postpone marriage until they finish secondary school or college.

Married women are more likely to use family planning if they have some formal education. A 1998 survey in the Philippines indicated a contraceptive use rate of 50% for married women of reproductive age who had at least some secondary education. Only 15% of their counterparts with no formal education used a contraceptive method.

In most societies, total family size declines as education increases. In the early 1990s, Peruvian women with at least some secondary education had nearly four fewer children, on average, than women with no formal education. A similar gap was recorded in a 1998 survey in Togo, West Africa. Togolese women with a secondary or higher education had 2.7 children on average, whereas women with no education had an average of 6.5 children.

Education usually expands employment options, and educated women may delay marriage and childbearing to earn income. Also, school may introduce young women to new ideas or values that could influence the number of children they want and also their use of family planning.

Women's education is also associated with better child health. Children of mothers with some education have fewer risk factors for infant mortality. Infants are at a higher risk of dying if they are born to adolescents or to mothers older than age 40, if they are born into large families, or if they are born less than 2 years after an older sibling.

By delaying marriage and childbearing, education reduces high-risk births to teenage mothers. In Indonesia, for example, 52% of women ages 20–24 who completed less than 7 years of education had a baby by age 20, whereas only 13% of those with 7 or more years of education had a baby by age 20. The gap was less pronounced in Kenya, as shown in Fig. 5, but greater in Peru and Egypt in the early 1990s. Women who have completed some formal education tend to wait longer between pregnancies and births and to stop childbearing at a younger age than do less educated women. Consequently, they have smaller families and have fewer births after age 40.

In most societies, children of mothers with some education have a lower risk of dying than children whose mothers had no education. As shown in Fig. 6, in Zambia, the IMR was 133 for the children of mothers with no education, whereas it was 82 for children of women with a secondary or higher education. The

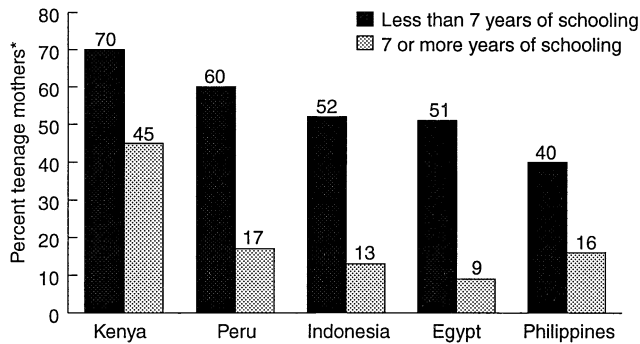


FIGURE 5 Mother's education and teenage childbearing in selected countries, early 1990s. \*Percentage of women ages 20–24 who had a child by age 20 (reproduced with the permission of The Alan Guttmacher Institute from The Alan Guttmacher Institute (AGI), *Hopes and Realities: Closing the Gap between Women's Aspirations and Their Reproductive Experiences*, p. 18, AGI, New York, 1995).

twentieth century brought enormous improvements in literacy and educational levels. The recent improvements in literacy rates reflect the expansion of educational services throughout the world. The United Nations Educational, Scientific, and Cultural Organization (UNESCO) reported that 77% of people over age 15 were literate in 1995 compared with only 56% in the 1950s. Basic literacy is nearly universal among populations of Europe, North America, and other industrialized regions, but the range is substantial throughout the rest of the world. In 1995, an estimated 50% of the

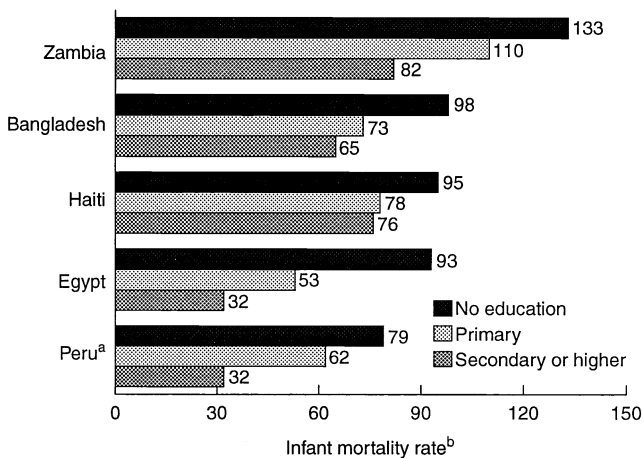


FIGURE 6 Mother's education and infant mortality in selected countries, mid-1990s. <sup>a</sup>Excludes women with higher than secondary education. <sup>b</sup>The infant mortality rate is the number of deaths to children less than 1 year of age per 1000 live births in a given year (source: Demographic and Health Surveys).

populations of south Asia were literate, as were 57% of the populations in sub-Saharan Africa and the Middle Eastern Arab states. More than 83% of the populations are literate in east Asia and Latin America and the Caribbean.

Rapid population growth in some countries is undermining improvements in educational attainment. In the sub-Saharan African countries of Angola, Benin, and Togo, for example, economic difficulties and burgeoning numbers of young people caused school enrollment ratios to level or decrease in the 1980s and 1990s (Bledsoe and Cohen, 1993). In the mid-1990s, about 67% of girls and 81% of boys in sub-Saharan Africa were enrolled in primary school, according to UNESCO estimates.

## B. Economic Development and Environment

In most societies, poor families have higher mortality and fertility than affluent families. Some of the association between poverty and population reflects the lower educational levels and rural residence of poor households. However, the relationship among demographic variables, poverty, and affluence is highly complex, and it is tied to the broader question of how population size and the pace of population growth are linked to economic development. The issue is further complicated by questions about whether economic growth and human activity are causing irreversible damage to the natural environment.

The research into these questions has yielded contradictory results. The extremes of these differences are characterized by two opposing camps: "pessimists" and "cornucopians" (Marquette and Bilsborrow, 1997). The theoretical foundation of the pessimist view can be found in writings published in 1798 by the economist Thomas Malthus. He suggested that the potential population size is limited by the amount of cropland, and therefore food, available for human consumption. Malthus assumed (based on his observations of eighteenth-century English society) that if population growth continued unchecked, population would outstrip the food available and cause widespread famine and death. He also described a natural feedback mechanism: When the population grew too large for the available food supply, elevated mortality would reduce the population to the level that could be sustained by the amount of food produced.

A neo-Malthusian view of the relationship between population, economic growth, and resources gained

credence between the 1940s and the 1960s, a period of unprecedented population growth and economic development. In a landmark study in the 1950s, Ansley Coale and Edgar Hoover found that population growth slowed economic development and held down per capita incomes. These researchers also assumed that the supply of some natural resources and capital was fixed, or that supply would grow more slowly than population. Other researchers during this period expanded the idea that rapid population growth will eventually reach some absolute limit on resources and that population growth accompanied by the environmental stresses associated with economic development could cause irreversible damage to the basic natural systems that sustain life. Such concerns were popularized by books such as *The Population Bomb* by Paul Ehrlich, *The Limits to Growth* by Dennis Meadows, and also by other authors.

The ideological basis of the “cornucopian” approach owes much to the writings of Ester Boserup in the 1960s and 1970s, who argued that the need for more food coupled with the synergy created by the concentration of intellects and flow of ideas in dense settlements can stimulate the adoption of better farming techniques or sharing of higher yield plant varieties. Economist Julian Simon, in *The Ultimate Resource* and other writings, also rejected the idea that population growth was a threat to the welfare of humans or the environment.

The scientific evidence concerning the effects of population size and growth on economic development was still inconclusive in the early 1990s at the national and regional level, but it was less ambiguous at the individual level. An extensive review of research completed in 1994 by the Overseas Development Council observed that “the clearest evidence of negative effects of population growth under high fertility are at the individual and household levels.”

In the late 1990s, several studies provided a clearer picture of the relationship between population and development at the national level and the links between poverty and demographic factors at the household level. Several studies suggested that a rapid transition from high to low fertility contributed to the economic miracles in South Korea and other east Asia countries (Mason, 1997). The rapid fertility decline increased the share of working-age people in the population relative to young and old dependents, which created a “demographic bonus.” The working-age population adds more to the economy than it consumes in services and generates taxes and savings that can be invested in education and further economic growth. This demographic bonus may last several decades; it recedes as the bulge of

working-age men and women reach retirement age and the proportion of the dependent population relative to the working-age group population increases again.

Research shows that countries can benefit from this bonus only if they increase the value of their human capital—especially the youth entering the labor force—through education, and if governments adopt policies favoring international trade and industrialization. The newly industrializing Asian countries capitalized on their demographic bonus by making these investments. As researcher Andrew Mason noted, these Asian countries “raised millions of people from abject poverty and transformed some of the poorest economies in the world to some of the richest.” Research on the relationship between population change, economic development, and environmental systems remains much less clear, plagued by the complexity of the relationships and the difficulty of measuring such factors as environmental quality. At a symposium on population change and economic development in 1998, scientists assigned highest priority to research on population–development–environment linkages.

### III. POPULATION PROSPECTS: 2000–2050

In the past century, the world’s population has undergone a sweeping change in both its total numbers and its distribution across regions. The twenty-first century is likely to see the second phase of this transformation—lower fertility and an even more dramatic redistribution of population among the more developed and less developed countries. Nearly all future world population growth will take place in less developed countries. In other words, the earth is reinventing itself demographically.

Because mortality is relatively low, fertility levels and trends will determine future population size. A common issue (and a common assumption) is when, or whether, a country will reach the “magic” replacement-level TFR of about 2.1 children per woman. With fertility at replacement level, a population eventually will cease growing and “stabilize” at a given size.

Every 2 years, the UN Population Division produces a set of population projections for every country. By 2050, the UN suggests that total world population will increase to between 7.3 and 10.7 billion. In the high projection, world population will still be increasing in 2050; under the low projection series, it will have begun



a gradual decline. All three projections assume continuing declines in fertility. They differ in terms of fertility levels attained in 2050, from 1.57 for the low projection to 2.5 for the high, illustrating the major differences in total population size that can result from seemingly very small differences in childbearing during the next 50 years.

Regardless of the projection used, the UN indicates that at least 1.3 billion people will be added to the world's population during the next 25 years. There are three reasons for this inevitable growth. First, fertility in less developed countries is on average twice as high as in more developed countries. Second, the young age structure of less developed countries constitutes momentum for population growth for several decades, no matter what future fertility trends may be. Third, continuing improvements in mortality will contribute to additional growth, particularly in countries in which life expectancy remains comparatively low.

#### IV. RESPONSES TO HUMAN POPULATION GROWTH

Anxiety about the negative effects of rapid population growth and excessive population numbers has a long history. Long before Malthus, ancient Greeks and Egyptians voiced concern about "overpopulation" in lean times. They also promoted population growth in times of plenty.

In the 1930s and 1940s, scientists and intellectuals in some less developed countries, such as Egypt, India, and Mexico, began to express concern that rapid population growth would hinder development in their countries. Widely publicized food shortages and famines in certain less developed areas in the 1960s were also linked to rapid population growth.

These concerns sparked many actions throughout the world directed at lowering fertility and slowing population growth. India initiated a national policy to slow population growth in 1952. The International Planned Parenthood Federation, the largest private sector organization devoted to family planning, was founded in the same year. United Nations involvement in population issues also increased. The first UN meeting on global population was convened in 1954, in collaboration with the International Union for the Scientific Study of Population. United Nations agencies, including UNICEF and the World Health Organization, incorporated reproductive health services into their

missions. In 1969, the UN Fund for Population Activities became a separate entity. Beginning in the 1960s, governments of some wealthier countries, most notably the United States, supported efforts to strengthen family planning programs in less developed countries.

#### A. Population Policies

The idea that couples should limit their family size went against cultural mores in many societies, and some governments were loathe to support a potentially unpopular policy. Many governments embraced the more acceptable idea that fertility would decrease and that population growth would slow as living standards increased through economic development. This view was expressed at the 1974 UN World Population Conference when an Indian delegate declared that "development is the best contraceptive."

During the late 1970s and 1980s, concern about the negative effects of population growth on economic development broadened. Increasing numbers of countries accepted the idea that government actions could slow population growth.

National efforts to influence population growth include incentives to have more or fewer children and disincentives for having more than a given number of children. These efforts have met with mixed success. Some argue that China's population policies initiated in the 1970s were a success from a demographic perspective. China's TFR decreased from about 6.0 in the 1960s to less than 2.0 in the 1990s, in part because of government policies and programs. However, China's stringent "one-child family" policy introduced in 1979 was widely criticized for violating human rights. Between 1975 and 1977, Indira Gandhi's government in India promoted male sterilization campaigns that sometimes led to coercion. Public outrage about the reported abuses contributed to the downfall of Gandhi's government and created a backlash against family planning programs in India that took years to overcome.

Women's rights activists and others have generally opposed a demographic rationale for family planning as an infringement on individual rights. They have argued that women's rights and well-being should take precedence over national interests. Many have criticized the family planning programs' lack of integration with other health services.

During the 1970s and 1980s, women throughout the world began forming small nongovernmental organizations (NGOs) to lobby for improvements in their social, economic, and political circumstances. By the 1990s,

women's NGOs in less developed countries were advocating for improvements in family planning programs by better informing clients about various contraceptive methods, expanding the range of methods available, and encouraging service providers to treat clients with greater respect.

As of 1997, 155 countries subsidized family planning services, and 68 stated explicitly that they wanted to slow their population growth. In Africa, the world's fastest growing region, 40 countries viewed their fertility levels as too high and 36 had policies to lower fertility. A few countries, in contrast, viewed their fertility rates as too low and welcomed faster population growth. In 1997, 23 countries reported to the UN that they had explicit policies to increase birth rates. Many governments in Europe and the former Soviet Union are concerned that continued low levels of fertility will cause rapid population aging and an eventual decline in population size. Some small oil-rich countries in the Persian Gulf also want to increase, or at least maintain, current levels of population growth. They view population growth as a way to spur socioeconomic development and reduce their reliance on foreign labor.

### B. A New Vision: The 1994 International Conference on Population and Development

In 1994, the world community redefined the world's view of population growth and the best way to address this growth. The Program of Action adopted at the International Conference on Population and Development (ICPD) in Cairo, Egypt, agreed to the need to stabilize global population growth. One hundred and eighty countries agreed to a strategy placing population within the context of sustainable development and calling for investments in human development, especially improvements in women's status, as key to stabilizing population growth. It rejected the use of demographic targets by family planning programs and it integrated family planning into a broader women's health agenda.

Participation in the meeting by NGOs was a critical factor in achieving such widespread consensus. Twelve hundred NGOs participated as delegates and observers, and for the first time conference deliberations on global population were informed by a wide range of interests, from the grassroots to the highest levels of government. Women's groups were a driving force behind the strong emphasis on women's empowerment as part of human

development. This focus was also driven, however, by research from the past 30 years that linked fertility declines with reductions in infant mortality, increased use of family planning, and improvements in women's education and other aspects of women's status contributing to their empowerment.

Despite the consensus, the ICPD engendered dissent and debate. Ideological and religious tensions characterized discussions before the conference, deliberations during the conference, and follow-up after the conference. Abortion generated the most highly publicized ideological differences. There was also debate regarding definitions of reproductive health and family and adolescent reproductive rights and responsibilities. None of the 179 nations rejected the central premises and goals of the ICPD, despite the range of political structures, cultures, and religions they represented. This marked the first time in the history of UN conferences that no official delegation rejected the entire document.

The final ICPD document defined reproductive health to encompass a broad range of services, including family planning, prenatal and postnatal care, medical attention at birth, cancer screening, and protection from sexually transmitted diseases. It also supported access to safe abortion where it is legal, but it stated that abortion should not be used as a method of family planning.

The ICPD Program of Action specified five goals for 2015 to improve individual and family well-being and enhance women's status, including universal access to family planning and other reproductive health services; universal access to primary school education; increased access by girls and women to secondary and higher education; and reductions in infant, child, and maternal mortality. The ICPD document also called for government and private sector actions to alleviate poverty, protect the environment, encourage greater male involvement in the family, and address the specific health and education needs of adolescents.

A unique feature of the agreement was that it called for specific financial commitments to achieve these goals. Countries agreed that less developed countries would pay for two-thirds of the costs of reproductive health programs, including family planning, and that more developed countries would pay for the remaining costs, estimated to be \$17 billion per year by the year 2000. The historic agreements reached at the ICPD were reaffirmed at subsequent UN conferences in the 1990s, including the World Summit for Social Development, in Copenhagen in 1995; the Fourth World Con-

ference on Women, in Beijing in 1995; the UN Conference on Human Settlements (or Habitat II), in Ankara in 1996; and the World Food Summit, in Rome in 1996.

## V. CONCLUSION

Population change has been one of the most significant events of the twentieth century. Since 1900, the world population has more than tripled in size and average life expectancy has increased by two-thirds. Declines in childbearing and shifts in population distribution were more striking than at any other time in history. Along with these population changes, the world has witnessed extraordinary improvements in technology, communication, education, and agriculture. These changes have undermined the dire predictions of Thomas Malthus and his successors that population growth would lead to worldwide famine and disease. However, these predictions may come true for some areas of the world. According to the UN, more than one-fifth of today's population lives in poverty, subsisting on less than \$1 a day. The HIV/AIDS pandemic threatens the health and well-being—and the very survival—of large portions of the population in many countries.

Under all likely scenarios, the twenty-first century will see continued population increases, at least during the first few decades. This is because of the built-in momentum of growth associated with the very young age structures of most less developed countries. The growth will also be fueled by childbearing levels that are still higher than replacement levels in most areas of the world. Not all countries will experience this growth, but they will all be affected by it.

At the end of the twentieth century, the world community made financial and program commitments to continue investments in family planning and other health programs, in education, and in greater social and economic opportunities, especially for women, arguing that these are key to future population stabilization.

## See Also the Following Articles

ECONOMIC GROWTH AND THE ENVIRONMENT • ENERGY USE (HUMAN) • HUMAN EFFECTS ON ECOSYSTEMS, OVERVIEW • SUSTAINABILITY, CONCEPT AND PRACTICE OF

## Bibliography

- Ashford, L., and Makinson, C. (1999). *Reproductive Health in Policy and Practice: Case Studies from Brazil, India, Morocco, and Uganda*. Population Reference Bureau, Washington, D.C.
- Baudot, B. S., and Moomaw, W. R. (Eds.) (1997). *The Population, Environment, Security Equation*. Macmillan, New York.
- Birdsall, N., and Sinding, S. (1999). Chairman's report on the Symposium on Population Change and Economic Development, Bellagio, Italy, November 2–6, 1998.
- Bledsoe, C. H., Casterline, J. B., Johnson-Kuhn, J. A., and Haaga, J. G. (Eds.) (1999). *Critical Perspectives on Schooling and Fertility*. National Academy Press, Washington, D.C.
- Castles, S., and Miller, M. J. (1998). *The Age of Migration*, 2nd ed. Guilford, New York.
- Cohen, J. E. (1995). *How Many People Can the Earth Support?* Norton, New York.
- Cohen, M. A., Ruble, B. A., Tulchin, J. S., and Garland, A. M. (1996). *Preparing for the Urban Future, Global Pressures and Local Forces*. The Woodrow Wilson Center Press, Washington, D.C.
- Gelbard, A., Haub, C., and Kent, M. M. (1999, March). World population beyond six billion. *Population Bull.* 54.
- Haub, C., and Cornelius, D. (1998). *1998 World Population Data Sheet*. Population Reference Bureau, Washington, D.C.
- Jain, A. (Ed.) (1998). *Do Population Policies Matter?* Population Council, New York.
- Livi-Bacci, M. (1992). *A Concise History of World Population*. Blackwell, Cambridge, MA.
- Lutz, W. (Ed.) (1996). *The Future Population of the World. What Can We Assume Today?* International Institute for Applied Systems Analysis, Laxenburg, Austria.
- Malthus, T. R. *An Essay on the Principle of Population as It Affects the Future Improvement of Society. With Remarks on the Speculations of Mr. Godwin, M. Condorcet, and Other Writers*. Reprinted in *On Population* (G. Himmelfarb, Ed.). Modern Library, New York (1960).
- Population Reference Bureau (1997). *Improving Reproductive Health in Developing Countries*. Population Reference Bureau, Washington, D.C.
- Sen, G., Germain, A., and Chen, L. C. (Eds.) (1994). *Population Policies Reconsidered: Health, Empowerment, and Rights*. Harvard Univ. Press, Cambridge, MA.
- United Nations (1996). *Levels and Trends of Contraceptive Use as Assessed in 1994*. United Nations, New York.
- United Nations (1998). *World Population Prospects: The 1998 Revision*, Vol. 1. United Nations, New York.
- The World Bank (1994). *Averting the Old Age Crisis*. Oxford Univ. Press, New York.



# POPULATIONS, SPECIES, AND CONSERVATION GENETICS

David S. Woodruff  
*University of California San Diego*

---

- I. Genetic Variation and Its Significance
  - II. Levels of Interest to Conservation Geneticists
  - III. Evolutionary Processes of Interest to Conservation Geneticists
  - IV. Genetic Management: Examples
  - V. Genetics, Evolvability, and the Future of Biodiversity
- 

## GLOSSARY

**evolution** Biological and physical processes leading to heritable change in characteristics of populations or species over time. Key agents of evolutionary change are mutations, genetic drift, gene flow, and natural selection. The ability of a species to adapt to environmental changes is controlled genetically and is here termed evolvability.

**extinction** The final disappearance of a species unable to evolve, to adapt to changes in its local physical or biological environment, or to shift its geographic range to avoid such changes. Extirpation is the local extinction of a population on a habitat patch that may subsequently be recolonized by dispersal from other populations of the same species. Risk of extinction is codified, and vulnerable, endangered, and critical imply a significant probability of extinction within years or decades rather than centuries; threatened is used here as a general term.

**fitness** Life time reproductive success of an individual

with a particular genotype relative to another individual with a different genotype. Natural selection typically favors "survival of the fittest" and thus facilitates the future evolvability of a population.

**genetic enhancement** Management actions taken to increase the genetic variability and viability of a population; includes translocations and reintroductions.

**genetic erosion** A population viability-threatening process in small, isolated populations whereby random genetic drift and inbreeding diminish the population's innate genetic variability. Increasingly, it is associated with habitat fragmentation.

$N_e$  Symbol for the genetic effective size of a population, an important indicator of future evolvability or extinction risk; usually much less than the observed census size ( $N$ ).

**variation** The abundant normal genetic differences between individuals in a population. Such variation is ultimately due to changes at the DNA base pair sequence level but is typically monitored at higher levels of expression. Overall genetic variation is associated with population viability and future evolvability. Conservation genetics is focused mainly on the protection and maintenance of genetic variation.

---

**CONSERVATION GENETICS** is concerned with population genetic variation, population viability, and the future evolution of species. Conservation genetics, ecol-

ogy, and habitat management together provide the technical underpinnings of conservation biology, a crisis-oriented science of biodiversity management. Still in its infancy, conservation genetics focuses on the characterization of variation in populations and species and on the management of innate levels of variation in evolutionarily significant units in nature and in their captive or managed analogs. Conservation genetic methods are borrowed from evolutionary biology and molecular genetics and are under development. Although some genetic management principles flow directly from current evolutionary theory, several key scientific problems remain to be solved before we can effectively deal with the issues presented by the biodiversity crisis. Although single-species ecological methods have dominated conservation management practice, it is clear that maintaining the future evolvability of species will require greater genetic intervention in the future. Conservation genetics is thus a cornerstone of biodiversity conservation.

## I. GENETIC VARIATION AND ITS SIGNIFICANCE

Until the 1960s, it was widely thought that genetic variation was unusual. Biologists regarded populations as being composed of many very similar "wild-type" individuals and a few rare mutant individuals. Following the introduction of allozyme electrophoresis it quickly became clear that most plant and animal populations were highly variable. Subsequently, it has become possible to examine DNA sequences directly and large numbers of base pair mutations have been discovered in functional genes of most living organisms studied. Even higher levels of variation occur in the introns and highly repetitive stretches of noncoding "junk DNA" that lie between eukaryote genes. High levels of variability characterize most natural populations of plants and animals and this variability is thought to be both a product of evolutionary agents and a determinant of the future evolvability of a population. Genetics became a cornerstone of conservation science in 1981 with the publication of Frankel and Soulé's *Conservation and Evolution*. Since then, geneticists have studied the effects of the ongoing global reduction of genetic variability on population viability and species persistence. Key reviews of the progress made include Schonewald-Cox *et al.* (1983), Loeschcke *et al.* (1994), Olney *et al.* (1994), Frankham (1995), Avise and Hamrick (1996), Smith and Wayne (1996), and Landweber and Dobson (1999).

Estimates of variability in a population vary with the feature examined and analytical method used. Karyotypic variation is usually low within a species, allozyme variation is high, and DNA sequence variation may be very high in some parts of the genome. The principle of high natural variability is perhaps best illustrated by the published surveys of allozyme patterns in plants and animals (Schonewald-Cox *et al.*, 1983). Geneticists typically sample a population of 20 individuals and determine each individual's genotype at 10–20 loci. This allows them to calculate the mean number of alleles per locus ( $A$ ), the mean individual heterozygosity ( $H$ ), and the proportion of loci in the population that are polymorphic ( $P$ ). A few thousand species have been examined to date and variation is typically in the following range:  $H = 0.05\text{--}0.15$  and  $P = 0.20\text{--}0.50$ . Still higher levels of natural variation are found at the DNA sequence level, especially in areas of the genome which are apparently free to mutate. Short repeat sequences (microsatellites) of nuclear DNA, for example, are often 10 times more variable than allozyme loci. A polymorphic allozyme may have 2 or 3 alleles segregating in a population, whereas a dinucleotide repeat microsatellite may have 10–15 alleles (differing in the number of times the motif is repeated) in the same individuals. Biochemical and molecular genetics have shown that most living organisms are richly variable.

Much of the cryptic variation in natural populations that has been discovered in the past 30 years appears to be selectively neutral or near neutral in its effect on the phenotype. That is, the individuals carrying these various allelic variants appear perfectly healthy. It is rare to find a single locus genetic trait in which a deleterious condition such as albinism is controlled by a single allelic variant. In fact, there is strong circumstantial evidence that much of this genetic variation is actually beneficial. Although the relationship between genetic variability and individual fitness is not well understood, it is clear that variability is associated with viability. Experiments and field observations with a few species have shown that there is a positive relationship between genetic variability and individual growth rate, size at maturity, symmetry of body parts, fecundity, and health as measured by parasite load. Extrapolating from the individual level to the population level, it is a fundamental maxim of evolutionary biology that genetic variation is positively related to adaptability or evolvability. Fisher's fundamental theorem holds that additive genetic variation in fitness is positively related to a population's ability to respond to natural selection. Evolutionary success, the ability of a species to persist despite changes in climate and exposure to new diseases, predators,

and competitors, is somehow related to innate genetic variability. In a world of unpredictable change, alleles that are selectively neutral for thousands of generations may suddenly become lifesavers for the individuals that carry them. If evolution is largely dependent on genetic variability, then the conservation of species will sooner or later depend on the conservation of their innate genetic variability (Woodruff, 1989).

Of course, there are many apparently successful genetically invariant plants and animals. There are many known clonal organisms whose descendants are genetically identical to their mothers. Such clonal plants, snails, fish, and lizards are genetically invariant even though they may exhibit great ecological success. However, they are more likely to go extinct when their environment changes than are closely related sexually reproducing species. Although meiosis ensures the maintenance of high variability in the sexual species, their asexually reproducing daughter species gradually lose their initial low variability and over time become evolutionary dead ends.

There are many different ways of measuring genetic variability, including the examination of karyotypes and single-locus markers (allozymes, mitochondrial and chloroplast DNA restriction fragment length polymorphisms), DNA minisatellite fingerprints, random amplified polymorphic DNA, mtDNA sequences, nuclear DNA sequences, and nDNA microsatellites (simple sequence repeats). Benirschke and Kumamoto (1991) and Smith and Wayne (1996) provide numerous examples of the application of karyotypes and molecular genetic markers, respectively, to conservation. The various methods differ in their resolution of pedigree, population, and species-level questions and there is no single correct technique. One method, allozyme electrophoresis, is informative with mammals but frustratingly uninformative for birds. The methods also differ greatly in their cost and technical difficulty. Some relatively inexpensive methods are satisfactory for one-time analyses but the results cannot be built on in subsequent studies. Other methods, in contrast, provide genotypes which can be archived in expandable permanent databases such as GenBank for future comparative study of samples collected at other times or places. Although a given method will give comparable results in a study of closely related populations or species, results across unrelated taxa may not be comparable.

Currently, nuclear and mitochondrial sequence data are the most informative methods for characterizing variability at or above the level of populations. For studies of variation within populations, hypervariable nuclear microsatellite loci are ideal markers. Sequence

data and microsatellite genotypes can now be determined directly from minute DNA samples amplified many times by the polymerase chain reaction. Both methods require fully equipped laboratories, trained personnel, and considerable time and money for developing the synthetic DNA primers to amplify the gene sequences of interest in species that have never been studied before.

Noninvasive genotyping methods involving the extraction of DNA from shed hair and feathers were introduced in 1989 and are now widely used (Morin and Woodruff, 1996). Noninvasive (shed tissues, feces, urine, and scent markings) and nondestructive (toe, tail, and ear clips and fish scales) genotyping permits the study of wildlife populations that previously were almost impossible to sample. Not surprisingly, the DNA extracted from some types of samples may be degraded and very difficult to work with. Nevertheless, with technical care and patience it is possible to genotype some animals without actually seeing or handling them.

Conservation geneticists can also genotype museum specimens and determine patterns of variability over periods of decades and sometimes centuries. Several species that have gone extinct recently, including the dodo, moa, thylacine, and quagga, have been characterized genetically. DNA can also be extracted and sequenced from some fossil remains, but DNA degradation rates are such that fossils more than a few million years old are unlikely to yield reliable sequences. Although enormously interesting to evolutionary biologists, such ancient DNA cannot be used to recreate extinct organisms. Conservation geneticists must concentrate on saving existing biodiversity; they cannot fall back on the idea of being able to recreate extinct species from the tiny fragments of DNA (typically less than 0.00001% of an organism's genome) currently under scrutiny.

Genetic data, once acquired, are used by conservation geneticists to quantify within- and between-population variability. Variability within populations is used to establish pedigree relations, mating system, sex ratio, and genetic effective population size ( $N_e$ ). Interpopulation comparisons reveal spatial structuring and historical patterns of gene flow. Geographic variation is normal within a species and the new field of phylogeography uses genetic information to infer the historical relationships among populations. Single and multilocus genetic differences between kin, populations, subspecies, and species are expressed as genetic distances. These metrics and their interpretations are beyond the scope of this review, but it is important to note that the absolute values of genetic distances will vary between different

groups of plants and animals and increase over geological time. Within a group of related species a major difference in the genetic distances observed within and between taxa can be used to define species and other evolutionarily significant units for conservation management purposes.

Although the previous discussion focused on molecular genetic variation, the growing field of quantitative trait genetics promises to provide a new means of measuring evolutionarily significant population variables (Storfer, 1996). Quantitative genetics is concerned with characters such as morphology, behavior, parasite resistance, and physiology that are controlled by several to many genes that work additively, in dominance/recessive relationships, or epistatically. Such oligogenic and polygenic control, involving quantitative trait loci (QTL), is characteristic of many traits of interest to conservationists—those that effect long-term population persistence and evolvability. QTLs affecting body size, hatching date, and predator avoidance behavior (escape speed), for example, are ecologically important and are arguably of greater significance to conservationists than allozyme polymorphism. Conservationists are therefore interested in the heritability of such traits that have a direct impact on fitness. High heritabilities of a QTL indicate that a population has great potential for trait evolution and low heritabilities indicate a more limited ability to respond to environmental change. Unfortunately, such heritability is difficult to measure because it requires pedigree-level studies conducted over several generations or long-term manipulative experiments such as controlled garden plots. Heritability ( $h^2$ ) is the ratio of the variance of the genetically inherited proportion of a trait (the additive genetic variance,  $V_A$ ) to the total phenotypic variance  $V_P$  measured in the population.  $V_A$  is one component of genetic variance ( $V_G$ ) which also includes nonadditive genetic variance due to dominance ( $V_D$ ) and epistasis ( $V_I$ ); only  $V_A$  responds to directional selection. Estimating  $V_A$  is further complicated by the need to estimate variance due to the environment ( $V_E$ ) as well as the other genetic components. Nevertheless, the preparation of studbooks for captive populations of endangered animals and the comparison of laboratory-raised seedlings of rare plants to their parents in the wild have provided opportunities for the first limited applications of QTL analyses to conservation. Future technical advances may permit the inclusion of QTL in the conservation geneticist's tool kit of predictors of a population's risk of extinction.

A major issue requiring resolution is the validity of the widely held relationship between molecular genetic variation and ecological viability and evolutionary po-

tential. Are single gene markers (those typically surveyed by geneticists) useful indicators of variation at quantitative traits? Are multilocus allozyme surveys unreliable as predictors of a population's viability since heterozygosity may be only weakly correlated with the additive genetic variation associated with QTLs? In desert topminnows of the genus *Poeciliopsis*, Vrijenhoek (in Avise and Hamrick, 1996, pp. 367–397) found that rapid losses of heterozygosity in small, isolated populations were associated with a decline in fitness that was manifested as poor competitive ability, growth rate, developmental stability, and resistance to parasites. O'Brien (1994), in a remarkable series of genetic studies of captive and wild cheetahs (*Acinonyx jubatus*), demonstrated a powerful association between low levels of genetic variability and susceptibility to viral diseases.

There is no "normal" level of variation for a population—even as determined with a specific method. Cheetah, northern elephant seal (*Mirounga angustirostris*), and European badger (*Meles meles*) are ecologically successful despite low absolute levels of genetic variation. Different types of genetic variability will respond differently to natural selection, inbreeding, population collapse, and range fragmentation. Conservation geneticists can identify cases in which variation has or is being lost, establish the causes of the loss, and make recommendations to counter its ultimate effects. Because currently available genetic markers are only proxies for fitness determinants, this underscores the importance of using different markers. It is not the purpose of this review to discuss the technical problems associated with each method or the applicability of the different methods to different groups of plants and animals, but it is important to emphasize that each method has unresolved technical problems (e.g., null alleles, allelic dropout, pseudogenes, and nonreplicable patterns). Until molecular and QTL genotyping become routine, conservation geneticists must guard against the natural tendency to overstate the statistical power of their results.

## II. LEVELS OF INTEREST TO CONSERVATION GENETICISTS

### A. Genes

Long before the term conservation genetics was coined, the phrase "genetic conservation" was introduced to describe the science of managing specific genes or phenotypic traits in crop plants, land races and cultivars, bacteria and fungi used in food production, and domes-

ticated animals. Genetically modified organisms (GMOs) are simply special cases that require even more intensive management for their perpetuation. The methods of gene discovery and artificial selection developed for managing microorganisms, plants, and animals are relevant to the far more broadly focused field of conservation genetics, but very few wild species have received such intensive effort.

Discussions of the need to save this or that desired gene in a population are in fact arguments for saving a particular allelic form (variant) of a gene and not the gene itself. Some alleles are common and others are rare. Deleterious alleles (e.g., alleles responsible for albinism or other genetic “defects”) are typically very rare and have frequencies of less than 0.0001. Conservation geneticists are often asked to devise breeding plans that will further reduce or even eliminate such alleles from a population. On the other hand, it has been argued that conservation geneticists should strive to save rare alleles in threatened populations because they may prove vital for a population’s adaptation to future environmental changes. Although this argument is reasonable, the maintenance of desirable rare alleles, even if they were identifiable, requires very large population sizes ( $N_e > 5000$ ) and is simply not possible in most management programs. Rare alleles contribute very little to variation in fitness among individuals and are less likely than alleles at relatively high frequency to be the basis of adaptive response to environmental change.

It has been suggested that conservationists should focus on saving diversity at major histocompatibility complex (MHC) genes because they play a role in recognition of infectious agents, disease susceptibility, and defense. This recommendation was well intended but rejected because the functional (fitness-related) significance of the large number of alleles at each of the many MHC genes is unknown. Managing them as a single linkage group would require very large populations or the inevitable loss of variation at other potentially important loci.

## B. Populations

Interest in the concept of minimum viable populations (Gilpin and Soulé, 1986) spawned the development of new methods of pedigree analysis and population viability analysis. Populations have both a census size ( $N$ ) and a genetic effective population size ( $N_e$ ); the latter is one of the most important concepts in theoretical conservation biology and is defined in Section III,G and by Lande and Barrowclough (1987). The characterization of genetic variation within and among popula-

tions enables geneticists to help set conservation priorities. Comparative levels of variation and gene flow (or lack of either) provide clues to a population’s viability and extinction proneness. Data on genetic relationships among populations guide translocation decisions and identify well-defined clusters of populations for management as separate entities.

Pedigree analysis refers to a suite of genetic models for understanding processes in small populations. First developed for assessing management practices in captive populations, pedigree analysis is also applicable to wild populations of individually monitored organisms. It is used to establish kinship and individual founder contributions, to identify genetically desirable and undesirable individuals and their descendents, to minimize inbreeding, to describe population structure and mating system, and to choose individuals for reintroduction or translocation. Examples of pedigree analysis with the Gene Drop computer program include Haig’s (1998) study of the red-cockaded woodpecker, *Picoides borealis*. Pedigree management programs based on mean kinship or equalizing founder contributions seek to minimize inbreeding in local subpopulations and in the metapopulation as a whole. Captive breeding programs have been successful in slowing the loss of genetic variation and preventing inbreeding depression.

Population viability analysis (PVA) is the methodology used to assess the ecological and genetic risks facing a wild or captive population and to develop a conservation management plan. PVA refers to a suite of mathematical models that seek to predict the probability of a population’s extinction by some time in the future, e.g., 20 or 100 years. Early models considered demography (growth rate, present population size, and birth rate) and environmental stochasticity, but Gilpin and Soulé (1986) broadened the definition to include genetic factors. Genetic factors, including genetic drift and fixation of deleterious mutations, are expressed through demographic factors that affect population dynamics. The genetic factors thus contribute to extinction probabilities through a very complex and little understood series of interactions affecting fitness. Geneticists can use population variation to provide estimates of the various parameters of interest to modelers and managers. Unfortunately, genetic models with linkages to ecological factors are still insufficiently developed to yield the type of statistically powerful PVAs managers seek. Most PVAs, in fact, have not included genetic parameters, but this is changing as the significance of genetics in the long-term survival of small populations becomes more widely appreciated (Beisinger and McCullough, 2000).



Metapopulations are populations of subpopulations within some defined area, in which dispersal from one local population (subpopulation) to at least some other habitat patches is possible. There is significant turnover of local populations, local extinction, and recolonization by dispersal. The metapopulation concept is central to much ecology and conservation theory and single- and multiple-species metapopulation dynamics are reviewed by Hanski and Gilpin (1996). The genetic effective size of a metapopulation is affected by the carrying capacity of the habitat patches, the rates of extirpation and recolonization, the number and source of the founders, the number of local populations, and the rate of gene flow between patches. It is difficult to establish metapopulation effective size using genetic data because it is strongly affected by extirpation and recolonization dynamics. As in the case of genetic effective population size of single populations, the metapopulation effective size is 10–100 times less than the census size in many species. Metapopulation dynamics, with frequent local extinction and recolonization of habitat patches by few founders, can reduce  $N_e$  to a small fraction of  $N$ , with a resulting loss of genetic variability like that associated with a demographic bottleneck. Detailed studies of metapopulations are reviewed by Hanski and Gilpin (1997) and include the Glanville fritillary butterfly (*Melitaea cinxia*) and the red bladder campion (*Silene dioica*), both in Sweden; the checkerspot butterfly (*Euphydryas editha*) restricted to serpentine outcrops in California; and the pikas (*Ochotona princeps*), a small mammal restricted to isolated talus slopes in alpine areas. Although the theory of metapopulations is well developed, and the relevance of metapopulation theory to the management of small, semi-isolated populations of threatened species is clear, the empirical testing of the ecological and genetic predictions is only just beginning.

Using molecular markers to characterize mating system, population structure, and phylogeography permits recognition of management units (MUs; sets of populations with shared distinctive alleles frequencies) and evolutionarily significant units (ESUs; sets of populations distinguished by strong phylogenetic structure based on multilocus mtDNA or nDNA variation). Moritz (in Smith and Wayne, 1996, pp. 442–456) describes Australian examples of MUs (the yellow-footed wallaby, *Petrogale xanthopus*, which is patchily distributed on isolated rock outcrops in southwest Queensland) and ESUs (four ghost bat ESUs previously treated as a single widespread species, *Macroderma gigas*, whose evolutionary divergence was not apparent until genetic methods were applied). Although the criteria used to define these two categories are not agreed upon, ESUs repre-

sent deep phylogenetic subdivisions typically within a species or occasionally, as in the case of local endemics, the entire species, and MUs reflect shallower subdivisions within a species which for practical reasons become the focus of management activities. Translocations between ESUs are typically undesirable.

### C. Subspecies

Subspecies or local geographic races have been the focus of intense conservation efforts. Genetics is useful in establishing whether such groups of populations are sufficiently different to warrant separate conservation efforts. The conservation of every local race or subspecies is difficult to justify if they are genetically almost identical. Although we may recommend trying to preserve every variety of wild tomato or maize in a seed bank, it is difficult to use the same justification to try to conserve all local variants of geographically widespread organisms.

Many biologists have abandoned the subspecies concept and the associated trinomial nomenclature. To them, the subspecies is an evolutionarily insignificant artificial taxon typically based on a few superficial morphological features. There are numerous cases in which genetic studies provide no support for the traditional subspecific taxonomy. Nevertheless, there is much to be learned from the geographic patterns of variation in nature. Subspecies were typically defined as geographical races with allopatric or parapatric distributions. The observation of significant hybridization without introgression, in the latter cases, led to the development of the semispecies concept which is of relevance to conservation. Semispecies are typically parapatrically distributed and hybridize where their ranges meet, but they show very limited introgression. They are effectively isolated groups of populations evolving independently of one another. Collectively, a group of semispecies comprise a superspecies and each semispecies is treated taxonomically as a full species. In this case, geographically defined taxa that used to be treated as subspecies are actually independent evolutionary lineages and therefore as worthy of conservation as other “good” species.

Conservation geneticists are often asked to advise on proposals to pool individuals from allopatric populations on the argument that it is better to save a generic species than no species. Although poorly differentiated subspecies may often be mixed without genetic harm, the pooling of individuals of well-differentiated semispecies or species is likely to have negative genetic consequences. Unfortunately, management decisions are of-

ten made before the appropriate genetic data are available.

Geneticists are often involved in two other situations involving subspecies and local variants. The first involves populations on either side of national or state boundaries that may be assigned to different subspecies and can receive radically different levels of protection; a species-wide conservation plan may allocate resources differently and be biologically preferable. The second situation occurs when peripheral populations of a widespread species become the focus of conservation activities. Although such peripheral populations may be at high risk of local extinction, their conservation may not be warranted, especially in cases in which reintroduction is practical. On the other hand, some peripheral populations may be critical to a species' long-term survival. Some peripheral populations may be better adapted to changing climatic conditions than central populations even though the latter may be more genetically variable. At a time of global warming, poleward peripheral populations of a species may be more important than those closest to the equator.

#### D. Species

As it is thought that half the species of larger vertebrates are at risk of extinction in the next 100 years, most discussions of the conservation of biodiversity focus on species. Species are fundamental units of evolution and classification (taxonomy). However, despite their centrality to the field, conservation geneticists rarely work at the level of whole species and concentrate instead on infraspecific levels of organization.

Ernst Mayr introduced the biological species concept in 1942; species were defined as groups of actually or potentially interbreeding natural populations that are reproductively isolated from other such groups. This concept stimulated an enormous amount of research in the second half of the twentieth century. Despite its impact, problems with Mayr's original definition, with its overemphasis on reproductive isolation and its focus on sexually reproducing outcrossing populations over short time spans, led to the development of at least 18 alternative concepts by the end of the century. Although some of these have limited operational utility, three newer biological species concepts are of particular relevance to conservation geneticists: Wiley's evolutionary species concept, Templeton's cohesion species concept, and Mallet's genetic cluster species concept. All seek to recognize discrete groups of populations with a shared evolutionary future. All three concepts, like Mayr's, seek to capture the essential genetic relatedness within, and

the genetic distance between, biological species. All allow that the absolute values of observed relatedness and distance will vary with the geological age of a species and will vary in different groups of plants and animals.

Mayr's biological species concept forced researchers to search for reproductive isolating mechanisms between species in nature and to investigate their potential significance in laboratory and greenhouse hybridization experiments. Such work typically took years and the results were often compromised by methodological limitations. It became clear that morphology was not always a good indicator of species boundaries and not surprisingly that traditional taxonomy was often a poor guide for conservation decision making. The introduction of protein electrophoresis and molecular genetic methods of measuring genetic variation has dramatically changed the approach to defining species. It is now possible to quickly establish whether populations exchange genes or whether they are effectively reproductively isolated. It is possible to estimate genetic distances between groups of populations and gauge their significance in comparison to within-population variation. It is possible to estimate the time since a speciation event and the historical patterns of gene flow within and between taxa. Our newly found ability to characterize and recognize species genetically does not diminish the value of field and laboratory studies of behavioral ecology, but it does permit geneticists to make powerful contributions to the practice of conservation.

Although it is fortunately still unusual for geneticists to work on the conservation of entire species, there are an increasing number of cases in which every individual in a species has been genotyped to some degree for management purposes. Such cases include Przewalski's horse (*Equus przewalski*), San Clemente loggerhead shrike (*Lanius ludovicianus mearnsi*), whooping crane (*Grus americana*), and Catalina mahogany (*Cerocarpus traskiae*).

#### E. Communities

Although it is unusual to think of geneticists working at higher levels of organization than species, regional multispecies phylogeographic surveys are useful in defining the historical interactions of whole communities of organisms. Studies of regional phylogeographic structure, as for example in the southeastern United States or savanna ecosystems of Tanzania, are relevant to the design and maintenance of biodiversity sanctuaries (Avice, 2000). Elsewhere, populations in ecotonal regions have been found to have higher gene diversity

and are thus recommended for higher conservation priority (Smith and Wayne, 1996).

### III. EVOLUTIONARY PROCESSES OF INTEREST TO CONSERVATION GENETICISTS

#### A. Mutation

Mutations, the occurrence of heritable changes in the genetic material, are typically very rare processes that ultimately provide the raw material of genetic variation upon which the other agents of evolution operate. Mutations span a wide array of phenomena; from single base pair changes in the genetic code to accidental doublings of the number of chromosomes in a gamete. Many mutations are deleterious or lethal, some are near neutral, and a few may be beneficial to the carrier. The vast majority of mutations are completely invisible in the phenotype and can only be detected with an array of genetic techniques.

Mutations become of concern to conservationists in a couple of circumstances. First, the presence of a normally very rare allele of major effect in a remnant or closed captive population can have serious consequences. Deleterious traits discussed by Ryder and Fleischer (1996) include hairlessness in red-ruffed lemurs (*Varecia variegata ruber*), funnel chest in black-and-white ruffed lemurs (*Varecia variegata variegata*), and congenital diaphragmatic hernia in golden lion tamarin (*Leontopithecus rosalia*). Second, artificial selection for rare alleles is sometimes the goal of captive breeding programs. White tigers (*Panthera tigris*), homozygous for a recessive allele, are beautiful but suffer genetic disease with metabolic, nervous, and developmental consequences. Selection for such traits in the context of the conservation of an endangered species is unjustified.

As most populations of conservation concern lose genetic variability, the question arises as to whether new mutations will replace variability lost by population extirpation and genetic erosion. The answer is yes and no. Given that mutation rates are typically on the order of one per  $10^5$  cell divisions, the time for the accumulation of new variants in a population is on the order of tens of thousands of years. Therefore, conservation geneticists are more concerned with the deleterious effects of mutations in small populations than with their very long-term benefits.

Extinction due to genetic causes is almost unknown, but their contribution to the process should not be

ignored. Although natural selection purges deleterious alleles from populations almost immediately, mildly deleterious, near-neutral mutations will gradually increase in frequency and become serious problems when their frequencies exceed 0.05 or  $1/(2N_e)$ . Fortunately, this process takes hundreds of generations in all but the smallest isolated populations. Therefore, although such mildly deleterious mutations are rarely considered by wildlife conservationists, they will ultimately diminish the long-term viability of many threatened taxa. On the other hand, managers of captive populations have to recognize this threat from the outset. If the goal of a breeding program is to return a captive population to the wild, then managers should maximize genetic variation including mildly deleterious mutants. Alternatively, if a population cannot be returned to the wild and must be sustained in captivity for many generations, then managers will need to purge deleterious mutations as they are identified.

Mutation rates at near-neutral genes controlling quantitative characters set a lower limit on the population size necessary for future evolution. Harmful mutation rates set lower limits for population sizes for avoiding inbreeding depression and for preventing genetic erosion of fitness by the accumulation of mildly deleterious mutations. The suggestion that small populations ( $N_e < 100$ ) may decline in fitness with the accumulation of mildly deleterious mutations, termed mutational meltdown, is under theoretical and experimental study.

#### B. Mating System

Although much population genetic theory is premised on "random mating," such behavior is rarely observed in nature. Even related species may have very different mating systems with very different genetic consequences. Self-fertilization in hermaphroditic organisms and obligate outbreeding in dioecious organisms are the two extreme modes. Conservation managers have to be aware of these differences if they are to mimic a species' natural history.

The most extreme examples of conservation problems involving mating systems involve cases in which the last surviving individuals in a sexually reproducing population all belong to the same sex. The last passenger pigeon (*Ectopistes migratorius*) was a female; the last member of one of the 11 surviving subspecies of Galapagos tortoise (*Geochelone elephantopus*) is a male, Lonesome George. It is likely that cloning technologies currently under development will be applicable to saving such lineages in the future.

### C. Inbreeding

Inbreeding refers to the mating of close relatives in species that are normally outbreeding. Matings between father and daughter, brother and sister, or first cousins are examples of inbreeding. Many species of plants and animals have evolved devices to minimize close inbreeding. Species vary greatly in their tolerance of inbreeding; some trees and dioecious plants are obligate outcrossers. The genetic underpinnings of inbreeding depression are best understood in *Drosophila*, in which recessive lethal mutations and mildly deleterious mutations are major causes. Gradual inbreeding permits natural selection to purge the former but the partially recessive near-neutral mutations continue to increase in frequency and significance. Outcrossing populations that suddenly decline in numbers usually experience reduced viability and fecundity known as inbreeding depression. Inbreeding produces increased homozygosity of recessive partially deleterious mutants and by chance in small populations these alleles become fixed. In the simplest genetic situation of a trait under the control of a lethal recessive allele, there is an increased risk that the offspring of two related healthy but heterozygous individuals will inherit the harmful allele from each parent and die. Although the risk in this case is only one in four, this amounts to a powerfully strong fitness differential on which natural selection will act. Generalizing from this simplest single-locus case, geneticists speak of inbreeding depression as the manifestation of the whole genomic effects of mating of close relatives. These effects may involve outright genetic disease (congenital abnormalities) but are more often subtle and appear as decreased growth rate, behavioral abnormalities, and reduced fertility and fecundity. Therefore, inbreeding is rare in typically outbreeding populations but becomes a serious problem in small isolated populations. In small fragmented populations in nature and in captive populations, inbreeding depression can threaten population viability. Animal breeders learned these lessons from centuries of experience with artificial selection, and they limit inbreeding rates to less than 2% per generation. The consequences of very close inbreeding are well illustrated by experience with establishing "inbred" strains of laboratory mice; the majority of inbred lines die out within 10 generations.

There is abundant evidence that captive wildlife populations suffer inbreeding depression. Ralls (in Schonewald-Cox *et al.*, 1983, pp. 164–184) was the first to show that even well-intended captive breeding programs subjected small populations to inbreeding depression. She reviewed empirical records for 40 species, mainly ungulates, in zoos and found in-

breeding to be a problem in most cases. Inbreeding is also associated with decreased growth rate and blindness in a captive wolf population in Sweden. In the wild, the final decline of the heath hen (*Tympanuchus cupido*) on Martha's Vineyard island in 1932 involved inbreeding effects. Other better documented cases (Hedrick in Smith and Wayne, 1996, pp. 459–477; Lacy, 1997) involving declining or threatened wild populations involve the middle spotted woodpecker (*Dendrocopos medius*) in Sweden, the Florida panther (*Puma concolor coryi*), Barrow Island populations of the black-footed rock wallaby (*Petrogale lateralis*), common shrews (*Sorex ananeus*) in England, deer mice (*Peromyscus polionotus*), and Glanville fritillary butterfly (*Melitaea cinxia*) in the Aland Islands, Finland (Saccheri *et al.*, 1998).

In 1980, Franklin and Soulé independently showed, based on theory and experiments, that inbreeding depression can be avoided in the short term if  $N_e > 50$ . The inbreeding coefficient  $F$  increases by  $1/2N_e$  per generation and centuries of animal breeding experience shows that a 1% increase in  $F$  per generation is tolerable; thus, an  $N_e = 50$  is necessary to avoid inbreeding depression. Franklin and Soulé further concluded that an  $N_e > 500$  was necessary to enable a population to continue to evolve in the long term. Although this 500 number has been revised upwards, the simplicity of the Franklin–Soulé numbers caught the attention of managers and legitimized the role of genetics in conservation. The theory behind the 50 number is still accepted (Lande, 1999), but it is important to realize that its derivation was based on controlled laboratory experiments; larger  $N_e$ s are required in nature, where environmental fluctuations are more severe and stressful.

### D. Outbreeding

Outbreeding, or the crossing of unrelated individuals, is widespread in nature. It is widely believed that sexual reproduction evolved in part because chromosomal crossing over and recombination facilitated by outbreeding produces more genetic variability than do other mating systems. Many species of plants and animals have effective immunological and behavioral mechanisms to favor outbreeding. The latter include sex-biased dispersal of young adults from their natal population and elaborate courtship behaviors. Even in plants with both male and female flowers, outbreeding is ensured by asynchronous maturation of male and female gametes and the evolution of various self-incompatibility systems.

Outbreeding depression occurs when very distantly

related conspecific individuals are mated or when members of two different but related species hybridize. The male and female genomes are sufficiently different to produce a hybrid with genetic disorders. Conservation geneticists encounter outbreeding depression in inadvertently mixed captive populations. Sterility, or partial sterility in one sex, and high neonate mortality are commonly observed manifestations.

### E. Hybridization

Outbreeding depression occurs in nature in some hybrid zones between semispecies and species of plants and animals. Hybrids are interesting because they show that the evolution of many groups of plant and animal species involves both lineage splitting and lineage anastomosis. Fertile interspecific hybrids permit gene flow between species (introgression). Hybrids call into question species definitions that overemphasize reproductive isolation. The notion that "species" are somehow "purebred" and always reproductively isolated from their close relatives is not borne out by observations of some animal groups and many plants, in which low rates of hybridization between congeners often occur in nature.

In the past, it was argued that hybrid populations did not qualify for legal protection under the U.S. Endangered Species Act. However, hybrids are very much a normal part of nature. Rare or very rare in some groups, and more common in others, they present conservationists with a dilemma because their occurrence appears to diminish the value of a taxon. Should one save Florida panthers if they are known to harbor genes of introduced South American panthers, a different subspecies? Do Texas red wolves (*Canis rufus*) merit conservation if they are gray wolf-coyote hybrids? Should one save the remaining Przewalski's horses if it is shown that historical mismanagement resulted in a large fraction of the surviving animals being tainted by the genes of domestic horses, a karyologically distinct species? Whether hybrids should be afforded the same priority as nonintrogressed populations or species will remain controversial; in the previous cases, the answer was yes and geneticists contributed to pedigree management.

Habitat disturbance can result in increased opportunities for hybridization between species that would not normally interbreed. Fragmentation of the recently continuous Pacific Northwest old-growth forest has led to hybridization between the declining northern spotted owl (*Strix occidentalis*) and the barred owl (*S. varia*), which favors disturbed sites.

Hybridization is more common in plants than in

animals; therefore, not surprisingly it is in plants that there are numerous examples of rare species being hybridized into extinction (genetic assimilation) by hybridization with a more common sympatric congener (Soltis and Gitzendanner, 1999). This is the case for the Catalina Island mahogany, in which 5 of the remaining 11 adult trees are actually hybrids with the more common mountain mahogany (Rieseberg and Swensen in Avise and Hamrick, 1996, p. 305–334). Other cases involve plants (*Asteracea: Argyranthemum*) in the Canary Islands. The Simien jackal (*Canis simensis*) of Ethiopia is at risk of being introgressed into extinction by hybridization with domestic dogs. It is now recognized that restocking rivers with genetically uniform hatchery-bred salmon has contributed to the collapse of the Pacific Northwest salmon runs. Hatchery fish show reduced fitness in the wild (they are not locally adapted) and compete and hybridize harmfully with the remaining wild salmon (Lande, 1999).

### F. Gene Flow

Gene flow is a fundamental agent of evolution based on the dispersal of genes between populations of a species. It involves the active or passive movement of individual plants, animals, gametes, or seeds. Gene flow involves not just dispersal but also the successful establishment of the immigrant genotypes in the new population. Gene flow is often confusingly referred to as migration, but the latter term is best reserved to describe dispersal behaviors involving a seasonal or longer term round-trip. Gene flow tends to homogenize linked populations and lack of gene flow permits interpopulation differentiation. It is of interest to geneticists and managers in that to conserve a population one needs to establish the historical patterns and rates of gene flow. This is typically estimated from allele frequency data and reported in terms of the number of "migrants" per generation. In theory, one migrant per generation between two populations will ensure that they remain genetically homogeneous. Inbreeding depression can be ameliorated by the artificial translocation of one reproducing migrant per generation between populations.

Gene flow is often gender biased and limited to certain phases of the life cycle. It may be accelerated under certain climatic conditions that occur at frequencies of many years or at irregular intervals many years apart. Interspecific gene flow results in introgressive hybridization (discussed previously). The translocation of individual organisms results in gene flow if they reproduce at the release site. In the future, genetically depauperate populations will be enhanced by transloca-

tion of individuals from more secure areas. Unfortunately, such genetic enhancement carries risks associated with the introduction of pathogens that could harm the target population or completely unrelated species. Furthermore, the introduction of individuals from genetically well-differentiated source populations may result in outbreeding depression in the threatened population of conservation concern (discussed previously). Gene flow can thus erode the genetic basis of adaptation to local conditions.

If previously continuous populations become fragmented, historical patterns of dispersal and gene flow may be disrupted with potentially serious consequences for population viability. For example, if young female chimpanzees can no longer emigrate from their natal social group because of habitat destruction in the surrounding countryside, their isolated natal population will experience increased inbreeding.

### G. Genetic Drift

Genetic drift involves the loss of alleles from a population by chance. Random fluctuations in allele frequencies in small populations reduce genetic variation, leading to increased homozygosity and loss of evolutionary adaptability to change.

The rate at which alleles are lost from a sexually reproducing population by genetic drift can be predicted. Sewall Wright (1969) developed the basic theoretical model in 1931 and showed analytically how the rate varies with population size. Actually, it is not the census size ( $N$ ) that is important but rather the genetic effective population size ( $N_e$ ). This parameter takes into account the fact that closely related individuals will share alleles by common descent. Monozygotic twins are genetically identical and therefore should be counted as one individual rather than two. Sibs share half their genes with each other and half with each of their parents and are therefore not equivalent to two genetically unrelated individuals. The genetic effective number of individuals in a population is therefore almost always less than the number of individuals counted by an ecologist.  $N_e$  can, under some breeding systems, be one or two orders of magnitude less than  $N$ . Consider, for example, the number of adults in a sexually reproducing population: In a monogamous species the census count of adults is useful, but in a harem species only 1 of the 10 males may be contributing to  $N_e$ .  $N_e$  can be variously defined in terms of unequal sex ratios among breeders, fluctuations in population size over several generations, and variance in family size (Lande and Barrowclough, 1987).

Wright (1969) defined the variance effective population size ( $N_e$ ) as the number of individuals in an ideal population that would experience genetic drift at the same rate as the actual population.  $N_e$  can be defined and estimated in various ways using temporal ecological data, DNA sequences, and various methods of estimating migration rate. Some methods of estimation have theoretical value but little operational utility—it is almost impossible to determine the values that some algorithms require. Nevertheless, by estimating  $N_e$  one can assess the effects of different population management strategies. Unequal numbers of males and females, increased variance in family size, and temporal fluctuations in  $N$  all cause  $N_e$  to be much less than the census size,  $N$ . In many endangered populations  $N_e$  is only 10–30, and at such levels genetic variation becomes significant for a population's viability.

### H. Population Bottlenecks

Sudden population declines followed by recovery in numbers are referred to as population or demographic bottlenecks. They can have an immediate impact on variability at molecular genetic loci as genetic drift robs the population of its innate variation. The evidence of a bottleneck may persist for hundreds of thousands to millions of generations in low levels of variation at allozyme and molecular genetic marker loci. However, large populations that are almost isogenic at such loci may maintain high heritable variance in QTL, low inbreeding depression, and high heterozygosity for simple repetitive microsatellite DNA because variation at QTL may return to outbred levels in  $10^3$  to  $10^4$  generations. Furthermore, bottlenecks can actually result in a short-term increase in population variation because epistatic variation (due to interactions among genes controlling a trait) is converted into additive variation. Whether such release of previously hidden variation is beneficial or harmful to population viability is unknown.

Sudden reduction in  $N$  results in a loss of fitness unless there is a rapid and sustained recovery. Gradual reduction, on the other hand, permits natural selection to purge recessive lethal mutations and avoid a substantial part of inbreeding depression. The best advice a geneticist can give the manager of a collapsed population is to increase  $N$  as fast as possible and then worry about genetics.

Very low variability is known for many sexually reproducing species whose currently large populations have recovered from one or repeated brushes with extinction. If a large variable population collapses, then

the few individuals that survive the catastrophe carry only a fraction of the original population's genetic variability through the demographic bottleneck. By chance, some individuals and the alleles they carried are lost to genetic drift. Only one of six mtDNA haplotypes survived the severe bottleneck ( $N = 14$ ) in the whooping crane in 1938. Genetic drift becomes a significant agent of evolutionary change in small populations. Drift may account for the very low levels of variability observed in African cheetahs and northern elephant seals. Cheetahs were not known to be genetically less variable than other cats or at genetic risk until half of a large captive breeding colony died soon after being exposed to a common domestic cat coronavirus (feline infectious peritonitis virus). Although the northern elephant seals are known to have recovered after having been overhunted to near extinction in the late nineteenth century, the low levels of variation in the cheetah may be attributable to metapopulation dynamics rather than a classic population collapse. Metapopulation structure, with frequent extirpation and recolonization of subpopulations, can reduce metapopulation  $N_e$  orders of magnitude below the census population size and mimics the genetic effects of a demographic bottleneck.

### I. Genetic Erosion

Genetic erosion, the decrease in population variation due to random genetic drift and inbreeding, is both a symptom and a cause of endangerment of small isolated populations. Population genetic theory shows that variation will be lost by genetic drift with an almost clock-like regularity (Wright, 1969). In closed populations, in the absence of factors promoting genetic variation (mutation and gene flow) the expected rate of loss of heterozygosity, or rate of loss of genetic variance in quantitative characters or selectively neutral variation, is  $1/2N_e$  per generation. Little variation is lost in any one generation but small  $N$  sustained for several generations can severely deplete variability. Most variability is lost within  $2N_e$  generations. An effective population of 10 is predicted to lose heterozygotes five times faster than a population of effective size 100; 50% of its heterozygosity will be lost in approximately 20 generations. Therefore, in theory, small isolated populations have a higher rate of loss of heterozygosity and are expected to have lower levels of genetic variation than large continuously distributed populations. Because variability is inherently related to evolvability, genetic erosion in small, recently fragmented populations may contribute to their endangerment.

Barrett and Kohn (1991) and Young *et al.* (1996) review the growing literature of the population genetic consequences of habitat fragmentation for plants. There are numerous examples of a positive relationship between  $N_e$  and population genetic variation at allozyme loci for remnant populations. Ouberg and colleagues conducted experimental investigations of genetic erosion with *Scabiosa columbaria* and *Salvia pratensis* plants under common garden conditions and found positive correlations between variance for adaptive traits related to growth rates and population size. Recently, others studied *Clarkia pulchella* and found an increased probability of extinction associated with decreased  $N_e$ . Such experimental studies indicate the potential significance of genetic erosion in natural populations.

The phenomenon of genetic erosion has long been understood in terms of population genetic theory, but the critical early stages of the process in nature have gone undocumented because the changes are rapid and difficult to monitor. I developed a new molecular genetic method for monitoring genetic erosion and provided evidence for its commencement in mammal populations isolated recently in small forest fragments (Srikwan and Woodruff, 2000). An opportunity to study genetic erosion empirically arose when 165 km<sup>2</sup> of lowland rain forest were flooded in 1987 following construction of a hydroelectric dam on Khlong Saeng, southern Thailand. Former hilltops became permanent islands in Chiew Larn reservoir and retained their original fauna of 12 species of small mammals. During years 5–8 postfragmentation, the demographic collapse of these communities and genetic erosion in three common species [a forest rat (*Maxomys surifer*), a tree mouse (*Chiropodomys gliroides*), and a tree shrew (*Tupia glis*)] whose populations were effectively isolated on some islands were monitored. Nuclear microsatellite markers are sufficiently variable to be used to monitor the process of genetic erosion in nature. As expected, small populations lost variability by genetic drift faster than large populations, and allelic variability is a better indicator of the onset of genetic erosion than heterozygosity. Interestingly, in one of the three species studied, genetic erosion commenced before detectable demographic decline.

This demonstration that the process of genetic erosion can be monitored in free-ranging natural populations provides numerous research opportunities because habitat fragmentation is a very ubiquitous phenomenon. Furthermore, the method can easily be upscaled to larger mammals of conventional concern to conservationists. Although monitoring genetic ero-

sion in long-lived species may not be practical, much can be learned immediately by comparing isolated populations to those still more continuously distributed.

The importance of such rapid genetic erosion on population viability remains unclear because there are so few studies of the process in nature. Two larger questions remain to be answered: At what point (in terms of  $N$  and  $N_e$ ) does genetic erosion threaten a population's viability? and What level of natural or artificial gene flow can protect a population from the negative effects of genetic erosion? Answers to such questions may emerge from the 35-year study of the decline and assisted recovery of an isolated population of greater prairie chicken, *Tympanuchus cupido* (Westemeier *et al.*, 1998). Unfortunately, there are very few studies of this duration.

Genetic enhancement, the introduction of selected individuals into a threatened population with the intent of maintaining or increasing its genetic variability and hence viability, is a conservation method in its infancy. Although the genetics may seem straightforward, the translocation of new individuals carries a significant risk of introducing diseases into the threatened population. Furthermore, ill-planned genetic enhancement may lead to a breakdown of local adaptation (outbreeding depression) and actually decrease a threatened population's viability.

## J. Natural Selection and Adaptation

Natural selection, the differential survival and reproduction of some genotypes over others, is the major agent of microevolutionary change. It is of interest to conservation geneticists for two reasons. First, human activities can radically alter selection coefficients in both natural populations and managed ones. Such human-influenced evolutionary change is termed artificial selection whether or not it is intentional. Intense harvesting based on size or gender can cause rapid changes in behavior and natural history and reduce fitness. Examples include reduced body size in game fish and the impact of hunting only male horned or tusked mammals on social behavior. Second, one of the major challenges facing geneticists this century will be assisting species to adapt to ongoing global climatic changes. In the past, in the absence of humans, natural selection favored individuals adapted to change and many species shifted their ranges to accommodate major changes. Unfortunately, in the twenty-first century, the pace of environmental alteration and destruction is too fast for many species to respond. Conservation managers will have

to intervene on behalf of many species if they are to survive.

Quantitative characters are typically under stabilizing selection and show some optimum phenotype from which they can evolve in response to environmental changes. This optimum balances current fitness with the need for future flexibility or adaptability. The maintenance of this variability imposes a fitness cost or genetic load on a population—the price for long-term evolvability.

The rate of directional selection that a population can manage, in response to some environmental change, is determined in part by its innate variability. Rapid anthropogenic changes, such as those associated with global warming, place a premium on genetic variation and adaptability, especially in fragmented populations. To maintain variability in quantitative characters (long-term adaptability), the Franklin–Soulé number for an effective population size of at least  $N_e = 500$  is often cited. Revisionary work by Lande (1999) has shown, however, that an upward revision to  $N_e = 5000$  is required. Such numbers are larger than those found in many endangered and threatened populations and underscore the need for genetic vigilance in their management and the importance of keeping numbers as high as possible.

The risk of extinction due to fixation of mildly deleterious mutations is comparable in importance to environmental stochasticity and could substantially decrease the long-term viability of populations with  $N_e$  of less than a few thousand. The current recovery goals for many threatened species are inadequate to ensure their long-term viability if this requires  $N = 10,000$  individuals. Genetic and demographic factors, acting synergistically, require that minimum viable populations be  $>10^4$ .

## IV. GENETIC MANAGEMENT: EXAMPLES

Until a textbook on conservation genetics is written, the reader must consult conference proceedings and the primary literature for examples of the application of genetics to conservation management. The journals *Conservation Biology* and *Molecular Ecology* are especially useful in this regard. Without making reference to the specific methods employed (because these continuously change), the following examples, in addition to those mentioned previously, are illustrative of the type of contributions conservation geneticists have



made. Details may be found in Loeschcke *et al.* (1994), Olney *et al.* (1994), Frankham (1995), Avise and Hamrick (1996), Smith and Wayne (1996), Lacy (1997), Landweber and Dobson (1999), and the specific references cited in the following sections.

### A. Pedigree Management in Very Small Populations

Olney *et al.* (1994) describe examples of multi-institutional breeding plans with genetic components for numerous species, including Przewalski's horse, Père David's deer (*Elaphurus davidianus*), Hawaiian goose (*Branta sandvicensis*), California condor (*Gymnogyps californianus*), and Tahitian *Partula* tree snails, all of which were extirpated in the wild. Cases of more intensive genetic management, including the establishment of relationships, founder representation, and breeding to maximize  $N_e$ , include captive populations of lion-tailed macaque (*Macaca silenus*) (San Diego Zoo), Speke's gazelle (*Gazella spekei*) (St. Louis Zoo), Waldrapp ibis (*Geronticus eremita*) (Zurich Zoo), Guam rail (*Rallus owstoni*), Micronesian kingfisher (*Halcyon cinnamomina*), and Mauritius pink pigeon (*Nesoenas mayeri*). Genetic sex determination of juvenile California condors enabled recovery program managers to pair birds efficiently.

Geneticists have identified low genetic variability as a concern in wild and captive populations of many species, including cheetah, Californian Channel Island fox (*Urocyon littoralis*), Newfoundland black bear (*Ursus americanus*), Gir Forest Asian lions (*Panthera leo*), southern koalas (*Phascolarctus cinereus*), European bison (*Bison bonasus*), Arabian oryx (*Oryx leucoryx*), Père David's deer, and Torrey pine (*Pinus torreyana*). A loss of self-incompatibility alleles may pose a threat to reproduction in plants with genetically determined self-incompatibility systems such as the rare lakeside daisy (*Hymenoxys acaulis*) in Illinois.

Geneticists identified inbreeding as a probable cause of reproductive failures in populations of Ngorongoro lions (*Panthera leo*), Florida panther (*Puma concolor coryi*), Barrow Island black-footed rock wallaby (*Petrogale lateralis*), bighorn sheep (*Ovis canadensis*), Puerto Rican parrot (*Amazona vittata*), and the Isle Royale gray wolf (*Canis lupus*).

### B. Mating System

Geneticists have discovered that the mating systems of many species differ from expectations based on direct

observations, often with profound implications for captive management and for management of "wild" populations. In some species, females preferentially mate with males outside their social group, e.g., Atlantic salmon (*Salmo salar*), blue tits (*Parus caeruleus*), and pilot whales (*Globicephala* sp.). Preferential interpod mating in long-finned pilot whales indicates the importance of conserving as many pods as possible. Genetics shows that highly gregarious black vultures (*Coragyps atratus*) are in fact monogamous and that Galapagos hawks (*Buteo galapagoensis*) are polyandrous. Despite behavioral observations suggesting a high frequency of matings between sibs in Australian splendid fairy wrens (*Malurus splendens*), genetics shows that outcrossing is the norm. In other birds, including stripe-backed wrens (*Campylorhynchus nuchalis*), geneticists found that subordinate males reproduce. The mating system of wild gray wolf, chimpanzees, and other species have been established and used to improve captive breeding programs or management in reserves. Establishing the mating system of the red-cockaded woodpecker led to improved estimates of  $N_e$  and changes in the recovery program for a small population in South Carolina.

### C. Problems of Hybrids

Geneticists have elucidated the hybridity of some taxa in both the wild and in captivity, including Przewalski's horses (many of which are domestic horse hybrids), Asian lions (most animals in Western zoos were hybrids between Asian and African lions and were removed from the breeding program), and the red wolf of Texas [shown to be primarily a natural coyote (*Canis latrans*)  $\times$  gray wolf hybrid]. Genetic data revealed the threat by hybridization to indigenous Scottish deer (*Cervus elaphus scoticus*) by introduced Japanese sika deer (*C. nippon*). Genetics was used to identify hybrids among the remaining Catalina Island mahogany and establish seedlings of pure *Cercocarpus traskiae* at several sites to protect the species from extinction by introgression. Genetic markers are being used to conserve remnant cutthroat trout populations by identifying and removing populations introgressed by hybridization with nonnative species.

### D. Genetic Censusing

Because every individual in most species is genetically distinct, it is possible to census populations by counting unique multilocus genotypes in an area. Geneticists have censused very difficult to count animals such as

African forest elephants and the largely fossorial northern hairy-nosed wombats (*Lasiorhinus krefftii*) by individually genotyping dung samples. Fecal genotyping (including sexing) has also been used to provide more accurate census data of animals such as coyotes and bears than could be obtained from long-term ecological surveys. The genetic data also provided information on home ranges and pedigree relatedness without requiring that the animals be seen or disturbed.

### E. Phylogeography, Gene Flow, and Population Structure

Awise (2000) reviews many examples of phylogeographic studies that provide managers with essential data on population structure and on the characterization of MUs and ESUs. Geneticists have also provided estimates of historical gene flow between populations that would be impossible to obtain by direct observation (e.g., chimpanzees, humpback whales, and green turtles). Comparative phylogeographic studies of four sympatric species of East African bovids underscore the dangers of extrapolating results from one species to another and support conservation efforts that take species-specific differences into account.

### F. Defining Species, ESUs, and MUs

Karyological differences distinguish Bornean and Sumatran orangutans (*Pongo pygmaeus*) and captive breeding programs are now managed to prevent hybridization of the two. Dik-dik species (*Madoqua*) can also be distinguished karyologically and neonate mortality in captive populations has been reduced following the sorting of animals by karyotype. MtDNA sequences have been used to sort sibling species and subspecies of gibbons (*Hylobates* sp.) of unknown geographic origin into correct ESUs, to distinguish sibling species of chimpanzees (*Pan troglodytes* and *P. verus*) that have been mixed in captivity, and to identify subspecies of black rhinoceros, *Diceros bicornis*. Genetics was used to show that the conservation of the living fossil tuatara (*Sphenodon*) of New Zealand depends on the management of not one but two genetically recognizable species. There are many cases in which genetic data justify conservation efforts for isolated subspecies or varieties that are shown to be genetically well differentiated—for example, Darwin's fox (*Dusicyon fulvipes*) of Chile, Kemp's ridley turtle (*Lepidochelys kempi*), San Clemente Island loggerhead shrike, and several "subspecies" of Hawaiian Amakilu honeycreeper (*Hemignathus virens*). The Mexi-

can wolf (*Canis lupus baileyi*) was found to be a genetically distinct ESU, untainted by hybridization with gray wolf, coyote, or dog.

The endangered San Clemente Island shrike is illustrative of several of the points made previously because it is technically a subspecies of a widespread mainland bird and might accordingly be written off as a peripheral population, local variant, or "just a subspecies." Taxonomic practice and field observations notwithstanding, my genetic survey revealed no evidence that it has hybridized with the neighboring mainland subspecies since 1915 despite repeated opportunities. It is genetically differentiated and apparently reproductively isolated and merits management as a separate ESU (Mundy *et al.*, 1997).

Geneticists have also provided data questioning the justification of other taxon-focused conservation efforts. The 27 original subspecies of leopard (*Panthera pardus*) were found to be referable to only 8 genetically defined subspecies or ESUs. Similarly, genetic variation provides no justification for conserving all 30 described subspecies of puma (*Puma concolor*). The dusky seaside sparrow (*Ammodramus maritimus nigriscens*), after considerable and unsuccessful efforts to save it, was shown to be only a marginally distinct local race of a common widespread species.

Taxonomic status (species or subspecies) does not automatically lead to the justification of conservation efforts. Proposals to reduce the eastern Pacific black sea turtle, *Chelonia agassizii*, to a subspecies of the widespread green turtle, *C. mydas*, are supported genetically but do not justify the abandonment of this taxon as a high conservation priority.

A phylogeographic survey of 14 subspecies of the songbird bananaquit (*Ceoreba* sp.) from 15 West Indian islands and the mainland of South and Central America showed that if it were necessary to restock the small and vulnerable populations on the northern Lesser Antilles Islands, the birds from nearby Puerto Rico or from Jamaica would be genetically inappropriate. Similar studies of orioles (*Icterus* sp.) are relevant to the conservation of the Montserrat oriole, *I. oberi*, which is under threat of catastrophic (volcanic) extinction.

Phylogeographic analyses have also helped to define natal homing patterns in marine turtles on foraging grounds. It was found that both green turtle and loggerhead turtle rookeries are demographically autonomous and that low levels of interrookery matrilineal exchange suggest that extirpated colonies are unlikely to recover by natural recruitment of nonindigenous females. Similar methods have been used to define stocks in whales and dolphin and elucidate the migratory strategies of

different groups of humpback whales intermingled at the breeding ground near Hawaii.

Hierarchical phylogenetic analyses have been used to suggest conservation priorities among species of cranes. Within a group of related species the evolutionarily oldest monophyletic clades are considered to represent a greater genetic heritage than recently originated clades (Forey *et al.*, 1994). Similarly, it has been argued that areas with a disproportionate number of evolutionary ancient genotypes are more valuable than areas populated by recent colonists.

### G. Reintroductions, Translocations, and Genetic Enhancement

Two genetic generalizations complicate the prospects for successfully moving organisms around in the wild or for returning them to the wild after a period in captivity. First, the chances for successful reintroduction are diminished by evidence for rapid genetic adaptation to captivity in fish, plants, and *Drosophila*. The same problem applies to wildlife brought into captivity for reintroduction at some future time. Second, in plants, cryptic local adaptation results in the fitness of transplants being about half that of residents even when environments are apparently similar. This makes it difficult to reestablish populations once they are extirpated.

Genetic criteria were used to choose founders for a new population of the extirpated Guam rail on the island of Rota, for sea otters (*Enhydra lutris*), and for a Western Australian shrub (*Corrigan grevillea*) reduced to 27 plants. Genetic data were also used to influence the choice of source population for a Gila topminnow reintroduction program (Vrijenhoek in Loeschcke *et al.*, 1994, pp. 37–53). Fish from a population with high allozymic diversity were selected to successfully replenish the diversity and viability of a declining and nearly monomorphic population rather than fish from a less variable but adjacent population.

Genetic study of red-cockaded woodpeckers in the southeastern United States led to the recommendation that translocations be made between nearby populations rather than moving birds over great distances. In another study, it was shown to be genetically appropriate to use gray wolves from British Columbia as a source of animals for reintroduction into Yellowstone National Park. Genetic variation and multilocus genetic similarity were used to justify the introduction of panthers from Texas (traditionally regarded as a different subspecies) to counter severe inbreeding depression (low het-

erozygosity, poor sperm quality, and cryptorchidism) in the remnant Florida panther population.

Genetic criteria have also been used to argue against certain types of translocations. For example, it was found that Tasmanian eastern barred bandicoots (*Parameles gunnii*) should not be used as a source population for enhancing the endangered mainland Australian population. South African wild dogs (*Lycaon pictus*) were genetically inappropriate for reintroduction into Kenya. Isolated northern and southern populations of Brazilian muriqui (*Brachyteles arachnoides*) are well differentiated genetically, so threatened northern populations should perhaps not be translocated to larger southern reserves.

An example of the hazards of implementing a transplantation program without first considering genetic factors involves the endangered Hawaiian silversword. In this case, the outplanted individuals were all descendants of only one or two maternal plants and therefore retained only a small fraction of the genetic diversity of the remaining populations of the species they represented. Furthermore, because they were so closely related to one another, they had significant reproductive problems associated with self-incompatibility and a seed set of <20% (Rieseberg and Swensen in Avise and Hamrick, 1996, pp. 305–334).

### H. Conservation Management

Genetic methods are used in forensic identification of tissues of endangered species in illegal or misrepresented trade (e.g., abalone, “caviar,” cage birds, and primates including chimpanzees). Misrepresentation occurs when a wild-caught parrot or falcon is claimed to be legally captive bred. Geneticists showed that some whale meat legally on sale in Japan was actually meat of endangered and allegedly protected species including humpback whale. Sequence data revealed that loggerhead turtles from Caribbean nesting beaches are threatened by a Mediterranean fishery off Spain.

Geneticists are playing an increasingly important role in the management of mixed stocks of threatened and commercially harvested fish. Consider the conservation management of salmon (*Oncorhynchus*) in the U.S. Pacific Northwest. Salmon with their precise homing behavior present a major problem because each local spawning population should be managed as a genetically distinct taxon. Should managers give every natal stream-adapted salmon stock equal priority? Geneticists have developed allelic frequency marker systems for stock identification that are sufficiently sensitive to permit real-time regulation of mixed stock fisheries

involving both hatchery and wild salmon. The latter can be partially protected because they return from the sea to the rivers a few weeks later than the former and the fishery can be terminated when they are detected.

Using the same approach, geneticists developed markers to identify endangered and protected winter-run chinook salmon in the Sacramento River in California. Fall- (hatchery) and spring-run chinook salmon do not enjoy protection, but the fishery could not be managed from a conservation perspective until the different races could be identified during the downstream migration of smolts to the ocean.

Genetic tracking of movements of migratory birds permits sorting arctic and Mexican falcons (*Falco peregrinus*) that mix when the former reach their wintering grounds. Similar genetic tracking permitted the identification of the wintering grounds of several species of declining arctic shorebirds and led to changes in conservation focus from the breeding grounds to the wintering grounds.

Finally, geneticists have been able to monitor the loss of variability in translocated populations of wild turkey (*Meleagris gallopavo*), white-tail deer (*Odocoileus virginianus*), and alpine ibex (*Capra ibex*).

### I. Other Applications

Space does not permit presentation of all the caveats, corrections, and revised recommendations made in most of the previous cases as more data became available and as various tests were repeated. Many of the classic examples are less clear-cut than originally proposed. Space does also not permit mention of all the examples in which genetics showed populations were not genetically depauperate or sufficiently distinct to warrant priority conservation efforts. Such contributions are of equal importance to biodiversity conservation because they free up limited resources for other investigations.

Although most of the previous examples involved captive or wild populations of threatened species, much useful conservation genetics can be done using laboratory animals as model organisms. Valuable experimental tests of conservation genetic principles have been completed using *Drosophila*, *Tribolium*, mosquitofish (*Gambusia holbrooki*), the butterfly (*Bicyclus anynana*), *Mus*, and *Peromyscus* (Frankham, 1995; Leberg in Smith and Wayne, 1996, pp. 87–103). For example, using laboratory populations of *Drosophila* it has been shown that equalizing family size can double  $N_e$  and more intensive pedigree management can increase the  $N_e/N$  ratio 40-fold. Similarly, in most of the examples dis-

cussed previously, investigators also examined the genetics of a close relative of the taxon of interest. It is standard practice to develop genetic methods using a common relative as a surrogate before commencing work on a highly endangered taxon. This approach also has the advantage of providing comparative data useful in interpreting the results of a study of an endangered taxon.

## V. GENETICS, EVOLVABILITY, AND THE FUTURE OF BIODIVERSITY

The magnitude of the task facing conservation geneticists is daunting. There are on the order of 10 million species living on the planet today and about 2 million of these are recognized and named in a formal taxonomic sense. Describing a new species requires little more than that a scientist know what it looks like and where it is found; unfortunately, this constitutes our state of knowledge for most of biodiversity. Closer to  $10^4$  than to  $10^5$  species have been characterized ecobehaviorally, and only on the order of  $10^3$  species have been examined by geneticists. If sound conservation is based on a knowledge of ecology, behavior, and genetics, then we must admit that we are currently capable of scientifically managing the evolution of less than 1000 species. However, the number of species requiring individual management to prevent their extinction in the next 100 years is in excess of 10,000.

Conservation geneticists have devoted their efforts disproportionately toward the charismatic megavertebrates. Whales and cats have received more attention than bats and rats. Given the enormous number of species requiring attention, one might inquire as to how priorities are set. First, most research has gone into species that were favored for utilitarian reasons; they provide us with food, clothing, medicines, recreation, or companionship. Most of this research has been aimed at stock improvement rather than whole genome or species conservation. Second, as already noted, the charismatic megavertebrates and a few groups of flowering plants have received inordinate attention. Perhaps not surprisingly, rare species have also been studied more than common ones. The same applies to phylogenetically unique species, living fossils, and evolutionary relics. However, because the real goal is to save functional ecosystems, conservation geneticists are rethinking their priorities. Although some rare species clearly merit genetic management, it would be better to focus more attention on ecological keystone species whose

activities are critical to the maintenance of entire communities. We also need to know more about the genetics of ecologically successful colonizing species and of clades of species that have evolved very recently (the cichlid flocks of Rift Valley lakes) because their study may teach us how to manage apparently less successful taxa.

Conservation geneticists rarely advocate bringing plants and animals into captivity "to save them." Species are typically better managed in their natural communities than in isolation. Existing institutions concerned with conservation, however, are not equipped to deal with the magnitude of the task they face (Woodruff, 1989). Parks and wildlife reserves are the preferred approach to both species and community conservation. Zoos and botanic gardens are extremely limited in what they can accomplish and can at best serve only to shelter a few critical cases that require intensive care. Germplasm frozen storage systems are valuable adjuncts for researchers, but no credible geneticist has yet proposed that we will be able to awaken these frozen tissues and recreate the animals from which they were derived. Although frozen tissue banks are extremely valuable for geneticists, the revitalization of mammoths, quaggas, thylacenes, and dodos is still science fiction.

In its first two decades, conservation genetics was perceived by some wildlife ecologists as an unnecessary intrusion into their field. It was argued that demography and behavior are far more important than genetics in saving endangered species. Others have argued that the genetic threat to population viability has been overstated (Lande, 1988). Genetics was viewed as too theoretical and contributing too little and too slowly to the day-to-day efforts to save populations in nature. Furthermore, molecular genetics studies, which are relatively expensive, compete for the limited funds available to the traditional conservationists. Some of the criticisms were justified and some were not. It is easy to disparage the potential contribution of genetics to saving a particular population or species if genetics is defined very narrowly as, for example, the determination of heterozygosity in a remnant population. In the case of cheetahs in Africa, it is clear that predation by lions and humans is more significant today than low variability. Similarly, in captivity, different husbandry practices in different zoos are more significant than poor sperm quality. However, if one takes a longer term view, the answer is different: Genetics is and will be increasingly important. As this review shows, geneticists have a great deal to offer managers. It is incorrect to suggest that ecology and genetics are alternative approaches. Although there are clearly times when genetic

studies will be lower priority in a multifaceted conservation strategy, it is undeniable that increasingly more populations will need genetic management. Genetics, ecology, and behavior are all necessary parts of biodiversity conservation.

Although conservation geneticists focus on populations and species, their ultimate goal is the conservation not of things but of a process, evolution, that produced them. The ultimate goal of conservation biology is to preserve the processes of organic evolution—to maintain the ability of populations and species to evolve and communities to function and provide ecosystem services. The basic science is still not equal to the task conservation geneticists are expected to perform. The relationship between genetic variation and "genetic health" is illusive and needs clarification. Society's expectations of conservation geneticists also need to be specified or we will forever be accused of treating the symptoms and not the causes of the biodiversity crisis. Typically, species are not afforded legal protection until their populations have fallen into the hundreds, 10–100 times below the level at which their genetic integrity and viability are reasonably secure. Geneticists need to point out that current standards of endangerment are far too low, that recovery from previous mass extinctions took on the order of 10 million years, and that we have not thought through the global implications of a 50% decrease in the number of remaining larger plant and vertebrate species (Myers, 1996).

### See Also the Following Articles

BIODIVERSITY, EVOLUTION AND • CONSERVATION BIOLOGY, DISCIPLINE OF • GENETIC DIVERSITY • INBREEDING AND OUTBREEDING • POPULATION DIVERSITY, OVERVIEW • POPULATION VIABILITY ANALYSIS (PVA) • SPECIES DIVERSITY, OVERVIEW • SUBSPECIES, SEMISPECIES

### Bibliography

- Avise, J. C. (2000). *Phylogeography*. Harvard Univ. Press, Cambridge, MA.
- Avise, J. C., and Hamrick, J. L. (Eds.) (1996). *Conservation Genetics: Case Histories from Nature*. Chapman & Hall, New York.
- Barrett, S. C. H., and Kohn, J. R. (1991). Genetic and evolutionary consequences of small population size in plants: Implications for conservation. In *Genetics and Conservation of Rare Plants* (D. A. Falk and K. E. Holsinger, Eds.), pp. 3–30. Oxford Univ. Press, New York.
- Beissinger, S. R., and McCullough, D. R. (Eds.) (2000). *Population Viability Analysis*. Univ. of Chicago Press, Chicago.
- Benirschke, K., and Kumamoto, A. T. (1991). Mammalian cytogenetics and conservation of species. *J. Heredity* 82, 187–191.
- Forey, P. L., and Humphries, C. J., and Vane-Wright, R. I. (Eds.)

- (1994). *Systematics and Conservation Evaluation*. Oxford Univ. Press, Oxford.
- Frankel, O. H., and Soulé, M. E. (1981). *Conservation and Evolution*. Cambridge Univ. Press, Cambridge, UK.
- Frankham, R. (1995). Conservation genetics. *Annu. Rev. Genet.* **29**, 305–327.
- Gilpin, M. E., and Soulé, M. E. (1986). Minimum viable populations: Processes of species extinction. In *Conservation Biology: The Science of Scarcity and Diversity* (M. E. Soulé, Ed.), pp. 19–34. Sinauer, Sunderland, MA.
- Haig, S. M. (1998). Molecular contributions to conservation. *Ecology* **79**, 413–425.
- Hanski, I. A., and Gilpin, M. E. (Eds.) (1996). *Metapopulation Biology*. Academic Press, San Diego.
- Lacy, R. C. (1997). Importance of genetic variation to the viability of mammalian populations. *J. Mammal.* **78**, 320–335.
- Lande, R. (1988). Genetics and demography in biological conservation. *Science* **241**, 1455–1460.
- Lande, R. (1999). Extinction risks from anthropogenic, ecological, and genetic factors. In *Genetics and the Extinction of Species* (L. F. Landweber and A. P. Dobson, Eds.), pp. 1–22. Princeton Univ. Press, Princeton, NJ.
- Lande, R., and Barrowclough, G. F. (1987). Effective population size, genetic variation, and their use in population management. In *Viable Populations for Conservation* (M. E. Soulé, Ed.), pp. 87–124. Cambridge Univ. Press, New York.
- Loeschcke, V., Tomiuk, J., and Jain, S. K. (Eds.) (1994). *Conservation Genetics*. Birkhäuser Verlag, Basel.
- Morin, P. A., and Woodruff, D. S. (1996). Non-invasive genotyping for vertebrate conservation. In *Molecular Genetic Approaches in Conservation* (T. B. Smith and R. K. Wayne, Eds.), pp. 298–313. Oxford Univ. Press, Oxford.
- Mundy, N. I., Winchell, C. S., Burr, T., and Woodruff, D. S. (1997). Microsatellite variation and microevolution in the critically endangered San Clemente loggerhead shrike (*Lanius ludovicianus mearnsi*). *Proc. R. Soc. London B* **264**, 869–875.
- Myers, N. (1996). The biodiversity crisis and the future of evolution. *Environmentalist* **16**, 37–47.
- O'Brien, S. J. (1994). Genetic and phylogenetic analyses of endangered species. *Annu. Rev. Genet.* **28**, 467–489.
- Olney, P. J. S., Mace, G. M., and Feistner, A. T. C. (Eds.) (1994). *Creative Conservation. Interactive Management of Wild and Captive Animals*. Chapman & Hall, London.
- Ryder, O. A., and Fleischer, R. C. (1996). Genetic research and its application in zoos. In *Wild Animals in Captivity* (D. G. Kleiman et al., Eds.), pp. 255–262. Univ. Chicago Press, Chicago.
- Saccheri, I., Kuussaari, M., Kankare, M., Vikman, P., Fortelius, W., and Hanski, I. (1998). Inbreeding and extinction in a butterfly metapopulation. *Nature* **392**, 491–494.
- Schonewald-Cox, C. M., et al. (Eds.) (1983). *Genetics and Conservation: A Reference for Managing Wild Animal and Plant Populations*. Benjamin-Cummings, Menlo Park, CA.
- Smith, T. B., and Wayne, R. K. (Eds.) (1996). *Molecular Genetic Approaches in Conservation*. Oxford Univ. Press, New York.
- Soltis, P. S., and Gitzendanner, M. A. (1999). Molecular systematics and the conservation of rare species. *Conserv. Biol.* **13**, 471–483.
- Srikanth, S., and Woodruff, D. S. (2000). Genetic erosion in isolated small mammal populations following rain forest fragmentation. In *Genetics, Demography and Viability of Fragmented Populations* (A. Young and G. Clarke, Eds.). Cambridge Univ. Press, Cambridge, UK.
- Storfer, A. (1996). Quantitative genetics: A promising approach for the assessment of genetic variation in endangered species. *TREE* **11**, 343–348.
- Westemeier, R., Brawn, J. D., Simpson, S. A., Asker, T. L., Jansen, R. W., Walk, J. W., Kershner, E. L., Bouzaty, J. L., and Paige, K. N. (1998). Tracking the long-term decline and recovery of an isolated population. *Science* **282**, 1695–1698.
- Woodruff, D. S. (1989). The problems of conserving genes and species. In *Conservation for the Twenty-First Century* (D. Western and M. Pearl, Eds.), pp. 76–78. Oxford Univ. Press, New York.
- Wright, S. (1969). *Evolution and the Genetics of Populations*, Vol. 4. Univ. Chicago Press, Chicago.
- Young, A., Boyle, T., and Brown, T. (1996). The population genetic consequences of habitat fragmentation for plants. *TREE* **11**, 413–418.





# POPULATION VIABILITY ANALYSIS

Hugh P. Possingham,<sup>\*†</sup> David B. Lindenmayer,<sup>‡</sup> and  
Michael A. McCarthy<sup>†‡</sup>

<sup>\*</sup>National Center for Ecological Analysis and Synthesis, <sup>†</sup>The University of Adelaide,  
<sup>‡</sup>Australian National University

---

- I. What Is Population Viability Analysis?
  - II. Stochastic Population Processes and Extinction
  - III. Why Do PVA?
  - IV. Methods and Tools for PVA
  - V. Testing PVA
  - VI. Criticisms of PVA: What Makes a Good PVA?
- 

## GLOSSARY

**stochasticity** A stochastic process is one in which the state of the system cannot be precisely predicted given its current state and even with a full knowledge of all the factors affecting that process. In a population context, we may know the detailed life history parameters of a species. However, various unpredictable (stochastic) processes, such as the chance nature of birth and death (demographic stochasticity), year-to-year variation in climate (environmental stochasticity), and catastrophes, means that accurately predicting the precise size of a population in the future is not possible. Despite this, we can use probability theory and simulation models to make probabilistic predictions. A stochastic population model is one in which each possible future population size has an associated probability. This approach to prediction is the same as stating that the chance of getting a head with the next toss of a fair coin is 50%. Our

prediction is accurate but we cannot say if the outcome will be a head or a tail.

**viability (and related concepts)** A viable population is one deemed to have a reasonable chance of long-term persistence. Such a definition begs two questions: What is reasonable and what is long-term? It has been stated that a viable population is one that has a 99% chance of persisting 1000 years. Different managers, researchers, and policymakers use different measures that range from a 99% chance of persisting 1000 years to a 95% chance of persisting 100 years. The population size that just meets the definition of being viable is called a minimum viable population (MVP). One of the key questions of reserve design is how big does a reserve, or network of reserves, need to be to contain an MVP. This size is called a minimum viable habitat area for a particular species.

---

**POPULATION VIABILITY ANALYSIS (PVA)** is a process in which the extinction probability of a population is assessed. This article discusses the processes that interact to cause extinction, and this discussion is then used as the basis of explaining how the extinction process can be modeled. An emphasis is placed on recent attempts to provide quality control for PVA and use the models as a decision support tool for conservation planning.



## I. WHAT IS POPULATION VIABILITY ANALYSIS?

Traditionally, population viability analysis (PVA) is a process in which a stochastic population model is used to assess the viability of a population. The population may be a species, subspecies, metapopulation, or an isolated subpopulation of a single species. Classical PVA uses a demographic population model to derive a result such as the following: Given current circumstances the probability of extinction of species  $X$  over the next  $Y$  years is  $Z\%$ . For example, in his seminal 1983 paper, Shaffer used a demographic population model to determine the probability of extinction of grizzly bears, *Ursus arctos*, in Yellowstone National Park within the next 100 years.

Most PVAs use a demographic population simulation model which incorporates all the processes likely to affect the dynamics of the population. PVA differs from other population modeling exercises in its focus on predicting the probability of extinction and the inclusion, typically, of stochastic processes that are believed to have a strong impact on the probability of extinction. For convenience, we categorize these stochastic processes into four types; each can be thought of as a way in which variability influences the dynamics of a species. Given that these processes are central to our thinking about PVA, they are described in detail. In a broader context PVA can be thought of as just one branch of the larger field of population modeling, in which the primary focus is on an assessment of risk. Risk is the probability of an unfavorable event; in this case the unfavorable event is the extinction of a species.

PVA is unique in the context of conservation biology in that it enables us to assess the adequacy of conservation efforts. Given an acceptable risk of extinction for any species, then, in theory we should be able to manage every species so that its extinction risk does not exceed this acceptable level. If we can do this, our conservation strategy would be deemed adequate.

During the past decade, we have realized that the simple view of PVA as a process using a demographic simulation model to assess extinction risk is too narrow. First, the restriction of PVA to the use of demographic population modeling software ignores other tools for assessing viability. We now view PVA as a broader process that includes various sorts of theory, analytic and computer modeling, and forms of data analysis. Second, our ability to accurately assess the extinction probability of a species is questionable. The broader roles of PVA are explored later. It is this management-

based approach to PVA that is now broadly accepted and considered practically useful.

## II. STOCHASTIC POPULATION PROCESSES AND EXTINCTION

### A. Types of Demographic Population Fluctuation

#### 1. Demographic Stochasticity

Demographic stochasticity is the chance nature of birth and death. It causes populations to fluctuate because populations are composed of individuals that are units. Each unit counts as one and is born and dies as a unit, so populations can only increase and decrease on the set of integers. Each individual has a probability of dying and a probability, each year, of finding a mate and producing offspring. The dynamics of five separate populations, fluctuating as a result of demographic stochasticity alone, are shown in Fig. 1.

Each population has its own trajectory even though they all had the same life history parameters and started from the same population size. In a sense, demographic stochasticity is the most fundamental phenomenon that causes populations to fluctuate because it occurs in all populations. In Fig. 1 each population should, on average, increase 5% per year, but one becomes extinct. The importance of demographic stochasticity for extinction and PVA diminishes in larger populations because the minor “wobbles” caused by the unitary nature of birth and death are overshadowed by the more dramatic environmental stochasticity and catastrophes (Fig. 2).

A rarely discussed type of demographic stochasticity is the chance nature of sexual determination in organisms with more than one sex. For example, a population may be known to have six individuals, which may appear to have some chance of persisting if conditions are good. However, there is a 1 in 32 chance that all the individuals are the same sex and the population is doomed if each has an equal chance of being male or female. When populations are small, a skewed sex ratio can have a major impact on the chance of extinction.

#### 2. Environmental Stochasticity

As biotic and abiotic conditions vary from year to year, there will be variation from year to year in vital rates. This sort of variation tends to affect the entire population; therefore, there are years in which breeding suc-

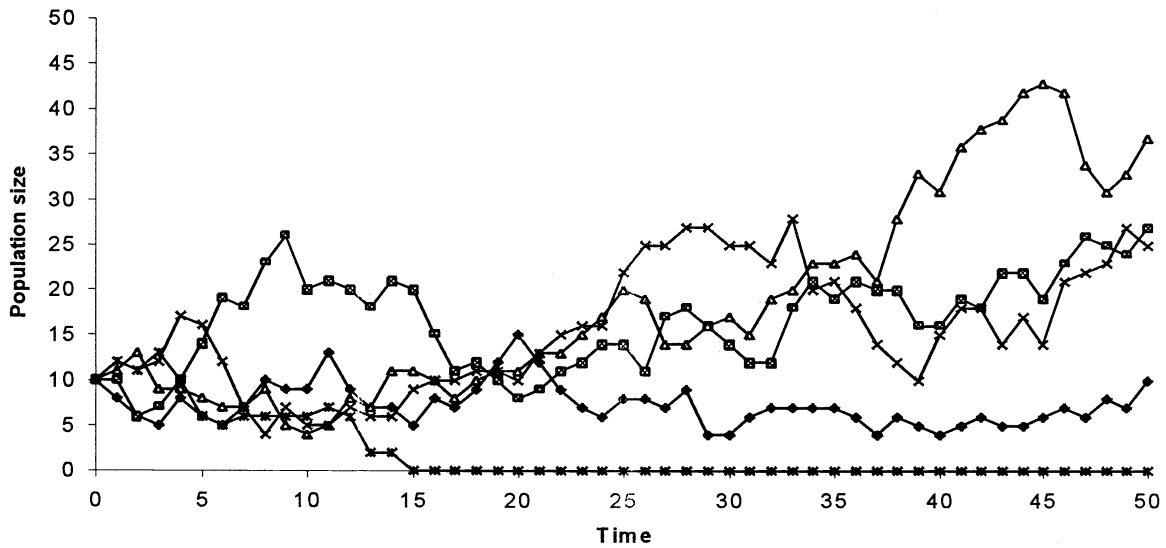


FIGURE 1 Five sample trajectories for a population in which demographic stochasticity is operating. For each individual, each year there is a 30% chance of dying. Those that survive have a 50% chance of giving birth to one more individual. Despite identical conditions the populations diverge rapidly, with one becoming extinct after 15 years and another increasing to more than 40 individuals. With these parameters, in the absence of stochasticity, the population should on average increase at a rate of 5% (=70% of 1.5) per year.

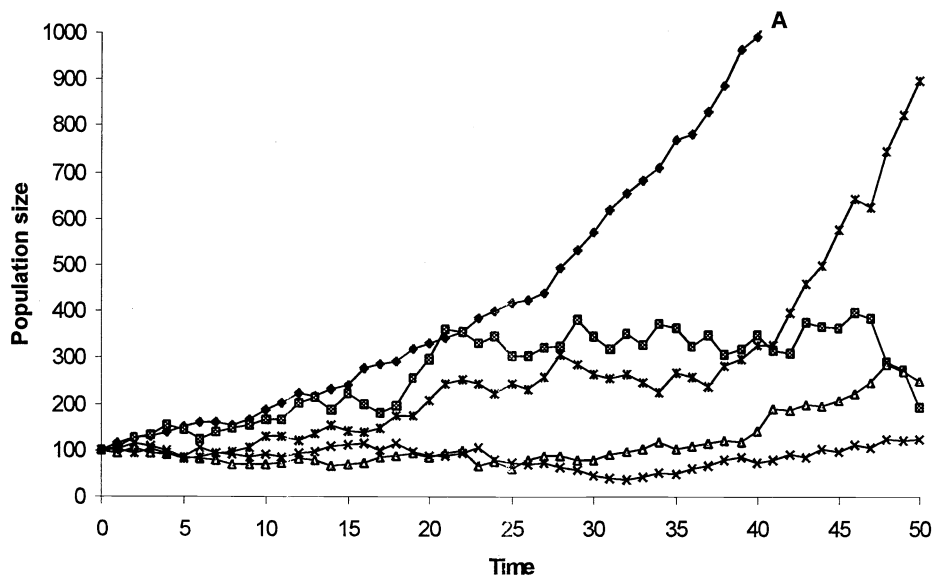


FIGURE 2 The impact of environmental stochasticity on the dynamics of five populations. For trajectory A, the parameters are the same as those used for Fig. 1 with just demographic stochasticity, except the population started with 100 animals. Note the fairly consistent increase of about 5% per year with minor fluctuations caused by demographic stochasticity. For the other four populations, the death rate is not fixed at 30% but varies randomly between 20 and 40% from year to year. Also, the probability of an offspring is not fixed at 50% but varies from 40 to 60% from year to year for the whole population. Note the much wider fluctuations in population size, even when the populations are larger than 100. Also note the differences between the trajectories and dynamics even though the parameters are the same on average—the consequence of stochasticity.

cess may be good or bad. Because this variation tends to operate on the entire population, it causes more dramatic fluctuations in population size (Fig. 2).

### 3. Catastrophes

Some natural events cause precipitous declines in the sizes of populations, such as fire, flood, disease, and frost. Although these events can be seen as an extreme form of environmental variation, they are often considered separately as catastrophes. We define a catastrophe as being a single event that causes the death of a substantial fraction of a population. Recent analysis implicated catastrophes in the extinction, or near extinction, of many species. The Puerto Rican parrot was almost wiped out by a hurricane, the Guadeloupe oriole by a volcanic eruption, black-footed ferrets by canine distemper, and a population of heath hen by fire. Species with populations that are spatially restricted, such as those on islands, are particularly vulnerable to catastrophes. However, disease appears to be able to decimate continental populations over huge areas. Disease is implicated as a factor contributing to declines of once common species, such as the passenger pigeon and many Australian carnivorous marsupials such as the Tasmanian tiger. Figure 3 shows the trajectory of four

populations for which there is a 2% annual probability of 90% of the individuals being killed in a single catastrophe.

Including catastrophes is an important part of any PVA (Mangel and Tier, 1993) but a difficult component to include in models because of limited data on the frequency and impact of catastrophes. Although we often have good data on catastrophes such as fire, and these have been successfully included in models of Australian forest fauna (Lindenmayer and Possingham, 1996), we rarely have good data on catastrophes such as disease that are more difficult to detect but probably just as important.

### B. Genetic Stochastic Processes

Genetic stochastic processes change the frequencies of genes within a population. When a population is small and isolated, genetic variation is typically lost from generation to generation. This process is called genetic drift. As frequencies of alleles change at random with each successive generation, genetic variability is reduced because rarer alleles are lost by chance through the stochastic nature of birth and death. A consequence of genetic drift is increased homozygos-

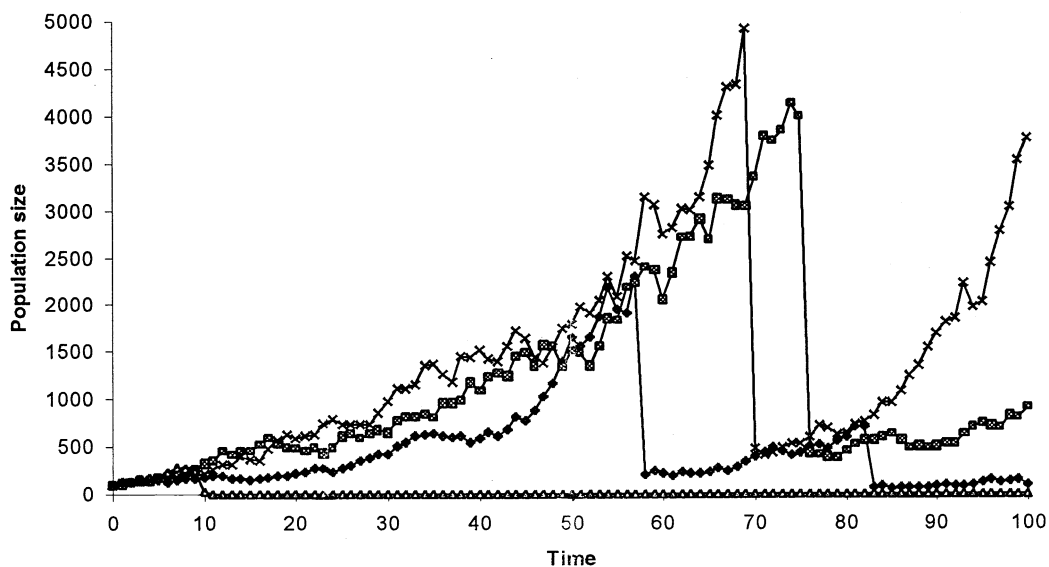


FIGURE 3 The impact of catastrophes on the dynamics of four populations. The same parameters as in Fig. 2 are used and include both demographic and environmental stochasticity. However, there is a 2% chance per year that 90% of the population dies. These catastrophes cause enormous fluctuations in the population sizes that can take very healthy large populations to the brink of extinction whereupon demographic and environmental stochasticity are sufficient to cause extinction.

ity, which can in turn lead to juvenile mortality and lower rates of fecundity and growth (inbreeding depression). Inbreeding depression is defined as mating between individuals that are related by ancestry and is more likely in populations that are, or have been, small. Inbreeding depression results in the selective removal of inbred animals and the genes carried by such animals (Lacy, 1993). Morphological defects in threatened species such as the Florida panther (*Felis concolor coryi*) and lack of breeding success in the Puerto Rican parrot (*Amazona vittata*) appear to be a result of inbreeding depression. The restoration of genetic variability is slow and believed to be between 1 and  $10 \times 10^{-6}$  genes per generation per 100 years. Thus, genetic mutations have only a minor effect in reversing the loss of genetic variation during the time frame typical for PVA. A longer term consequence of loss of genetic variability is a reduced ability of a population to adapt to environmental changes. The long-term consequences of lost genetic variability are not well understood.

### C. Overview of the Role of Stochasticity on Time to Extinction

Each kind of stochasticity has quite distinct effects on the relationship between extinction probability and population size (or carrying capacity). Analytic and simulation studies have shown that although the mean time to extinction, and hence the viability of a population, grows nonlinearly with population carrying capacity when only demographic stochasticity is operating, the other two forms of stochasticity can make even large single populations quite vulnerable to extinction (Fig. 4). With only demographic stochasticity the mean time to extinction increases very rapidly above species-specific thresholds (usually approximately 50–100 individuals). Therefore, in the absence of environmental variability, catastrophes, and genetic stochastic processes, quite small populations could persist for very long times. Adding environmental variation means that the benefits of increasing population size are diminished and there are no simple thresholds above which populations are likely to be completely secure. Because catastrophes can reduce a population to a small size very quickly, a larger population size brings diminishing returns in terms of viability. When this occurs, the best way to improve population viability may be to ensure that a single catastrophe cannot affect the whole population—a risk-spreading approach. This issue will be discussed later.

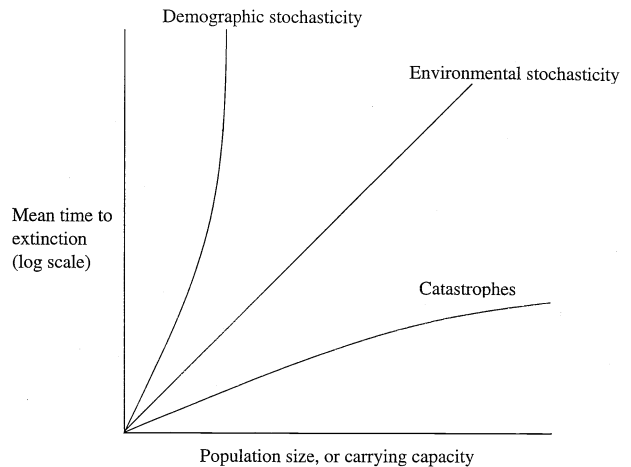


FIGURE 4 The three different types of population stochasticity have different impacts on the relationship between the mean time to extinction and the size (or carrying capacity) of a population.

### D. Deterministic Extinction Processes and Extinction Vortices

A major contribution of PVA to conservation biology is that it focuses attention on stochastic extinction processes in small populations. This should not distract our attention from equally important, more deterministic processes that lead to extinction (Caughley, 1994). Habitat loss and associated fragmentation, increases in mortality rates from introduced predators, and reduced reproduction through habitat modification all have significant impacts on the carrying capacity and rates of increase of populations. Although the distinction between deterministic and stochastic extinction processes is useful conceptually, it is worth noting most important impacts on species have deterministic and stochastic components.

The deterministic and stochastic extinction forces need to be integrated. We can think of the interaction of some of these processes as extinction vortices, in which they nonadditively contribute to the demise of a species. For example, a catastrophe may take a population close to extinction, resulting in a period of inbreeding depression which keeps the population sufficiently low so that demographic stochasticity finally causes extinction. Another example is the way in which habitat loss causes population fragmentation where each isolated population is small enough to fall victim to the interaction of genetic and demographic stochastic processes.

Understanding the process of extinction is funda-

mental to our ability to model it. The obvious question arises: What use is an extinction model?

### III. WHY DO PVA?

#### A. Applications of PVA

Originally, PVA had two general purposes. Shaffer's (1981) original intent was to assess the viability of a population. This assessment of extinction risk is useful in that it informs us of the effectiveness of conservation strategies. In the case of the grizzly bear in the Yellowstone region of the United States, the assessment implied that the species was unlikely to persist in the long term. In the following years, PVAs were tied very closely to the idea of a minimum viable population (MVP). Scientists and managers used PVA to determine if the strategies for conserving a population were adequate. The goal was to use PVA to derive a target population that was large enough to be viable and then to recover the species to that level as quickly as possible.

When one of the primary tools for managing a threatened species includes setting aside an area of suitable habitat for that species, the idea of an MVP extends to that of a minimum viable habitat area (MVHA). A MVHA is a useful concept in which the main threatening process is habitat destruction. However, it has limited value when threats to a species are not simply removed by reservation, such as the continued activities of a predator (e.g., the western swamp tortoise in Western Australia). A MVHA can be predicted using a PVA by predicting extinction probabilities as a function of reserve area. Typically, the relationship between extinction probability and area conserved is sigmoidal (Fig. 5). The MVHA is simply derived by choosing an acceptable level of extinction probability and then conserving the corresponding area. In some cases, the possibility that a single habitat patch ensures viability is unlikely because of catastrophic events, and PVA models that allow for multiple populations (metapopulations) are necessary to determine an adequate reservation strategy. Such a strategy depends on the number, size, and spatial distribution of reserves and the nature of the catastrophic events.

Another application of PVA that relies on the accurate assessment of extinction probability for a species is that of ranking threatened species. Many countries and states are required, for legislative and/or resource allocation purposes, to maintain and update lists of threatened species. The International Union for the Conservation of Nature has explicit definitions of the different categories of threatened species: critically en-

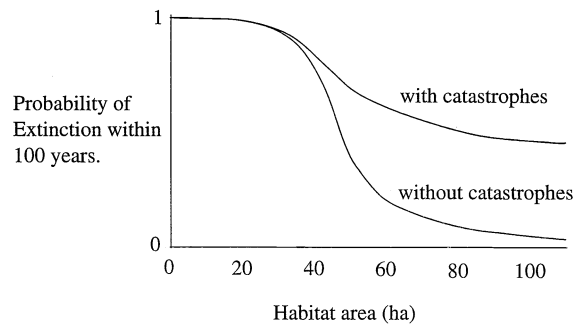


FIGURE 5 The relationship between old-growth patch size and extinction probability of Leadbeater's possum (Lindenmayer and Possingham, 1996) for high-quality habitat in the absence of any catastrophes. Reserves less than 20 ha are of almost no value in isolation, whereas reserves of 100 ha have considerable resilience to demographic and environmental stochasticity. When there are fires that destroy the entire patch, there is no single patch size that gives an acceptably low extinction probability.

dangered, endangered, vulnerable, rare, insufficiently known, and extinct. These criteria are becoming increasingly complex and they vary from country to country. One means of classifying species according to their level of threat would be to do a PVA on every species in the list. Although this has the merits of apparent objectivity, PVAs are rarely used for classifying threatened species because there is insufficient data to parameterize PVAs for so many species. One approach to dealing with the uncertainty associated with extinction probabilities is to take a decision analysis, or fuzzy set, approach in which each species is allocated a probability of being in a particular threatened species category.

In the United States, the ideas of MVPs, MVHAs, and species classifications are given a litigious dimension because of the federal Endangered Species Act. The Endangered Species Act and associated recovery plans and habitat conservation plans state in a variety of ways that species should be managed for viability. Indeed, any development should not compromise the viability of an endangered species. The idea of habitat conservation planning is to allow development where parties agree to the level of mitigation necessary to ensure viability. Because the notion of viability is central to this law and the planning processes that have arisen from the law, PVA has entered the courts in land-management and planning disputes.

#### B. PVA for Making Management Decisions

During the 1990s, scientists questioned the traditional uses of PVA and also questioned the ability of PVA

to accurately predict extinction risk for most species (Possingham *et al.*, 1993; Taylor, 1995; Beissinger and Westphal, 1998). Workers have found that the probability of extinction of an organism is very sensitive to small variations in key life history parameters, especially adult mortality rates. There appear to be adequate data to make very accurate estimates of extinction risk for only a very small fraction of threatened species.

An alternative role for PVA is as a decision support tool to help make management decisions. This role focuses on the robustness of different management decisions to deliver acceptable conservation strategies. PVA can be used to make decisions regarding the following:

1. Determining if, and how much of, a population should be brought in to captivity
2. Translocation strategies between populations or to new sites
3. Reintroduction strategies
4. Deciding how to set priorities between species
5. Determining optimal habitat management, including disturbance regimes such as fire and logging
6. Control strategies for the predators, competitors, and prey of threatened species
7. Optimal size, distribution, and spatial pattern of reserve systems

Probably the most widespread and frequent application of PVA is through the activities of the Captive Breeding Specialist Group (CBSG). This group aids zoos and conservation agencies throughout the world to set conservation priorities and actions by workshopping species groups region by region.

Some argue that PVA has been most useful in providing a framework within which experts can operate. The task of running a PVA draws the experts together and focuses their attention on estimating population dynamics and identifying threats. This is certainly true for the CBSG workshops and users must view PVA as a decision support tool and not as the decision maker.

### C. Other Measures of Viability

The probability of extinction of a population is not the only measure of population viability. Other common measures are the mean and median of the time to extinction. When the extinction probability for a species is fairly constant, the expected life span of the species can be thought of as being negatively exponentially distributed so the rate of extinction is simply the inverse of the mean time to extinction. All these measures are fairly closely related.

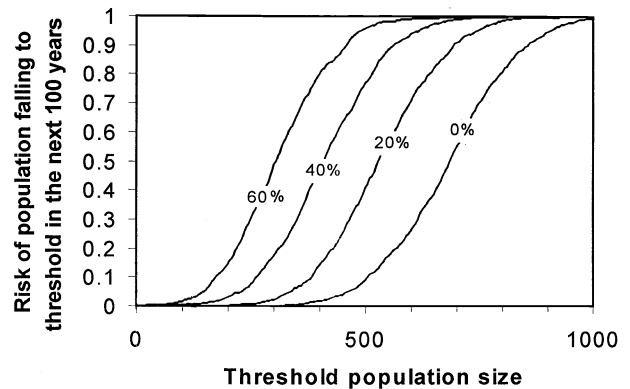


FIGURE 6 Hypothetical risk curves. The user can choose whatever combination of threshold population size and level of risk he or she wishes. The four lines are quasi-extinction risk curves for different levels of population harvesting.

Because of the uncertainties associated with modeling the dynamics of very small populations (Allee effects, genetic effects, etc.), one prudent approach to providing a measure of risk is to introduce the notion of quasi-extinction, which is the risk of the population decreasing in number to some preset size. For example, although the probability of extinction is the chance of decreasing to zero individuals, a manager may argue that a species has failed long before this and prefer to set a threshold population size of, for example, 20, below which the population is quasi-extinct. This approach eliminates some of the problems associated with modeling very small populations.

An even more general approach to assessing extinction risk that extends the idea of quasi-extinction is to develop risk curves. A risk curve (Fig. 6) shows the probability of a population decreasing below any population size, from zero (extinction) to the carrying capacity of the population. With a risk curve, the manager is free to pick whichever quasi-extinction threshold he or she wishes, or to assess alternative management strategies by investigating changes in the whole risk curve.

## IV. METHODS AND TOOLS FOR PVA

### A. Demographic Models for PVA

Generally, a PVA is a demographic model that includes some or all of the extinction processes discussed previously. There are several ways of constructing these models, and there are several publicly available simulation "packages" that people can use to do PVA. The

general categories of PVA model are discussed in the following sections.

### 1. Matrix Population Models

Basic demographic projection is one of the most powerful and important parts of PVA, although sometimes it is bypassed. By constructing an age- or stage-structured population projection matrix, one can determine both the overall growth rate of the population and the sensitivity of that growth rate to variation in life history parameters by finding the dominant eigenvalue of the matrix. More simply, the population can be projected forward in time using a simple spreadsheet. If the dominant eigenvalue of the population is less than 1 then we know the population is doomed to extinction (deterministic extinction). If there is uncertainty about the value of key life history parameters, then the risk of deterministic extinction can be calculated by carrying out a sensitivity analysis on the dominant eigenvalue. Although this approach does not allow for any of the stochastic forces described previously, it gives the manager an idea of how the population is faring in the absence of stochasticity. If the dominant eigenvalue is likely to be less than 1, then our sensitivity analysis allows us to target those life history attributes, such as birth rate or adult death rate, most likely to generate a positive long-term growth rate. Matrix population models can include stochasticity (indeed, RAMAS Age is based on the application of projection matrices for its predictions), but without a computer package the mathematics is quite complex.

### 2. Analytic Methods and Approximations

Although including stochasticity normally requires that we proceed to a computer simulation, there are analytic methods for calculating the probability of extinction, and in particular the mean time to extinction, of populations. One approach uses Markov chain theory to model a basic birth and death process. This model includes only birth and death and can be used to generate analytic expressions for the mean time to extinction.

Another analytic approach involves approximating the stochastic population dynamics with differential equations called "diffusion" equations. Diffusion equations can be analyzed to generate explicit expressions for the mean time to extinction of a population under different sorts of stochasticity; indeed, this approach was used to generate the relationships shown in Fig. 4.

Although analytic approaches have elegance and generality, they are restricted in the processes which can be included and they can be quite difficult to parameterize from field data. They have made significant con-

tributions to our general appreciation of the role of stochasticity in extinction and underpin a general theory of extinction, but they are rarely used in practice.

### 3. Time Series Analysis Methods

A relatively new method for exploring extinction risks utilizes analysis of time series of population abundance. During the past decade, a considerable amount of time series data on a wide variety of organisms have been obtained. If these time series of population size are sufficiently long (at least 15 years), then it is possible to statistically "fit" different models of stochastic population growth (such as a Ricker model or logistic model) to these time series. Such models are necessarily simple but have the advantage that they do not require any life history data—only a series of counts. Given the best fit model, it is a straightforward task to simulate the dynamics into the future and calculate extinction risks.

Although this approach will prove increasingly useful for making extinction risk predictions, and hence helping with tasks such as prioritizing species, the lack of mechanisms in these models limits their application for management purposes.

### 4. Numerical Methods

The approach of using a Markov chain (or stochastic population projection matrix) can be expanded to include other forms of stochasticity besides simply demographic. If transitions from any population state to any other population state can be estimated as a function of life history attributes and biotic and abiotic influences, then increasing these matrices to increasingly higher powers allows us to project the probability distribution of possible population sizes into the future. This method provides the foundation of the RAMAS extinction risk software.

The numerical method has the advantage of some mathematical elegance, and it is certainly powerful when coupled to front-end software such as that available in RAMAS. However, most people who build models from scratch find the mathematics involved too complex and tend to use the Monte Carlo simulation method.

### 5. Monte Carlo Simulation Models

The most common way of doing PVA is to use a Monte Carlo simulation model. These simulation models are either written from scratch or the user uses one of the available packages (Lindenmayer *et al.*, 1995). The most frequently used packages are VORTEX and RAMAS (several versions/types); however, others are available (e.g., ALEX and GAPPS). The basic structures of all

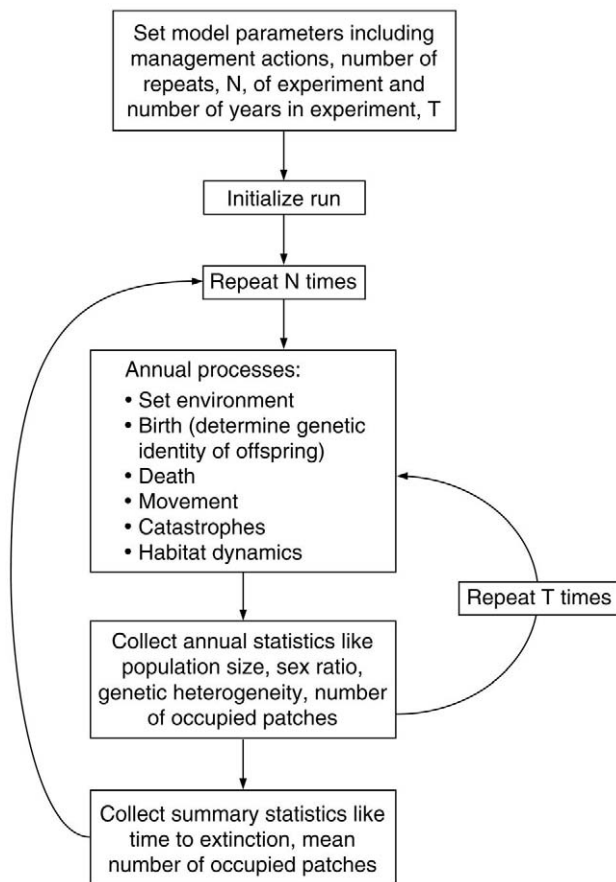


FIGURE 7 A flow chart showing the typical processes included in a Monte Carlo simulation PVA.

these models are fairly consistent and are best visualized in a flow chart (Fig. 7). However, the details of how each component is modeled and which processes can be included in the models vary considerably from model to model and species to species. Each has its strengths and weaknesses. For detailed application to a particular species it is advisable, although expensive and time-consuming, to construct models specific to that species that focus on relevant processes and management options.

For example, several different kinds of PVA models were built to determine efficient and adequate conservation strategies for the northern spotted owl (*Strix occidentalis caurina*). These ranged from simple demographic models to very detailed landscape models with habitat dynamics and very specific biological factors such as predation by another owl. The diversity and number of models for this species largely reflect the controversy regarding it: The spotted owl requires large

tracts of forested land which are very valuable for timber production. For species that have far less economic importance, there is neither the time nor the money to embark on detailed modeling exercises and so one of the existing packages is used.

Monte Carlo simulation models run on a PC have the advantage that they can incorporate many different parameters and processes; their complexity is limited only by the imaginations of the biologists and managers. However, the inclusion of detail and complexity can give the illusion of accuracy, whereas some consider that adding more poorly understood processes will make the results less accurate. The factors driving the extinction estimates from complex models are often very difficult to isolate.

## B. Sensitivity Analysis of PVA Models

Sensitivity analysis is an important component of modeling because one can use it to systematically investigate the complex interactions of a model. Sensitivity is usually measured by varying a parameter by a small amount from its estimated value. The resulting change in the state variable (e.g., the risk of extinction) provides an index of the sensitivity of the model to that parameter. Sensitivity analysis provides practical information for model builders and users by highlighting parameters that have the greatest influence on the results of the model. It can highlight model parameters that should be most accurately measured so as to maximize the precision of the model, give a general indication of the reliability of the model predictions, and highlight parameters and interactions that have the largest influence on the population to help determine effective management strategies.

The sensitivity of deterministic matrix population models can be determined analytically by eigen analysis. Extension of these techniques to PVA is not appropriate for most PVA models except for matrix models for at least three reasons: (i) PVA models are often complex, so obtaining solutions analytically is difficult if not impossible; (ii) in PVA, the result of interest is the risk of population decline and not the deterministic growth rate; and (iii) interactions between variables are largely ignored.

Any method of sensitivity analysis should be clearly defined, interactions between parameters should be distinguishable from single parameter effects, and the method should account for variability associated with parameter estimates. The simplest approach to sensitivity analysis of PVA models is to vary the model parameters in turn and investigate the effect on the risk of



population decline. Alternatively, such a sensitivity analysis can determine whether the relative efficacy of different management decisions changes with different parameter values. PVA models may be complex with numerous parameters, particularly when individuals are modeled, and understanding the relative importance of different parameters and interactions between parameters may be difficult. For example, if there are 10 parameters and three levels for each parameter are to be investigated, then 21 different parameter combinations would be necessary to assess each parameter independently. However,  $3^{10}$  (=59,049) different combinations are required to test all possible interactions.

When the risk of population decline is the important state variable, logistic regression may be useful for summarizing the effects of different parameters and interactions. Data for the regression are generated by using numerous parameter combinations. For each parameter combination, the PVA model is simulated to obtain a limited number of predictions of the incidence of decline. The regression analysis uses the model parameters as explanatory variables and the incidence of decline as the dependent variable. The regression equation provides a simple expression to approximate how the probability of decline is influenced by the model parameters.

### C. Problematic Issues in PVA Models

Although all models of populations are necessarily false because they cannot include all the factors and processes that influence population dynamics, we believe that some processes and factors in PVA modeling are important but are difficult to incorporate.

#### 1. Genetics in PVA Models

Early work on PVA paid considerable attention to the role of genetics, although the significance of genetics in short-term threatened species management has declined in the past decade. Two genetic issues are of particular concern: inbreeding depression and loss of heterozygosity. Inbreeding depression is of particular concern for very small populations, especially captive populations. Estimating the likelihood and effect of inbreeding depression on vital rates is problematic. The loss of genetic heterozygosity is a long-term concern. Populations with an effective size greater than 50 are generally thought to experience low losses of heterozygosity. VORTEX is a generally available package that allows the user to include genetic factors, hence its popularity with zoos and reintroduction programs.

Genetic data can be difficult and expensive to obtain and one needs to consider how the data will influence one's decision making. A useful role of molecular genetic data in PVA is to provide estimates and plausible bounds for demographic attributes that are otherwise impossible to obtain (such as social structure and dispersal rates).

#### 2. Density Dependence in PVA Models

Density dependence at both low and high densities is important for PVA. At low densities the possible consequences of positive (or negative) density dependence are critical to the ability of a species to recover. For populations with limited habitat, the way in which vital rates change as populations approach their carrying capacity, and whether the carrying capacity fluctuates greatly from year to year, is known to influence estimates of viability. Unfortunately, information on density-dependent processes for most threatened species is extremely limited. Different PVA models use different forms of density dependence, from logistic growth and a population "cap" to no density dependence, whereas other models allow the user to choose the appropriate form.

#### 3. Space in PVA models: Metapopulation Models and Risk Spreading

Space is an essential component of any PVA model for which the intention is to make statements about what management actions should take place at which particular geographic locations. For example, in determining habitat reconstruction plans for threatened species (linking corridors and new habitat patches), it is generally essential to represent space explicitly in the models.

Space can be incorporated into a PVA model either implicitly or explicitly. The implicit use of space is typical of many metapopulation models in which the interest is in the number of patches that are occupied, with no particular regard to exactly which patches are occupied. These models can be used in a general sense to report on the values of different reserve sizes or corridors. For detailed management for which the question of exactly where a corridor needs to be located, or which patch needs to be protected, might need to be answered, we normally require models in which real landscapes are explicitly represented.

#### 4. Habitat Dynamics

Early PVA models, and almost all of the analytic approaches, assume a static habitat. Given that habitat loss is one of the primary causes of species extinction,

the introduction of habitat dynamics in PVA models is of increasing interest. Even in the absence of habitat loss, most habitats are dynamics and some recent PVAs have focused as much on the dynamics of the landscape as on the dynamics of the population. This reflects an increasing interest in population ecology—not just in the role of space but also in the role of a changing habitat mosaic in population dynamics.

Combining habitat dynamics with spatial complexity has been facilitated by faster computers and the integration with geographic information system (GIS) software. Future wildlife managers will certainly have access to GIS systems that enable them to interactively explore the consequences of different habitat management actions on extinction probabilities. These complex new-age GIS/PVA tools will include the tools inherent in GIS systems, habitat suitability models, management models, stochastic catastrophes, and detailed population dynamics. To some this will be a powerful and compelling tool and to others a dangerous cascade of complex untested models interacting in ways that the user will never understand.

#### D. Other Methods of Assessing Viability

There are many measures of viability and many ways of assessing viability—from a wild guess to a complex computer model, and from a simple analytic model to extrapolation from empirical data. So far, we have discussed more traditional model-based methods. Here, we briefly mention two other ways of assessing the viability of a population

##### 1. Historical Data

The occurrence of populations in isolated habitat, especially on oceanic islands, gives us an idea of how large patches of habitat need to be for the long-term persistence of certain species. For example, in the Caribbean reptiles of certain species rarely occur on islands less than critical sizes, but they are consistently present on larger islands. This sort of island data can help us set MVHAs for particular taxa. As is known from island biogeography theory, islands are not isolated and their species compliment is a consequence of local extinction and colonization processes. In this context, the presence of a species on an island may not reflect its long-term persistence, only its recent arrival. However, where species have persisted on islands long enough to show variation from mainland populations, we can assume they have persisted for long enough to be termed viable. Although this general approach to PVA has been fruit-

ful, the species that occur on these islands are usually not the same species for which we are trying to design reserve systems on continents. The question arises, how relevant is it to use data on MVHAs from one species for another species?

##### 2. Extrapolation from Similar Species

Although we are unlikely to have data on the viability of, or MVHAs for, threatened species, it is possible that such information on similar, more common species can be used to assess viability. For example, we may know much about the fate of many small populations of a common species and use this information to determine the MVHA of a similar threatened species.

Expanding this line of thinking, some workers have begun to develop rules of thumb that relate extinction probability of MVHA to key life history attributes of a species, such as body size or home range size. We do know enough to say that predators can persist at much lower densities than herbivores of a certain size, and that larger animals and plants appear to be able to persist for longer periods in small populations than can smaller animals or plants. These fundamental patterns may enable us to draw broad conclusions about the viability of species and strategies for their protection with limited data and no predictive modeling. This crude approach may only be appropriate when data, time, and money are limited.

#### V. TESTING PVA

Uncertainty about parameter estimates and the structure of PVA models means that their predictions may be inaccurate. As with any model, PVA is an imperfect description of reality, and model testing with real field data helps to identify the limits of its accuracy. The primary prediction of many PVA models is the risk of extinction of an endangered species over several decades or centuries. In such circumstances, we do not have the luxury to test the models by monitoring numerous populations until they become extinct. A sufficient number of populations rarely exist, and endangerment demands immediate management decisions.

Instead, we can test PVA models by predicting population attributes prior to extinction. The kinds of predictions that can be tested include the annual population growth rate, the population size in a series of years, the local population sizes in patches, and the probability of local extinction of patches. It is not sufficient to test the models by comparing mean predictions with

observations because this ignores variability, which often is a critical factor influencing the risk of population decline. Tests of PVA models should investigate the distribution of the predictions—not only the average predictions but also factors such as the variance and predicted extremes such as local extinction. Replicates are necessary for such tests and may be obtained from annual observations over several years or from observations from several different locations. These tests rely on appropriate data obtained from the field, which may require a substantial commitment to field surveys. As noted previously, we may need to test PVA models on more common species and then, by analogy, assume that they will work for threatened species. This requires a certain leap of faith.

Correspondence between the predictions of the model and observed data does not imply that the model is correct. Tests of PVAs may not be sufficiently powerful to identify the errors, or multiple compensating errors may exist. Regardless of how many times a PVA model has been tested, new data or a new test will expose the model to a chance of being proved wrong. This is similar to the scientific process of hypothesis testing.

Discrepancies between the predictions of a PVA model and observed data should not lead to complete rejection of the model. Instead, efforts should be made to identify the source of errors, which may be rectified to provide improved predictions. Identification of model errors is a useful part of model development. Such an approach to model testing is consistent with the use of PVA models as decision support tools. Improved PVA models should contribute to improved management decisions.

## VI. CRITICISMS OF PVA: WHAT MAKES A GOOD PVA?

There is a healthy skepticism about the value of PVA. Although almost all authors acknowledge that PVA is valuable and a cornerstone of the science of conservation biology, PVA, like any tool, has been abused.

The following are major criticisms of PVA:

1. Not enough data
2. Black box model that no one understands
3. Sensitivity to parameters
4. Missing processes; not enough ecology

Because of the “black box” nature of PVA, there is concern that PVAs should be more explicit and detailed

in their reporting of how the model works and exactly how the parameters were chosen. This is just part of good science—that is, whatever we do should be repeatable and transparent. Despite the diversity of PVA models, there is consensus on what constitutes a good PVA. Attributes of a good PVA include

1. A description of why the PVA is being done (a motivation)
2. A fundamental understanding of the species' ecology, including what constitutes suitable habitat and the ability of the species to disperse between patches of habitat
3. Detailed demographic data and the way in which demographic rates change with time and habitat: Care needs to be taken in describing exactly what parameters in the model do and exactly how they were estimated from data
4. An understanding of the response of the species to threats
5. An assessment of the current state of the population, acknowledging inaccuracies in the estimate
6. A statement about how the PVA is likely to help managers make decisions
7. A description of constraints on decision variables and management options
8. An understanding of the environmental disturbances that directly threaten (or indirectly threaten through habitat alteration) a species
9. A detailed description of the model and data so that the PVA process can be repeated and extended
10. An honest assessment of uncertainties in the data, processes that have been ignored, and possible flaws in model construction and parameter estimation
11. A sensitivity analysis.

## Acknowledgments

This work was conducted as part of the Biological Diversity Working Group supported by the National Center for Ecological Synthesis and Analysis, a center funded by NSF (Grant No. DEB-94-21535), the University of California at Santa Barbara, and the State of California. Work by all three authors was also supported by an ARC SPIRT grant to HPP and DBL, partly funded by Environment Australia. We thank our numerous colleagues who have worked on PVA with us, especially Mark Burgman, Sandy Andelman, Bob Lacy, Drew Tyre, and Ian Ball.

## See Also the Following Articles

CARRYING CAPACITY, CONCEPT OF • FRAMEWORK FOR ASSESSMENT AND MONITORING OF BIODIVERSITY • MEASUREMENT AND ANALYSIS OF BIODIVERSITY • POPULATION DENSITY • POPULATION DYNAMICS

## Bibliography

- Beissinger, S. R., and Westphal, M. I. (1998). On the use of demographic models of population viability in endangered species management. *J. Wildlife Management* **62**, 821–841.
- Boyce, M. S. (1992). Population viability analysis. *Annu. Rev. Ecol. Syst.* **23**, 481–506.
- Burgman, M. A., Ferson, S., and Akçakaya, H. R. (1993). *Risk Assessment in Conservation Biology*. Chapman & Hall, London.
- Gilpin, M. E., and Soule, M. E. (1986). Minimum viable populations: Processes of species extinction. In *Conservation Biology: The Science of Scarcity and Diversity*, pp. 19–34. Sinauer, Sunderland, MA.
- Lande, R. (1988). Genetics and demography in biological conservation. *Science* **241**, 1455–1460.
- Lande, R. (1993). Risks of population extinction from demographic and environmental stochasticity and random catastrophes. *Am. Nat.* **142**, 911–927.
- Lindenmayer, D. B., and Possingham, H. P. (1996). Ranking conservation and timber management options for Leadbeater's possum in southeastern Australia using population viability analysis. *Conserv. Biol.* **10**, 235–251.
- Mangel, M., and Tier, C. (1993). A simple and direct method for finding persistence times of populations and application to conservation problems. *Proc. Natl. Acad. Sci. USA* **90**, 1083–1086.
- Possingham, H. P., Lindenmayer, D. B., and Norton, T. W. (1993). A framework for the improved management of threatened species based on population viability analysis. *Pacific Conserv. Biol.* **1**, 39–45.
- Shaffer, M. L. (1981). Minimum population sizes for species conservation. *Bioscience* **31**, 131–134.
- Taylor, B. L. (1995). The reliability of using population viability analysis for risk classification of species. *Conserv. Biol.* **9**, 551–558.





# POVERTY AND BIODIVERSITY

Madhav Gadgil

*Indian Institute of Science*

- 
- I. Of Ecosystems and People
  - II. Autonomous People
  - III. Subjugation: Political and Economic
  - IV. Market Forces
  - V. Land Use Changes
  - VI. Value Appropriation
  - VII. Ecological Refugees
  - VIII. Costs of Conservation
  - IX. Ecodevelopment
  - X. The Challenge Ahead
- 

**ecosystem people** People meeting the bulk of their resource requirements from a limited area near their habitation through gathering or low input agriculture and animal husbandry.

**resource catchment** Locality from which the resources consumed by a group of people are derived.

---

## GLOSSARY

**artificial ecosystems** Ecosystems dominated by human artifacts and human induced material and energy flows.

**biosphere people** People enjoying access to resources garnered from the entire biosphere and made available through markets.

**biosphere reserve** An international conservation program aiming to protect representative ecosystems throughout the world in ways compatible with development efforts.

**ecological footprints** Spatial coverage of ecological impacts consequent on the activities of a given group of people.

**ecological refugees** People that have lost access to their traditional base of natural resources yet have very limited access to resources through markets.

**WEALTH—AND ITS OBVERSE, POVERTY**—is today reckoned in terms of the capacity of a person or groups of persons to obtain goods and services through exchanges in the market-place. The bulk of such goods and services are partly or wholly products of intensively managed or artificial ecosystems; crop fields, plantations, shrimp ponds, mechanical fishing fleets, factories, towns, and cities. Most of these intensively managed or artificial ecosystems tend to harbor low levels of biodiversity. Rich people are then those with extensive access to produce of managed or artificial ecosystems supporting low levels of biodiversity; poor people have very limited access to such produce. Ecosystems harboring high levels of biodiversity are, on the other hand, natural or semi-natural ecosystems with low levels of human demands for their produce. Rich people rarely live in close vicinity of such ecosystems, though they may visit them for recreational purposes. Groups of poor people may, on the other hand, permanently live in their vicinity. In some cases, the poor control and serve as the stewards of such biodiversity rich ecosystems. More often though many such biodiversity rich ecosystems are subjected to overexploitation, primarily

to meet the large resource demands of the rich, often living far away, with the local poor serving as agents of the resultant destruction of biodiversity. Large numbers of poor also live permanently in the vicinity of biodiversity poor, intensively managed, or artificial ecosystems. There is then no simple relationship between poverty and biodiversity; the equations vary greatly from context to context.

## I. OF ECOSYSTEMS AND PEOPLE

To appreciate these complexities, one must tease out the many contexts. These may be viewed along two axes; one of the extent to which ecosystems have been transformed through human interventions, and the second of the manner in which people relate to the world of nature, to the world of manufactured artifacts, and to each other. The terrestrial and aquatic ecosystems may be classified into four major categories: (a) natural ecosystems, which are subject to very low levels of human demands because of their inaccessibility; (b) natural and semi-natural ecosystems, including low-input agricultural systems subject to higher levels of human demands; (c) ecosystems managed intensively for biological production; and (d) largely artificial ecosystems dedicated to industrial production and organized services. Building on Dasmann's (1988) pioneering work, we may, for our purpose, classify people into four major, largely inclusive categories; (a) autonomous ecosystem people, (b) subjugated ecosystem people, (c) biosphere people, and (d) ecological refugees. Ecosystem people have limited access to sources of energy other than human and livestock muscle power and to the more sophisticated artifacts. They gather or produce most of the resources they consume from their immediate surroundings, from the forest, scrub, rivers, or seas and from low-input cultivation. In some of the more inaccessible corners of the world, the ecosystem people are still autonomous. However, over most of the earth they have been subjugated by the biosphere people and have very limited control over their own resource base of natural and semi-natural ecosystems. They gather and produce little that can fetch value in markets and therefore have very limited access to products of intensively managed and artificial ecosystems. The biosphere people owe their dominant position to their extensive control over artifacts and additional sources of energy. They engage in energy-intensive agriculture, animal husbandry, aquaculture, or in organized services or industrial production and generate much of value in the markets. They have large ecological footprints thanks to their substantial purchasing power; their

resource catchments are vast encompassing all of the biosphere, bringing to them goods and services from all over the earth. This confers on them the ability to take over resources in demand by the ecosystem people, catalyzing transformation of natural and semi-natural ecosystems into those managed intensively to meet their own demands or, in some special cases, conserved as natural or semi-natural ecosystems for recreational purposes. In the process, they often deprive ecosystem people of access to their traditional resources; this may convert them into ecological refugees. Ecological refugees are then people with attenuated access to resources of natural and semi-natural ecosystems, with little purchasing power to access products of intensively managed and artificial ecosystems. They often end up constituting the unorganized labor force in tracts of intensive agriculture and urban settlements (Gadgil, 1995; Gadgil and Guha, 1995).

## II. AUTONOMOUS PEOPLE

I propose to explore the relation between poverty—or wealth—and biodiversity in this framework of different categories of ecosystems and of people. Table I summarizes the relationship and notes the examples that are discussed later at some length. All of these derive from my own fieldwork in India; I use them purely because I know them well. However, they do represent broader patterns encountered in other parts of the world. Not all 16 cells of the matrix are occupied; moreover our focus is on poverty, and we shall therefore not elaborate on the relationship of the biosphere people with biodiversity, except as they impact on the ecosystem people or ecological refugees. There are relatively few examples of truly autonomous ecosystem people, people fully in control of largely natural ecosystems with very light human demands, and in consequence with high levels of biodiversity. Inhabitants of Sentinelese island (11°30' N lat. 92°15' E long.) in the Andaman-Nicobar chain in the Bay of Bengal provide one such example. These relatively inaccessible islands harbor tropical rain forest biota with high levels of endemism (i.e., of species restricted to this chain of islands). They were inhabited by a number of hunter-gatherer tribal groups, without knowledge of metal tools, with the exception of Shompens of Nicobars who had a more advanced fishing economy. British colonized the islands in mid-nineteenth century, an attempt that was strongly resisted by the indigenous people. By 1870s the resistance was overcome and as a consequence many of the tribal populations either drastically declined or were exterminated.

TABLE I  
Variety of People-Ecosystem Contexts, with Specific Examples

Ecosystems	People			
	Autonomous ecosystem people	Subjugated ecosystem people	Biosphere people	Ecological refugees
Inaccessible natural/seminatural ecosystems	Poor in modern economic sense, with access to high levels of biodiversity, e.g., Sentinelese islanders	—	—	—
Accessible natural/seminatural ecosystems	—	Poor in modern economic sense, serve as agents for destruction of biodiversity, e.g., Gangtes of Manipur, sometimes as stewards, e.g., Village Forest Committees, protectors of Saranas	Visitors for recreational/commercial use	Poor in modern economic sense, serve as agents of destruction of biodiversity, e.g., Panshet peasants, Maldharis of Gir
Ecosystems managed intensively for biological production	—	—	Well off in modern economic sense, promote low biodiversity production systems	Poor in modern economic sense, alienated from biodiversity, e.g., Jharkhand tribals
Artificial ecosystems dedicated to industrial production/organized services	—	—	Well off in modern economic sense, promote high biodiversity systems for recreational purposes	Poor in modern economic sense, alienated from biodiversity, e.g., Jharkhand tribals

nated. During British times the islands primarily served as a convict colony, with many of the released convicts settling down to agriculture (Superintendent of Government Printing, 1909). After independence in 1947 the islands were used to create agriculture based settlements of many people displaced from then East Pakistan (now Bangladesh), as well as to supply forest resources for rapidly growing forest based industries, especially plywood (Saldanha, 1989). As these were progressively overused and exhausted leading to substantial erosion of biodiversity, pressure has built up to overcome the resistance of the tribal groups still holding out to open up their territories to commercial forest exploitation. Two of the tribal groups, however, do continue to retain hold over their territories; these are Jarwas and Sentinelese. Jarwas live on the larger South Andaman and Ba-

ratung islands (12°0' N–12°20' N lat. 92°45' E–92°55' E long), about two-thirds of which has been colonized by immigrants and subjected to forest exploitation. Their territory is therefore easily accessible; it is, however, stoutly defended by Jarwas with their bows and arrows. There are ongoing attempts to overcome this with the aid of a major road passing through their territory. This Jarwa territory remains much richer in biodiversity than the rest of the islands; Jarwas are, however, poorer in a modern economic sense with access to very few and simple artifacts. The Sentinelese live on a smaller island, which remains entirely under their control. This island too is very rich in biodiversity; the Sentinelese are also very poor in a modern economic sense. Presumably Jarwas and Sentinelese do not themselves have concepts of wealth and poverty or biodiver-



sity (Bhattacharyya, 1993; Gadgil, 1998). They apparently do have concepts of members of their own groups as opposed to aliens and desirability of excluding aliens from their territory. They remain poor in the modern sense because they have successfully maintained themselves in isolation; for the same reason their territory retains high levels of biodiversity.

### III. SUBJUGATION: POLITICAL AND ECONOMIC

The vast majority of the ecosystem people of the world are, however, no longer isolated in this fashion. In no case do they seem to have voluntarily sought to integrate with the larger society dominated by the biosphere people. Rather the biosphere people have sought to integrate them in order to access resources of their territories, as well as to utilize their labor. The process of such integration tends to proceed through several stages involving political subjugation, followed by economic subjugation. Such subjugation is accompanied by a reduction in levels of access of the people to the biodiversity resources of their own localities; though simultaneously they may have enhanced access to other resources through market channels, thereby implying a reduction in the levels of their poverty in modern economic terms. The process is also generally accompanied by an erosion of biodiversity driven by overexploitation of natural resources to meet outside demands with the subjugated ecosystem people often serving as agents of such erosion.

These processes are the dominant form of interplay of poverty and biodiversity in the world today, and we propose to illustrate them through a series of case studies. The most detailed case to be presented would involve Gangtes, one of the Kuki tribes of Churchandpur district (23°4'–25°50' N lat. and 93°–94°47' E long.) of the northeastern state of Manipur on the India-Myanmar-China border, since this illustrates the complete sequence from autonomy through political and then economic subjugation (Gadgil, Hemam, and Reddy, 1997; Hemam, 1997). Like many other Kuki and Naga groups, Gangtes continued head hunting well into nineteenth century. At that time, there were no roads so that people had to walk for several days to reach markets where they could exchange hand-woven cloth or honey for iron tools or rice. These journeys were hazardous as they could involve passage through territories of alien tribes. Each settlement was in consequence a largely self-sufficient and self-governing en-

tity. Their base of subsistence was shifting cultivation with long fallows of about 15 years. There were no fixed village boundaries and individual social groups presumably shifted around from time to time. The settlements were located on hill-tops; the valleys were considered unsafe and more vulnerable to raids from alien groups. Among Gangtes, hereditary village chiefs determined through primogeniture made all group-level decisions in consultations with a council of elders. They assigned lands for cultivation and annually received about 16 kg of grain from each family in the group as a tribute for use in entertaining any visitors. In the course of cultivation the densest, tallest forest tracts were left alone as requiring too much labor for conversion to fields; the more recent fallows were also avoided. The people worshipped many natural elements including mountain peaks, streams, plants, and animals. They strictly protected patches of sacred groves called gamkhal as also other spirit possessed lands called nungens. Also protected were bamboo groves called mavuhak from which bamboo may be extracted only for house construction, but the shoots, much relished as food, were left alone. This preserved a luxuriant vegetation, which led Captain Pemberton to remark in 1835: "I know no spot in India, in which the products of the Forests are more varied and magnificent but their utility is entirely local" (Pemberton, 1835).

The process of political subjugation of Gangtes began in the late nineteenth and early twentieth centuries with the development of incipient links with mainstream India, then ruled by the British. Initially, the British did not so much want access to resources of Churchandpur as of the much richer Myanmar. Hence they wanted to ensure safe movements of British troops in the hill tracts of Manipur bordering Myanmar. A people without fixed, well-defined villages and with traditions of killing aliens coming into their territories were an obvious threat. The British therefore concentrated on fixing village boundaries and assigning land ownership. Given the Gangte system of hereditary village chiefs having a major role in all decisions including assigning land for shifting fields, they decided to confer on the chiefs all ownership over land, converting others into tenant farmers from a legal perspective. This move then formally introduced wealth and poverty to the Gangte society. However, during the British reign this made little operational difference in most areas, with no acceptance of private property in land by Gangtes, except in the Churchandpur town area beginning 1930s. Elsewhere Gangte community members continued to pay the village chief a tribute of 16 kg of grain as before. The late nineteenth and early twentieth centuries also

witnessed the gradual spread of Christianity. Christianity rejects the attribution of sacred qualities to elements of nature and, consequently, taboos against killing of certain animals or felling of certain patches of forests like gamkhal. Conversion to Christianity therefore began to slowly erode the traditional belief system underlying these conservation practices.

In this phase of political subjugation there was little change in the economy. The road network remained extremely limited and the Gangte communities remained almost totally self-sufficient in terms of resource use. The fallow cycles for shifting cultivation were still long, and substantial areas of forest retained protection in forms of sacred groves. However, by introducing the legal concept of private ownership and a religion that questioned attribution of sacred qualities to nature, this phase did set the stage for the radical changes that followed independence; but unlike on the mainland, this political subjugation had limited impact in the absence of an access to markets and therefore of any attempt at economic subjugation.

#### IV. MARKET FORCES

The British were prompted by the march of the 'Free India Army' fighting in collaboration with the Japanese into Manipur through Myanmar to strengthen the road network of Manipur on a war footing, laying the foundation for the rapid development of transport, communication, and commodification that began on independence in the 1950s. This was consistent with the policies of economic development and national integration adopted at the time of independence. These development policies also encouraged forest based industries by offering them many resources, including wood, at highly subsidized rates (Gadgil and Guha, 1992, 1995). As the production of these industries grew at a rapid pace, the demands outstripped supplies leading to severe overexploitation. Such overexploitation was facilitated by the fact that there was no social group sufficiently motivated to ensure a more sustainable pattern of resource use with a secure enough control over the resource stocks. Access to markets meant that a community no longer suffered an inevitable shortage of resources if those from their own vicinity were overused and exhausted. Neither would such a difficulty be experienced by an industry drawing resources from a larger spatial scale. Market access therefore reduced the motivation of the various parties concerned, local communities, such as those of Gangtes, as well as other consumers such as plywood industry for restraints on levels of

harvests. At the same time conversion to Christianity eroded the conservation ethos of local communities grounded in attribution of sacred qualities to various natural elements. Linkages to the larger national society also diluted the control of any one agency over any given resource base by bringing a variety of actors into play. Thus in the autonomous phase where a person might even be killed while passing through alien territory, there was a clear-cut control by a local community. In the phase of political subjugation this was legally recognized as the control in terms of ownership by the local chieftains. But in the post-independence phase of economic subjugation the state Forest Department has stepped in to claim control over forest lands and forest resources. How this claim is to be reconciled with the claim over land ownership by the Gangte chiefs has not been clarified. This uncertainty has affected the security of resource control by both parties. At the same time, the traditional authority structure of a Gangte village community headed by a hereditary chief with a council of elders has also been affected by the institution of a democratically elected village council. There are continuing contradictions in this system as well since the hereditary chief remains at the head of the elected village council. Under these circumstances, forest resources of Gangte villages have been affected by demands greatly exceeding sustainable yield levels. These demands are no longer local, but reflect the markets not only of remaining parts of the state of Manipur, but the rest of the country as well. Indeed market pulls come from even beyond national borders as Indian plywood was over some periods exported in substantial quantities to the mid-east.

The traditional Gangte society is largely egalitarian. This situation has radically changed with the linkages to the market economy as the chiefs are now motivated to take advantage of the individual ownership of vast tracts of lands assigned to them. Their assertion of such ownership rights over the land or the tree resources standing on the land is against customary norms. The chiefs therefore press home such assertions to a variable degree, depending on the level of the expected monetary gain. The higher the level of expected gain, the more firmly is ownership asserted by the chiefs. In general, such gains are highest close to the market town of Churhandpur, they decline with distance from the town and from the roadhead. As Table II shows, this is reflected in the chief's forcing all cultivators to buy plots of land close to the town, while in remote areas all that the cultivators pay is the traditional annual tribute of 16 kg of grain. Similarly close to the town community members pay a royalty to the chief even

TABLE II  
Resource Ownership, Sharing, and Utilization Patterns among Different Sections of Gangte Tribe at Different Levels of Accessibility and Modernization

	Shifting cultivation area		Settled agriculture area	Town area
	Less accessible	Easily accessible		
1. Ownership	Village chief	Village chief	Private	Private
2. Commercialization of forest and current forest condition	Little commercial exploitation and less degradation	Under heavy commercial exploitation and degraded condition	No forest	No forest
3. Onset of commercial forest exploitation	Still untouched, except for domestic use	Early 1970s	Early 1950s	Early 1930s
4. Current status of exploitation	Little exploited, good forest cover	Heavily exploited, degraded forest	Completely exhausted	Completely exhausted
5. Collection of timber	Free access	Pay royalty to the chief	No collection	No collection
6. Collection of fuelwood	Free access	Free access for domestic use	Pay royalty to the chief	Pay royalty to the chief
7. Maintenance of VFR <sup>1</sup> and Bamboo reserve	Still intact	No more	No more	No more
8. Traditional rights of resource sharing and use	Still followed	All have disappeared except for free use of shifting fields and fuelwood collection for domestic use	No more	No more
9. Collection from forest	Fuelwood, wild vegetables, timber and other NTFP <sup>2</sup>	Fuelwood, wild vegetables, timber, and other NTFP	Fuelwood collected from forests of other villages	Occasional collection of fuelwood from nearby forests
10. Land use	Shifting cultivation	Shifting cultivation, terraced cultivation	Wet rice cultivation, pineapple plantation	Habitation
11. Commercial crop plantation	Chili and other vegetables	Banana, chili and vegetables	Pineapple, ginger, and other vegetables	Habitation
12. Energy use	Human muscle power, fuel-wood, and little use of fossil fuel	Human muscle power, fuelwood, and little use of fossil fuel	Human muscle power, animal power, fuelwood, fossil fuel, and electricity	Fuelwood, electricity, fossil fuel, and less use of human and animal muscle power

<sup>1</sup> Village Forest Reserve (VFR).

<sup>2</sup> Nontimber Forest Produce (NTFP).

for collection of fuelwood, while further afield they have free access to even more valuable timber.

## V. LAND USE CHANGES

This process of over-exploitation of forest resources has also shaped patterns of use of land. Since the forest resources bring in highest levels of net profits closest to the market town of Churhandpur, all tree stocks have been exhausted in this area. In response, settled

cultivation on terraced fields with very low levels of crop diversity such as monocultures of pineapple has replaced the much more biologically diverse shifting cultivation. In more remote areas too, commercial monoculture plantations have been taken up of species like Agor (*Aquilaria aquatocha*), the highly valuable wood that was the first to be exhausted even from areas far from market.

The traditional land-use pattern of Gangtes included leaving aside substantial areas as gamkhals, or so-called village forest reserves (VFR), as well as bamboo reserves

and other spirit lands. These practices have been completely abolished in villages close to market towns where land and its produce have acquired high commercial value. In some of the more remote villages the practice of protection of a VFR encircling the settlement has been revived following recognition of its value as a firebreak preventing the spread of fire to the settlement during the slash and burn operations (Gadgil, Hemam, and Reddy, 1997). The traditions of protection of such forest patches still continue in other more remote areas.

Links to markets has meant access to many new goods such as soaps and transistor radios for the Gangtes; goods that need cash for purchase. Sale of timber, fuelwood, and other forest produce such as cane is their only source of cash. But generation of this cash has led to a reduction in the return of nutrients to the fields in slash and burn cycles with a depletion of the standing tree stocks. This has reduced the levels of productivity of shifting cultivation (Ramakrishnan, 1992). The spurt in population growth, favored by introduction of modern health care systems, especially control of malaria, with a rate as high as 4% in some of the north eastern hill areas has also increased the pressure on land and led to a shortening of fallow cycle and a reduction in productivity of shifting cultivation. The net result is that today Gangtes depend on market for about 30% of their food requirements and over 50% of other needs (Hemam, 1997).

## VI. VALUE APPROPRIATION

These cash requirements are met from harvests of forest produce, either through sale or through wages earned as laborers in tree-felling operations taken up by contractors. With the assertion of land ownership and demand for royalty by chiefs, part of the profits go to them. However, the chiefs are unfamiliar with the skills of development of infrastructure such as roads, organizing transport, and marketing of forest produce. As a result outside contractors carry out most of the timber harvesting operations, retaining a bulk of the profits for themselves. Thus in the early 1990s in Nalwan village the contractor paid Rs. 50,000 (US \$ 1600) for rights to harvest timber from an area of over 1000 ha. A conservative estimate of the value of timber thus accessed is Rs. 60 million given a standing stock of 100 tonnes/ha and price of timber at Rs. 600/tonne. The chief was then paid less than 1.0% of the timber value. Of course, the contractor would have to invest in labor charges, making roads, and transport. Nevertheless, it is clear that bulk of the cash income flows to the con-

tractor, a part to the chief, and a very small fraction to members of Gangte community. The Gangte use this cash to meet their routine requirements. None of it is saved or invested productively in ventures such as developing horticultural plantations to replace shifting cultivation fields. In the long run then the whole process of forest resource use taking place today is leading to exhaustion of these resources with little gains for the Gangte community members.

Today the Gangtes have greater access than they had earlier to goods on the markets and are in that sense richer, though this access is still extremely limited and by all criteria they are quite poor. However, at an earlier time when they were poorer in terms of market access, they had more secure and more equitable rights over land and substantial access to biodiversity and other natural resources; this has drastically reduced. The erosion of biodiversity this transformation has accompanied has in part been driven by the desire of Gangtes to acquire cash; however, the real driving force behind the loss of biodiversity lies in the demand for forest produce by the biosphere people. Much of the wood of Manipur has been used for manufacture of plywood, and all the value added in this process has been appropriated by those already rather well off, controlling trade and industry of mainland India. The transformation has also witnessed a loss of the traditions of conservation of biodiversity, such as those of protection of sacred groves, by relatively poor people like Gangtes.

Gangtes thus provide an excellent case study of the different stages and processes involved in the transformation of autonomous societies of ecosystem people with access to high levels of biodiversity, but with little access to manufactured artifacts into societies that have lost control over their base of living resources, often turning into instruments of depletion of these resources. Thus subjugated, the ecosystem people may end up with some, though often insignificant, enhancement in their access to artifacts, but with significant losses in their access to biodiversity.

## VII. ECOLOGICAL REFUGEES

Gangtes have come under the influence of the state and the market economy relatively recently. True, most of them have lost all legal claims over land or forest resources, and the claims of Gangte chiefs over large tracts of land are being contested by the Forest Department; nevertheless, in fact, Gangtes still control extensive tracts of forested land. This level of autonomy is unusual on the Indian subcontinent where state authority has

been holding sway for several centuries. One such unusual area is that of Maharashtra Western Ghats, where the local peasants had retained a considerable measure of autonomy by virtue of their active role in the Maratha armies of seventeenth and eighteenth centuries. The British were reluctant to offend them on the defeat of Marathas in early nineteenth century. As a consequence they were permitted to continue their shifting cultivation, which was banned over much of the remainder of the subcontinent. Individual peasants, rather than any chiefs were assigned land ownership among the Marathas. So while poor in modern economic terms because their hilly terrain permitted only very restricted access to the markets, these peasants lived in tracts rich in forests and biodiversity with full rights over the resources till the time of independence. The tract was also rich in water resources with annual precipitation in excess of 3000 mm.

On independence the Indian nation state geared itself to tap these rich water resources to promote economic growth with the help of irrigation and hydroelectric projects. One such irrigation project was launched in the Panshet valley to the west of Pune, the old capital of Marathas (18°30' N lat.–73°40' E long.). The Indian state has assumed extensive powers of acquisition of land for public purposes and all the more fertile low lying paddy land of the Panshet valley was taken over at low rates of compensation. The peasants still retained ownership over their hill lands above the submersion level. This had been traditionally under long fallow shifting cultivation during which mango and *Terminalia chebula* trees, both producing fruit of considerable value, were left intact. However, as roads reached Panshet valley to facilitate the dam construction work, these trees caught attention of charcoal merchants. These merchants worked in collusion with the government officials in charge of land acquisition to persuade the peasants to sell the trees at exceedingly low prices with the assurance that the peasants would soon be resettled with alternative farmland far from Panshet valley. Most peasants sold the trees, and there was a massive spurt of deforestation in 1950s during the construction of the dam, the trees being primarily used to produce wood charcoal for consumption in Pune. The peasants, however, were never provided alternative agricultural land elsewhere and have continued staying on hill slopes above the dam for the past 40 years, while some of them have migrated to Pune, primarily to work as unskilled labor. The entire catchment has been devastated by the initial deforestation in 1950s followed by continual, ever more intensive cultivation of hill slopes by the impoverished peasants. Today the peasants are

poor in every sense and live in an environment depleted of biodiversity. They have served as agents of this destruction of biodiversity that has only benefited the charcoal merchants of Pune and provided citizens of Pune with a relatively inexpensive fuel for a period of few years. Thus subjugated, these ecosystem people may be said to have been turned into ecological refugees (Gadgil, 1979).

Our next case study is that of the tribal populations of the so-called Jharkhand tracts of the states of Bihar and West Bengal (22°0'–24°0' N lat. 84°0'–87°40' E long.). This region adjoins the old, thickly settled Gangetic plains that were the nucleus of the 2000 year old Magadha empire extending over much of the Indian subcontinent. Emperor Ashoka of Magadha is famous for his conquest of and subjugation of the tribals in the extensive hilly tracts that border the Gangetic plains (Gadgil and Guha, 1992). This part of the country also fell rather early to British colonialists, who quickly established a hold over the region during the second half of the eighteenth century, far more firm than that over the Maratha Kingdom discussed earlier. The forest of Jharkhand are rich in sal (*Shorea robusta*), an important timber for the production of railway sleepers. This was a resource much valued by the British who therefore acted firmly to ban all shifting cultivation by the tribal people in early nineteenth century. The forests were also taken over either as a property of the British government or that of private landlords. Where the tribal villages were a part of the government-controlled forest lands, the tribals were totally deprived of any rights to forest resources. Unlike with Gangtes, the landlords too did not come from among the tribals themselves. They came from other castes at a higher level in the hierarchy. These landlords then exploited the lower caste or tribal tenants on their lands quite ruthlessly. The "Jharkhand" tribals were therefore subjected to extreme political as well as economic subjugation as early as the first half of the nineteenth century. Their only earnings came from work as poorly paid tenant farmers or as forest laborers along with sale of some minor forest produce at extremely low rates. They then have a long history as poor people serving as agents of destruction of biodiversity.

These Jharkhand tribals are among the most striking example of ecological refugees of the country. They have served as the most mobile source of very inexpensive, unskilled labor in many contexts. One of the major economic enterprises under the British rule was the development of tea estates replacing the rain forest of Brahmaputra valley in northeastern India. This involved wholesale destruction of biodiversity. While the planta-

tion owners were British, the laborers responsible for actual destruction of biodiversity were mostly Jharkhand tribals working under conditions that have been described as being close to slavery (Gadgil, 1942).

Elsewhere, these tribals were resettled in so-called forest villages. The primary focus of the plantations they worked on was to replace the natural sal-dominated, fairly diverse humid forest with monocultures of teak (*Tectona grandis*). Teak is excellent timber resistant to termite and fungal attacks, highly valued for ship building and gun carriages in nineteenth and for furniture and house construction in twentieth century. But teak is hard wood, little used traditionally; it also does not provide any other product of local utility. On the other hand, sal leaves are used to make plates and sal seed has value as an oil seed. Many associates of sal, such as mahua (*Madhuca indica*) and tendu (*Diospyros melanoxylon*) have great local utility. So replacement of natural sal forest by teak plantations deprives local people of access to a diversity of biological resources of value. Uprooted impoverished tribals have served over years as primary agents of such an erosion of biodiversity (Gadgil and Guha, 1995).

Transformations such as these are particularly common today in tropical developing countries, transformations that are actively encouraged by nation states as well as transnational development agencies as development efforts. These development efforts then have adverse impacts both on biodiversity and on the ability of the weaker sections of the population to access biodiversity. Indeed, they also have adverse impacts on the biodiversity conservation practices of the ecosystem people, practices such as protection to sacred plants and animals and sacred sites such as sacred groves and ponds.

## VIII. COSTS OF CONSERVATION

Conservation practices of the ecosystem people with their limited resource catchments are organized on limited spatial scales. Thus, the sacred groves have sizes ranging from a fraction of a hectare, to tens, at best hundreds of hectares. The biosphere people have their own conservation practices. In keeping with their larger resource catchments, they tend to protect much larger areas as nature reserves (Gadgil, 1996). Agrarian states had their tradition of hunting preserves of aristocracy, ranging in size from few hundreds to thousands of hectares. Such hunting preserves of nobles are described in Kautilya's Arthashastra, the 2000-year-old Indian manual of statescraft (Kangle, 1969). They are also

known from medieval Asia and Europe. The hunting preserves were aimed at protecting major game animals such as lions or antelopes and their habitat for the entertainment of aristocracy. Their main focus was therefore on exclusion of access, especially to hunting in these areas, by all the commoners. In this, they would have tended to accentuate the poverty of people living within and adjacent to the hunting preserves.

The modern-day nature reserves are even larger in spatial scale, in keeping with the even larger resource catchments of modern-day societies of biosphere people. Thus the largest national parks of the day extend over hundreds of thousands of hectares. Many of them, as in India, have their roots in the hunting preserves of the aristocracy. Two notable examples of these are the national parks at Gir, the last stronghold of the lion in India, and at Keoladev, Ghana, a wetland that shelters enormous nesting colonies of herons, storks, pelicans, and cormorants during monsoons and huge populations of migrating waterfowl in the winter. The focus at both these reserves is on exclusion of resource access by local peasants and herders, leading to major controversies.

At Gir (21°40' N lat. 70°30' E long.), the controversy centers on relocation of traditional colonies of buffalo keepers, the Maldharis outside the national park. Their free-ranging buffaloes were earlier an important prey for the lions, especially the males. On being located outside the quality of grazing for the buffaloes has declined, and despite some reduction in the mortality of buffaloes from lions, the Maldharis are worse off. Within the national park the herbivore populations have increased with reduction of competition for grazing by buffaloes, but the lion population has not gone up. In fact, male lions continue to hunt Maldhari buffaloes who are now outside the limits of the national park. While Maldharis were earlier willing to accept some buffalo kills by lions as a compensation for access to grazing within the national park, they are no longer willing to do so. So they poison carcasses of buffaloes killed by lions; as a consequence the lions suffer (Johnsingh, *pers. comm.*)

In Keoledev national park (27°15' N lat. -77°35' E long.), famous for its water birds too, modern conservation attempts have involved exclusion of grazing in the extensive shallow wetlands. When grazing was thus banned in early 1980s without any alternatives being made available, there were protests leading to police firing, with some deaths. But the ban was implemented forcing local peasants to substantially reduce their livestock holdings, impoverishing them to a significant extent. But this cessation of grazing has also had an ad-

verse impact on wetlands as a water bird habitat. This is because the wetlands have now become choked by an excessive growth of a grass, *Paspalum*. This particular conservation measure has thus tended to impoverish people as well as the bird life, which was at the center of the conservation concern (Vijayan, 1987).

Jharkhand tribals too have suffered from being excluded in the interest of nature conservation. One of the major recreation areas for the wealthy from Calcutta city is the Betla National Park (23°40' N lat.–84°40' E long.) in Daltonganj district of Bihar. The tribals who were earlier settled in many forest villages in what now constitutes this national park are facing serious difficulties. With the constitution of the national park, the forestry operations employing them have been halted. Further restrictions are imposed on their collection of forest produce. Moreover, the tribals are now being asked to move out of the forest villages altogether, without any provisions for alternative livelihoods. So the tribals who were earlier instruments of destruction of biodiversity to help meet economic demands of the wealthy are now being further impoverished to meet their recreational demands for access to biodiversity (Gadgil, 1998).

## IX. ECODEVELOPMENT

But there are other positive developments as well. The difficulties experienced by Jharkhand tribals have triggered a variety of political protests, some non-violent, others violent. Among their demands is the creation of a separate Jharkhand state; since they see states of Bihar and West Bengal as being dominated by interests responsible for the political and economic subjugation of tribals. One of the slogans of this Jharkhand movement has an ecological content: Teak is Bihar, sal is Jharkhand. Their demand is for the retention of natural sal-dominated forest.

A notable development accompanying these protests is the program of participatory management of forest resources. Under this program villagers are encouraged to organize village forest committees (VFC) to promote natural regeneration of forests. The VFCs are given certain authority to control forests, an authority that was earlier monopolized by the Forest Department, and are expected to assume responsibility for forest protection. In return they are given full rights over non-timber forest produce such as sal leaves and oil seeds, as well as a share of any timber that may be harvested. This program, which was initiated in 1973 in the tribal villages of Midnapore district (22°20' N lat.–87°20' E

long.) of West Bengal, has now spread over many other parts of this tribal belt, as well as elsewhere in the country. The program has by and large been a success. It is an important instance of the poor getting organized to promote ecological restoration, and in the process enhancing their own standard of living. Such ecological restoration also tends to enhance levels of biodiversity in conjunction with production of a variety of forest produce of value to local people (Poffenberger and McGean, 1996).

The Jharkhand political struggle also incorporates an issue with significant implications for biodiversity conservation. The tribal religious practices throughout the country included protection of sacred plants, animals, ponds, and forests, as was mentioned earlier in the Gange case study. The system of sacred groves, called "saranas," has been an important ingredient of the traditional tribal religion of Jharkhand. These saranas are variable in size—mostly small, about 0.1 ha—but very numerous and serve to protect samples of original natural plant life. These were treated with contempt by Christian missionaries who have a significant influence in this tribal region, as also by proponents of high-caste Hinduism. Now protection of these saranas has emerged as an element of the move to restore tribal self-respect and to empower them (Gokhale *et al.*, 1998). Indirectly then, conservation of indigenous biodiversity is beginning to emerge as a part of the agenda of the poor to pull themselves out of the poverty trap.

UNESCO's Biosphere Reserve program is another endeavor to combine efforts at conservation of biodiversity with alleviation of poverty. However, it has met with limited success largely because of inadequate understanding of local ecology and people's needs. Thus, Nandadevi (–30°30'–30°40' N lat. and 79°44'–79°58' E long.) in Western Himalayas is one of India's eight Biosphere Reserves. In this region, the traditional livelihoods depended on summer grazing in these pastures supplemented by collection of medicinal herbs. This grazing has been banned in the core zone of the biosphere reserve, along with mountaineering expeditions, another important source of cash incomes. As a result, local people feel significantly impoverished. The few ecodevelopment programs such as fuelwood plantations that have been brought in as a part of the Biosphere Reserve activities have failed to create sufficient levels of additional incomes. At the same time local people believe that in the absence of grazing the diversity of medicinal herbs in the pastures are now being replaced by extensive growth of a few species such as *Rumex* (Negi, 1999).

The philosophy behind biosphere reserves, that of combining conservation with development efforts, also motivates a number of development programs that go by the name of "Integrated Conservation and Development Programs" (ICDP) and the many projects funded by Global Environment Facility (GEF). However, there has been little genuine progress so far in actually realizing these goals. As a review of a GEF program in New Guinea documents, the programs remain ineffective as they continue to be designed by outsiders committed to conservation but with little understanding of what would motivate local people, especially the poor, to participate in conservation efforts (Global Environment Facility, 1998).

There are also attempts to bring to local people financial benefits either through ecotourism or through sustainable harvesting of biodiversity resources. There are limited success stories of combining the two in parts of Zimbabwe where substantial incomes are generated though high-priced hunting licenses for large mammalian game animals, with levels of such harvests kept well within sustainable limits (McNeely, 1995). These incomes are shared with local, relatively poor people. Other attempts at combining poverty alleviation with biodiversity conservation have focused on promoting enterprises such as processing of medicinal plants or wild fruit. Programs in India involving processing of fruit of *Phyllanthus emblica*, a moist deciduous forest tree species by local Solliga tribals in B. R. hills of Karnataka or processing and marketing of wild mango by local women's groups in the Kangra district of Himachal Pradesh have met with varying degrees of success (Chopra, 1998). Providing alternative means of livelihood to communities depending on a threatened biological resource has also been attempted with success in the Indian ocean island of Seychelles where turtle shell artisans were provided financial and technical support to shift to occupations like coconut souvenir carving (Global Environment Facility, 1997).

## X. THE CHALLENGE AHEAD

Poverty thus relates to biodiversity in many, complex ways. But, by and large, the poor are today instruments of decimation of biodiversity in their attempts to earn livelihoods. There has been little success in addressing this problem though designing programs that would combine conservation and sustainable utilization of biodiversity with alleviation of poverty. This then is a major challenge that must be addressed in years to come.

## See Also the Following Articles

BIODIVERSITY AS A COMMODITY • BIODIVERSITY-RICH COUNTRIES • DEFORESTATION AND LAND CLEARING • ECONOMIC GROWTH AND THE ENVIRONMENT • ECONOMIC VALUE OF BIODIVERSITY, OVERVIEW • ETHNOBIOLOGY AND ETHNOECOLOGY • INDIGENOUS PEOPLES, BIODIVERSITY AND • LAND-USE ISSUES • MARKET ECONOMY AND BIODIVERSITY • SOCIAL AND CULTURAL FACTORS

## Bibliography

- Bhattacharya, S. (1993). *Ecological Organization of Indian Rural Populations*. Ph.D. thesis. Indian Institute of Science, Bangalore.
- Dasmann, R. F. (1988). Towards a biosphere consciousness. In *The Ends of the Earth: Perspectives on Modern Environmental History*. (D. W. Worster, Ed.), pp. 277–288. Cambridge University Press, Cambridge.
- Chopra, K. (1998). Economic aspects of biodiversity conservation. *Economic and Political Weekly* XXXIII(52), 3336–3340.
- Gadgil, D. R. (1942). *Industrial Evolution of India*. Oxford university Press, New Delhi.
- Gadgil, M. (1979). Hills, dams and forests: Some field observations from Karnataka Western Ghats. *Proceedings of the Indian Academy of Sciences* 2, 291–303.
- Gadgil, M. (1995). Prudence and profligacy: A human ecological perspective. In *The Economics and Ecology of Biodiversity Decline* (T. M. Swanson, Ed.). Cambridge University Press, Cambridge.
- Gadgil, M. (1996). Managing biodiversity. In *Biodiversity: A Biology of Numbers and Difference* (K. J. Gaston, Ed.), pp. 345–365. Blackwell Science, Oxford.
- Gadgil, M. (1998). Conservation: Where are the people? *The Hindu Survey of the Environment*, 98, 102–137.
- Gadgil, M. and Guha, R. (1992). *This Fissured Land: An Ecological History of India*. Oxford University Press, New Delhi and University of California Press, Berkeley, CA.
- Gadgil, M., and Guha, R. (1995). *Ecology and Equity: Use and Abuse of Nature in Contemporary India*. Routledge, London.
- Gadgil, M., Hemam, N. S., and Reddy, B. M. (1997). People, refugia and resilience. In *Linking Social and Ecological System* (C. Folke and F. Berkes, Eds.), pp. 30–47. Cambridge University Press, Cambridge.
- Global Environment Facility. (1997). *GEF Project Implementation Review*. GEF, Washington, D.C.
- Global Environment Facility. (1998). *Study of GEF Project Lessons*. GEF, Washington, D.C.
- Gokhale, Y., Velankar, R., Chandran, M. D. S., and Gadgil, M. (1998). Sacred woods, grasslands and water bodies as self-organized systems of conservation. In *Conserving the Sacred for Biodiversity Management* (P. S. Ramakrishnan, K. G. Saxena, and U. M. Chandrashekhara, Eds.), pp. 366–396. Oxford and IBH Publishing, New Delhi.
- Hemam, N. S. (1997). *The Changing Patterns of Resource Use and Its Bio-social Implications: An Ecological Study amongst the Ganges of Manipur*. Ph.D thesis. Calcutta University, Calcutta.
- Kangle, R. P. (1969). *Arthasasthra*. University of Bombay, Bombay.
- McNeely, J. A. (1995). Economic incentives for conserving biodiversity: Lessons from Africa. In *Conservation of Biodiversity in Africa*.



- Local Initiatives and Institutional Roles* (L. A. Bennum, R. A. Aman, and S. A. Crafter, Eds.), pp. 199–215. National Museums of Kenya, Nairobi.
- Negi, N. S. (1999). *Co-variation in Diversity and Conservation Value Across Taxa: A Case Study from Garhwal Himalaya*. Ph.D thesis, Indian Institute of Science, Bangalore.
- Pemberton, R. B. (1835). *Report on the Eastern Frontier of India*, Calcutta.
- Poffenberger, M., and McGean, B. (Eds.) (1996). *Village Voices, Forest Choices: Joint Forest Management in India*. Oxford University Press, Delhi.
- Ramakrishnan, P. S. (1992). Shifting agriculture and sustainable development: An interdisciplinary study from North-Eastern India. In *Man and Biosphere Series* (J. N. R. Jeffers, Ed.), Vol. 10. The Parthenon Publishing Group, Paris.
- Saldanha, C. J. (1989). *Andaman, Nicobar and Lakshadweep: An Environmental Impact Assessment*. Oxford and IBH Publishing, New Delhi.
- Superintendent of Government Printing, Calcutta. (1909). *Andaman and Nicobar Islands. Imperial Gazetteer of India*, Provincial Series.
- Vijayan, V. S. (1987). *Keoladeo National Park Ecology Study*. Bombay Natural History Society, Annual Report. Bombay.



# PREDATORS, ECOLOGICAL ROLE OF

James Estes,\* Kevin Crooks,<sup>†</sup> and Robert Holt<sup>‡</sup>

\*University of California Santa Cruz, <sup>†</sup>University of California San Diego, and

<sup>‡</sup>The University of Kansas

- 
- I. Introduction
  - II. Background and Definitions
  - III. Case Studies
  - IV. Summary and Synthesis of Case Studies
  - V. Theoretical Studies of Predation and Biodiversity
  - VI. Generality
  - VII. The Future
- 

## GLOSSARY

**apex predator** An organism that occupies a food web's highest trophic level.

**bottom-up forces** Population-regulating processes based on the availability of food, nutrients, and energy.

**carnivore** An organism that consumes other animals.

**competitionism** The view that competition regulates populations.

**food web** The interconnections among organisms based on diet.

**herbivores** Organisms that feed on plants and other photosynthesizers.

**keystone species** A strong interactor that is relatively rare.

**Lotka–Volterra model** An early equation relating rate of population change to the interplay between competition and predation.

**mesopredator** A small to mid-sized predator.

**nutritionalism** The view that bottom-up forces regulate populations.

**phytoplankton** Microscopic primary producers that live in water column habitats.

**piscivores** Predators that consume fish in aquatic habitats.

**planktivores** Predators that consume zooplankton in aquatic habitats.

**top-down forces** Population-regulating processes that originate from consumer limitation.

**trophic cascades** A chain reaction of top-down interactions across multiple trophic levels.

---

**PREDATORS** occur in all of the planet's ecosystems and initiate top-down forces and trophic cascades in many of these. Although evidence for trophic cascades is strongest for aquatic systems, there is increasing evidence that they occur in a variety of terrestrial ecosystems as well. Trophic cascades result in weak or strong plant–herbivore interactions when the respective number of trophic levels is odd or even. Indirect food web effects of trophic cascades, known for lakes and kelp forests, are unstudied in most systems. The loss of large, apex predators from several terrestrial systems has resulted in mesopredator release—the proliferation of moderate-sized predators that commonly reduce or eliminate the smaller vertebrate species. Many dysfunc-

tional ecosystems have developed because of the loss of apex predators. Thus, reserve design and other conservation strategies must be adequate to preserve the apex predators.

## I. INTRODUCTION

The science of ecology has undergone a succession of paradigms on the nature and importance of species interactions, including those between predators and their prey. The earliest view (henceforth termed nutritionalism) was that bottom-up forces (i.e., primary production and the efficiency of energy and material transport upward across trophic levels) regulate populations. Ecosystem ecology was built around this view of nature, which implicitly holds that apex predators, as the end points of energy and material flux, are of minor consequence to ecosystem function. Beginning in the late 1950s and early 1960s, the focus on species interactions changed to competition. In contrast with nutritionalism, competitionism holds that lateral forces within trophic levels regulate population abundance. By this view, predators are no more or less important than any other species. Recently, top-down forces have captured the attention of ecology, thereby legitimizing predators as important ecological entities. Ecologists now recognize that important species interactions follow all three pathways (Fig. 1), often simultaneously and sometimes

interactively. Thus, although our focus in this article is on top-down forces generated by apex predators, understanding the ways in which predators influence biodiversity requires a more eclectic view of food webs and species interactions than simply “bottom-up vs top-down.”

We begin with a discussion of who the predators are and how they affect populations, communities, and ecosystems. We then present a series of case studies demonstrating the wide range of systems in which predation is an important organizing process, including examples of the unifying concepts and explanations of how they were discovered. This discussion is followed by a theoretical exploration of predation and biodiversity. Next, we discuss the levels of biological organization at which predation can influence biodiversity and develop a conceptual model for how apex predators might influence the location and strength of bottom-up and competitive forces in systems under top-down control. We conclude by considering the needs and opportunities for further research on predators and how predators are likely to figure into the future of conservation biology.

## II. BACKGROUND AND DEFINITIONS

### A. What Are Predators?

Broadly defined, all consumers are predators, thus including all living things except photo- and chemosynthesizers. The carnivorous plants add a minor wrinkle to this dichotomy. Nonetheless, predators would be represented by such diverse functional groups as herbivores, parasites (including microbes and parasitoids), and the immense diversity of invertebrate and vertebrate consumers that hunt and kill their prey. Attempts to define or classify predators based on trophic status, consumer-prey size relationships, or just about any other scheme are similarly problematic. For example, herbivores exist at one end of the trophic-status spectrum and microbes at the other, but it is uncommon for consumers in either group to kill their prey outright. Plants may even benefit from being eaten and a single act of predation by most microbes is of virtually no consequence to their prey because of the prey's immensely greater relative body size. Although herbivores, microbes, and parasites are predators in this broad sense, our focus is on those that kill their prey. Even this restricted definition includes a vast array of species.

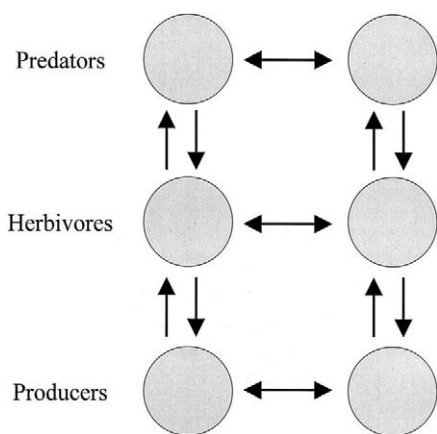


FIGURE 1 A simple stylized food web showing potential interaction pathways. The circles represent species, the downward-pointing arrows represent top-down forces, the upward-pointing arrows represent bottom-up forces, and the double-headed arrows represent competitive interactions. Although real food webs are far more complex, this figure shows the three main ways by which species interact with one another in nature.

## B. Species Interactions

The influence of predators on biodiversity depends first and foremost on direct predator–prey interactions. However, the consequence of predation to communities and ecosystems has less to do with these direct interactions than it does with their indirect effects. Although the direct effects of predation, by definition, are limited to impacts on prey populations, the nature of indirect effects is almost limitless, thus potentially causing populations to increase or decline anywhere in the food web. As will be shown in several of the case studies presented later, such indirect effects may involve long interaction chains, which in turn have broad impacts on associated ecosystems.

## C. Indirect Effects of Predators

An awareness of the indirect effects of predators can be traced back at least to the writings of Charles Darwin, who described an interaction chain leading from cats to mice to bumble bees to clover. Similar early examples were provided by such well-known ecologists as Charles Elton and G. E. Hutchinson. Hairston, Smith, and Slobodkin's (1960) now-classic paper (hereafter HSS) was perhaps the first effort to mold the indirect effects of predation into a conceptual model of trophic interactions and population regulation. HSS recognized four trophic groups—producers, decomposers, herbivores, and predators—and argued that although herbivores are commonly limited by predators, plants, decomposers, and predators are ordinarily limited by resources. The HSS model has weathered the test of time, along the way setting the stage for several conceptual advances, including the importance of top-down forces in population regulation and community organization, the ideas of keystone species and trophic cascades, and a generalized theory of food chain dynamics. Each of these is briefly explained in the following sections.

## D. Top-Down Forces

Bottom-up forces are those passing from producers to consumers, whereas top-down forces are those passing from consumers to producers. As previously mentioned, recognition of top-down regulation dates back to at least Darwin, although it was HSS that introduced the idea in ecology. A Special Features section in the journal *Ecology*, published in 1992, stimulated further interest in the issue, in part by pointing out that top-

down and bottom-up forces need not be competing processes, even though bottom-up forces are necessary for the function of all ecosystems. This realization freed ecologists to imagine a broad potential for the role of predation in nature.

## E. Keystone Species

HSS was followed in the mid-1960s by Robert Paine's highly influential paper on food web complexity and species diversity. Paine argued that predators often selectively consume and thus limit competitively dominant species, thus enhancing species diversity by releasing their subordinates from competitive exclusion. This argument was based on three essential premises: (i) Predators selectively consume the competitively dominant prey; (ii) in so doing, populations of the competitively dominant species are reduced; and (iii) in the absence of predation, the prey guild is limited by interspecific competition. Paine's work captured the interest of community ecologists because it linked the influence of predators to species diversity, and (perhaps most important at the time) it was supported by results from field experiments. His empirical studies of predation by sea stars on mussel bed assemblages were done in the temperate rocky intertidal zone where competition for space can be extreme. This also led to two important developments in ecology: the idea of keystone species and the intermediate disturbance model of species diversity.

The intermediate disturbance model, further refined and generalized by Joseph Connell, holds that species diversity is influenced by the intensity of disturbance (either physical or biological; Fig. 2). When the intensity of disturbance is very high or very low, species diversity is low because the most vulnerable species are eliminated in the former instance and excluded by their competitive dominants in the latter. These limiting conditions are relaxed at intermediate levels of disturbance, thereby elevating species diversity. The notion of keystone species, as envisioned by Paine, applied to cases in which predators were the agents of disturbance. Although the definition of keystone species has broadened, on the one hand, to include other kinds of interactions and grown more restrictive, on the other hand, to exclude the effects of common species, this idea is rooted historically with the indirect effects of predators.

## F. Trophic Cascades

A trophic cascade is the progression of indirect effects by predators across successively lower trophic levels.

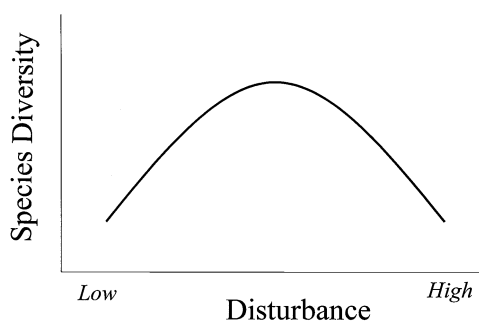


FIGURE 2 The intermediate disturbance model of species diversity. This model is based on the assumption that competitive exclusion occurs in benign systems. Thus, when the intensity of disturbance (including predation) is low, strong competitive interactions by the dominant species reduce species diversity. When the intensity of disturbance is high, species diversity is again low because those species that cannot cope are eliminated. Maximum species diversity occurs at intermediate levels of disturbance—strong enough to prevent competitive exclusion but not so strong as to directly eliminate species.

HSS's proposed relationship between predators, herbivores, and producers was a generalized trophic cascade. Stephen Carpenter and James Kitchell popularized this idea based on the striking influences of predatory fishes on the essential components of lake food chains—from minnows (the predatory fishes' prey) to zooplankton (prey of the minnows) and to phytoplankton (prey of the zooplankton).

### G. Generalized Food Web Theory

A generalized food web theory was developed by Stephen Fretwell to show how predation, trophic cascades, and food chain length combine to predict the strength of plant–herbivore interactions (Fig. 3). To understand this theory, first imagine an ecosystem with producers but no consumers. Lacking consumers, the producers are limited by competition for resources. Adding herbivores creates a two-trophic level system in which the plant populations become limited by herbivory. Adding predators limits herbivore populations, thus releasing the producers from limitation by herbivory and returning them to limitation by resource competition. The progressive increase of trophic complexity cascades downward through the food chain such that plant–herbivore interactions switch from being weak to strong as the respective number of trophic levels alternates between odd and even.

Next, we summarize many case studies that provide empirical evidence for these theories and concepts. Our

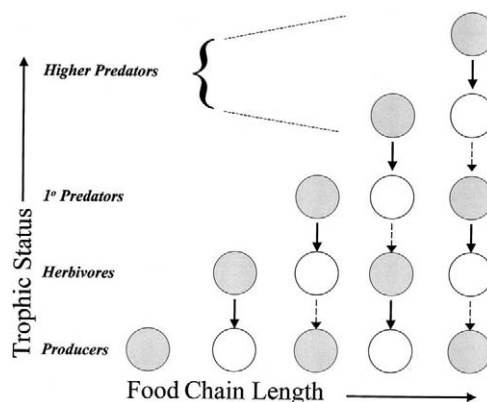


FIGURE 3 A graphical synopsis of Stephen Fretwell's theory of food chain length in systems under top-down control. The circles represent species or groups of species within particular trophic levels: ●, resource limitation; ○, consumer limitation. The solid arrows represent strong interactions, and the dashed arrows represent weak interactions. As food chain length becomes progressively longer, the plant–herbivore interactions alternate between being weak in odd-numbered systems and strong in even-numbered systems.

presentation is organized around the various systems in which the work was done.

## III. CASE STUDIES

### A. Rocky Shores

Studies of rocky seashores furnish the earliest and some of the most compelling evidence for the effects of predation on communities and ecosystems. The first well-known experimental studies were done by Joseph Connell in Scotland. Connell's work focused mainly on competition between the two barnacles (*Cthamalus stellatus* and *Balanus balanoides*) and predation on these species by the whelk (*Thais lapillus*). This research showed that the upper shore limit of *Cthamalus* was set by physical factors (weather) and the lower limit by competition for space with *Balanus* and predation by *Thais*.

Shortly thereafter, Paine began his studies of predation by the sea star (*Pisaster ochraceus*) in mussel beds along the outer coast of Washington. Paine hypothesized that sea star predation limited the lower distribution of mussels in the mid-littoral zone. This was subsequently confirmed by downward expansion of the mussel bed when the stars were removed. California mussels (*Mytilus californianus*) are the competitive dominants in this system. Predation by sea stars prevents mussels from controlling primary space (the rock

surface), the principal limiting resource in this system, thus permitting other species to coexist within mussel beds. In the absence of predation by sea stars, mussels dominate space, thereby excluding the competitive subordinates and reducing species diversity. Subsequent research has confirmed a similar role for *Pisaster* elsewhere in western North America and for other species of mussels and sea stars elsewhere in the world.

Other predators also influence rocky intertidal communities. Work by Philip Hockey and colleagues demonstrated a trophic cascade among African black oystercatchers (*Haematopus moquini*), herbivorous limpets, and intertidal algae. Succeeding studies of oystercatchers and limpets have confirmed similar interactions in South America, Australia, and western North America. Research in central and southern California further demonstrated how humans perturb the trophic cascade by exploiting owl limpets (*Lottia gigantea*, a large, territorial species) and by causing oystercatchers to abandon their breeding territories. The former effect causes a competitively subordinate guild of small limpets to replace owl limpets as the principal herbivore. The latter effect, induced simply by large numbers of humans being present along rocky shores, transforms the intertidal community from a three- to two-trophic level system. Small limpets come to dominate such areas, in turn reducing the algal cover. These human-caused perturbations probably are responsible for much of the modern-day character of rocky shores in central and southern California.

A final example of predation on rocky shores concerns the loco (*Concholepas concholepas*), a large muricid gastropod that consumes intertidal mussels and is exploited by humans in central and southern Chile. Juan Carlos Castilla excluded humans from a small stretch of shoreline at the Las Cruces Marine Laboratory near Santiago in order to better understand their influence on this system, and as expected loco abundance greatly increased. The more surprising result was a whole scale shift in the intertidal landscape, from one dominated by extensive mussel beds to one largely devoid of mussels. This particular example is noteworthy because it demonstrates (i) how humans can perturb predator-mediated interaction chains with landscape-level consequences, (ii) that reserves can be used effectively both to demonstrate and to mitigate such effects, and (iii) the power of experimental evidence.

## B. Kelp Forests

Kelp forest communities provide several examples of the ecological role of predators. One is that of the sea

otter, which was hunted to near extinction in the Pacific maritime fur trade. Following protection in the early 1900s, the process of recovery created a fragmented population distribution within what had been a continuously occupied range. Contrasts between areas with and without sea otters revealed striking differences in kelp forest communities. Areas with sea otters supported lush kelp forests, whereas those without otters were extensively overgrazed by sea urchins, the otter's principal prey. These patterns result from a trophic cascade, driven by sea otter predation on sea urchins, thus releasing kelp beds from sea urchin grazing.

In addition to contributing early empirical support for HSS, the sea otter–kelp forest system provides evidence for a wide range of predator-driven effects beyond those expected from simple trophic cascades. Sea otters influence numerous species by enhancing kelp abundance, thereby providing three-dimensional habitat and fueling increased primary production. This process is especially noteworthy because it shows how bottom-up processes can be altered by the top-down forces of apex predators. Other known or suspected consequences of sea otter predation in kelp forests are summarized in Fig. 4.

Understanding of the sea otter–kelp forest system has several interesting historical dimensions. Faunal remains in Aleut kitchen middens show that sea urchin size distributions during most of Aleut prehistory were similar to those of modern systems lacking sea otters, thus suggesting that aboriginal humans, by limiting sea otters, influenced coastal ecosystems long before modern humans arrived on the scene. Paleontological and biogeographical data provide an even longer time perspective. Because the distribution of sea otters and their recent ancestors was limited to the North Pacific basin, their influence on the evolution of plant–herbivore interactions has been inferred by contrasting plant defense and herbivore resistance between North Pacific and Australasian kelp forests. Australasian kelp forests apparently lacked predators of comparable influence to the sea otter, at least since the Pliocene–Pleistocene. In contrast with North Pacific kelp forests, strong coevolutionary forces between marine plants and their herbivores in Australasia were thus expected, thereby facilitating an arms race between plant defense and herbivore resistance. Marine algae use secondary chemicals as their principal defenses against herbivory, and for this reason the evolutionary hypothesis was put to an initial test by measuring the secondary chemistry of North Pacific and Australasian seaweeds. Phlorotannins (the principal chemical defenses in brown seaweeds) concentrations were approximately an order of

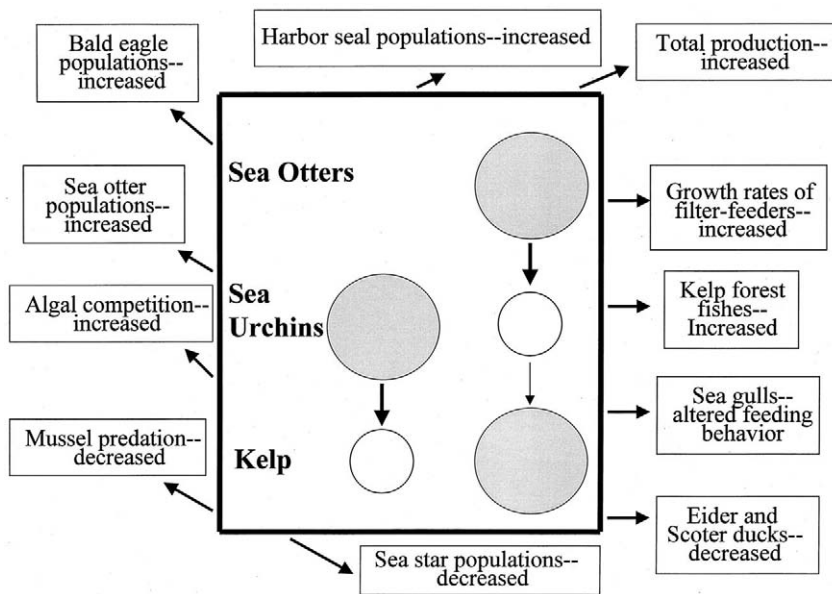


FIGURE 4 A conceptual representation of the direct and indirect effects of trophic cascades in the sea otter-kelp forest system. See the legend to Fig. 3 for explanations of circles and arrows. Some of the known or suspected indirect effects of these two alternate states of kelp forest community organization are shown around the periphery of the central box. See Estes (1996) for further discussion of specific cases (reproduced with permission from Peterson and Estes, 2000).

magnitude greater in Australasian algae and North Pacific herbivores were more strongly deterred by these compounds than were their Southern Hemisphere counterparts. These evolutionary responses to predation probably explain why Northern Hemisphere kelp forests have been so devastated by sea urchin grazing following decimation of their predators.

The sea otter-kelp forest system changed remarkably in recent years as killer whales entered the coastal ecosystem and began preying intensively on sea otters after their normal prey populations (seals and sea lions) declined. Since the early 1990s, killer whale predation has driven otter numbers downward by approximately an order of magnitude across large areas of western Alaska. The consequent reduction in sea otter predation has caused sea urchin numbers to increase and kelps to decline (Fig. 5). This example illustrates that predator-prey interactions, acting through trophic cascades, influence herbivore-plant interactions in a manner consistent with the predictions described earlier for oddvs even-numbered food chains (Fig. 4). It further indicates a role for predators in linking ecosystems over large areas.

Sea urchins have deforested kelp beds in the Gulf of Maine. Early reports attributed this to the overfishing of American lobsters (*Homarus americanus*), a pur-

ported ecological analog of the sea otter. However, this explanation is in doubt because the lobster fishery is apparently at an all-time high. Atlantic cod (*Gadus morhua*) also prey on a variety of benthic species, including urchins and lobsters, and the well-known collapse of cod populations may have influenced both lobsters and kelp forests in the Gulf of Maine.

Predation by sheephead (*Semicossyphus pulcher*, a benthic feeding fish) and spiny lobsters (*Panulirus interruptus*) is thought to limit sea urchins in warm-temperate kelp forests of southern California. Sea otters also occurred in this system, but deforestation events in this area did not occur until long after the otter's demise. The alternative urchin predators may explain the difference. As humans have progressively depleted these predators in recent decades through commercial and recreational fisheries, deforestation has become an increasing problem.

In the Southern Hemisphere, predation by rock lobsters (*Jasus lalandii*) in South Africa limits predatory whelks, in turn releasing subtidal mussel beds from limitation by whelk predation. A remarkable predator-prey role reversal occurred in this system following the extirpation of lobsters from several small islands. Whelk populations increased substantially in the lobsters' absence, thus transforming the reef from a mussel bed

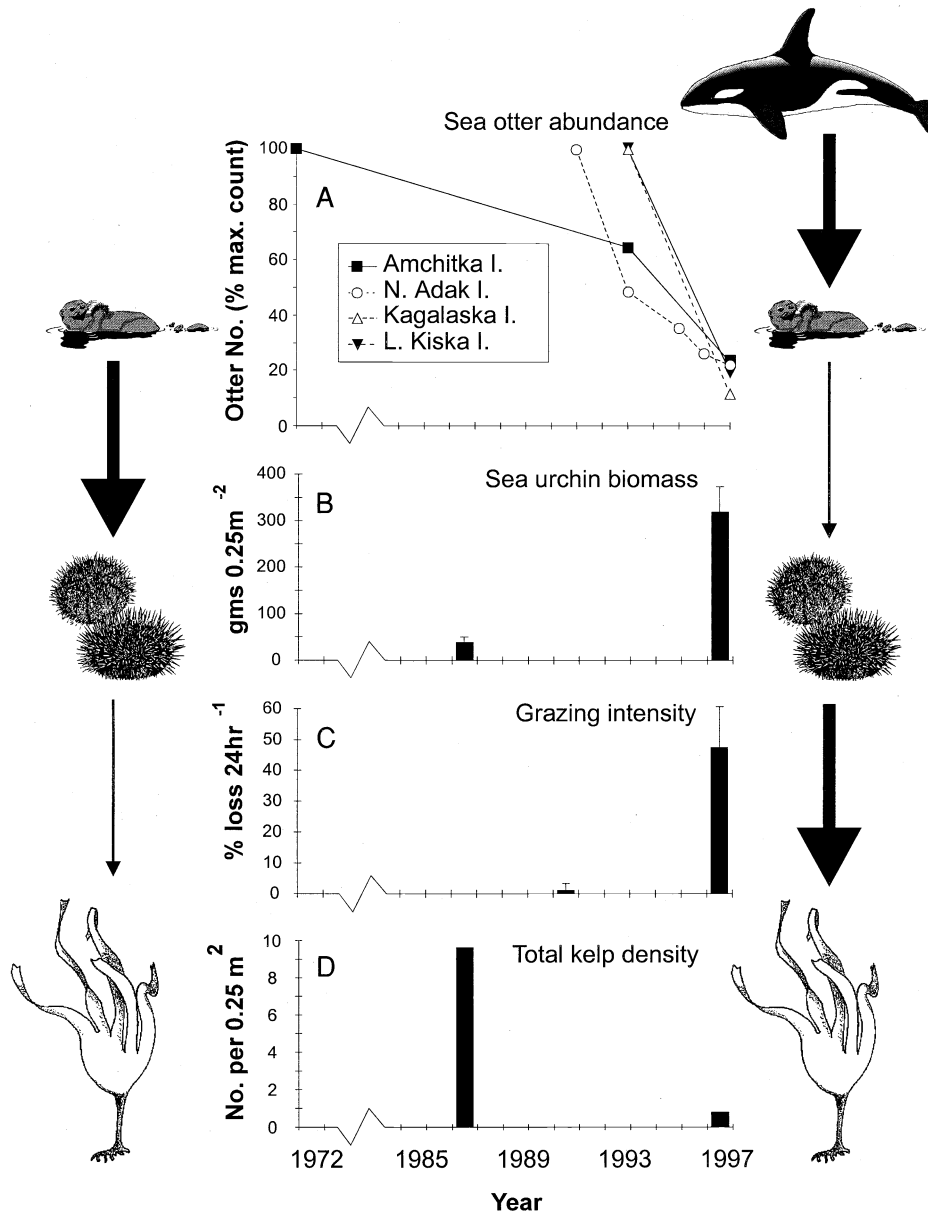


FIGURE 5 (A) Changes in sea otter abundance over time at several islands in the Aleutian archipelago and concurrent changes in (B) sea urchin biomass, (C) grazing intensity, and (D) kelp density measured from kelp forests at Adak Island, Alaska. Error bars in B and C indicate 1 SE. The proposed mechanisms of change are portrayed in the marginal cartoons: The one on the left shows how the kelp forest ecosystem was organized before the sea otter's decline and the one on the right shows how this ecosystem changed with the addition of killer whales as an apex predator. Thick arrows represent strong trophic interactions, and thin arrows represent weak interactions (reproduced with permission from Estes *et al.*, 1998).

into a kelp forest. In an effort to reestablish lobsters and their associated role as the system's dominant predator, a large number of lobsters were relocated to one of the islands. However, the whelks had become so abundant that they attacked the lobsters in mass, killing

all of them within hours of the translocation. This surprising case study demonstrates how a density-dependent role reversal between predator and prey can generate alternate stable-state communities.

These several examples from kelp forest systems



have shown that predators can shape populations and communities on ecological timescales and life history characters on evolutionary timescales. Comparative studies of sea urchins in tropical and temperate reef systems also suggest that predators influence their prey's behavior in complex ways. In warm temperate/tropical systems, sea urchins commonly retreat to protective cracks and crevices within the reef during daylight hours in order to avoid being eaten by diurnally active benthic predatory fishes. Fishes quickly attack urchins removed from their refuges during the day and placed on exposed habitats. Similar patterns have been shown for a variety of warm temperate and tropical systems in which benthic predatory fishes occur. However, the nature of urchin behavior appears to differ between species whose evolutionary histories are rooted in tropical vs temperate environments. Tropical species tend to display diel sheltering as a fixed behavior, regardless of ecological context, whereas the sheltering behavior is plastic in temperate species depending on whether predatory fishes are present or absent. The explanation for this difference in plasticity may lie in the fact that tropical urchins have long been subject to predation by diurnally active fishes, whereas temperate urchins have come into contact with benthic predatory fishes more recently, and then only at the warm margins of their geographical ranges.

### C. Lakes

Studies of freshwater lakes provide some of the clearest and best known evidence for trophic cascades. There are two main reasons for the quality of this evidence. Lakes, as discrete and recurrent entities, are well suited for comparative and experimental studies. Furthermore, the producers and herbivores (especially phytoplankton and zooplankton) have very short generation times, thereby making population-level responses to perturbations rapid enough for scientists to observe and document.

Some of the earliest evidence for the influence of predation in lake systems comes from Brooks and Dodson's analysis of New England lakes. These researchers showed that in the absence of planktivorous fishes, zooplankton assemblages were dominated by species with large body size because of their increased foraging efficiency and competitive superiority over small species. In lakes with planktivorous fishes, the composition of the plankton shifted toward small body size due to the influence of size-selective predation. This example was followed by Zaret and Paine's report on the cascad-

ing influences of introduced peacock bass (*Cichla ocellaris*) to Lake Gatun, Panama. Peacock bass, a cichlid native to the Amazon River, was first introduced to Lake Gatun in 1965 for sport fishing and consumption. These introduced predators are voracious piscivores and they caused a remarkable series of food web effects as the bass population grew and spread across Lake Gatun. The immediate influence was a rapid and extreme reduction of planktivorous minnows, thus causing zooplankton populations, including that of mosquito larvae, to increase. This example added two interesting dimensions to the understanding of lake systems. One is the strength of influence by an exotic predator on naïve prey, with broad-ranging indirect effects across the lake and surrounding terrestrial systems. In addition to the top-down effects described previously, the reduced populations of planktivorous minnows negatively impacted other apex predators, including several species of aquatic birds and predatory fishes. Another dimension is the potential impact on human health, in this case resulting from an increased threat of malaria because of increased mosquito populations. Similar examples of broad-ranging influences by exotic predators are known for many other lake systems throughout the world.

Numerous reports from various lake systems throughout the world show that altered populations of apex predators result in altered food webs. The essential players in these lake systems include four main groups of organisms: phytoplankton, herbivores, planktivores, and piscivores. The relationship of the first three of these to piscivore abundance, explained by cascading trophic interactions, is shown in Fig. 6. The evidence for these interactions comes from a variety of areas and approaches. Early insights were provided by contrasts among lakes in which piscivore populations varied serendipitously but for unknown reasons. There are many such examples from tropical and temperate lake systems in both the New and Old Worlds. Additional evidence that these patterns are caused by trophic cascades has come from the results of microcosm experiments, by tracking changes associated with the fortuitous extinction or reintroduction of piscivores into particular lakes through time, and recently by whole-lake experiments in which the piscivores were purposely added or removed. Although the details vary depending on such factors as food chain length and the nature of particular species, the overall view of food web dynamics in lake ecosystems is remarkably uniform, especially the importance of apex predators and trophic cascades.

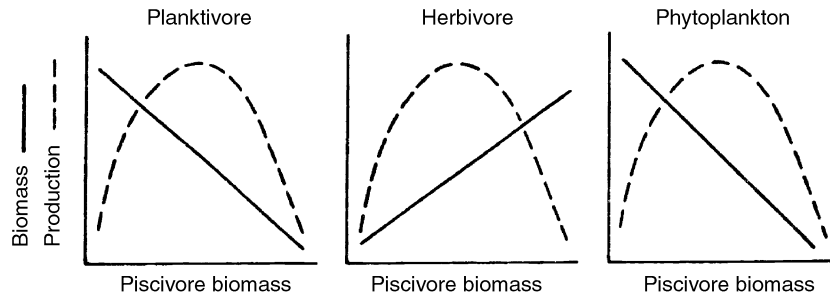


FIGURE 6 Piscivore biomass in relation to biomass (solid line) and production (dashed line) of vertebrate zooplanktivores, large herbivores, and phytoplankton in lake systems (reproduced with permission from Carpenter *et al.*, 1993).

### D. Rivers and Streams

Experimental work in rivers and streams has demonstrated important influences of predation on both food web structure and the life history of prey populations. The structure and dynamics of river food webs are grossly similar to those described for lakes, the main differences being that rivers are episodically disturbed by changes in water flow and they depend less on waterborne phytoplankton and zooplankton. Both fishes and birds are important apex predators in river food webs, and like lakes, many river food webs are strongly influenced by trophic cascades. The experimental exclusion of these predators from a variety of tropical and temperate river systems by Mary Power and colleagues provided several novel dimensions to the understanding of predation and trophic cascades. The manipulation of predatory fishes and other consumers provided consistent evidence for top-down forces and exclusion of birds has revealed depth-related gradients in the outcomes of trophic cascades. These findings show that the influence of predators on food webs can be strongly influenced by prey refuges, which in turn can vary across habitat gradients. Another important contribution of the riverine studies is that they have been done in systems that deviate in food chain length, thus providing the first experimental evidence that the strength of plant–herbivore interactions varies predictably between odd- and even-numbered food chains. Work by Sih and colleagues in streams of the eastern United States also shows how the risk of predation can influence the interplay between feeding and reproductive behavior in several prey species.

The pioneering work of Endler and Reznick on Trinidadian guppies provides some of the strongest and most comprehensive evidence for the selective effects of pre-

ation. The risk of predation to guppies by various larger fish species varies within and among streams, and a variety of life history characters, color patterns, and features of guppy mating systems covary accordingly. By manipulating predator populations and translocating guppies among habitats, these researchers demonstrated rapid selective responses to altered risks of predation.

### E. Oceanic Systems

Although the oceans dominate our biosphere and provide critical ecosystem services in such diverse forms as food production and climate control, little is known about the role of apex predators in the open sea. One reason is that the open sea and its associated seafloor habitat present serious logistical challenges to studies of any kind. Furthermore, many ocean ecologists still embrace the view that bottom-up processes are the main drivers of biological pattern in ocean ecosystems. Although bottom-up forcing in the sea is clearly important, this does not preclude top-down effects, which might be expected for several reasons. One is that strong predator-induced effects occur broadly in lakes and the general structures of ocean food webs (from phytoplankton to zooplankton to planktivores–piscivores) are similar to those of lakes. A second is that nowhere else on the planet are predators so abundant, as witnessed by the vast schools of marine mammals, seabirds, and predatory fishes. Despite this, we are aware of but one example of an oceanic trophic cascade. Pink salmon (*Oncorhynchus gorbuscha*) populations in the North Pacific fluctuate on a 2-year cycle. During years when pink salmon are abundant, zooplankton are depressed and phytoplankton are abundant, whereas during years when pink salmon are rare, zooplankton are abundant

and phytoplankton are relatively rare. Another potential example of predation in the open sea resulted when the blue, fin, sei, and minke whales were decimated by the whaling industry. This reduction in the great whales may have released Antarctic krill populations from limitation by predation, in turn elevating the carry capacities of other krill-feeders—pinnipeds, penguins, and perhaps additional groups of consumers. Increased growth rates and reduced age of first reproduction of seals and whales after the depletion of great whales from Antarctica have been interpreted as evidence for such effects.

Predation by gray whales (*Eschrichtius robustus*) and walrus (*Odobenus rosmarus*) has important effects on seafloor systems. Gray whales influence these systems by resuspending sediments and consuming amphipods. Furrows formed by the whales in the soft benthos are colonized by scavenging lysianassid amphipods, serve to accumulate detritus, and thus facilitate a local detritus-based food web. Walrus further impact these systems by consuming clams and other large infauna, in turn attracting predatory and detritivorous sea stars.

Although evidence from food web dynamics for a role by apex predators in the open sea is spotty at best, behavioral patterns of various prey species suggest strong predator–prey interactions. For example, krill and other large zooplankters typically undergo diel vertical migrations that take them beyond the foraging range of marine birds and mammals during daylight hours. Many species of forage fish and zooplankton form dense swarms, which probably reduce their likelihood of being consumed by predators that must search for and capture individual prey. Pagophillic (ice-loving) pinnipeds in the Arctic and Antarctica also provide a commanding case. In the Arctic, where polar bears and humans are both important predators, pinnipeds flee from the ice to water at signs of danger. In Antarctica, where the threat of predation is much greater in the water (from killer whales and leopard seals) than it is on the ice, the pinnipeds do not display such extreme flight behavior and often are nearly oblivious to potential disturbances when hauled out.

## F. Boreal/Temperate Forests

Although terrestrial biotas of the New World once contained numerous large mammalian carnivores, the potential ecological significance of these predators was unknown until recently. There are at least five reasons for the prolonged state of ignorance. One is that the largest of these creatures—gray wolves and grizzly

bears—were all but exterminated in the United States and Mexico well before modern ecological research had taken form. Second, even if large carnivores have been able to persist in the face of direct persecution, they are extremely difficult animals to study due to their low densities, nocturnality, secretive habits, aggressive behavior, and wariness of humans. Third, just as ocean ecologists have downplayed the importance of predators in the open sea, many wildlife ecologists have tended to be skeptical about the importance of predation in population regulation, and this topic has been hotly debated in the wildlife literature. In a famous example, Rasmussen, Leopold, and then HSS attributed an irruption of mule deer on the Kaibab Plateau, subsequent overgrazing, and the eventual mass starvation of the deer herd to the extermination of gray wolves and other large predators, but Caughley later attempted to debunk this explanation. Fourth, the long generation times of key players (decades to centuries for trees; multiple years to decades for ungulates and carnivores) and the large areas required for the measurement or manipulation of their representative populations make rigorous study of the top-down effect of apex predators very challenging. Finally, political, social, ethical, and legal issues have dissuaded many scientists from studying large mammals.

Despite these difficulties, there are indications of top-down effects by large predators in boreal forests. McLaren and Peterson used historical information on wolf and moose abundance, together with growth ring measurements from balsam fir, as evidence for a trophic cascade at Isle Royale in Lake Superior (Fig. 7). Wolf numbers have fluctuated substantially throughout the twentieth century, apparently in large measure because of demographic factors related to their small population size. Inverse changes in moose numbers followed wolf population fluctuations, thus suggesting regulation by wolf predation. Direct measures of herbivory were unavailable. However, the distance between annual tree rings in balsam fir showed that sapling growth rates were lower when moose were abundant than when moose were rare.

## G. Fragmented Coastal Scrub Habitats

Although experimental manipulation of terrestrial carnivores is exceedingly difficult, fragmented habitats can provide valuable, large-scale, ecological experiments to rigorously explore the top-down effects of mammalian predators. Large carnivores are particularly vulnerable to local extinction in fragmented landscapes due to large ranges and resource requirements, low population

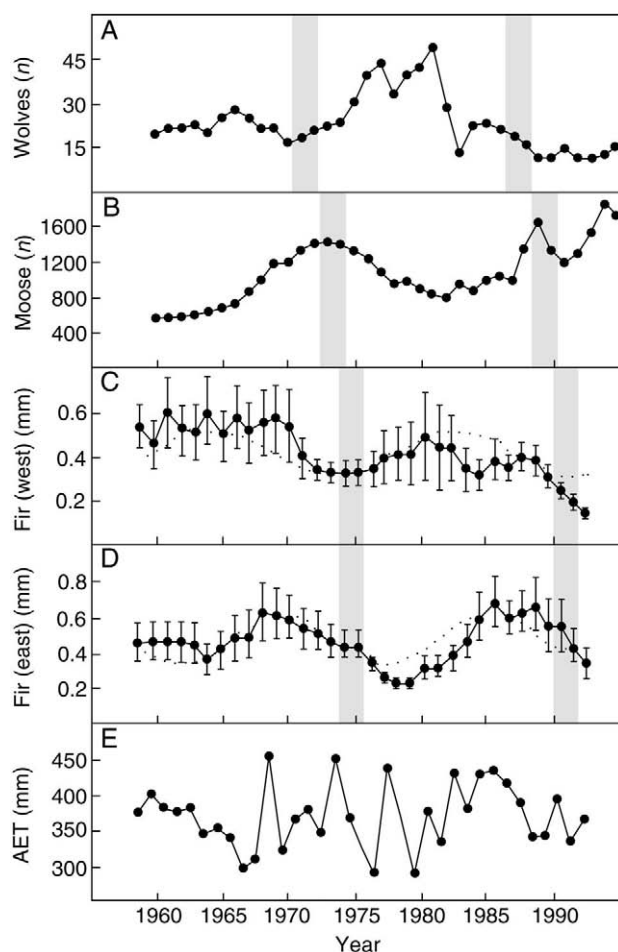


FIGURE 7 The trophic system on Isla Royale, reconstructed for 1958–1994. (A) Wolf abundance calculated from aerial surveys; (B) moose abundance calculated from skeletal remains and aerial surveys; mean ring-width indices for balsam fir from the west (C) and east (D) ends of Isla Royale; and (E) actual evapotranspiration from April to October, a measure of water availability during the growing season. The shaded areas highlight intervals of forage suppression that the authors believe are closely tied to periods of elevated moose density, which in turn follow periods of low wolf density (reproduced with permission from McLaren and Peterson, 1994).

densities, and direct persecution by humans. Larger carnivores can depress populations of smaller mammalian carnivores, or “mesopredators,” through direct predation, resource competition, and interference competition, including spatial and temporal avoidance. Consequently, the decline and disappearance of dominant carnivores in fragmented systems may lead to the ecological release of smaller predators that in turn threaten birds and other vertebrates. This process has been labeled “mesopredator release” and the phenome-

non has been implicated in the extinction of prey species worldwide.

In coastal southern California, intensive urbanization during the past century has destroyed most of the native coastal sage scrub and chaparral habitats, creating a “hot spot” of endangerment and extinction in the region and leaving undeveloped canyons dissecting coastal mesas to function as habitat islands immersed within a matrix of inhospitable urban habitat. Michael Soulé proposed the mesopredator release hypothesis as a possible mechanism to explain the rapid disappearance of scrub-breeding birds from the habitat fragments. He predicted that the decline of the most common large predator (coyote) would result in the ecological release of native (striped skunk, raccoon, and gray fox) and exotic (domestic cat and opossum) mesopredators, and that increased predation by these particularly effective avian mesopredators would result in higher mortality and extinction rates of scrub-breeding birds.

To test this prediction, Crooks and Soulé exploited a serendipitous ecological experiment—spatial and temporal variation in the distribution and abundance of coyotes among these urban habitat fragments—to investigate direct and indirect effects of this top predator on community structure. In accordance with the mesopredator release hypothesis, lower visitation rates of coyotes in small, isolated remnants resulted in elevated numbers and activity of urban mesopredators. Coyotes directly preyed on some mesopredator species; for example, domestic cats were found in approximately 20% of coyote scats in the fragments. Mesopredators temporally avoided coyotes as well. In fragments that coyotes visited episodically during the course of the study, mesopredator activity increased when coyotes were absent. As predicted, scrub bird diversity was lower in fragments with fewer coyotes and more mesopredators, even after accounting for the positive effect of fragment area and the negative effect of fragment age on bird persistence.

The top-down effect of coyotes on cats seems to have had the strongest impact on the decline and extinction of scrub-breeding birds in the urban fragments. Unlike wild predators, domestic cats are recreational hunters. Maintained well above carrying capacity by nutritional subsidies from their owners, they continue to kill even when prey populations are low. Using data on cat densities and predation rates, Crooks and Soulé estimated that cats surrounding a moderately sized fragment return approximately 840 rodents, 525 birds, and 595 lizards to residences per year. Such high levels of predation appear to be unsustainable for many small vertebrate populations. For example, existing population

sizes of some birds do not exceed 10 individuals in small to moderately sized fragments, so even modest increases in predation pressure from mesopredators may quickly drive native prey species to extinction. Extinctions of scrub-breeding birds are frequent and rapid; at least 75 local extinctions may have occurred in these fragments during the past century. Overall, this example illustrates that trophic cascades generated by the disappearance of an apex predator can combine with other fragmentation effects to influence species diversity in terrestrial systems.

### H. Midcontinent North American Prairies

The top-down effect of coyotes is evident in other systems as well. Historically, coyotes were generally confined to open plains and arid regions of western North America. The eradication of wolves from most of the continental United States in the late 1800s and early 1900s likely facilitated the expansion of coyote populations. Currently, predator control efforts in the United States are directed primarily at coyotes, and lethal and nonlethal measures have resulted in at least temporary coyote declines in some areas. In the Prairie Pothole Region of the North American midcontinent, coyote population reduction is considered one of the principal causes of increases in red fox populations starting in the 1930s; other carnivore species have been shown to increase in areas of coyote control as well. Although coyotes will kill red foxes, spatial and temporal avoidance, behavioral exclusion, and territorial shifting are likely the primary mechanisms by which coyotes reduce fox populations.

Red foxes are the most important predators on nesting ducks and their eggs and offspring in the prairie region. The ecological release of red foxes following coyote control has therefore resulted in increased predation on ground-nesting dabbling ducks, most notably mallards. Predation accounts for more than 70% of nest failures in these duck species, and in some areas intense predation on eggs, ducklings, and hens has been sufficiently intense to depress recruitment below replacement levels. Predation has resulted not only in population declines of duck species in the Prairie Pothole Region but also in altered population composition and skewed sex ratios.

Interestingly, since the mid-1970s, coyote populations have begun to rebound in parts of the Prairie Pothole Region due to restrictions in control and fur harvest methods and to reduced commercial value of fur. In these areas, expanding coyote populations have contributed to reduced red fox activity and higher duck

nest success; overall nest success in sites where coyotes are the principal canid is nearly twice as high as that in areas where red foxes dominate. Overall, the excellent series of studies by Sargeant and colleagues in the Prairie Pothole Region highlight the changes in the canid predator assemblage in North America during the past century and emphasize the top-down community-level effects generated by these dynamic canid–canid interactions.

### I. Tropical Forests

Some of the most dramatic evidence for top-down control by apex predators comes from research by John Terborgh and colleagues in New World tropical forests. Terborgh's vision of top-down control in this system stems from a contrast between Barro Colorado Island, in the Panama Canal, and Cocha Cashu Biological Station in Peru's Manu National Park. Although the two sites are similar in climate and native biota, Barro Colorado Island, because of its small size and isolation from other forest habitat, lost its apex predators (jaguars, pumas, and harpy eagles) shortly after construction of the Panama Canal. Barro Colorado Island currently supports notably higher densities of herbivorous mammals, such as agoutis, coatimundis, sloths, and howler monkeys, than does Cocha Cashu—differences attributed to the loss of predators from Barro Colorado Island. These ideas are now being put to a more rigorous test by using recently formed habitat fragments—the islands of Lago Guri in Venezuela—as a large-scale ecological experiment. The Caroni Valley of east-central Venezuela, once a vast, unbroken forest, was substantially altered by the 1986 creation of a hydroelectric impoundment. Within this 120-km-long by 70-km-wide reservoir, Lago Guri, the emergent hilltops became islands—isolated fragments of tropical forest that varied in size and distance from the shoreline border of unbroken forest. Although the larger islands retain nearly complete vertebrate faunas, the smaller islands lost up to 90% of the native vertebrate species, including all of the large vertebrate predators. Resulting changes in the forest system have been swift and sensational. Populations of herbivore species such as leaf-cutter ants, howler monkeys, iguanas, and rodents (all seed predators or herbivores) have increased by from one to three orders of magnitude. Indirect impacts on producers have been equally dramatic. Fewer than 5 of approximately 6070 native tree species are continuing to successfully reproduce, thus suggesting that highly impoverished floras will result from the loss of predators.

This particular example illustrates two important

points. First, predators often exert crucial roles in maintaining local species diversity. Terborgh's data indicate that the majority of tree species will eventually be lost from these islands systems, largely because of the loss of predators. This example, together with the Crooks–Soulé study of fragmented coastal scrub habitats, also serves to remind us of the power of ecological experiments and the importance of scale (space and time) in designing studies on the role of apex predators in nature. It is highly unlikely that any amount of study of unperturbed terrestrial systems could have demonstrated the magnitude and breadth of effect by relatively rare apex predators. However, anthropogenically disturbed systems, such as fragmented landscapes, offer unique opportunities to understand the complex trophic interactions generated by large carnivores.

## J. Exotic Predators on Islands

Perhaps nowhere is the top-down effect of predation on biodiversity so apparent as with the introduction of nonnative predators onto islands. Islands typically support few large predators and grazers, and apex predators such as mammalian carnivores are often absent. Consequently, insular endemic species regularly evolve in the absence of predation and thus lack adequate antipredator defenses. For example, many insular animals exhibit tame or fearless behavior that increases their vulnerability to introduced predators, and some island birds are flightless ground nesters. Similarly, many island plants do not produce the same noxious chemicals or physical defenses found in their closest mainland relatives—features that discourage herbivory.

Consequently, the introduction of nonnative predators can be catastrophic for sensitive island communities, to the point of driving insular prey species to extinction. The examples are numerous. Worldwide, of all the species that have gone extinct since 1600, 90% of the 30 species of reptiles and amphibians, 81% of the 65 mammal species, and 93% of the 176 species and subspecies of birds have been insular forms. Predation by introduced animals has been a primary cause for about 40% of the extinctions of birds on islands, and alien predators are endangering about 40% of currently threatened insular bird species. Introduced rats and domestic cats are notorious killers. Introduced rats have successfully invaded at least 80% of the world's 123 major island groups and are known to prey on a variety of insular vertebrates, including amphibians, reptiles, mammals, and birds. Indeed, predation by Pacific rats (*Rattus exulans*), black rats (*Rattus rattus*), and Norwegian rats (*Rattus norvegicus*) has been documented on

at least 15, 39, and 53 insular bird species, respectively, and introduced rats are thought to be responsible for 54% of insular bird extinctions caused by predators. For example, the introduction of black rats on Big South Cape Island, New Zealand, in 1964 resulted in the rapid extinction of 5 species of land bird, 1 species of bat, and an unknown number of invertebrates, including a species of large flightless weevil. On the Galapagos Islands, introduced black rats have reduced populations of the giant tortoise and the dark-rumped petrel by preying on eggs.

Mongoose and domestic cats have also been introduced to islands, at times deliberately to control nonnative rodents. Unfortunately, they instead often eradicate native prey species. On Hawaii, introduced mongoose had little impact on rodents but decimated flightless rail populations. Domestic cats have been accidentally or deliberately introduced to at least 65 island groups and are thought to be responsible for 26% of insular bird extinctions caused by predators. Incredibly, 375 cats on Macquarie Island near Australia killed an estimated 56,000 rabbits and 58,000 ground-nesting seabirds each year. On subantarctic Marion Island, 5 cats were introduced as pets in 1949 and by 1975 about 2000 cats were killing 450,000 burrowing petrels annually and were suspected in driving another petrel species to local extinction. In the mid-1900s about 5 cats were introduced to Kerguelen Island in the sub-Antarctic and their descendants have killed more than 3 million petrels per year and are responsible for the extinction of several bird populations. In the most infamous and perhaps most extreme example known, the lighthouse-keeper's pet cat on Stephen Island, a tiny island off New Zealand, arrived in 1894 and within a single year this one cat exterminated the flightless Stephen Island wren. The indirect effects of these changes, although likely important in some cases, are largely unstudied.

The effects of introduced predators extend far beyond their prey species and can include modification of ecosystem-level processes. For example, the New Zealand flatworm (*Artioposthia triangulata*) was accidentally introduced to the British Isles in the early 1960s and, with no natural predators, has spread rapidly. The flatworm is a voracious predator of native earthworm species. Earthworms, through their burrow excavation and casting activity, provide an invaluable ecosystem service by shaping the structure and hydrological patterns of soils. Flatworm infestations and the consequent depletion of earthworms alter both soil structure and hydrology. The ramifications are far-reaching and directly impact human welfare. By depleting native earthworms, the introduced New Zealand flatworm increases

the risk of surface runoff and therefore the potential of soil erosion, agrochemical pollution, and flooding.

Threatened prey species often recover, sometimes rapidly, when alien predators are controlled or eradicated on islands. However, the removal of alien predators can also yield unexpected results. For instance, a recent eradication of introduced black rats on Bird Island in the Seychelles has resulted in a population explosion of exotic crazy ants (*Anoplolepis longipes*). Ironically, these ants are now threatening those bird colonies that the rat eradication was intended to protect. Furthermore, simulation models predict that on islands colonized by both cats and rats, elimination of cats may release rat populations, and that increased numbers of rats may actually increase predation pressures on island birds. In essence, this cascade represents another example of the mesopredator release phenomenon, with cats on islands as top predators and rats as the mesopredators. On a variation of the theme, controlling cats on islands that also support exotic rabbits may result in more rabbits, excessive grazing by these prolific herbivores, and severe impacts to insular vegetation and associated animal species; this example, therefore, directly follows the predictions from the models of HSS and Fretwell. Clearly, alien predators on islands represent a complex, unpredictable, and occasionally dramatic example of the relationship between predators and biodiversity.

#### IV. SUMMARY AND SYNTHESIS OF CASE STUDIES

It is abundantly clear from the preceding examples that the manifestations of predation in nature are dramatic and diverse, occurring at organizational levels ranging from the behavior of individuals to the dynamics of ecosystems and on timescales ranging from ecological to evolutionary. Numerous studies show or suggest that predators influence the abundance, distribution, and population structure of their prey. Indirect effects of predation are less appreciated by scientists and the public, despite the fact that they occur broadly in nature, are important to ecosystem function, and often result in processes that benefit human welfare. Trophic cascades are the most common of known indirect effects. These may be nearly as ubiquitous in nature as the transfer of material and energy upward through food webs. In any case, ecologists should be more surprised by the absence of such top-down effects than by discoveries of new ones.

Despite extensive evidence for trophic cascades from

many of the planet's major ecosystems, we know little else about their influences on overall food webs. The lake studies provide an important exception because here a relationship between the top-down effects of predation and the bottom-up effects of production has been shown. These findings indicate that predators may fuel production in odd-numbered food chains, and that maximum production across all trophic levels should be realized at intermediate intensities of predation. A corollary to this hypothesis is that the length of food chains under top-down control is necessarily limited not by production and the efficiency of energy transfer but by the population constrictions that occur at consumer-regulated trophic levels. Such constrictions may be common in nature because intermediate intensities of predation probably occur rarely, except in highly managed ecosystems.

The theory of trophic cascades also provides guidance on where in a food web one might expect to find strong competitive interactions. Competition should be most strongly manifested within trophic levels in which populations are resource rather than consumer limited. These occur at the odd trophic levels in odd-numbered systems and even trophic levels in even-numbered systems. Although this prediction requires further analysis, it provides hope for an integrated theory of what previously has been viewed and treated as the largely unrelated processes of top-down, lateral, and bottom-up species interactions.

What humans perceive as "dysfunctional ecosystems" are often consequences of the recently altered roles of predators (e.g., losses of native species or introductions of exotics). Local extinctions and invasions are increasingly common, but unless these changes are observed or known from historical records, their significance to extant ecosystems may be difficult to understand. This point is illustrated by the following example. In a recent essay, Paine argued that the HSS model is generally correct for herbivorous insects but incorrect for herbivorous mammals—a contention supported by numerous examples of plant damage by mammalian grazers. The purported difference between insects and mammals as agents of herbivory has at least three possible explanations: that vegetation is intrinsically more vulnerable to mammalian than insect herbivores, that mammalian herbivores are less vulnerable to predation than their insect counterparts, or that the predators of mammals have been lost in disproportionately large numbers. The known reductions of large carnivores in North America make the latter mechanism a strong possibility.

Extinct interactions are difficult to infer from historical records; therefore, how might these alternatives be

assessed? As seen in some of the preceding examples, predation often leaves a mark on species-level characters, especially behavior, life history, and morphology. Vermeij's analysis of shell damage and morphological change (ostensibly from crushing predators) in Mesozoic marine gastropods provides a good example of one such record. The sudden increase in crushing predators was responsible for what Vermeij termed the "Mesozoic marine revolution." Using this perspective, Richard Palmer confirmed that shell structures indeed reduce the incidence of attack from crushing predators by experimentally removing spires from gastropod shells. Behavioral patterns also provide clues about the role of predators, as described in the case studies for sea urchins, pagophillic pinnipeds, and North American pronghorn. Other examples could be cited, but those listed are sufficient to make the point that morphology and behavior, when thoughtfully and cautiously interpreted, frequently reflect evolutionary response to predation.

Although these examples provide insight into species-level responses to predation on historical timescales, they afford little insight into the food web effects of predators. For this purpose, one might profitably examine the characteristics of producers. Plants can deter herbivores by modifying their morphology, demography, and chemistry. The degree to which these defensive characters exist among plant species and populations sometimes indicates the intensity of herbivory on historical timescales. Examples of this approach are provided by studies of differences in meristem location in steppe vegetation between the eastern and western slopes of the northern Rocky Mountains, variation in the resistance of birch trees to insect herbivores in boreal forests, the susceptibility to grazing damage in marine algae across coral reef habitats, and the evolution of reduced chemical and physical defenses in insular plants in the absence of herbivory. Mismatches in extant communities between the intensity of herbivory and the degree of plant resistance sometimes can be taken as evidence for recent changes in top-down regulation, as suggested from the previously described contrast between Northern and Southern Hemisphere kelp forests. Similar approaches might be taken to discern the evolutionary importance of apex predators in other ecosystems.

## V. THEORETICAL STUDIES OF PREDATION AND BIODIVERSITY

The topic of predation and biodiversity involves interactions among multiple species at different trophic levels

that interact in nonlinear, complex ways. A full understanding of this topic requires the analysis of mathematical and computer models, which permit one to keep track of multiple forces influencing the dynamics of interacting populations. A rich theoretical literature exists exploring impacts of predation on the dynamics and structure of ecological communities. Here, we summarize highlights of this work, emphasizing conceptual insights rather than mathematical details.

A predator can influence whether or not a particular species is present in a community either by facilitating its persistence (i.e., predators can enrich species composition) or by preventing it from invading (i.e., predators can constrain species composition). If removing an apex predator greatly increases the abundance of a particular prey, and this prey is a predator on species at lower trophic levels, predator removal could indirectly lead to shifts in competitive interactions and thus persistence of species several trophic levels removed (as in a trophic cascade). Even if a predator does not dramatically affect composition, it may strongly influence relative abundances of resident community members. Finally, predators can influence the existence and magnitude of temporal fluctuations in abundance. Mathematical models help one understand all these effects.

To analyze mechanisms influencing species composition, we consider the growth of species when rare. If each species in a community increases when rare, diversity is maintained in the face of perturbations. In theoretical studies, one writes equations describing the dynamics of each species and then analyzes this set of equations (e.g., with and without a top predator). To illustrate the basic approach, we discuss a simple model in detail and then briefly discuss results from other models. Theoretical studies suggest that there is no single relationship between predation and biodiversity but instead many relationships, depending on numerous contingent details of systems.

### A. Predation as a Density-Independent Mortality Factor: Effects on Biodiversity

As discussed previously, predator removal can unleash competition among prey and induce a wave of additional extinctions. To understand this effect, we express the growth rate of each species as a function of three factors:  $dN/dt = [\text{inherent growth}] - [\text{effect of resident competitor}] - [\text{mortality from resident predator}]$ , where  $N$  is abundance. Predation may both facilitate coexistence (e.g., by reducing the abundance and impact of competitors) and hamper coexistence (e.g., by direct mortality).



The most basic effect of predation is increased prey mortality. The simplest form of predation is that described by a fixed, density-independent mortality term. Generalist predators can occasionally act in this manner. Assume that two species experience strong, direct competition, described by the classic Lotka–Volterra competition model, with added mortality due to predation:

$$\frac{dN_1}{dt} = N_1 r_1 (K_1 - N_1 - \alpha_{12} N_2) / K_1 - m_1 N_1, \quad (1)$$

where  $N_1$  is the density of species 1,  $N_2$  is that of species  $j$ ,  $r_1$  and  $K_1$  respectively are the intrinsic growth rate and carrying capacity of species 1, and the competition coefficient  $\alpha_{12}$  is the effect of an individual of species 2 on species 1 (compared to the effect of 1 on itself). The second term expresses predation as density-independent mortality at a constant per capita rate  $m_1$ . (A comparable equation for species 2 completes the model.)

Using the criteria that each species should increase when rare, after some algebra the following condition for species coexistence emerges:

$$\frac{1}{\alpha_{21}} > \frac{K_1(1 - m_1/r_1)}{K_2(1 - m_2/r_2)} > \alpha_{12}, \quad (2)$$

where the term  $K_1(1 - m_1/r_1)$  is the effective carrying capacity of species 1 in the face of predation. This inequality implies several interesting conclusions. If  $\alpha_{12}\alpha_{21} > 1$ , no pattern of imposed, density-independent mortality leads to coexistence. (This in effect says that interspecific competition is stronger than intraspecific competition.) Likewise, if competition coefficients are unity, one will not observe coexistence, regardless of the pattern of mortality. If the two competitors have the same intrinsic growth rate, and predation is uniform (or more generally,  $r_1/m_1 = r_2/m_2$ ), mortality drops out, so there is no effect of predation on coexistence.

However, if  $\alpha_{12}\alpha_{21} < 1$  and the two competitors differ in the ratio  $m/r$ , predation can occasionally facilitate coexistence. Consider a case of a competitive hierarchy, such that  $\alpha_{12} = 0$ , but  $\alpha_{21} > 0$  and species 2 is competitively excluded. In the absence of the predator, species 2 is excluded if  $K_2 < K_1\alpha_{21}$ ; with the predator, coexistence is permitted if  $K_2(1 - m_2/r_2) > K_1(1 - m_1/r_1)\alpha_{21}$ . Comparing these two inequalities, it can be seen that the predator facilitates coexistence only if  $m_2/r_2 < m_1/r_1$ . In other words, the dominant competitor must either experience higher mortality or have a lower intrinsic

growth rate. However, if predation is too low (low  $m$  for both species), there will still be competitive exclusion. Moreover, if predation is too intense, there will not be coexistence because one (or both) species is directly eliminated by predation.

This model illustrates several general points that are applicable to a wide range of models. First, predation will not always facilitate coexistence. Second, if a keystone predator effect is possible, the effect occurs only within a particular range of population parameter values. Typically, predator-mediated coexistence requires an intermediate level of predation. Very intense, generalized predation will almost always reduce species richness. This is particularly likely in systems in which prey species have not had a shared evolutionary history with predators (e.g., the brown rat snake as a predator on birds on Guam), the physical structure of the environment makes it easy for predators to encounter prey (e.g., no refuges, as in open lakes), or the species in question have low intrinsic growth rates. Third, predator-mediated coexistence requires a tradeoff: The species that is the superior competitor needs to be more vulnerable to predation. When this occurs, predator removal will endanger the persistence of inferior competitors (as in Paine's *Pisaster*). Finally, the appropriate measure of a species' vulnerability to predation combines mortality rates ( $m$ ) and the ability to replenish losses ( $r$ ). High, uniform rates of mortality tilt the balance of competitive interactions toward species with high intrinsic growth rates.

The previous model structure assumes that predation influences competition via changes in abundance. However, prey can also show behavioral shifts when faced with predators—for instance, spending more time in refuges and less on foraging. Such “higher order interactions” can either make coexistence more difficult or weaken competition, depending on the detailed nature of the behavioral changes.

The Lotka–Volterra model most literally applies to systems with strong, direct interference interactions, in which the only dynamical variables are each competing species density. In models of exploitative competition for a single limiting resource, predation that leads to density-independent mortality can influence which species wins, but it will not lead to coexistence. Predation is more likely to promote (or occasionally to destroy) coexistence when mortality rates are dynamical variables responding to prey abundance. We next explore several modifications of this model which illustrate the rich repertoire of dynamical behaviors made possible when predation varies dynamically in response to prey abundances.

## B. Numerical and Functional Responses

The rate of mortality imposed by a predator reflects both predator numbers and the attack rate per predator. The total rate of mortality is  $mN = fP$ , where  $P$  is predator density and the parameter  $f$  is the number of prey (of the focal species) captured per individual predator. Because predators consume prey, the rates of the demographic parameters (birth, death, and movement) of predators will often vary as a function of the abundance of prey. Thus, the number of predators should depend on prey numbers. This is the numerical response. The rate at which an individual predator captures prey of a given species should also depend on the number of prey that are available, typically (although not always) increasing with prey abundance but saturating at high prey numbers. This is the functional response. It is useful to express the functional response as  $f = aN$ , where  $a$  is the attack rate per predator per prey—the risk of mortality an individual prey faces from an individual predator. Different predators have distinct numerical and functional responses and therefore will have different impacts on species coexistence.

## C. Specialist Predators and Biodiversity

Specialist predators (whose diets are restricted to single prey species) typically reduce abundance of their favored species, freeing up resources for nontarget species. This can facilitate coexistence if dominant competitors attract more specialist predators or parasites than do subordinates. This diversifying effect of specialized predation can be mimicked in Eq. (1) by letting the attack rate on each species be a function of the abundance of a specialist predator,  $m_i N_i = a_i P_i N_i$ , where  $P_i$  is the abundance of predator species  $i$ , which is specialized in its foraging just to prey species  $i$ . To complete the model, we need an equation for the dynamics of each predator:  $dP_i/dt = P_i[g_i(N_i, P_i)]$ , where  $g$  is a function which increases with  $N_i$  (e.g., because predators convert prey consumption into births) but may decrease with  $P_i$  (e.g., because predators interfere with each other). For any biologically reasonable system, at low numbers of their required prey specialist predators must decrease (i.e.,  $g_i < 0$ ). Hence, the growth rate of prey species  $i$  when it is rare and other prey are resident will involve only the intrinsic growth of prey  $i$ , discounted by competition with the resident species. Because the residents continue to be attacked by their own specialist predators, their numbers will be depressed below carrying capacity, reducing competition imposed on the

rare focal species. Thus, specialized predators are expected to facilitate prey coexistence.

An interesting twist is that specialist predator–prey interactions are often unstable when the predator is effective at limiting prey numbers because of time lags in the numerical response of predators to prey. This has two consequences for species coexistence. First, as shown by Peter Abrams and others, with saturating functional or numerical responses instability tends to depress average predator numbers and thereby increase average prey numbers. This increases competition, making coexistence more difficult. Second, with unstable dynamics between a resident specialist predator and its prey, there will be times when that prey is rare and an inferior competitor can invade and temporarily persist, only later to be competitively excluded when the predator is rare and the dominant competitor has rebounded in numbers. Thus, unstable specialist predator–prey dynamics induces instability in community composition as well.

These effects are believed to be particularly important when considering impacts of insect herbivores on plant communities. However, most predators in the examples discussed in this article tend to have generalized diets.

## D. Generalist Predators and Biodiversity

### 1. Switching and “Enemy-Free Space”

The expectation that specialist predation helps competing species to coexist depends on the very general and reasonable assumption that specialist predators will have numerical responses to their prey. Generalist predators can have a wide range of effects on prey communities. A single, effective generalist can act like a whole suite of specialists in promoting prey coexistence if the predator ignores whichever prey species is temporarily the rarest, concentrating attacks on common prey—the mode of foraging behavior called switching. As shown by Roughgarden and Feldman, switching predators can readily prevent competitive exclusion. In our formulation for mortality due to predation, for example, on prey species 1,  $m_1 N_1 = a_1 P_1 N_1$ , we can represent switching with an attack rate  $a_1 = a_1(N_1, N_2, \dots)$ , where  $a_1$  declines toward zero as  $N_1$  approaches zero but increases if the other  $N_i$  decrease. In effect, a prey species may persist because of a refuge in relative rarity, as defined by the predator’s behavioral responses. Because the predator is also reducing the abundance of potential competitive dominants, this is a potent mechanism for predator-mediated coexistence. If such a predator is removed, numerous prey species may risk extinction.

This idea is intuitively appealing, but there are surprisingly few demonstrations of it in natural systems. Those that do seem to involve prey species that are found in different habitats. A behavioral rule that predators leave patches with few prey, and aggregate in patches with numerous prey, leads to switching and will tend to foster prey coexistence. If prey species are spatially segregated, they are not likely to be strongly competing in any case. Other potential cases of switching seem to involve prey species with very different strategies for blending into the background, different activity times, or different behavioral tactics for escaping predation. John Lawton suggested that such species differences promoting coexistence be viewed as partitioning of enemy-free space, an aspect of niche differentiation.

In any case, several examples of species exclusion caused by predation discussed elsewhere in this article show that predators often do not tend to ignore rare prey species but rather continue attacking even to the point of extinction.

## 2. Saturating Functional Responses

All predators have a maximal capacity for attacking prey that is set by limited time or gut capacity. Time or effort expended in attacking one prey will be unavailable for attacking other prey. If predator numbers are fixed, an increase in abundance of one prey species may reduce attacks on another. As with switching, we can represent this as  $a_i = a_i(N_1, N_2, \dots)$ , but now attacks decline with the abundance of each species. This inverse density dependence has two consequences. First, considering the effect of a prey species on itself, its mortality will decline with increasing abundance. This inverse density dependence can lead to alternative stable states for a given prey species—one at low and another at high densities. The low-density equilibrium may even be at zero density. Second, an increase in any prey species can reduce attacks on other prey species; this in effect makes alternative prey indirect mutualists. Such mutualism is most likely when considering predators that are constrained in their numerical responses (e.g., due to long generation lengths relative to those of their prey).

Unlike switching, saturating functional responses are universal. The existence of indirect mutualisms arising because predator functional responses can be swamped may explain many natural phenomena, such as herding, mixed-species flocks of birds and schools of fish, and synchronized mass emergences and migrations. One consequence of importance to biodiversity is that if some prey species are reduced in abundance,

there is an immediate increase in predation on other, alternative prey, which thus risk extinction.

## 3. Numerical Responses and Apparent Competition

The potential for indirect mutualism via the functional response is often offset by a kind of indirect competition via the numerical response. Just as consumer species can reduce each other's abundance by depleting a shared limiting resource, alternative prey species can indirectly depress each other's abundance by increasing the abundance of a shared predator. This indirect interaction is known as apparent competition. It is particularly likely if predator population growth rates increase with the abundance of each prey type in the predator's diet and predator numbers are not strongly limited by other factors (e.g., territoriality or higher order predators). It is also more likely for predators with short generation lengths, not greatly exceeding those of their prey, or for predators which are highly mobile and can quickly aggregate into habitats with unusually high prey densities.

When any given prey species is rare, an increase in predator abundance will usually lead to an increase in its mortality. Predator abundance is expected to increase with the productivity and availability of alternative prey. The rate of predation on a focal prey is determined by the indirect, cumulative impact of alternative prey, sustaining the predator population at densities higher than allowed by the focal prey. Particularly when prey species are not strongly competing, the negative indirect interaction of apparent competition can limit prey species diversity.

When a prey species is rare, its rate of population growth can be represented as  $r - aP$ . As noted previously, often the presence of alternative prey reduces the attack rate on a rare species. However, these same prey determine the magnitude of  $P$ . The net effect of prey on each other can only be determined by analyzing specific models. However, some general points are worth making. A given prey species is excluded by predation if  $0 < r < aP$ . Prey species with low  $r$  are particularly vulnerable to exclusion by shared predation, as are species which have high attack rates. Another way of stating the exclusion criterion is that the maximal predator density which this prey can tolerate is  $r/a$ . All else being equal, prey with low values for  $r/a$  are vulnerable to exclusion (note that a prey species with a high value for  $r/a$  can sustain a high abundance of a generalist predator, which can then with impunity overexploit alternative prey with lower  $r/a$ ). The upshot of these observations is that there is a tendency toward

the exclusion of prey species in the diet of polyphagous predators. The maintenance of high diversity requires mechanisms which offset this effect. Such mechanisms might include predator switching, constraints on the predator numerical response, or prey adaptations that reduce predation at low density (e.g., spatial refuges).

Often, apparent competition will be strongly asymmetrical. A productive prey, which can only be successfully attacked by predators when it is young, ill, or aged, may not be strongly limited by predation. However, this species can sustain a high population of the predator, which can then severely depress species more vulnerable through their life history.

If different prey species occupy different habitats, and predators have limited mobility, this too can prevent exclusion via shared predation. However, mobile predators can be sustained by productive prey in one habitat and with impunity can overexploit prey in low-productivity habitats. A serious effect of habitat fragmentation is that it exposes species in habitat remnants to predation from generalist predators sustained by alternative prey in the surrounding landscape. Most examples of dramatic prey limitation by predation seem to depend on the availability of alternative prey, which permit predator numbers to remain high. The brown rat snake on Guam can persist on a diet of rats and lizards, which permits it to eat out of existence every native bird species.

Previously, we discussed the cascading effects of top predator removal. Such removals shift the factors regulating intermediate mesopredators or herbivores, which will increase and become more regulated by food availability than they were in the past. This can unleash strong apparent competition effects at lower trophic levels. Prey species harmed by polyphagous predators via "mesopredator release" are victims of apparent competition.

### E. Generalist Predators and Community Stability

Generalist predators can have many different effects on the overall stability of communities. Here, we discuss a few interesting effects.

Richard Vance examined a Lotka–Volterra model akin to Eq. (1) but with a predator showing numerical responses to each prey species. Even if the pairwise interactions were all stable, the entire ensemble could show large-amplitude cycles, or even chaotic dynamics.

If generalist predators are mobile and seek out patches with high prey abundance, this can lead to switching. If these responses are rapid, then generalist

predators help stabilize prey dynamics. However, there can be time lags in these responses, which in turn can be destabilizing. Recent theoretical explorations suggest that switching behavior in patchy environments leads to systems which persist but which have bounded oscillations. Moreover, in changing landscapes, mobile generalist predators can concentrate in habitat remnants, leading to transient spikes of high predation and extinction risk for prey species residing in these remnants.

Generalist predators usually have a mixture of a few strong interactions and many weak interactions with the species of prey in their diet. Theoretical studies have recently demonstrated that weak interactions can help reduce inherent instabilities in strong predator–prey interactions. However, the effect depends on detailed assumptions made about predator–prey feedbacks; for instance, if prey flow into a given habitat and contribute to the diet of a resident predator, this predation does not feed back to influence prey numbers in the source habitat. If the predator instead is highly mobile, its ability to feed in multiple habitats may permit severe overexploitation of prey in some habitats.

Given saturating functional responses, in unstable systems apparent competition can be reduced relative to that in stable systems because the nonlinearity in the functional response means that predators are harmed more by times of low prey abundance than benefited by times of high prey abundance. This reduces the number of predators which can be sustained, thereby weakening apparent competition effects. The dominant effect may then be competitive exclusion (if the prey are strong competitors) or indirect mutualism (with noncompeting prey and saturating functional responses).

The theoretical studies summarized here suggest that there is no universal effect on predation on biodiversity but rather many effects. More important, they help emphasize potential surprising effects of predator removal and highlight the range of information a conservationist needs to gauge the likely impacts of management alternatives. Most theoretical studies of predation have been limited to interactions across two trophic levels. More complex theoretical studies of trophic interactions are needed to investigate such phenomena as trophic cascades.

## VI. GENERALITY

There is much evidence for the influences of predators on species, populations, communities, and ecosystems. However, how predictable and widely occurring are

these effects? To what degree do predators regulate the structure and function of the planet's ecosystems relative to other biological interactions and physical forces?

These important questions can be asked at two levels—within and between ecosystems. Within systems, the question of generality usually concerns the geographical extent of an effect that has been demonstrated at one or several sites. Michael Foster raised the issue for the influences of sea stars in mussel beds and sea otters in kelp forests, claiming that the interactions were less common than generally believed. Bruce Menge has begun to deal with the problem for sea stars and mussel beds by identifying some of the conditions along rocky shores under which sea star predation effectively limits or regulates mussel populations. Not surprisingly, sea star predation is important at many sites but not everywhere. For sea otters and kelp forests, Estes and Duggins approached the question by evaluating how consistently predictions of the otter–kelp forest paradigm played out at many sites with and without sea otters. In this case, the predictions held up (i.e., otter-dominated sites supported kelp forests and otter-free sites were deforested by sea urchin grazing) nearly everywhere they examined from southeast Alaska to the western Aleutian archipelago (Fig. 8), and similar results have been obtained from British Columbia. However, as mentioned previously, these patterns are less general in southern California kelp forests.

Across ecosystems, the question of generality for top-down influences of apex predators becomes one of both relative importance and variation in process. We know that trophic cascades are not limited to aquatic systems, as earlier suggested by Donald Strong. Nonetheless, understanding whether ecosystem function is controlled by trophic cascades involving a few key species (the HSS and keystone species models) or by a greater complexity of food web interactions within and across systems (a view espoused by Polis and Strong) remains one of ecology's most daunting challenges.

## VII. THE FUTURE

Although predators have long concerned conservationists and resource managers, the future will bring heightened attention to this group. As a general rule, apex predators are more vulnerable to local extinctions than are lower trophic-level species. The rapid demise of predators has occurred in part because they have been (and in many areas still are) treated as competitors of humans for fish, wildlife, and agricultural resources; thus, they have been persecuted rather than conserved.

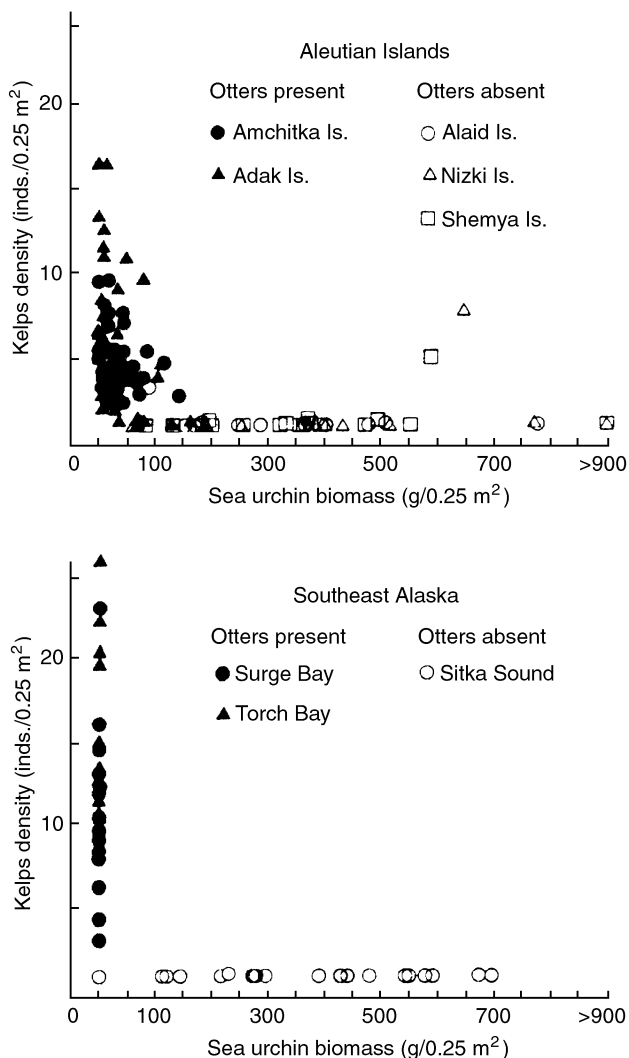


FIGURE 8 Epibenthic kelp density plotted against estimated sea urchin biomass from locations with and without sea otters in the Aleutian Islands and southeast Alaska. Points represent averages of 20 randomly selected plots from each location. These data show that kelp forest community structures vary predictably depending on the presence or absence of sea otters (reproduced with permission from Estes and Duggins, 1995).

Habitat fragmentation has also hastened local extinction rates of large apex predators because their typically low densities and large home ranges render the smaller fragments incapable of maintaining viable populations. These facts, together with the increasing realization that predators are often essential for maintaining native ecosystems, is leading to a paradigm shift in conservation biology, especially in the area of reserve design. Earlier approaches to conservation planning focused on preserving representative habitats. Many ecologists now believe that this approach, although necessary, is not

sufficient. Large reserves, or a series of smaller connected reserves, are necessary to maintain viable populations of predators, which in turn are essential for maintaining the functional integrity of these systems. This view has important implications to restoration ecology, which until recently has focused mainly on the reintroduction of native plants and the elimination of exotic species.

If the maintenance or restoration of native predators is important to conservation biology, so is the elimination of exotic predators. Exotic predators have devastated many natural biotas, both because of their ability to reduce or exterminate native prey species and the many indirect effects of these prey throughout their food webs. As discussed previously, island biotas provide the most obvious and poignant examples of such effects. The removal of exotic species is usually expensive, time-consuming, and wrought with technical challenges. Nonetheless, exotic predators, because of their great mobility and low population density, are often easier to remove than other invasive species.

The conservation and management of predators requires more and better information than is currently available. As recently noted by George Schaller, field studies have been conducted on less than 15% of the species of mammalian carnivores. Although this is indeed a feeble record, there are even fewer successful efforts to understand the ecological importance of this group.

Are all or most of the predators important players in prey population regulation and ecosystem dynamics? Some clearly are important, but are these examples exceptions or a general rule? The fact that so many species already have been depleted or eliminated hampers the pursuit of answers. The depleted status of so many predators raises the additional question of how their influences vary with population density. Most apex predator populations probably were once regulated by competition for food, which in turn must have strengthened predator-prey interactions. However, humans have changed this, so we now must wonder how the strength of these interactions varies with distance below equilibrium density. The several possible functional relationships, shown in Fig. 9, have very different management implications. Conservationists might hope for a relationship like that depicted by B in Fig. 9 because this would mean that the maintenance or restoration of the ecological roles of predators requires only that they be present in low numbers. In contrast, fisheries and wildlife managers might hope for a relation more like that depicted by C, in which the influence of predators is of little consequence at population levels below carrying capacity.

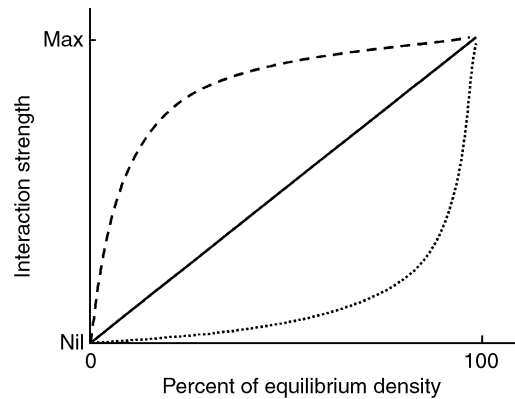


FIGURE 9 Several population relationships between predator-induced interaction strengths and predator population abundance relative to equilibrium density. In A (solid line), the relationship is linear. In B (dashed line), the main influences of predators occur at a wide range of population densities, whereas in C (dotted line) these influences occur only at high population densities.

A final need is for greater understanding of the full range of food web effects by predators. Most prior studies of predator effects focused on predator-prey interactions, thus creating the overly simplistic mind-set still held by many wildlife and fisheries managers. Most studies of indirect effects take this perspective only a step further by focusing on trophic cascades. However, as discussed previously, trophic cascades are expected to mediate competitive interactions and influence the strength of bottom-up forces by altering production levels at the base of the food web. Little is known about such interactions, although several examples of long and complex chain reactions among species show that they can be tremendously important to ecosystem function.

For the most part, food webs and ecosystem dynamics have been studied by one of two approaches. The oldest of these involves descriptions of food web structure and estimating the transfer of materials and energy through their various linkages, the critical assumption being that flux rate reflects interaction strength. The fallacy of this assumption is evident in the fact that strong top-down forces necessarily reduce prey abundance to such low levels that the interaction may no longer be apparent in a static food web. The second approach is to observe these dynamics directly by perturbing the system. Unfortunately, most experimental studies have been limited to processes acting on small spatial and temporal scales (for obvious logistical reasons) and to invertebrates and heterothermic fishes (for social and political reasons). Our general lack of understanding of the larger apex predators (especially large

mammals) is due in large part to the almost complete absence of experimental evidence.

Creative approaches are needed to put the ecological influences of large apex predators into a dynamical context. The use of historical information and opportunities provided by "natural experiments" holds promise in this regard. The historical record is a rich source of information that has barely been utilized. Assessments of the wolf/moose/boreal forest system from tree ring analysis and the sea otter-kelp forest system from faunal remains in Aleut kitchen middens illustrate the utility of this approach. Fortuitous change is another potential means of assessing the ecological roles of large apex predators. Studies by Terborgh on the recently formed islands of Lago Guri, by Crooks and Soulé on mesopredator release in urban habitat fragments, and by Estes and Palmisano on sea otters and kelp forests demonstrate the value of this approach. As these and other examples show, ecologists and resource managers must remain mindful of appropriate scales of time and space for both learning about predators and applying that knowledge to the conservation of biodiversity.

### Acknowledgment

We are deeply grateful to Stacey Reese for helping prepare the manuscript.

### See Also the Following Articles

CARNIVORES • COMPETITION, INTERSPECIFIC • FOOD WEBS • KEYSTONE SPECIES • PARASITISM • POPULATION DYNAMICS • SPECIES INTERACTIONS • TROPHIC LEVELS

### Bibliography

- Byers, J. A. (1997). *American Pronghorn: Social Adaptations and the Ghosts of Predators Past*. Univ. of Chicago Press, Chicago.
- Carpenter, S. R., and Kitchell, J. F. (Eds.) (1993). *The Trophic Cascade in Lakes*. Cambridge Univ. Press, New York.
- Carpenter, S. R., Kitchell, J. F., and Hodgson, J. R. (1985). Cascading trophic interactions and lake productivity. *Bioscience* 35, 634–639.
- Castilla, J. C., and Duran, L. R. (1985). Human exclusion from the rocky intertidal zone of central Chile: The effects on *Concholepas concholepas* (Gastropoda). *Oikos* 45, 391–399.
- Crooks, K., and Soulé, M. E. (1999). Mesopredator release and avifaunal extinctions in a fragmented system. *Nature* 400, 563–566.
- Estes, J. A. (1996). Predators and ecosystem management. *Wildlife Soc. Bull.* 24, 390–396.
- Estes, J. A., and Duggins, D. O. (1995). Sea otters and kelp forests in Alaska: Generality and variation in a community ecology paradigm. *Ecol. Monogr.* 65, 75–100.
- Estes, J. A., Tinker, M. T., Williams, T. M., and Doak, D. F. (1998). Killer whale predation on sea otters linking coastal with oceanic ecosystems. *Science* 282, 473–476.
- Fretwell, S. D. (1987). Food chain dynamics: The central theory of ecology? *Oikos* 50, 291–301.
- Hairston, N. G., Smith, F. E., and Slobodkin, L. B. (1960). Community structure, population control and competition. *Am. Nat.* 94, 421–425.
- Holt, R. D. (1977). Predation, apparent competition, and the structure of prey communities. *Theoretical Population Biol.* 12, 197–229.
- McLaren, B. E., and Peterson, R. O. (1994). Wolves, moose, and tree rings on Isle Royale. *Science* 266, 1555–1558.
- Paine, R. T. (1966). Food web complexity and species diversity. *Am. Nat.* 100, 65–75.
- Peterson, C. H. and Estes, J. A. (2000). Conservation and management of marine communities. In *Marine Community Ecology* (M. Bertness, M. Hay, and S. Gaines, Eds.) Sinauer (in press).
- Power, M. E. (1990). Effect of fish in river food webs. *Science* 250, 411–415.
- Reznick, D., and Endler, J. A. (1982). The impact of predation on life history evolution in Trinidadian guppies (*Poecilia reticulata*). *Evolution* 36, 160–177.
- Terborgh, J., et al. (1999). The role of top carnivores in regulating terrestrial ecosystems. In *Continental Conservation: Scientific Foundations of Regional Reserve Networks* (M. E. Soulé and J. Terborgh, Eds.), pp. 39–64. Island Press, Washington, DC.
- Vermeij, G. J. (1977). The Mesozoic marine revolution: Evidence from snails, predators and grazers. *Paleobiology* 3, 245–258.



# PRIMATE POPULATIONS, CONSERVATION OF

Russell A. Mittermeier and William R. Konstant  
*Conservation International*

---

- I. Introduction
  - II. Overview of Global Primate Diversity and Conservation
  - III. Threats to Primates
  - IV. Conservation Status of Primates
  - V. Primate Conservation During the Past 20 Years
  - VI. Outlook for the Future
- 

## GLOSSARY

- anthropoids** “Higher” primates such as marmosets, tamarins, monkeys, and apes.
- diversity** The variety of species within a given taxonomic group.
- endemic** Unique to or occurring only in a specified geographic location or region.
- prosimians** “Lower” primates such as lemurs, lorises, and galagos.
- 

At the beginning of the twenty-first century, we contemplate environmental changes that may be wrought by global warming and ozone layer depletion and brace for mass extinctions of plants and animals predicted by many of the world’s leading biologists. As primatologists, we contemplate the subjects of lifelong study and wonder how many of the world’s prosimians, monkeys, and apes will survive into the next millennium since

efforts to stem the loss of their tropical forest habitats have met with only minimal success.

## I. INTRODUCTION

During the past few decades, interest in nonhuman primates has increased significantly on many fronts and has helped build support for conservation. Pioneering and long-term field studies of the great apes (Schaller, 1963; Goodall, 1968; Fossey, 1983; Galdikas, 1995) have dispelled age-old myths about mankind’s closest living relatives, greatly narrowing the gaps between these species and our own, and provided new insights into human origins and behavior. The continuing search for drugs to treat global maladies such as malaria, cancer, and AIDS has required large numbers of nonhuman primates as experimental subjects (Mack and Mittermeier, 1984). For some species, this use has contributed to serious declines of wild populations and ultimately forced issues of conservation and captive breeding as part of the long-term strategy for biomedical research. In other cases, little known and formerly obscure primate species of the neotropics, Africa, and Asia have emerged as prominent “flagships” for conserving their tropical forest habitats, which biologists agree are the richest natural terrestrial ecosystems on the planet.

Although interest in nonhuman primates is increasing, the threats to their survival persist. Varying combinations of habitat destruction, hunting, and live capture have driven dozens of primate species to the brink of



extinction, to the point that several taxa now number only in the low thousands and a few no more than a few hundred individuals. Such populations are doomed without long-term protection, monitoring, and a heightened understanding of their plight by local human populations.

In the face of continuing threat, however, primate conservationists can also reflect upon the past century and realize with some degree of pride that, to the best of anyone's knowledge, not a single primate taxon went extinct during that period. The next few years could conceivably witness the loss of such primates as the Tonkin nub-nosed monkey (*Rhinopithecus avunculus*) of Vietnam or Miss Waldron's red colobus (*Procolobus badius waldroni*) of West Africa, both critically endangered and the latter has not been located during several recent surveys. In the new millennium, the survival of these and other threatened taxa will depend on the continuation of existing conservation programs and the establishment of new ones according to a global strategy for the most endangered. Fortunately, support for

global primate conservation increased in the latter half of the 1990s after having suffered something of a dry spell, and the expertise is at hand to direct available resources to the highest priority species, habitats, and projects.

## II. OVERVIEW OF GLOBAL PRIMATE DIVERSITY AND CONSERVATION

The order Primates represents one of 21 mammalian orders that together total at least 4675 species. It includes 13 families, 63 genera, and approximately 630 taxa worldwide. Two suborders of primates are recognized, the Prosimii (prosimian or lower primates) with 7 families and the Anthropoidea (higher primates) with 6 (Table I).

Living prosimians occur only in the Old World, despite the fact that North America once represented a major center of their early evolutionary history. Of the

TABLE I  
Primate Diversity by Family

Family	Common name(s)	Distribution	Genera	Species	Taxa
Suborder Prosimii					
Cheirogaleidae	Dwarf, mouse, and fork-marked lemurs	Madagascar	5	9	12
Megaladapidae	Sportive lemurs	Madagascar	1	7	7
Lemuridae	Ring-tailed, gentle, brown, crowned, red-bellied, mon-goose, and ruffed lemurs	Madagascar	4	10	19
Indriidae	Indri, avahi, and sifakas	Madagascar	3	6	12
Daubentonidae	Aye-aye	Madagascar	1	1	1
Lorisidae	Angwantibos, lorises, pottos, and galagos	Africa, Asia	8	20	44
Tarsiidae	Tarsiers	Asia	1	5	11
			23	58	106
Suborder Anthropoidea					
Callitrichidae	Marmosets, tamarins, and Goeldi's monkey	Neotropics	5	37	58
Cebidae	Owl; titi; squirrel; capuchin; saki; bearded saki; howling; spinder and woolly monkeys; and uakaris and muriquis	Neotropics	11	62	147
Cercopithecidae	Macaques, baboons, mangabeys, guenons, colobus, leaf monkeys, langurs, and shub-nosed and proboscis monkeys	Africa, Asia	19	101	273–283
Hylobatidae	Gibbons and siamangs	Asia	1	11	28
Pongidae	Chimpanzees, pygmy chimpanzees, gorillas, and orangutans	Africa, Asia	3	5	9
Hominidae	Man	Global	1	1	1
			40	217	516–526
Total			63	275	622–632

seven extant prosimian families, five (Lemuridae, Cheirogaleidae, Megaladapidae, Indriidae, and Daubentonidae) occur naturally only in Madagascar, where they are represented by 14 genera and at least 33 species and 51 taxa. The family Tarsiidae is only slightly more geographically widespread, represented in Indonesia and the Philippines by 1 genus, 5 species, and 11 taxa. The 8 genera, 20 species, and 44 taxa of the Lorisidae are distributed throughout mainland Africa, India, and Southeast Asia.

The anthropoid primates are a much more diverse group than the prosimians, having almost twice the number of genera and four times the number of species and total taxa. Of the six anthropoid families, two (Callitrichidae and Cebidae) occur only in the New World tropics, two (Cercopithecidae and Pongidae) occur throughout much of Africa and Asia, one (Hylobatidae) is restricted to Asia, and the last (Hominidae), represented by our own species, is global in its distribution. New World nonhuman primates comprise 16 genera, 99 species, and 205 taxa. By comparison, the Old World monkeys (subfamilies Cercopithecinae and Colobinae), although much more widely distributed, are only somewhat more diverse, comprising 19 genera, 101 species, and 273–283 taxa overall. The lesser apes, which include the gibbons and siamangs, are a relatively small group that includes a single genus, 11 species, and 28 taxa. The great apes are even less diverse with 3 genera, 5 species, and only 9 taxa overall. Finally, the human primate is the sole representative of a morphologically variable but monotypic family. Accordingly, for the remainder of this article, we use “primate” only in reference to nonhuman taxa unless otherwise specified.

Examining global primate diversity from a regional perspective, it can be seen that it is by no means evenly distributed. For example, the remaining tropical forests of Madagascar are dwarfed by the extensive tropical forests of Africa, Asia, and Central and South America, but Madagascar has by far the densest concentration of primate diversity anywhere on Earth within an area of 587,045 km<sup>2</sup>. Although usually considered part of the Ethiopian zoogeographical region, Madagascar is in many respects a zoogeographic region by itself, and especially so with respect to its unique primate fauna. All of Madagascar's primates are endemic, with the exception of the mongoose lemur (*Eulemur mongoz*) and the brown lemur (*Eulemur fulvus*), which also occur on the nearby Comoro Islands but were almost certainly introduced there by man.

With the exception of the Barbary macaque (*Macaca sylvanus*), all primates of sub-Saharan Africa and nearby islands (e.g., Zanzibar and Bioko) occur within the Ethi-

opian zoogeographical region. They include 20 genera, at least 64 species, and 190–200 taxa overall. If we consider Madagascar in combination with the African continent, the entire region harbors a significant 38% of global primate diversity at the species level but an even more impressive 54% of higher order (beta) diversity at the generic level.

Within the neotropics, primates occur from southern Mexico through Central America and northern South America and as far as southern Brazil, northern Argentina, and Paraguay (but not Chile). Native primate populations also occur on Trinidad; there are several introduced populations of African primates on the islands of St. Kitts, Nevis, Barbados, and Grenada; and both New World and Old World species have been introduced to the island of Puerto Rico. Sixteen genera and 205 taxa are found within the neotropics—amounts comparable to those for Africa. However, the 99 primate species of the neotropics are the most for any single major region, account for 36% of global primate diversity, and are approximately equivalent to the number of species inhabiting both Africa and Madagascar.

Asian primates are found mainly in the Oriental zoogeographical region and in the southeastern portion of the Palearctic and in Wallacea (the transition zone between the Oriental and Australian regions). In southern Asia primates are widely distributed from the Indian subcontinent and the island of Sri Lanka, throughout Southeast Asia as far as the Philippines and the Indonesia islands of Halmahera and Sulawesi, to central and north Asia from Afghanistan through southern China (including the islands of Hainan and Taiwan) to Japan. In contrast to Africa (excluding Madagascar) and the neotropics, where primates are basically continental, Asian primates are found in large numbers on islands as well. The region is home to 13 genera, 71 species, and 176 taxa, a slightly lower level of generic diversity by comparison to the neotropics and Africa but a level of species diversity approximately equal to that of Africa.

Ninety-two of the world's 192 sovereign nations have wild primate populations; the 7 countries with the highest primate species diversity are listed in Table II. Brazil, with 77 species, is by far the leader and accounts for slightly more than three-fourths of all Neotropical primate species. Together, the top four countries (Brazil, Democratic Republic of the Congo, Indonesia, and Madagascar), which also represent the world's four major primate habitat regions, account for 181 or 182 species, or approximately two-thirds of all living primates.

Furthermore, three of the top four countries for primate diversity (Madagascar, Brazil, and Indonesia) also are at the top of the list of the world's top countries

TABLE II  
World's Top Seven Countries for Primate Species Diversity  
(>30 Species)

Country	Species	Genera
Brazil	77	16
Indonesia	36	9
Madagascar	34	14
Democratic Republic of the Congo	33–34	18
Peru	32	12
Cameroon	31	18
Colombia	31	12

for primate endemism (Table III). Madagascar is by far the international superstar with 34 unique species, 14 unique genera, and five unique families, all representing 100% levels of endemism. Primate faunas for the following two countries on the list, Brazil and Indonesia, are approximately 50% endemic. Although Brazil has more endemic species (39) than Madagascar, it covers an area almost 15 times as large, and only 2 (*Leontopithecus* and *Brachyteles*) of its 16 primate genera (13%) are endemic (one-seventh the number for Madagascar). Indonesia is a distant third on the list with 19 endemic species and 1 endemic genus (*Simias*), after which the number of endemic species decreases precipitously and no other country can claim an endemic primate genus.

The key point is that, within the broad geographic regions that provide critical habitat for wild populations of nonhuman primates, there are a handful of countries that harbor a disproportionately large share of the world's primate fauna. The four "megadiversity" countries—Brazil, Democratic Republic of the Congo, Indonesia, and Madagascar—must therefore rank among the highest global priorities for conserving primates.

TABLE III  
World's Top Six Countries for Primate Endemism

Country	Endemic species	%	Endemic genera	%
Madagascar	33	100	14	100
Brazil	39	51	2	13
Indonesia	19	53	1	11
Colombia	4	13	0	0
Peru	3	9	0	0
Democratic Republic of the Congo	3	8	0	0

Furthermore, if we extend the analysis to consider subspecies, we find that other countries will rise to the top of the priority list. Consider India for example. With only 15 primate species and three endemics, it is well behind many other countries on the previous lists. However, of India's 34 or 35 primate taxa, 22 are endemic, and the additional 19 endemic subspecies represent important wild populations that should be considered when establishing conservation priority rankings.

### III. THREATS TO PRIMATES

Threats to the survival of nonhuman primates are easily divided into three major categories: habitat destruction, hunting for food and a variety of other purposes, and live capture for export or local trade (Mittermeier *et al.*, 1986). The effects of these threats vary significantly from species to species and from region to region and are influenced by the extent of remaining habitat, human population densities, the presence of roads and large rivers, the nature and degree of human activity within the range of a particular species, local hunting traditions, the size and desirability of different species as food items or as sources of other products useful to man, the demand for a given species in research or the pet trade, enforcement of existing wildlife laws, and regulation of commercial animal dealers. However, one or more of the three major threats affect almost all primate populations.

#### A. Habitat Destruction

On a global scale, habitat destruction is the principal factor contributing to the disappearance of wild primate populations. The continuing growth of the human population and its ever-expanding need for natural resources have contributed greatly to the destruction or alteration of natural habitats on an almost unimaginable scale, and nowhere has this problem been more acute than in the tropical regions of the world. More than 90% of all nonhuman primates inhabit the tropical forests of Africa, Asia, and South and Central America, and these forests are being cut at a rate of more than 10 million ha per year (Bryant *et al.*, 1997).

The immediate effects of habitat destruction on nonhuman primates vary significantly from one region to the next. For example, in Madagascar and the Atlantic forest region of eastern Brazil, so little suitable forest habitat remains that any further loss constitutes a grave threat to primates and other wildlife. In contrast, in the vast forest regions of Amazonia and the Congo Basin,

which along with the island of New Guinea represent two of the three remaining major tropical wilderness areas of the planet, the effects of habitat destruction are only starting to be felt.

## B. Hunting for Food and Other Purposes

The hunting of primates by human populations occurs for a variety of reasons, but by far the most important is to acquire food. Although primate hunting is prohibited by law in many countries, enforcement of such protective legislation is typically rare and sometimes nonexistent in the remote areas where this activity almost always occurs.

Hunting of primates as a source of food is a significant threat in at least three parts of the world: the Amazon region of South America, west Africa, and central Africa. In each region, primates are among the animals most frequently hunted and they are regularly sold in markets, except where this is prohibited by law. However, even in areas in which primate hunting is common, it by no means affects all species equally. In Amazonia, for example, the larger monkeys, such as *Lagothrix*, *Ateles*, *Alouatta*, and *Cebus*, are heavily hunted and among the most desirable food species, whereas smaller monkeys such as *Saguinus* and *Saimiri* are rarely shot for food because they barely provide enough meat to recompense the hunter for the cost of his or her shotgun shell. As a result of better law enforcement in recent decades in countries such as Brazil, bushmeat hunting is no longer considered a serious threat to wild primate populations. The situation in central and west Africa, in contrast, has quickly reached a critical stage because bushmeat hunters in these regions are being paid by logging companies to shoot large quantities of primates and other forest wildlife to feed work crews.

In areas where the hunting of primates for food is common, it can sometimes represent a threat even more severe than forest destruction. For example, in some parts of Amazonia there are large tracts of primary forest remaining in which populations of *Lagothrix*, *Ateles*, *Alouatta*, and *Cebus* have effectively been exterminated by excessive hunting (Mittermeier and Coimbra-Filho, 1977; Soini, 1982). This is sometimes referred to as the "empty forest syndrome" and appears to be exactly what is currently happening in central and west Africa, where it includes both monkeys and apes. Of course, in areas where food hunting and deforestation are both prevalent, populations of all forest primates and other game species can disappear quickly.

It is important to note that, in some parts of the

world, religious restrictions or other cultural factors prohibit (or inhibit) the killing and eating of primates. In India, for example, primates are rarely hunted for food because they are linked to the monkey god Hanuman, which occupies an important role in the Hindu religion, whereas in strictly Muslim countries primates are not eaten because their flesh is considered unclean and unfit for human consumption. Indeed, in India, Hindu people refuse to kill rhesus monkeys and resist translocating them even when populations become so high that they constitute a menace to humans. In other countries, such as Madagascar, local taboos may exist against eating certain primates (e.g., *Indri*), whereas other species (e.g., *Eulemur* and *Varecia*) may be the most popular food items for a given tribe or village.

Primates are also hunted to supply many other products in addition to food: traditional medicines, bait, body parts for ornamentation, and trophies. Primate hunting to supply medicinal products may be nothing more than a by-product of food hunting in most cases, and it usually involves the use of specific body parts for their supposed medicinal value. In south India, for example, the meat of the Nilgiri langur (*Presbytis johnii*) and the endangered lion-tailed macaque (*Macaca silenus*) is regarded as an aphrodisiac and thought to contain other medicinal properties. The blood of leaf monkeys, such as Phayre's langur (*Presbytis phayrei*) in Thailand, is believed to impart vigor to the drinker, especially when mixed with local whiskey. Also, in various South American countries, drinking from the cup-shaped hyoid apparatus of an adult male howling monkey (*Alouatta*) is reported to cure goiters and stuttering as well as to ease a mother's labor pains during childbirth. Although the hunting of primates for medicinal purposes is considered a relatively minor factor overall in the global decline of wild primate populations, when it involves endangered species, such as in the case of India's lion-tailed macaque or some of the Southeast Asian colobines (*Trachypithecus* and *Pygathrix*), it can be a serious threat.

Primates are also shot to provide bait for capturing and killing other animals, mainly in remote areas of the Amazon region. There, spotted cat hunters preferentially shoot larger monkeys such as *Lagothrix* and *Ateles* to bait crude wooden traps set for jaguars and ocelots. Any number of Amazonian primates may also be shot for fish or turtle bait, and in Sri Lanka monkeys often serve as bait for crocodiles. Although the use of primates for bait is a relatively minor threat, it can and does add to the pressures faced by overexploited, large-bodied taxa such as *Lagothrix* and *Ateles*.

In some countries, primates may be killed for their

skins or to provide other body parts used in ornamentation. Perhaps the most striking case of this is in Africa, where the skins of black-and-white colobus (*Colobus guereza*) and related species have been used to fashion cloaks and headdresses for native African peoples but have also figured significantly in the international fur trade. For example, in 1899 a reported 223,599 monkey skins were auctioned in London, and at least 2.5 million were probably exported to Europe between 1880 and 1900, especially to Germany, where they were used to make capes, muffs, and rugs. As recently as the early 1970s, colobus monkey rugs were common in East African tourist shops and colobus coats were being sold in Europe and Japan.

Throughout much of Amazonia, tourist shops still offer stuffed monkeys, monkey skulls, monkey-skin hats, monkey-tail dusters, and necklaces fashioned from monkey teeth, bones, hands, feet, and tails. However, these products are typically available on a small scale and almost always as a by-product of hunting for food.

Nonetheless, the demand for primate body parts for sale to tourists can be a serious matter if it involves endangered species. The most striking example of this is the slaughter of mountain gorillas (*Gorilla gorilla beringei*) in Rwanda and the Democratic Republic of the Congo (formerly Zaire) which produces hands and skulls for sale to European tourists (Fossey, 1983). Although rare, this practice still occurs despite effective, long-term conservation programs in this region.

Hunting primates for sport is fortunately rare and a minor threat to wild populations. It appears to be most prevalent near logging camps and within military zones in remote areas of developing countries, where arms are plentiful and law enforcement is basically nonexistent. Children armed with slingshots and air rifles are often among the worst offenders. More prestigious trophy hunting has also played a role (albeit a minor one) in primate decline. Species such as the gorilla were especially desirable quarry for nineteenth-century and early twentieth-century trophy hunters, and the tales of their exploits are recounted in many books. On the whole, however, sport hunting must be considered a very minor factor unless an endangered species is involved, in which case the activity is almost always illegal as well.

A final reason for hunting primates considered here is because they are sometimes considered agricultural pests; for some African and Asian species, this can represent a significant drain on wild populations. The most striking example is that of government-sponsored "monkey drives" that were common in Sierra Leone several decades ago. Eleven of the country's 14 primate

species were routinely shot or driven into nets and clubbed to death during such drives; only 3 species were considered harmless to farm crops. According to government records, approximately 250,000 monkeys were destroyed in such drives between 1949 and 1952, and these were only the ones actually counted. Bounties were paid for primate heads or tails, and there was no control over the species killed.

The major crop raiders are usually the more adaptable and widespread species, such as the savanna baboons (*Papio* spp.) in Africa and the macaques (*Macaca* spp.) in Asia, but there are also instances of orangutans being killed for raiding fruit trees and gorillas being killed for destroying crops. The only Neotropical species regarded as agricultural pests are the capuchins (*Cebus* spp.), whose common names sometime reflect their crop-raiding habits. For example, the common name for the tufted capuchin (*Cebus apella*) in Colombia is maicero and one of the Surinamese names for the weeper capuchin (*Cebus olivaceus*) is nyan-karu mongi, both of which translate as "corn-eater."

It is difficult to assess how much damage primates actually do to crops in different parts of the world. It is equally difficult to determine how effective pest control efforts have been or to what degree they have contributed to the decline of wild primate populations. However, as primate habitats continue to be encroached upon, resulting in shortages of other food sources, it is likely that the more adaptable primate species will continue to raid crops and perhaps become more dependent on them as a food source. This, unfortunately, will likely result in increased conflict between man and nonhuman primates.

### C. Live Capture of Primates

Primates have routinely been captured alive for export (the international trade to supply zoos and for biomedical research and pharmaceutical testing) or to serve local pet trades. The height of the international primate trade began at the end of the 1950s and continued through the early 1960s, during which time hundreds of thousands of monkeys were taken from the wild each year (Mack and Mittermeier, 1984). The trade consisted largely of rhesus macaques (*Macaca mulatta*) exported from India and used in laboratory tests as part of the effort to develop a vaccine for polio and squirrel monkeys (*Saimiri sciureus*) imported by the United States from several Amazonian countries. Subsequently, the imposition of export bans by habitat countries, import restrictions by user countries, and a decreased demand from biomedical research and zoological parks contrib-

uted to a significant decline in the international traffic in primates.

In 1982, in recognition of the serious effect that live capture for export can have on wild primate populations, the International Union for Conservation of Nature and Natural Resources (IUCN)/ Species Survival Commission (SSC) Primate Specialist Group prepared a *Policy Statement on the Use of Primates for Biomedical Purposes*, which includes the recommendation that endangered, vulnerable, or rare species be considered for use in biomedical research projects only if they are obtained from existing, self-sustaining captive breeding colonies. This policy statement was subsequently adopted by the World Health Organization and the Ecosystem Conservation Group of the United Nations, which includes UNESCO, the Food and Agricultural Organization (FAO), UNEP, and IUCN. It is still valid today.

#### IV. CONSERVATION STATUS OF PRIMATES

The most recent, comprehensive conservation status assessment of the world's primates is included in the *1996 IUCN Red List of Threatened Animals* (Baillie and Groombridge, 1996), a collaborative effort of the IUCN Species Survival Commission, the World Conservation Monitoring Centre, and BirdLife International. This document differs significantly from past red lists in its use of new categories and criteria for threat. All primate taxa were included in this assessment and have been identified as Threatened (a designation which includes the categories Critically Endangered, Endangered, and Vulnerable), Lower Risk: Conservation Dependent, Lower Risk: Near Threatened, Lower Risk: Least Concern, Extinct, Extinct in the Wild, and Data Deficient. In general terms, a taxon is defined as

1. Critically Endangered if the extent of its occurrence is estimated to be less than 100 km<sup>2</sup>, if its population is estimated to be less than 250 individuals, and/or quantitative analysis indicates the probability of extinction in the wild is at least 50% within 10 years or three generations.

2. Endangered if the extent of its occurrence is estimated to be less than 5000 km<sup>2</sup>, if its population is estimated to number less than 2500 individuals, and/or if quantitative analysis shows the probability of extinction in the wild is at least 20% within 20 years or five generations.

3. Vulnerable if the extent of its occurrence is estimated to be less than 20,000 km<sup>2</sup>, if its population is estimated to number less than 10,000 individuals, and if quantitative analysis shows the probability of extinction in the wild is at least 10% within 100 years.

As a result of this assessment, 204 (approximately one-third) of the world's approximately 620 primate taxa are currently considered Critically Endangered, Endangered, or Vulnerable. Of these, 104 taxa (approximately 17%) are listed as Critically Endangered or Endangered—34 in the neotropics, 29 in Africa, 17 in Madagascar, and 24 in Asia (Tables IV and V). Nine genera (*Allocebus*, *Varecia*, *Indri*, *Daubentonia*, *Leontopithecus*, *Brachyteles*, *Simias*, *Pan*, and *Gorilla*) are considered Endangered, as is the monotypic family Daubentonidae.

#### V. PRIMATE CONSERVATION DURING THE PAST 20 YEARS

The IUCN was established in 1948 to promote and carry out scientifically based action for the conservation and sustainable use of living natural resources. IUCN enrolls sovereign states, governmental agencies, research institutions, and nongovernmental organizations to conserve the world's natural heritage. The SSC was founded in 1949 and is the largest of IUCN's six commissions, with more than 7000 volunteer member scientists, field researchers, government officials, and conservation leaders from 88 countries. SSC works principally through its more than 100 specialist groups, of which the Primate Specialist Group is one of the largest.

The founding mission of the Primate Specialist Group is to maintain the current diversity of the order Primates, with dual emphasis on (i) ensuring the survival of endangered and vulnerable species wherever they occur and (ii) providing effective protection for large numbers of primates in areas of high primate diversity and/or abundance.

Although activities under way in many parts of the world make it inevitable that a portion of the world's forests and primates living in them will disappear, the role of the Primate Specialist Group is to minimize this loss wherever possible by

1. Setting aside special protected areas for critically endangered, endangered, and vulnerable species
2. Creating national parks and reserves in areas of high primate diversity and/or abundance

TABLE IV  
Critically Endangered and Endangered Primates

Critically Endangered	Endangered
Hairy-eared dwarf lemur ( <i>Allocebus trichotis</i> )	White-collared lemur ( <i>Eulemur fulvus albocollaris</i> )
Slater's lemur ( <i>Eulemur macaco flavifrons</i> )	Black-and-white ruffed lemur ( <i>Varecia variegata variegata</i> )
Alaotran gentle lemur ( <i>Hapalemur griseus alaotrensis</i> )	Indri ( <i>Indri indri</i> )
Golden bamboo lemur ( <i>Hapalemur aureus</i> )	Diademed sifaka ( <i>Propithecus diadema diadema</i> )
Broad-nosed gentle lemur ( <i>Hapalemur simus</i> )	Milne-Edwards' sifaka ( <i>Propithecus diadema edwardsi</i> )
Red ruffed lemur ( <i>Varecia variegata rubra</i> )	Coquerel's sifaka ( <i>Propithecus verreauxi coquereli</i> )
Silky sifaka ( <i>Propithecus diadema candidus</i> )	Aye-aye ( <i>Daubentonia madagascariensis</i> )
Perrier's sifaka ( <i>Propithecus diadema perrieri</i> )	Buffy-tufted-ear marmoset ( <i>Callithrix aurita</i> )
Tattersall's sifaka ( <i>Propithecus tottersalli</i> )	Buffy-headed marmoset ( <i>Callithrix flaviceps</i> )
Crowned sifaka ( <i>Propithecus verreauxi coronatus</i> )	Golden-headed lion tamarin ( <i>Leontopithecus chrysomelas</i> )
Black-faced lion tamarin ( <i>Leontopithecus caissara</i> )	Bicolored tamarin ( <i>Saguinus bicolor bicolor</i> )
Black lion tamarin ( <i>Leontopithecus chrysopygus</i> )	Cotton-top tamarin ( <i>Saguinus oedipus</i> )
Golden lion tamarin ( <i>Leontopithecus rosalia</i> )	Coiba Island howling monkey ( <i>Alouatta coibensis coibensis</i> )
Red-handed howling monkey ( <i>Alouatta belzebul ululata</i> )	Night monkey ( <i>Aotus lemurinus griseembra</i> )
Coiba Island howling monkey ( <i>Alouatta coibensis trabeata</i> )	White-bellied spider monkey ( <i>Ateles belzebuth brunneus</i> )
Northern brown howling monkey ( <i>Alouatta fusca fusca</i> )	Hybrid spider monkey ( <i>Ateles belzebuth hybridus</i> )
Brown-headed spider monkey ( <i>Ateles fusciceps fusciceps</i> )	Grizzled spider monkey ( <i>Ateles geoffroyi griseus</i> )
Azuro spider monkey ( <i>Ateles geoffroyi azuerensis</i> )	Panamanian spider monkey ( <i>Ateles geoffroyi panamensis</i> )
Southern muriqui ( <i>Brachyteles arachnoides</i> )	White-whiskered spider monkey ( <i>Ateles marginatus</i> )
Northern muriqui ( <i>Brachyteles hypoxanthus</i> )	Bald uakari ( <i>Cacajao calvus calvus</i> )
Northern Bahian brown titi ( <i>Callicebus personatus barbarabrownae</i> )	Bald uakari ( <i>Cacajao calvus novaesi</i> )
White-fronted capuchin ( <i>Cebus albifrons trinitatus</i> )	Red uakari ( <i>Cacajao calvus rubicundus</i> )
Margarita Island tufted capuchin ( <i>Cebus apella margaritae</i> )	Black-bearded saki ( <i>Chiropotes satanas satanas</i> )
Buffy-headed tufted capuchin ( <i>Cebus xanthosternos</i> )	Central American squirrel monkey ( <i>Saimiri oerstedii oerstedii</i> )
Yellow tailed woolly monkey ( <i>Lagothrix flavicauda</i> )	White-collared mangabey ( <i>Cercocebus atys lunulatus</i> )
Colombian woolly monkey ( <i>Lagothrix lagotricha lugens</i> )	Sanje mangabey ( <i>Cercocebus galeritus sanjei</i> )
Central American squirrel monkey ( <i>Saimiri oerstedii citrinellus</i> )	Tana River mangabey ( <i>Cercocebus galeritus galeritus</i> )
Miss Waldron's red colobus ( <i>Procolobus badius waldroni</i> )	Roloway monkey ( <i>Cercopithecus diana roloway</i> )
Mentawai macaque ( <i>Macaca pagensis</i> )	Red-eared monkey ( <i>Cercocebus erythrotis erythrotis</i> )
Tonkin snub-nosed monkey ( <i>Rhinopithecus avunculus</i> )	Golden monkey ( <i>Cercopithecus mitis kandti</i> )
Delacour's langur ( <i>Trachypithecus delacouri</i> )	Preuss's monkey ( <i>Cercopithecus preussi insularis</i> )
Silvery gibbon ( <i>Hylobates moloch</i> )	Preuss's monkey ( <i>Cercopithecus preussi preussi</i> )
Nigerian gorilla ( <i>Gorilla gorilla?</i> )	Slater's guenon ( <i>Cercopithecus sclateri</i> )
Mountain gorilla ( <i>Gorilla gorilla beringei</i> )	Drill ( <i>Mandrillus leucophaeus leucophaeus</i> )
	Drill ( <i>Mandrillus leucophaeus mundamensis</i> )
	Bouvier's red colobus ( <i>Procolobus badius bouvieri</i> )
	Niger Delta red colobus ( <i>Procolobus badius epieni</i> )
	Uhehe red colobus ( <i>Procolobus badius gordonorum</i> )
	Zanzibar red colobus ( <i>Procolobus badius kirkii</i> )
	Pennant's red colobus ( <i>Procolobus badius pennanti</i> )
	Preuss's red colobus ( <i>Procolobus badius preussi</i> )
	Tana River red colobus ( <i>Procolobus badius rufomitatus</i> )
	Temminck's red colobus ( <i>Procolobus badius temmincki</i> )
	Japanese macaque ( <i>Macaca fuscata</i> )
	Moor macaque ( <i>Macaco maura</i> )
	Sulawesi black macaque ( <i>Macaca nigra</i> )
	Lion-tailed macaque ( <i>Macaca silenus</i> )
	Grizzled leaf monkey ( <i>Presbytis comata</i> )

continues

Continued

Critically Endangered	Endangered
	Douc langur ( <i>Pygathrix nemaeus</i> )
	Yunnan snub-nosed monkey ( <i>Rhinopithecus bieti</i> )
	Guizhou snub-nosed monkey ( <i>Rhinopithecus brelichi</i> )
	Pig-tailed snub-nosed monkey ( <i>Simias concolor</i> )
	Golden-headed langur ( <i>Trachypithecus poliocephalus</i> )
	Black gibbon ( <i>Hylobates concolor</i> )
	Western lowland gorilla ( <i>Gorilla gorilla gorilla</i> )
	Grauer's gorilla ( <i>Gorilla gorilla graueri</i> )
	Eastern chimpanzee ( <i>Pan troglodytes schweinfurthi</i> )
	Central chimpanzee ( <i>Pan troglodytes troglodytes</i> )
	Western chimpanzee ( <i>Pan troglodytes verus</i> )

3. Maintaining parks and reserves that already exist and enforcing protective legislation in them

4. Determining ways in which human and nonhuman primates can coexist in multiple-use areas

5. Establishing conservation-oriented captive breeding programs for threatened taxa

6. Ending illegal and otherwise destructive traffic in primates

7. Ensuring that research institutions using primates are aware of conservation issues and the status of species they use, that they use primates as prudently as possible, and that they make every attempt to breed most or all of the primates they require

8. Creating public awareness of the need for primate conservation and the importance of primates as a natural heritage in the countries in which they occur

In late 1977, the chairman of the Primate Specialist Group, in collaboration with group members, wrote *Global Strategy for Primate Conservation* (Mittermeier, 1978). This document was an attempt to organize primate conservation activities based on the highest international priorities and to ensure that limited funds avail-

able for primate conservation were put to the best possible use. The first draft of the *Global Strategy* included 65 projects in Africa, Asia, and South and Central America. Each project was categorized as highest priority, high priority, priority, and desirable based mainly on the status of focal species and how likely the project would be to bring about the desired conservation action. The *Global Strategy* quickly led to a substantial increase in funding for primate conservation activities, and in 1979 it led to the establishment of a special Primate Program and Primate Action Fund by the World Wildlife Fund–U.S. In addition to major projects supported as a result of this program, the Primate Action Fund provided rapid support for small primate conservation projects (ranging from \$500 to \$3000). The Primate Action Fund functioned for more than a decade, contributing several hundred thousand dollars to more than 100 projects. Other key institutions that contributed significantly to primate conservation during this period include the New York Zoological Society (now the Wildlife Conservation Society), the Fauna and Flora Preservation Society (now Fauna and Flora International), the Rare Animal Relief Effort, the Jersey Wildlife

TABLE V  
Threatened Primates by Region

Region	Taxa	Threatened	%	Critically Endangered and Endangered	%
Neotropics	202	69	34.2	34	16.8
Africa	190	41	21.1	29	14.7
Madagascar	51	35	68.6	17	33.3
Asia	176	60	34.1	24	13.6



Preservation Trust, Wildlife Preservation Trust International, and the National Geographic Society.

Almost a decade after the *Global Strategy* was launched, the first regional primate conservation action plans were prepared by the IUCN/SSC Primate Specialist Group. First to be published was the *Action Plan for African Primate Conservation: 1986–90* (Oates, 1986), which was quickly followed by the *Action Plan for Asian Primate Conservation: 1987–91* (Eudey, 1987) and several years later by *Lemurs of Madagascar, An Action Plan for Their Conservation: 1993–1999* (Mittermeier *et al.*, 1992). The last plan to appear was *African Primates: Status Survey and Conservation Action Plan* (Oates, 1996), which is an update of the 1986 document. These action plans have effectively focused conservation activities in three or four major regions in which primates occur, and they are useful measures with regard to the success of proposed strategies.

The first vehicle for regular and effective communication among the world's primate conservationists was the *IUCN/SSC Primate Specialist Group Newsletter*, which was launched in 1981. Changed to *Primate Conservation* in 1985, it has appeared on more or less an annual basis ever since. In addition, the four regional sections of the Primate Specialist Group subsequently began publishing their own periodic newsletters to meet the increasing need for more timely information. *Asian Primates* appeared in 1991, *Neotropical Primates* and *Lemur News* in 1993, and *African Primates* in 1995. In combination, these newsletters have significantly increased the amount, quality, and timeliness of information available to primate conservationists throughout the world.

## VI. OUTLOOK FOR THE FUTURE

In the future, there is a need to sustain conservation activities based on recommendations of the original *Global Strategy for Primate Conservation* and subsequent regional action plans as well as to increase the focus on those primate taxa most seriously threatened with extinction. The twentieth century ended without a single primate taxon being lost—an enviable record indeed considering the number of reptiles, birds, and other mammals known to have disappeared during this period. However, several species and subspecies are in serious jeopardy.

Much of the groundwork for developing a more focused conservation strategy has already been done. The *1996 IUCN Red List of Threatened Animals* provides a

good starting point for identifying the highest priority taxa. To those that are considered Critically Endangered and Endangered, we expect that the Primate Specialist Group will add several more species and subspecies and then consider factors such as taxonomic uniqueness to establish priority rankings for conservation action. The results of this analysis are expected to be presented in the Year 2000 in the form of a global action plan for the world's most endangered primates.

With such a plan in hand, serious work can begin to amass both the human and the financial resources needed for implementation. Fortunately, several new sources of support for primate conservation have materialized during the past decade. Although the World Wildlife Fund–U.S. Primate Program no longer exists, many other traditional nongovernmental sources still offer grants for field, captive, and laboratory programs; academic institutions continue to provide funds for primate field studies that have significant conservation impact; government-supported efforts such as the Indo–U.S. Primate Project provide excellent models for international cooperation; and an increasing number of zoos have joined forces to focus on regional primate faunas, generating funds not only for captive breeding programs but also for *in situ* projects. In addition, at least two new significant sources of philanthropic support dedicated to primates were established in the 1990s: Primate Conservation, Inc., and the Margot Marsh Biodiversity Foundation. Together, these organizations and agencies represent the core of funding necessary to move ahead with a global action plan for the most endangered primates. It is hoped that such a plan will help uncover new sources of support as well.

## See Also the Following Articles

CAPTIVE BREEDING AND REINTRODUCTION •  
 CONSERVATION EFFORTS, CONTEMPORARY • ENDANGERED  
 MAMMALS • MAMMALS, CONSERVATION EFFORTS FOR •  
 NATURAL RESERVES AND PRESERVES • ZOOS AND  
 ZOOLOGICAL PARKS

## Bibliography

- Baillie, J., and Groombridge. (1996). *1996 IUCN Red List of Threatened Animals*. IUCN Species Survival Commission, the World Conservation Monitoring Centre and BirdLife International, Gland, Switzerland, Cambridge, UK.
- Bryant, D., Nelson, D., and Tangle, L. (Eds.) (1997). *The Last Frontier Forests: Ecosystems and Economies on the Edge*. World Resources Institute, Washington, DC.

- Eudey, A. A. (1987). *Action Plan for Asian Primate Conservation: 1987–91*. IUCN/SSC Primate Specialist Group, Gland, Switzerland.
- Fossey, D. (1983). *Gorillas in the Mist*. Houghton Mifflin, Boston.
- Galdikas, B. M. F. (1995). *Reflections of Eden: My Years with the Orangutans of Borneo*. Little, Brown, Boston.
- Goodall, J. (1968). The behaviour of free-living chimpanzees in the Gombe Stream area. *Anim. Behav. Monogr.* 1(3), 161–311.
- Mack, D., and Mittermeier, R. A. (1984). *The International Primate Trade (Volume 1)*. TRAFFIC, Washington, DC.
- Mittermeier, R. A. (1978). *A Global Strategy for Primate Conservation*. IUCN/SSC Primate Specialist Group, Washington, DC.
- Mittermeier, R. A., and Coimbra-Filho, A. F. (1977). Primate conservation in Brazilian Amazonia. In *Primate Conservation* (Prince Rainier of Monaco and G. H. Bourne, Eds.), pp. 117–166. Academic Press, New York.
- Mittermeier, R. A., Oates, J. F., Eusey, A. E., and Thornback, J. (1986). Primate conservation. In *Comparative Primate Biology, Volume 2A: Behavior, Conservation and Ecology* (G. Mitchell and J. Erwin, Eds.), pp. 3–72. A. R. Liss, New York.
- Mittermeier, R. A., Konstant, W. R., Nicoll, M. E., and Langrand, O. (1992). *Lemurs of Madagascar: An Action Plan for Their Conservation (1993–1999)*. IUCN/SSC Primate Specialist Group, Gland, Switzerland.
- Oates, J. F. (1986). *Action Plan for African Primate Conservation: 1986–1990*. IUCN/SSC Primate Specialist Group, Gland, Switzerland.
- Oates, J. F. (1996). *African Primates: Status Survey and Conservation Action Plan*. IUCN/SSC Primate Specialist Group, IUCN–The World Conservation Union, Gland, Switzerland.
- Schaller, G. (1963). *The Mountain Gorilla: Ecology and Behavior*. Univ. of Chicago Press, Chicago.
- Soini, P. (1982). Primate conservation in Peruvian Amazonia. *Int. Zoo Yearbook* 22, 37–47.





# PROPERTY RIGHTS AND BIODIVERSITY

Susan S. Hanna  
*Oregon State University*

---

- I. The Context of Property Rights
  - II. Forms of Property Rights
  - III. Functions of Property Rights
  - IV. Evolution of Property Rights
  - V. Property Rights and Biodiversity Protection
  - VI. Conclusions
- 

## GLOSSARY

**discount rate** Rate at which future benefits and costs are discounted.

**economic rent** Payment to resources in excess of costs.

**ecosystem services** Contributions of ecosystem components through genetics, information, and reproduction.

**externality** An effect—either a cost or a benefit—external to the generating activity.

**free rider** One who enjoys the benefit of a good or service without paying the cost.

**institutions** Organizational structures that shape human interactions.

**path dependence** The influence of previous actions on the present state.

**prisoner's dilemma** Uncertainty about others' behavior leading to choices that are individually rational but collectively irrational.

**property rights systems** Bundles of property rights with their associated rules and obligations.

**public goods** Goods and services which if supplied to one person can be supplied to additional people at no extra cost.

**rent seeking** The attempt to gain advantage through claims on resource surpluses.

**time horizon** Time period over which future benefits and costs are considered.

**tragedy of the commons** Overuse of resources resulting from open access and the incomplete accounting for costs.

**transactions costs** Costs other than price associated with the trade of environmental goods and services: information, negotiation, decisions, and enforcement.

---

**PROPERTY RIGHTS** to natural resources define privileges and responsibilities in the use of environmental goods and services. They specify the way people are to behave toward one another as they use environmental resources. This chapter describes the form and function of property rights in general and discusses the relation of property rights to biodiversity in particular. This discussion summarizes what is known about the potential and limitations of property rights to protect biodiversity. It also examines the considerable uncertainty that exists with respect to the design of property rights for biodiversity protection.

## I. THE CONTEXT OF PROPERTY RIGHTS

Property rights define the conditions that guide and control the human use of the natural environment. They establish the terms under which people use and sustain the capacity of the environment to generate a continuing flow of goods and services. Property rights are a means by which people interact with their environment and control their behavior toward one another. They embody the expectations people have about natural resources and influence the way people make resource use decisions. Property rights link humans to each other and to natural systems through these expectations and decisions.

### A. Scope

Property rights to natural resources control both use and conservation. Their scope may be either individual species or areas of land and water in which species live. The scope of property rights also includes different types of use. At present, most systems of property rights are designed for direct uses, such as catching fish for food, but they may also include indirect uses, such as the right of the public to enjoy populations of whales or to protect endangered species like bald eagles. Property rights are almost never defined for unused species or for communities of species. The idea of using property rights for the protection of not only single species but also biodiversity is a new and broader application of their accustomed use.

### B. Values

If resources are used, they have value. The values of ecosystems that derive from direct uses are most easily recognized—for example, the harvest of shellfish and seaweed from intertidal marine ecosystems. Their value is expressed in markets or in the subsistence they provide. But indirect use also reflects values that, although not well quantified, are important and well articulated as economic concepts. Resources have option values for the future in terms of their potential uses—for example, the future food or pharmaceutical uses of marine species. These potential uses are valued as insurance against the unknown.

Resources also have quasi-option values in future uses that may exist through preservation or be lost by irreversible actions, such as the localized genetic information in subraces of salmon that are lost as habitat is destroyed.

Finally, resources have existence values in their aesthetic or spiritual attributes, such as the enjoyment of watching an osprey catch a fish or the satisfaction of just knowing that ospreys are nesting nearby. Existence values can also be thought of as bequest values, in that the unused goods and services are available to be passed on to future generations.

An important contributor to the option value of ecosystems is ecological resilience. Resilience is the elasticity of the ecosystem; the ability of an ecosystem to absorb perturbations and continue its essential functions. Although much remains unknown about the necessary conditions for resilience, what is known is that there are keystone species that require protection and conservation to enhance both species diversity and critical ecosystem functions.

The option values of direct uses of ecosystems are relatively easy to understand and measure, because they are linked to tangible goods like seafood and timber that are traded through markets. Considerable uncertainty exists, however, about the value of services or goods that are not currently used. Measuring and accounting for the option values of ecosystem services and unknown goods is difficult because there is no market in which people can express their willingness to pay for them or be compensated for their loss.

### C. People and Biodiversity

Biodiversity can be viewed in a number of dimensions, including the diversity of species, the diversity of genetic material, and the diversity of functional roles in the ecosystem. Many of the benefits of biodiversity are public goods, in which value accrues to all. The question for humans and biodiversity is one of managing the direct and indirect effects of activities in ecosystems in order to maintain the diversity of genetic, species, and functional components.

One challenge is how to provide appropriate incentives so that people find it in their interest to promote and maintain the public good of biodiversity. A second challenge is how the rights, rules, and responsibilities that constrain resource use can be expanded from single species or physical areas to multiple species that may be distributed over wide areas. A third challenge is how to make the transition from traditional single-species commodity production types of use to new types of use that accommodate the protection of species diversity through the maintenance of ecosystem services.

### D. Ownership

The idea of ownership is well understood for goods produced by ecosystems. Overall, ownership is better

TABLE I  
A Classification of Pure Forms of Property Rights

Property rights	Rights holder	Privileges	Responsibilities
Open access	None	Capture	None
State property	Citizens	Designate management authority	Stewardship
Common property	Collective	Exclude nonmembers	Participation, maintenance
Private property	Individual	Socially acceptable uses; control of access; transfer	Compliance with laws

defined for land-based resources than for marine resources. But for both, the full range of property rights described in Table I applies. Property rights to goods from land ecosystems are usually clearly defined and range from private to public. Property rights to goods from marine ecosystems are similarly mixed and often still include open access. Property rights to goods from marine ecosystems vary from private/community/state in nearshore areas, to state property in offshore zones, to open access in international waters.

The idea of owning ecosystem services, in contrast, is less familiar. At present, ecosystem services are either considered the property of the public or of the owner of the area in which they exist. Ecosystem services are public goods, but in neither land-based nor marine ecosystems are rights to services specifically defined. One particular ecosystem service—the genetic information embodied in species—is particularly important to the question of biodiversity. Property rights to the informational properties of resources are in only the very early stages of development and application.

## II. FORMS OF PROPERTY RIGHTS

Property rights take many forms. Differences in form derive from the scope of the rights, the type of rights owner, and the privileges and responsibilities of the rights owner. The most common categories are private, public, state, and common (Table I). The types of property rights are ordered loosely along a spectrum, where the owner ranges from an individual person to no one.

### A. Open Access

Open access, *res nullius*, is the absence of property rights and has no ownership assigned. It is open to all, with

resources becoming owned only at the point of capture. The absence of rights is often, but mistakenly, referred to as “common property.” The “tragedy of the commons” metaphor is a description of the outcome of open access. It describes the use of a common-pool resource from which it is too costly to exclude people from use.

In the tragedy of the commons, individuals make choices about resource use based on their own private costs and benefits. And while the benefits of their actions accrue to them alone, the costs, in terms of the effect on the resource, are social costs spread among all users. Eventually and inevitably without some control over access, the failure to account for social costs locks people into behavior that leads to resource degradation and overuse.

### B. Common Property

Although the “tragedy of the commons” is a metaphor that is widely used, it does not really describe common property. Common property, *res communes*, is owned by an identified group of people who have the right to exclude nonowners, the duty to participate in decisions about use, and the responsibility to act as resource stewards. For example, nearshore fishing territories may be owned and managed by the residents of the adjoining coastal community. Community members as owners decide how many people have rights to use the resource, what kind of rights these may be, and what objectives they have for resource productivity.

### C. State Property

State property, *res publicae*, is owned by citizens who assign management authority and stewardship responsibility to an agency of the state. For example, U.S. fish and wildlife resources are the property of the citizens of individual states or of the country as a whole. The citizens designate the fish and wildlife agencies of individual states and the federal government to manage these resources in their name. The agencies may grant rights of use to individual users or communities of users, but the resources remain owned by the citizens at large.

### D. Private Property

Private property, *res privatae*, assigns ownership to named individuals including legal individuals, such as corporations, guaranteeing to those owners a bundle of rights about access and use. Although individuals have the greatest autonomy under private property, this form of property rights is also stunted by prohibitions against

unacceptable uses, such as activities that pollute. For example, individuals may own land and the fish and wildlife on that land, but the types of use to which they can put either the land or the wildlife may be constrained by law.

### E. Property Rights over Goods

The property rights listed in Table I have one attribute in common: they have, in practice, been applied almost exclusively to environmental goods. We think of resources as natural capital that create value through the size of standing stock and through the flow of resources from that stock. We are familiar with property rights over the stock and flow components of natural capital. For example, in marine ecosystems rights can be assigned to the resource only on capture (competitive fishing), to the standing stock itself (territories for sedentary species), or to the flow of goods from the standing stock (individual fishing quotas). These are the tangible components of ecosystems.

### F. Property Rights over Services

There are other intangible components of ecosystems that do not fall under these definitions of rights. For example, species provide services to the ecosystem and to humans through genetic information, reproduction, and contribution to critical ecosystem functions through roles such as predator or prey. Species and groups of species also provide additional value to humans through the information they contain and the pleasure they provide.

Property rights are poorly defined for ecosystem services such as genetic information. The idea of intellectual property rights over discoveries leading from information and ideas is a familiar one. Nations have patent and copyright systems that protect rights of ownership over goods and services that flow from the application of intellectual capital; these include designs, published works, and medicines. But the focus of intellectual property rights is on the discovery or creation that results from information, not on the information itself.

The lack of systems of property rights for genetic information is a critical issue for its protection. The information stored in biodiversity could lead to discoveries and patented products, but for the potential investor, the necessary protections to encourage investment are not in place. Since rights to the use of genetic information are unspecified, incentives for its protection are also absent. Biodiversity is treated as an open access

resource, with the familiar “tragedy of the commons” result.

### G. Optimal Forms of Property Rights

Biodiversity is a public good at both national and international scales. Developing the means to protect its value is a critical challenge to property rights. Some of the same conclusions that apply to property rights and the protection of single resources apply as well to multiple resources and biodiversity. The evidence of the applied property rights literature indicates that no single type of property rights can be a remedy for all needs of resource protection. It depends on the resource management objectives and on the context.

People often advocate a particular form of property rights as being best suited to manage and sustain natural resources. For example, arguments have been made for private property rights to natural resources to provide incentives for maintaining the flow of resource goods and services into the future. The idea is that private property rights would keep people from the “prisoner’s dilemma,” in which ignorance over the behavior of others leads to the collectively irrational outcome of resource overuse. The argument for universal private property ignores the differences in context in which they might be applied. It also disregards possibilities for cooperation and collective action that provide assurance about others’ behavior and that exist under other types of rights systems.

In recent years counter-arguments have arisen for the superiority of collective ownership over private ownership. These arguments are often framed in the context of community-based resource management. The idea behind this argument is that collective resource ownership and management is superior in terms of providing incentives for stewardship, providing collective access to resource benefits, and fairness in resource outcomes.

Disputing each of these arguments for the superiority of a single type of property right is a large body of research that has demonstrated that there is no particular form of property right that is superior in all cases. Each type of property right—with the exception of open access, which is the absence of rights—can either succeed or fail in sustaining resources. What matters is how well the property right system fits within the ecological, economic, and social context and how well the form of the right reflects the type of use.

The fundamental problem for biodiversity protection is that much of the biodiversity value falls outside the realm of direct use. Direct uses are easiest to monitor

and measure, control and trade. Indirect uses such as genetic information and aesthetic appreciation are less visible because their contribution is more diffuse over space and time. Because these resource values are seldom owned, bought, or sold, they are often at a competitive disadvantage in a market economy, leaving them undervalued and overused. This leads to the common approach of protecting biodiversity not through property rights but by removing them from human influence in sanctuaries, refuges, or reserves.

### III. FUNCTIONS OF PROPERTY RIGHTS

Property rights have several functions. They delineate the population of legitimate owners. They specify the allowable actions of these owners and their associated responsibilities so that expectations are consistent and enforcement of rules is possible. And by setting consistent expectations they reduce uncertainty about others' behavior. In the larger sense, they connect the pieces of the natural system to the pieces of the human system.

The basic functions of managing resources—coordinating users, enforcing rules, and adapting to changing environmental conditions—cannot be met without a system of property rights. The way that property rights function in any particular context determines whether the natural and human system will interact in compatible or conflicting ways. To be compatible with long-term biodiversity protection, systems of property rights must function in a way that deals with uncertainty, externalities, transactions costs, and scale.

#### A. Uncertainty

Uncertainty is endemic in natural systems. We lack knowledge about ecosystem structure and ecosystem condition. We often lack assurance about how others will behave, or we lack confidence about the future.

For example, there are large gaps in basic knowledge about biodiversity; about threshold levels of protection, the measurement of ecosystem function, and the definition of goals for biodiversity. In marine ecosystems relatively little is known about ecosystem composition, links between species, food, and reproductive requirements, critical ecological processes, and trade-offs among species.

These uncertainties affect how well property rights perform. Uncertainty about ecosystems may limit the ability to define goals and objectives for biodiversity. Uncertainty about the behavior of others will encourage intensified use to capture as many benefits as possible

while they are available. Uncertainty about the future will shorten the time horizons over which decisions are made, removing the incentive to invest in long-term protection. All of these forms of uncertainty work against the protection of biodiversity.

In an uncertain environment, decision making takes place through trial and error. The uncertainty creates a natural tension between the individual and the group, and between people and ecosystems. Property rights address some, but not all, of the components of uncertainty. They define and sanction ownership, resolving the question of future access. They provide assurance about the behavior of others, because they define appropriate types of use. They do not in themselves increase knowledge about the ecosystem, although they may provide an incentive for owners of property rights to produce this information.

#### B. Externalities

Another function of property rights is to resolve the problem of externalities, in which one action affects another. When property rights to resources do not exist or are incomplete, people do not take full account of the costs of their actions because there is no corresponding "owner" to lay counter claims. In open access fisheries, for example, people compete for resource benefits by fishing as fast as they can and catching as much as they can. The result is that one person's behavior affects the amount of fish available to others over time, unless their rights are correspondingly protected. Similarly, destructive fishing practices can eventually affect the functioning of the ecosystem if rights to habitat or non-fished species are not in place.

In some cases property rights may be defined but unenforceable, and the lack of enforcement then becomes equivalent to removing the right. For example, if fishing rights are expressed in terms of areas and people without rights are not excluded from those areas, their encroachment will render the right meaningless and externalities will continue. The outcome of missing or unenforceable property rights is biodiversity loss, as only some components of the ecosystem are protected, leaving others vulnerable to external effects.

The idea behind using property rights to protect biodiversity is to assign claimants to the full spectrum of ecosystem components and services, so that external effects are accounted for, or "internalized." If the owners of biodiversity are citizens, the state can then act in their behalf to protect biodiversity from various damages over space and time. The ability of the state to represent claims over biodiversity depends on a clear



definition of the goods and services of ecosystems, an articulation of the relation between ecosystem components and ecosystem function, and an active constituency for protection.

### C. Transactions Costs

How well a system of property rights functions both affects and is affected by transactions costs. Transactions costs are the costs of doing business, which in the resource management context include costs of gathering information, coordinating users, organizing decision making, and enforcing rules. Some transactions costs remain fixed regardless of the type of process used to make decisions. Others vary with the way decisions are made—the amount of data collected, analyses done, and the process used to make decisions.

Transactions costs are also influenced by the condition of the ecological system. As resources become depleted, a system of property rights must account for more and more externalities that increase the costs of management program design and enforcement. It is possible to create a system so costly to design or enforce that potential benefits are outweighed by the costs. This cost effect, particularly in consideration of actions to change property rights, is a particularly important factor in the consideration of the transition to property rights for protecting biodiversity.

### D. Scale

The functioning of property rights is also affected by the scale of the area over which they apply. There is a disconnect between the geographic scale over which species are distributed and the scale at which species habitat is owned. Property rights to goods are relatively easy to define and enforce because they are associated with geographic space and their value is embodied in the goods themselves. But property rights to services are much more elusive. For example, rights to the information value of species would involve several owners, even nations, over wide geographic space.

For rights to resources to have meaning in terms of providing incentives for their conservation, those rights must have value to the owners. The value rests in part on their uniqueness and exclusivity. When the resource is the genetic information encoded in a plant and the plant species is distributed over a range larger than that areas encompassed by the property right, the information embodied in that plant is also available to others, leaving the rights owner without the power to exclude or to capture the potential financial value from the

information. This scale mismatch limits the potential for property rights to provide the appropriate incentives for investing in conserving the genetic information.

The scale question also means that biodiversity protection is often an international issue. There are global benefits to biodiversity that are unaccounted for in national systems of property rights. Species distributions are independent of national boundaries. In addition, the loss of biodiversity is often an international externality, where the impact of one nation's actions is felt by another.

## IV. EVOLUTION OF PROPERTY RIGHTS

Property rights to resources tend to evolve incrementally over time in response to changes that alter the costs and benefits of particular forms. The current impetus for considering property rights for the protection of biodiversity is the scarcity resulting from declines in the number, range, and diversity of species that enhances biodiversity's value. The evolutionary path to this point, marked by a gradual expansion of property rights over ecological goods and services, is one that has reflected the relative changes in the benefits and costs of property rights as conditions of resource use change.

### A. Single-Species Use

When resources exist in surplus to human needs, conservation actions are unnecessary and so rights, rules, and responsibilities have little meaning. In an environment of surplus, the costs of developing protective mechanisms exceed the benefits gained. As resource abundance declines, either from human pressures or natural events, the increasing scarcity raises the benefit of protection. When the benefits of protection exceed their costs, property rights in some form develop. The first controls are on how the resource is taken or how much is taken. For example, an open access, no-rules fishery becomes restricted by how people can fish—for example, by seasons or by limits on gear—and how many fish they can take. Eventually, property rights of access to the fishery are developed to restrict who can fish.

### B. Ecosystem Use

As ecosystems are further exploited they reach a point where the focus on extracting and conserving individual species erodes the functioning of the ecosystem as a whole. Exploitation of species is interconnected, the

sources of impacts are diffuse, and effects of exploitation are system-wide. Species richness declines, community interactions are altered, and resilience erodes. It is at this point that the focus of managing human effects shifts from single species to ecosystems.

This shift in focus takes place against a background of using resources for the production of goods: lumber from trees and seafood from fish. The rights, rules, and responsibilities are all directed toward this end. They are not designed to control effects on unused ecosystem components or to protect the flow of ecosystem services. The system of property rights is poorly adapted to these interactive effects.

Biodiversity decline creates scarcity in services provided by species richness and genetic information. But the same scarcity in ecosystem services that provides an incentive to create property rights also presents a difficulty to developing ecosystem-level property rights. Ecosystem management is typically proposed when exploitation levels of single species are too high and showing signs of stress. Human competition and conflict over access to individual species complicate the shift to a broader focus and cause users to challenge the legitimacy of an ecosystem approach.

### C. Expectations about Use

Property rights systems, once in place, create expectations about what is normal. They start in motion a new path of development. At each stage in the path, the condition of the resource, the definition of rights, and the expectations about those rights influence action. The further along the path the more embedded are the property rights and the more vested are people in their continuance. When faced with the necessity of expanding those rights to the protection of biodiversity, it must be done against the legacy of the path to this point.

A property rights transition to the protection and conservation of biodiversity is affected by resource conditions at the intervention point. Protecting biodiversity means losing access to some of the goods and services of the ecosystem, which, under the old property rights rules, were at the disposal of the owner. Many resources have competing uses with known market values. Unknown future uses of these resources will be hard pressed to compete against known present uses.

## V. PROPERTY RIGHTS AND BIODIVERSITY PROTECTION

A number of emerging issues illustrate the challenges to the use of property rights to protect biodiversity.

Four are particularly relevant: uncertainty, exclusivity, distribution of benefits, and the alignment of private and social goals.

### A. Uncertainty

Protecting biodiversity requires more than a set of rules and responsibilities for property rights set within institutions that accommodate the attributes of the ecosystem and the people who use it. How to design property rights and institutions for this accommodation is what is at issue. Although there is general understanding of the importance of biodiversity to the stability, function, and sustainability of ecosystems, there is poor understanding of specifically what to protect. For example, knowledge is generally lacking of the thresholds at which biodiversity loss irreversibly changes ecosystems. Knowledge of the role played by individual species in contributing to critical ecosystem functions is also absent. These uncertainties create a corresponding uncertainty in the objectives and design of property rights to reflect the full range of values involved.

### B. Exclusivity

A problem with many property rights that apply to natural resources is that they do not specify claims to the full range of goods and services provided by an ecosystem. In failing to fully specify property rights claims, they fail to protect exclusive use. If property rights were defined for all components of an ecosystem, users and decision makers would take all the consequences of their actions into account. This would be the first step in making biodiversity conservation profitable. But under current systems of property rights, this is rarely the case. The history is to apply property rights to natural resources as commodities but not to the services they provide or to their existence value. The lack of full specification means that it is unclear who can claim and control rights of use.

### C. Distribution of Benefits

How benefits will be distributed is the core debate of natural resource policy. The same debate is at the heart of biodiversity protection. Biodiversity values have the potential to be protected through a number of different types of property right. But whatever type of right is used will create winners and losers. It is difficult to design an effective system of property rights without addressing the questions of what the objectives are, how progress toward those objectives will be measured,

and the time frame over which they will be met. How to promote equitable distributions is a difficult question in all resource policy, but it is particularly so for the complex issue of biodiversity. Equity questions extend into the international arena, where the distribution of benefits between rich and poor nations is at stake.

Effective resource protection depends on legitimacy—on the acceptance of rules and procedures by participants. Many natural resource systems are difficult to monitor, and the possibilities for circumvention of rules are many. When people doubt the legitimacy of the system of property rights because they cannot accept its distributional outcomes, their incentives are to undermine rather than promote its evolution to a new form. Scarcity compounds the erosion of legitimacy by creating greater incentives and opportunity for rent seeking that is characteristic of resource competition.

#### D. Alignment of Private and Social Goals

For property rights to bring private preferences into line with public preferences for biodiversity conservation, the tensions between private and social goals must be resolved. Individuals have private goals for productivity of ecosystems goods that may not be compatible with social goals for biological production. In addition, biodiversity is a public good, so it is subject to the potential for free riders to enjoy the benefits without paying the costs. Biodiversity protection must be accomplished in concert with existing rights and rules that favor direct uses of ecosystem goods.

How to realign private and social incentives to conserve biodiversity is an important and to a large extent unresolved question. Most current systems of rights, because they are still unspecified for ecosystem services, favor the conversion of ecosystems into goods. Options for change include the expansion of the scope of property rights, payment of compensation to owners for conserving rather than using resources, or developing prohibitions against certain uses.

## VI. CONCLUSIONS

Property rights outline the conditions under which resources can be used and the interactions of people in using them. They exist in particular contexts that are combinations of ecosystems and people. There are many forms and permutations, none of which is superior to others. What determines the effectiveness of property rights is the extent to which they are able to perform the functions of reducing uncertainty, removing exter-

nalities, containing transactions costs, and accommodating scale.

Property rights evolve in response to external conditions. Scarcity is the driving force for their existence, and their contribution of benefits in excess of costs is the means of their continuance in a particular form.

Despite the increasing scarcity of biodiversity, particular challenges face the evolution of systems of property rights to meet these conservation needs. To protect biodiversity, a transition must be made between current systems of rights that protect direct uses of ecosystems to expanded systems of rights that also protect indirect ecosystem services such as genetic information and aesthetic enjoyment.

The application of property rights to protect biodiversity is challenged by forces that are both internal and external to ecosystems. Internally, the increasing scarcity brought on by overexploited resources means that the range of options has declined. Fewer adaptations are possible. Formidable information needs for keystone species and critical functions of ecosystems exist.

Both internally and externally, forces are competing to shape the time horizon of resource management. Biodiversity protection and notions of sustainable use—values held by the public at large—are long-term concepts, requiring a long time horizon for management decision making. At the same time, the overexploited levels of many resources combined with uncertainty about their sustainability leads to internal pressures to shorten the time horizon and make decisions for the short run.

As these external and internal pressures illustrate, the very conditions in ecosystems that require biodiversity protection are also those which are causing difficulties in its implementation. Increased knowledge of ecosystems is needed as are innovative designs for property rights that work within and across national boundaries.

### See Also the Following Articles

BIODIVERSITY AS A COMMODITY • COMMONS, THEORY AND CONCEPT OF • ECONOMIC VALUE OF BIODIVERSITY, OVERVIEW • ENVIRONMENTAL ETHICS • LAND-USE ISSUES

### Bibliography

- Bromley, D. W. (1991). *Environment and Economy: Property Rights and Public Policy*. Basil Blackwell, Oxford, UK.
- Ciriacy-Wantrup, S. V., and Bishop, R. (1975). Common property as a concept in natural resources policy. *Natural Resources Journal* 15, 713–728.

- Eggertsson, T. (1990). *Economic Behavior and Institutions*. Cambridge University Press, Cambridge.
- Hanna, S. S. (1998). Institutions for marine ecosystems: Economic incentives and fishery management. *Ecological Applications* 8(1) Supplement, S170–S174.
- Hanna, S., Folke, C., and Mäler, K.-G. (Eds.) (1996). *Rights to Nature*. Island Press, Washington, D.C.
- Hardin, G. (1968). The tragedy of the commons. *Science* 162, 1243–7.
- North, D. C. (1992). *Transactions Costs, Institutions, and Economic Performance*. Institute for Contemporary Studies, San Francisco.
- Ostrom, E. (1990). *Governing the Commons*. Cambridge University Press, Cambridge.
- Perrings, C., Maler, K.-G., Folke, C., Holling, C. S., and Jansson, B.-O. (Eds.) (1995). *Biodiversity Loss: Economic and Ecological Issues*. Cambridge University Press, Cambridge.
- Schmid, A. A. (1995). The environment and property rights issues. In *The Handbook of Environmental Economics* (D. Bromley, Ed.), pp. 45–60. Blackwell, Cambridge, MA.
- Sedjo, R. A., and Simpson, R. D. (1995). Property rights, externalities and biodiversity. In *The Economics and Ecology of Biodiversity Decline* (T. M. Swanson, Ed.), pp. 79–88. Cambridge University Press, Cambridge.
- Swanson, T. (1995). The international regulation of biodiversity decline: Optional policy and evolutionary product. In *Biodiversity Loss: Economic and Ecological Issues* (C. Perrings, K.-G. Maler, C. Folke, C. S. Holling, and B.-O. Jansson, Eds.), pp. 225–259. Cambridge University Press, Cambridge.
- Swanson, T. M. (Ed.) (1995). *The Economics and Ecology of Biodiversity Decline*. Cambridge University Press, Cambridge.
- United Nations Environmental Programme (UNEP). (1992). Convention on biological diversity.





# PROTOZOA

Bland J. Finlay

*Centre for Ecology and Hydrology—Windermere*

---

- I. Functional Roles
  - II. The Nature of Protozoan Species
  - III. Protozoa and Ecosystem Function
  - IV. Biological Diversity and Global Species Richness
- 

## GLOSSARY

- benthos** The substrate at the bottom of the sea or fresh waters. Most benthic protozoa live in the spaces between sediment particles.
- commensal** Of a protozoon—one that is loosely but not obligately associated with another organism (e.g., the ciliates attached to the external surfaces of crustacean zooplankton).
- endosymbiont** An organism living in a long-term association inside a host organism to their mutual benefit (e.g., the endosymbiotic methanogenic bacteria that live inside anaerobic protozoa and utilize waste  $H_2$  produced by the host).
- eukaryote** An organism with membrane-bounded nuclei in its cells. Protozoa are unicellular eukaryotes.
- heterotrophic** Mode of nutrition in which carbon is obtained from the organic compounds made by autotrophic organisms.
- phagotroph** An organism that ingests solid food particles (e.g., bacteria).
- planktonic** Living in the water column of lakes and/or the sea.
- 

**PROTOZOA** are microscopic organisms with animal-like features. Each protozoon typically exists as a single, independent cell, and all free-living protozoa fit within the definition of phagotrophic microbial eukaryotes. In some species, independent cells coalesce to form plasmodia (e.g., slime molds) or unite to form colonies (e.g., colonial choanoflagellates). There has never been unanimous agreement as to where the boundaries of the protozoa should lie, largely because of the extraordinary diversity of their lifestyles. Many species cause disease (e.g., malaria): Others thrive as commensals in the digestive tracts of ruminants and wood-eating insects, and many plant-like flagellates (e.g., *Euglena*) have at various times been referred to as protozoa, algae, or plants. Here, we are concerned principally with free-living protozoa, and the definition of this group is one that is based on its key function in the natural environment: Protozoa are capable of phagotrophy—the ability to capture and ingest food particles. Many also have functional chloroplasts or endosymbiotic algae, and the resulting consortia are known as “mixotrophs” because they are capable of both phagotrophy and phototrophy. Like most microorganisms, protozoa have very large population sizes, and they are the most abundant group of phagotrophic organisms in the biosphere. Biodiversity at the level of protozoa has characteristics that are not shared by macroscopic animals and plants. Most protozoan species are probably globally ubiquitous, so the global number of species is relatively small. A significant proportion of local protozoan species richness, at any moment in time, is rare or cryptic and awaiting

the arrival of conditions suitable for growth and reproduction.

## I. FUNCTIONAL ROLES

All of the important functional roles of free-living protozoa derive from their small size. The smallest flagellates are 2–4  $\mu\text{m}$  and most are  $<20 \mu\text{m}$ , most amoebae are  $<50 \mu\text{m}$ , and most ciliates are  $<200 \mu\text{m}$ . Exceptionally, some amoeboid protozoa, such as the radiolarians and foraminiferans, and the agglutinated foraminiferans of the deep-sea benthos, may reach 2 mm or more, especially if they have spiny extensions. Because protozoa are so small, most suitable prey items are other, smaller microbes. Protozoa are the principal consumers of the immense natural resource of bacteria and other microbes, and because they have population growth rates that are similar to those of the microbes on which they feed (doubling times in the order of 1 to several days), they are usually able to control microbial abundance. Flagellated protozoa can probably consume all bacterial production in the plankton. In the benthos, protozoa overlap in their niche requirements with nematodes, rotifers, tardigrades, turbellarians, and gastrotrichs, but because of their great abundance protozoa are indeed quantitatively the most important grazers in the freshwater and marine (including deep-sea) benthos. Also, just as microbes achieve astronomical abundance on a global scale, so too are the protozoa that graze on them represented by species populations with proportionately smaller but still unimaginably large global abundances.

Protozoan grazing on microbes also stimulates activity of the whole microbial community, in both aerobic and anaerobic environments. The process involved is not fully understood, although it may operate by increasing the rate of turnover of essential nutrients that would otherwise remain “locked up” in bacterial biomass. The net effect is that grazing by protozoa stimulates the rate of decomposition of organic matter.

The variety of shapes, sizes, and relative abundances of microbial food items has driven the evolution of a comprehensive suite of methods to capture them and a considerable diversification of protozoan morphologies. In general, the size of a protozoon relative to its prey dictates the most efficient food-capturing mechanism. Where the predator is typically much larger than its prey, filter-feeding prevails, and where the size difference is less the protozoon is more likely to be a raptorial feeder—one that seeks out relatively large individual food items. Thus, the planktonic choanoflagellate feed-

ing on a dilute suspension of tiny bacteria 20 times smaller than itself does so with a very fine filter, and the ciliate feeding on dinoflagellates half its size seeks out and captures each one individually. There are, of course, many exceptions, such as the marine heterotrophic dinoflagellates that use a feeding veil to trap and digest diatoms much larger than themselves.

The third main feeding type in protozoa is termed “diffusion feeding.” This is particularly common in planktonic amoeboid protozoa (radiolarians, foraminiferans, and heliozoans) and in the suctorians ciliates. It works when prey items collide with the sticky spines, tentacles, or axopods that radiate from the protozoon. Unlike the other two main modes of feeding, the protozoon simply waits for the arrival of its prey, much as a spider waits for an insect to be snared in its web.

Thus, there is a close link between protozoan morphology (especially of the food-capturing organelles) and the way in which a protozoon functions as a grazer. Therefore, when we classify the free-living protozoa into broad morphological groups, they are simultaneously allocated to broad functional groups. The three broadest morphological–functional groups are the amoeboid, the flagellated, and the ciliated protozoa, and each has its own strengths as a phagotroph. Representatives of all three may feed on the same type of microbes in the same place (e.g., in an aquatic sediment), but they will differ in the mechanics and efficiency of capture of any particular food particle. A filter-feeding flagellate will have a relatively large filter area, a high volume-specific clearance and competitive superiority over filter-feeding ciliates when grazing on planktonic bacteria. A helioflagellate and a suctorian ciliate will both practice diffusion feeding, but the former will be adapted for snaring bacteria, whereas a diffusion-feeding suctorian will specialize in trapping flagellates and ciliates.

Many protozoa are microaerophilic: They seek out habitats with a low level of dissolved oxygen that is just sufficient to drive their aerobic respiration and low enough to exclude metazoan competitors and predators. Microaerobic habitats are common in aquatic sediments and in oceanic “oxygen minimum zones.” These are zones in which the raw materials for microbial growth arrive from opposite directions (e.g., where oxygen and light arriving from above meet carbon dioxide and sulfide from below) and in which there is therefore an elevated abundance of microbial food. Therefore, microaerophily is an adaptive behavior: It brings protozoa into contact with high abundances of microbial food. It also stimulates the growth of nutritional symbionts such as sulfide-oxidizing bacteria and endosymbiotic algae. Many microaerophilic protozoa are also tem-

porary anaerobes, but unlike the “true” anaerobes that live permanently in the absence of oxygen, their metabolism is fundamentally aerobic. The true anaerobes—those that complete their entire life cycle in the absence of oxygen—live principally in aquatic sediments. There are many species, but none is ever abundant. Most use hydrogen-evolving fermentations for energy generation, and the hydrogen is used by anaerobic bacteria, especially endosymbiotic methanogens. Thus, methane is released from these protozoan consortia. The anaerobic protozoa are probably the only phagotrophic organisms capable of living permanently in the absence of dissolved oxygen.

The real diversity of symbiotic associations involving protozoa is poorly known. In some cases, complex interactive behaviors have evolved between the partners. In the marine, sand-dwelling ciliate *Kentrophoros*, the entire dorsal surface of the ciliate is a coat of sulfide-oxidizing bacteria that can grow only in the narrow layer within the sediment where oxygen and sulfide overlap. The ciliate host's innate microaerobic behavior enables it to seek out the habitat that the symbiotic bacteria need for growth. The ciliate then invaginates its dorsal surface and digests the bacteria because it does not have a mouth, and this is its only source of nutrition.

It is clear that many of these symbiotic consortia involving protozoa represent tightly integrated functional units; indeed, the symbionts may be almost as deeply embedded functionally in the consortium as the protozoon's other organelles. Two points must be noted. First, it is the combined phenotype of the consortium, rather than that of any individual consortium partners, on which natural selection will operate, and there are examples of how the fitness of a protozoon in a particular habitat can be improved by the acquisition of endosymbionts (e.g., the algal symbionts of ciliates living in the metalimnia of freshwater lakes). Second, the biodiversity of protozoa, when quantified simply in terms of protozoan species richness, will fail to take account of the large supplementary microbial diversity with which the protozoa are necessarily associated.

Therefore, the diversity of free-living protozoa may be classified into broad morphological–functional groups: amoeboid, flagellated, and ciliated protozoa. Almost any free-living protozoon can be placed without difficulty in one of these groups. It must be stressed, however, that these groups are not concordant with any system of classification of protozoa published in recent years; nor are they in most cases aligned with the independent lineages that are emerging in the molecular

phylogenies (e.g., those based on sequence variation in ribosomal RNAs) which reflect the main episodes in the history of eukaryotic evolution. Heterotrophic flagellate groups, such as the diplomonads and trichomonads, appear in the early emerging lineages, but other flagellates (e.g., choanoflagellates, euglenids, chryomonads, dinoflagellates, and haptomonads) are classified within recently diverging lineages. The amoeboid protozoa too are scattered across many lineages. The naked amoebae without mitochondria (the pelobionts) diverge early and close to the diplomonads, whereas the vahlkampfiids, slime molds, and various other groups of naked and testate amoebae appear in other independent lineages that are evolutionarily quite distant from each other. The morphological–functional group of the ciliates is the only one which remains intact, as a monophyletic group, in current molecular phylogenies.

The process of distilling the vast quantity of molecular information generated in recent years has generated some entirely new phyletic assemblages, including the stramenopiles, a group containing organisms as morphologically and functionally dissimilar as chryomonad flagellates and diatoms, and the alveolates—a group that embraces the dinoflagellates, the ciliates, and a large group of exclusively intracellular parasites (the apicomplexans). In the next section, I focus on the broad morphological–functional groups, define what is meant by species, and quantify the species within these groups.

## II. THE NATURE OF PROTOZOAN SPECIES

There is no universal agreement on what constitutes a protozoan “species.” The most widely used concept is the morphospecies because it is relatively easy to discriminate a great diversity of protozoa using body form alone. This concept is especially useful because morphology is closely related to ecological function. In many protozoa, the structure of the feeding apparatus and the size and shape of the cell determine how the protozoon functions in the natural environment; therefore, the form largely determines the ecological niche that the protozoon occupies. Discriminating species on the basis of form might then be equivalent to discriminating them according to the ecological niches they occupy. We could say that if a protozoon looks the same in different places, then it is the same in different places. However, there are problems with such a simple concept. It is known that a morphospecies can be com-



posed of ecologically distinct populations, such as those adapted for maximum growth rate at different temperatures. Therefore, although it is possible to collect apparently identical representatives of a morphospecies in different corners of the world, and one might be encouraged by the belief that they were filling exactly the same niche in these different places, in reality they could be phenotypically quite different. The same problem applies with respect to genetic differences. It is possible to find genetic differences, especially sequence differences in ribosomal RNAs, in morphologically identical isolates collected from different places. The significance of these genetic differences is not known; nor is it known if genetic and phenotypic divergence are correlated. It appears, however, that there is no correlation between genetic divergence and geographic distance.

Protozoa spend most of their time as asexual organisms, but some do reproduce sexually, if only periodically, and in some well-studied cases (e.g., the ciliate genera *Tetrahymena* and *Paramecium*) morphologically indistinguishable biological species (reproductively isolated gene pools, also known as sibling species) have been studied thoroughly. Different sibling species within a morphospecies may be genetically identical to each other or extremely divergent (at least with respect to ribosomal RNAs), and there is no apparent correlation between genetic isolation and genetic divergence. Again, there is no good evidence for biogeography of protozoan sibling species, and members of the same sibling species can be found on different continents. It is possible that sibling species carry unique phenotypic traits that equip each species for a particular niche, but this has not been demonstrated. The real problem with using a biological species concept for protozoa is that it is not practical for all but a few easily cultivated protozoa. The majority of protozoan species have never been cultivated, and neither the frequency nor the character of their sexual behavior (if any) are known or are ever likely to be known. Most people who study protozoa in the natural environment use the morphospecies concept because it is practical, it embodies the close link between form and function, and it is the morphospecies that fills the niche, if not always the minutest niches, that can be discerned by the human observer. The morphospecies may contain much genetic variation and be capable of expressing a wealth of phenotypic variation, but it is the best tool that is available for ordering the diversity that lies within the protozoa.

One particular problem that is particularly acute when dealing with organisms the size of protozoa is

that the quality of one's perception of morphology is very definitely a function of the tools that are available. Thus, better tools (e.g., quality microscopes) enable the discrimination of finer detail. The logical extension of this is that even individual protozoa in a population might be discriminated one from the other and could therefore be referred to as separate morphospecies. Modern optics have undoubtedly lent impetus to the business of describing many new species, some of which are undoubtedly legitimate, whereas many have been established on morphological criteria that are trivial or have no conceivable functional significance.

An additional problem with the morphospecies is that it is often difficult to decide exactly where a species begins and ends. Although most observations are made of the vast number of individuals that cluster around the central tendency of any population, the individuals at the tail ends of the distributions abut and often overlap those of other species. Morphological variation then appears to be continuous across many species. Examples of this have been described: In some of the large spongioid spumellarian radiolarians, the major features (e.g., skeletal and cytoplasmic morphology) that are used to discriminate species intergrade to such an extent that it is not possible to unambiguously ascribe individual radiolarians to nominal species.

However, the evolutionary process in radiolarians and other protozoa probably works in the same way as it does for other organisms such that it maintains phenotypically discrete species in niche space that is essentially continuously variable. The phenotypic traits that sustain these discrete species in protozoa are presumably quite diverse in character and possibly not totally comprehensible. Among the more accessible of these traits are the morphological characters that we use to separate species. However, the likely limitation of these is that they enable us to perceive only the fairly coarsely resolved, sometimes overlapping entities that we call morphospecies.

### III. PROTOZOA AND ECOSYSTEM FUNCTION

Protozoa are abundant. One gram of soil typically contains  $10^3$ – $10^7$  naked amoebae,  $10^5$  planktonic foraminiferans can often exist beneath 1 square meter of oceanic water, and almost every milliliter of fresh water or seawater on the planet supports at least 100 heterotrophic flagellates. When expressed in global terms, these numbers are very large, and an inevitable consequence of

the persistence of such a large number of very small organisms is that migration rates will be relatively high. It follows that rates of speciation and extinction must be low, as will the consequent global number of species. It also follows that protozoa are unlikely to have biogeographies, and “endemic” species probably do not exist. We might expect that the local diversity of protozoa would account for a significant proportion of global diversity, even if at any moment in time much of this diversity is represented by rare or inactive individuals (e.g., cysts awaiting the arrival of suitable conditions). There is indeed good evidence for the global distribution of protozoan species, including the morphologically distinctive flagellate *Rhynchomonas nasuta* that has been found in most aquatic and terrestrial environments worldwide; marine foraminiferans and ciliates found living in slightly salty water of desert oases, hundreds

of kilometers from marine coasts; the same radiolarian species living in high northern and southern oceanic latitudes; the same pond-dwelling ciliates living in Australia and northern Europe; and the cosmopolitan distribution of the same species of agglutinated foraminiferans in the deep-sea benthos. In general, protozoan morphospecies are ubiquitous and apparently cosmopolitan if the habitats to which they are adapted are distributed in different parts of the world. In accordance with this, the global number of protozoan species is indeed relatively modest (Table I), and the number of species that can be retrieved from a local area (e.g., a pond), in both “active” form and from a “passive” state is a significant proportion (usually at least 10% for various morphological-functional groups) of the global total. This fact may not be obvious from short-term ecological sampling programs because only a limited

TABLE I  
Estimates of Global Species Richness of Extant Free-Living Protozoa<sup>a</sup>

			Marine	Non-Marine	Total
Amoeboid protozoa	Slime molds	Dictyostelids	—	60	60
		Myxomycetes	—	550	550
	Rhizopod amoebae	Naked	180	220	400
		Testate	—	200	200
		Foraminiferans			
	Actinopod amoebae	Planktonic	40	—	40
		Benthic, inshore	4000 <sup>b</sup>	—	4000
		Benthic, deep sea	250 <sup>b</sup>	—	250
		Acantharians	150	—	150
	Flagellated protozoa	Excluding heterotrophic dinoflagellates <sup>c</sup>	Radiolarians, solitary	750	—
Radiolarians, colonial			50	—	50
Heliozonas			—	120	120
Marine plankton			420	—	420
Marine benthos			330	—	330
Ciliated protozoa	Heterotrophic dinoflagellates	Freshwater and soil	—	350	350
		Other mixotrophic flagellates	—	—	150 <sup>e</sup>
			1400	1660	3060
Total					11890

<sup>a</sup> Compiled from numerous published and unpublished sources. The more problematic estimates are highlighted.

<sup>b</sup> There is considerable uncertainty attached to these estimates. The figure of 4000 is generally accepted as a working figure for extant species but is probably inflated by synonyms, especially those of shallow-water benthic species. There is no firm information for species richness of deep-sea benthic foraminiferans. The estimate of 250 is approximately double the typical figure for local richness of species, most of which may be cosmopolitan, but there are probably many undescribed deep-sea soft-shelled foraminiferans that have in the past been ignored by geologists.

<sup>c</sup> The total given here for heterotrophic flagellates is 1100 species. This includes some synonyms and mixotrophs. It is believed by some that the real global total is closer to 3000 species.

<sup>d</sup> Assuming there are 1800 marine and 220 freshwater species, and that 50% of these are heterotrophs.

<sup>e</sup> This estimate is simply derived by doubling some recent estimates. Note that these species may be only temporarily phagotrophic.

number of microbial niches are available at any moment in time.

Protozoa and other microorganisms have other special properties. Microbial activities interact strongly with physical and chemical factors in the natural aquatic environment (e.g., light transmission or the concentrations of essential nutrients) to create a continuous turnover of microbial niches. These niches are quickly filled from the locally available diversity of rare and dormant microbes, and the activities of the latter create further reciprocal interactions. Therefore, the diversity of active protozoan species in a pond, at any moment in time, is the result of preceding reciprocal interactions involving many biological and nonbiological factors, and the biodiversity of protozoa and other microbes is an integral part of ecosystem functions such as carbon fixation and nutrient cycling.

#### IV. BIOLOGICAL DIVERSITY AND GLOBAL SPECIES RICHNESS

##### A. Amoeboid Protozoa

Slime molds may be regarded as protozoa because they spend most of their active lives as amoebae or as naked amoeboid (and often macroscopic) plasmodia (Fig. 1). They are known by many names, including mycetozoa ("fungus animals"), because although they never develop hyphae they produce fruiting bodies supported on cellulose-rich stalks. There are two large groups—the dictyostelids (cellular slime molds, e.g., *Dictyostelium*) and the myxomycetes (acellular slime molds, e.g., *Physarum*). Two smaller groups are the acrasids (e.g., *Acrasis*) and the protostelids (e.g., *Cavostelium*). The dictyostelids are typically phagotrophic amoebae, but when starved they aggregate, form a migrating slug, and then produce a fruiting body from which spores are released. These later germinate to produce amoebae. In the myxomycetes, individual amoebae with or without flagella coalesce to form a distinctive (often brightly colored) multinucleate plasmodium which gives rise to fruiting bodies. Slime molds are common in damp forest soils, tree bark, and dead or dying wood, and abundance and local species richness are greater in deciduous than in coniferous forests. There is extreme patchiness in the distribution of dictyostelid species in forest soils, and coexistence of multiple species in the same small soil samples is rare. Most myxomycete species are believed to be cosmopolitan. The phagotrophic amoeboid stages of slime molds may be quantitatively important grazers of bacteria, fungi, and the other primary decom-

posers of organic matter in soil. Other ecological interactions of slime molds may be at least as complex as their life cycles (e.g., the migrating slugs of dictyostelids appear capable of repelling grazing nematodes).

Rhizopod amoebae use pseudopodia for locomotion and feeding. There are two large groups—the "naked amoebae" (e.g., *Acanthamoeba*, *Vannella*, *Amoeba*, and *Vampyrella*) and the shelled "testate amoebae" (e.g., *Arcella*, *Nebela*, and *Euglypha*). Most rhizopod amoebae feed non-selectively by engulfing diatoms and other algae, unicellular and filamentous cyanobacteria, detritus, and bacteria. However, there are notable variants: *Vampyrella* dissolves a hole in the cell wall of a green alga or desmid and then enters through the hole to digest the cytoplasm of the prey. Some are opportunistic pathogens of man (e.g., in the genera *Acanthamoeba* and *Naegleria*).

Naked amoebae occur in the plankton, but these species are usually very small (<10  $\mu\text{m}$ ) and typically associated with suspended flocs (e.g., "marine snow"). However, it is likely that all rhizopod amoebae need to be attached to a surface if they are to feed, and much larger numbers are found in aquatic sediments. Amoebae are particularly abundant in soils ( $10^3$ – $10^7$  per gram dry weight), where together with heterotrophic flagellates they probably control bacterial abundance. A few testate amoebae can be planktonic (e.g., *Difflugia*), but most live in soils and in *Sphagnum* bogs. The prey size of testates must be limited by the shell aperture size.

Naked and testate amoebae offer different challenges when attempting to identify species. The naked forms are distinguished by the forms of their bodies and pseudopods in locomotion, but many species constantly change shape and the absence of fixed characters often makes it difficult to identify them without resorting to an analysis of ultrastructure. The testates construct tests which have species-specific architectures, but these are often modified according to nutritional status and local environmental factors, such that the variety of tests described is extraordinarily large and the task of deciding where one nominal species begins and the other ends is at least as difficult as it is for the naked amoebae. One practical solution to the problem of identifying living naked amoebae in natural samples is to allocate them to one of several broad morphotypes, e.g., those that extend lobose pseudopodia, slug-like forms with eruptive or non-eruptive locomotion, and fan-shaped forms. No clear relationship has been established between morphology and ecological function in naked amoebae, although it is likely that the limax (slug-like) amoebae are predominantly benthic species.

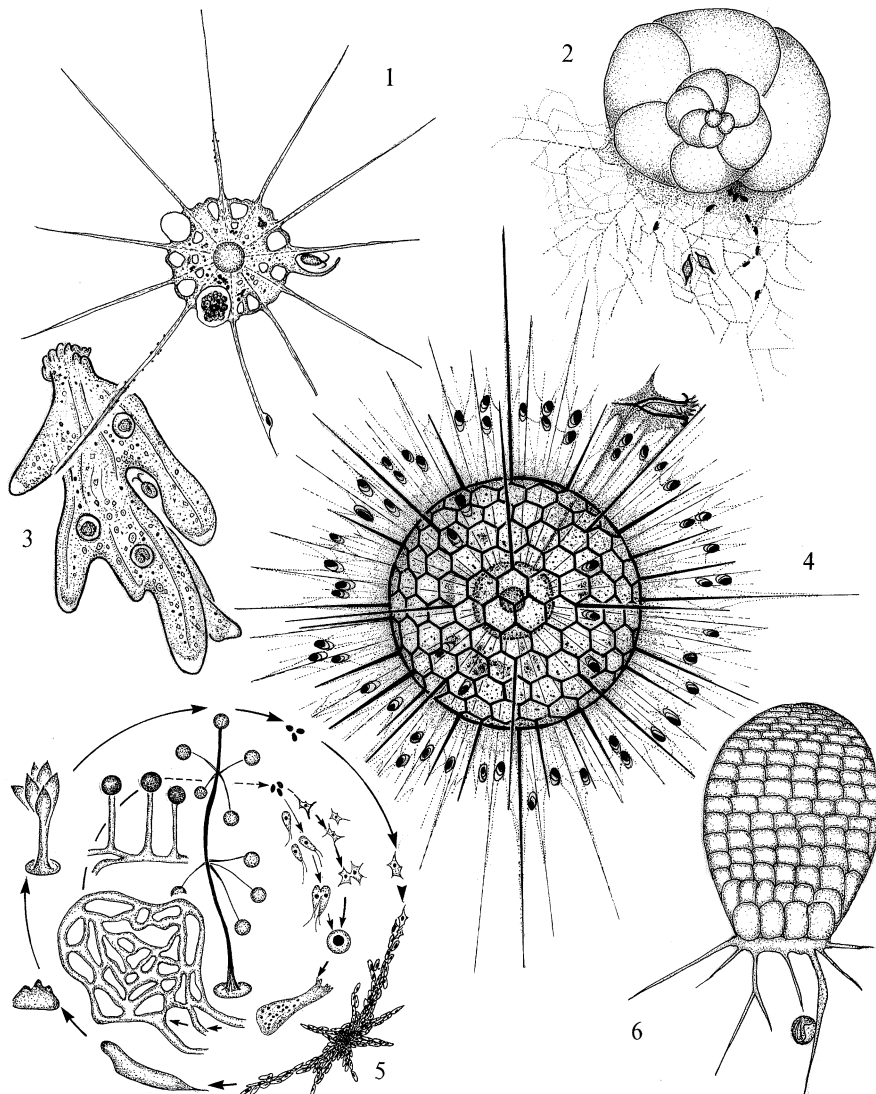


FIGURE 1 A selection from the variety of form and function in amoeboid protozoa. 1, an actinophryid heliozoon (*Actinophrys*; diameter about 0.1 mm) ingesting a flagellate; 2, a benthic foraminiferan (*Rotalia*) trapping diatoms and bacteria in its reticulopodial net; 3, a naked amoeba (*Amoeba*) using pseudopodia to trap a flagellate; 4, a polycystine radiolarian (*Heliosphaera*; spherical body ~ 0.3 mm) with symbiotic dinoflagellates and an entrapped tintinnid ciliate; 5, polymorphic life cycles of dictyostelid slime molds (e.g., *Polysphondylium*; outer circle) and myxomycete slime molds (e.g., *Physarum*; inner circle); 6, testate amoeba (*Assulina*; ~0.08 mm) with its prey (an algal cell) caught on a sticky filopod.

Some naked amoebae (especially those that form extensive anastomosing plasmodial networks) resemble hybrids of shell-free foraminiferans and slime molds. Some can reach several centimeters in diameter. They are typically very slow moving and apparently perfectly adapted for grazing extensive areas of soil surface (*Leptomyxa*) or diatom mats in the marine benthos (*Corallomyxa*). Some free-living naked amoebae are anaerobic

and do not have mitochondria. As a group, these are referred to as pelobionts, rhizomastigids, archamoebae, or karyoblasteans (e.g., *Mastigamoeba*, *Mastigella*, and *Pelomyxa*). They are relatively common in organically enriched anoxic freshwater and marine sediments. The very large (up to several millimeters in diameter) *Pelomyxa* supports several types of endosymbiotic bacteria, including methanogens, and probably lives exclusively

in freshwater sediments. It appears to ingest any organic particles with which it makes contact. It can be locally abundant and the dominant grazer in some anaerobic lake sediments. Most species of naked amoebae have been described from fresh waters and soil, and most marine species have been collected from inshore areas. It is not clear to what extent freshwater species can also live in seawater. Some genera (e.g., *Mayorella* and *Acanthamoeba*) are well represented in both environments.

Foraminiferans live in marine or brackish water environments, and most have calcareous shells. Cytoplasm emerges through the aperture and pores in the shell to form a repeatedly branching and anastomosing reticulopodial net in which prey are trapped. The “spinose” planktonic species (e.g., *Globigerinoides ruber*) have calcareous spines, and familiar Cretaceous chalk deposits (e.g., the white cliffs of Dover, UK) consist largely of the sedimented shells of these species together with the coccoliths of haptomonad flagellates. Some naked amoebae that produce similar pseudopodial networks live in fresh waters (e.g., *Biomyxa*) or soil (e.g., *Reticulomyxa*).

The shells of planktonic foraminiferans are generally 0.1–1 mm in diameter, but the pseudopodial network can reach 10 mm and the spines may increase the size to more than 20 mm, making them larger than many marine metazoans. Planktonic species are open-water organisms. They are most abundant in tropical and subtropical waters, in which there is typically one per liter (or  $\sim 10^5 \text{ m}^{-2}$ , depth integrated to 150 m). Each species tolerates only small variations in temperature and salinity, and distinct assemblages of species are associated with five “biogeographical provinces”—the surface water masses of equivalent temperature on either side of the equator (polar, subpolar, temperate, subtropical, and tropical). However, species in these assemblages are not isolated from each other: Polar species have been recorded in upwelling zones at 10°N, and species characteristic of high latitudes are also found in the tropics, but in very deep and cool waters.

A typical planktonic foraminiferan produces a mass of reticulopodia (the “halo”) that radiates outward from the shell. This is used to snare prey and to support algal symbionts. Spinose species appear to prefer zooplankton prey, and nonspinose species (e.g., *Globorotalia menardii*) prefer phytoplankton. These preferences may affect the local distribution of species: In the Red Sea north of 20°N, the dominating spinose species appear to reflect the greater local abundance of zooplankton.

Many planktonic foraminiferans, especially the spi-

nose species (e.g., *G. ruber*), contain algal symbionts that are carried along the spines by cytoplasmic streaming. There are probably only two types of symbionts in planktonic species: the dinoflagellate *Gymnodinium beii*, which is morphologically identical in many host species, and a small yellow-green (chryomonad-like) symbiont. The symbionts are protected from digestion by the host. The carnivorous foraminiferan *Hastigerina pelagica* envelopes itself in a cytoplasmic bubble capsule which is probably a flotation device but is also used for trapping and digesting prey. *Hastigerina pelagica* does not have symbionts but commensal dinoflagellates (*Pyrocystis* spp.) that live in the capsule. The role of symbionts and commensals is unclear. High rates of photosynthesis are maintained within the consortium, the symbionts probably transfer significant amounts of fixed carbon to the host, their photosynthetic activity may be necessary for calcification of the host’s shell, and the symbionts may use the host’s (nitrogenous) waste.

Benthic foraminifera have a greater diversity of symbiotic partners, including diatoms (*Nitzschia*), dinoflagellates (*Symbiodinium microadriaticum*), red algae, and chlorophytes (*Chlamydomonas*). Some also sequester chloroplasts from ingested algae. Although many benthic foraminiferan species have been described, information on their life cycles and ecology is available for only about 20 shallow-water species (e.g., *Allogromia*, *Rosalina*, *Spiroloculina*, and *Sorites*). These typically drag themselves slowly over surfaces, continually extending reticulopodia to trap prey (diatoms, algae, and bacteria). The abundance of benthic species is correlated with local biological productivity. In the fertile Mississippi Delta, numbers may reach several thousand in an area 10 cm<sup>2</sup>. Close to sewage outfalls, they are typically very abundant but low in species diversity. The spatial distribution of all benthic species is very patchy. This may be related to asexual reproduction (known only for benthic species) and the consequential limited range of migration of offspring from the parent. Benthic foraminiferans become relatively more important with increasing water depth, possibly due to a competitive advantage they have over metazoans in their ability to rapidly exploit seasonally deposited “phytodetritus.” In oxygen minimum zones along continental margins of tropical oceans, they can, because of their relatively small size, tolerate low oxygen conditions better than can metazoans. Also, certain taxa (e.g., *Bolivina* and *Uvigerina*) seem to be adapted for life in oxygen-depleted waters.

Foraminiferans in the deep-sea benthos are often abundant (typically hundreds per 10 cm<sup>2</sup>) and diverse (species richness may exceed that of nematodes), and

they are now regarded as major components of the fauna in this zone. Because they consume bacteria and detritus, they probably act as a link in the food chain to higher trophic levels in the benthos. Very small species (30–63  $\mu\text{m}$ ) are common, but the dominant fauna >500  $\mu\text{m}$  in central oceanic regions (and particularly at depths greater than the carbonate dissolution depth of  $\sim 4000$  m) are large agglutinated foraminiferans, especially the komokiaceans, which have complex anastomosing networks of tubules. In the benthos underlying the oligotrophic waters of the central North Pacific, the biovolume of komokiacean tests may greatly exceed the biovolume of metazoans. However, the amount of protozoan biomass in komokiacean tests is relatively small, and it may be even less than that of the metazoans that also occupy the test. *Edgertonia floccula* consists of a network of branching agglutinated chambers buried in a large (2- to 8-mm) inorganic mudball that it shares with cohabiting metazoans. This foraminiferan is one of the most abundant meiobenthic animals in some areas of the abyssal northeast Atlantic. Other deep-sea agglutinated foraminifera (e.g., *Bathysiphon* and *Hyperammina*) may even reach several centimeters in length, and some species stand vertically in the sediment and project into the overlying water, in which they presumably feed. Many of the common deep-sea foraminiferan species are probably cosmopolitan in their distribution.

The xenophyophoreans are another deep-sea group about which relatively little is known. They live in agglutinated, often fragile tests (typically 1–10 cm in length) and they too probably feed on sedimented detritus. Abundances of hundreds per square meter have been reported. The multinucleate protoplasm is enclosed within a labyrinthine organic tube encrusted with foreign particles. The organism also contains numerous crystals of barium sulfate, which have no obvious function. Growth is episodic, with intervening periods of months of inactivity. They have no confirmed fossil record, although they often use the sedimented shells of the planktonic foraminiferan *Globigerina* to strengthen their tests (as in *Homogammina maculosa*).

Foraminiferan fossils are useful for biostratigraphy because (i) their sedimented remains are abundant (e.g., in the “*Globigerina* ooze” which covers much of the abyssal plain beyond the continental shelves), (ii) individual shells are identifiable to species level, and (iii) their remains allow determination of sea surface “paleotemperature.” Seawater and calcite differ in their  $^{18}\text{O}/^{16}\text{O}$  ratio. This difference increases with temperature, so the ratio is used to estimate the water temperature at the time the calcite was deposited in the fossil foraminiferan shell. This estimate can then be compared

with the established correlation between the temperature of a water mass and the extant species it typically supports. Much of this work is supported financially by the oil exploration industry.

Actinopod amoebae all have axopods, which are stiffened ray-like pseudopodia that radiate from the cell body. They also have a peripheral network of filopodia, which are long, thin, sticky pseudopodia used for trapping food organisms that make chance contact. There are three principal groups: (i) radiolarians, consisting of the polycystines with perforated spiny lattice skeletons of silica and the phaeodareans with skeletons of hollow tubes and spines; (ii) acantharians with skeletal spines of strontium sulfate radiating from the center of the organism; and (iii) heliozoans, which often lack an internal mineral skeleton (although some species produce skeletal spicules which are either organic or siliceous) and superficially resemble polycystine radiolarians. All actinopod protozoa, apart from a large number of heliozoans, are exclusively planktonic in deep oceanic water.

Radiolarians and acantharians range in size from  $\sim 30$   $\mu\text{m}$  to several millimeters in diameter. Colonial radiolarians (e.g., *Collozoum longiforme*), which occur as gelatinous cylinders with diameters of several centimeters, can reach 3 m in length. Typical food items include diatoms, tintinnids and other ciliates, crustacean larvae, and other zooplankton, all of which are trapped in the peripheral network of axopods and filopodia. The smallest species may eat bacteria.

Dinoflagellate (*Symbiodinium*), chrysoomonad, and prasinomonad symbionts are associated with many polycystine radiolarians. They probably contribute to host carbon metabolism because the consortium can last many weeks without exogenous food. The symbionts are carried about in the host's cytoplasmic streaming, gathering in the peripheral filopodia at dawn and withdrawing inside the host shell at sunset. The phaeodareans that live near the ocean floor do not have algal symbionts. Radiolarians are found at all water depths down to  $\sim 8000$  m, although they are most abundant and diverse between 200 and 2000 m. No benthic species are known. Like planktonic foraminiferans, each radiolarian species inhabits an oceanic water mass that lies within a particular temperature range. Moreover, because of the immense depths to which living radiolarians sink, those species living in surface waters at high latitudes are also found in the tropics but in very deep and cool waters. As a consequence, the species assemblages found in surface waters give a misleading impression of radiolarian “biogeography.” It is believed that radiolarians may be transported in very deep oceanic

waters between high northern and high southern latitudes.

Radiolarian abundance and productivity increase in warmer waters. Abundance may be almost nil in the surface waters of the Antarctic, up to 50 per m<sup>3</sup> in the Caribbean and Gulf Stream, and up to several thousand per m<sup>3</sup> in the Gulf of Mexico. Colonial radiolarians are much less abundant. In the Sargasso they are considered abundant if represented by more than one colony per 100 m<sup>3</sup>. Acantharians can reach relatively high abundances (up to 25–35 per liter) and they are often the most abundant of the planktonic shelled amoebae, but relatively little is known about living organisms because as soon as they are collected the fragile cell membrane breaks and the strontium sulfate skeleton dissolves. Radiolarians are important stratigraphic tools, and skeletal morphology is the main characteristic used to identify species; hence, much information is available on the diversity of radiolarian skeletons.

The heliozoans are predominantly a freshwater group. They resemble functionally the radiolarians, especially with respect to their “diffusion” feeding. They span a wide size range. Some (e.g., *Actinosphaerium*) can be >1 mm, but most species are ~40 μm (*Actinophrys*). Depending on their size, different species feed with little apparent selectivity on algae, flagellates, ciliates, and rotifers. Small heliozoan species account for most of the biomass of amoeboid protozoa in the freshwater plankton. Most, however, are attached to or loosely associated with sediment and other submerged surfaces.

## B. Flagellated Protozoa

There is little consensus on how to classify the flagellates, and they are here divided into broad functional groups (Fig. 2). Heterotrophic flagellates are fundamentally important because they are abundant (there are seldom less than 1000 per milliliter, even in the plankton) and because their grazing activities are largely responsible for controlling the abundance of bacteria in aquatic environments. In some taxonomic groups, all species are exclusively heterotrophic (e.g., choanoflagellates and bodonids); others contain many mixotrophs (e.g., the euglenids and chrysomonads), whereas the haptomonads and cryptomonads are dominated by phototrophs and only a minority are capable of phagotrophy. In the past 15 years, a large diversity of heterotrophic flagellates has been discovered. Most of these are choanoflagellates, chrysomonads, euglenids, or bodonids. Some of the more easily recognized species (e.g., *Rhynchomonas nasuta*) have been recorded from a wide

range of habitat types in marine, freshwater, and terrestrial environments.

Many heterotrophic flagellates are anaerobes, including intestinal parasites of man (e.g., *Giardia intestinalis*), but free-living “diplomonads” (*Hexamita* and *Trepomonas*) and *Retortamonas* are relatively common bacteria feeders in organically enriched anoxic waters and sediments. Some anaerobic flagellates have hydrogenosomes, the anaerobic derivatives of mitochondria. Almost all of these are endosymbionts or internal parasites of animals. They are also known as parabasalans and include many of medical or economic importance, e.g., the human parasite *Trichomonas vaginalis*, and *Trichomitopsis*, which degrades cellulose in the hindgut of termites and other wood-eating insects and also supports endosymbiotic methane-producing bacteria (termites may be responsible for a globally significant flux of methane to the atmosphere).

Most choanoflagellates are small (<10 μm), with a feeding filter that forms a “collar” around the single anterior flagellum. The flagellum creates the water current that flows through the filter. Choanoflagellates are exclusively phagotrophic and either solitary (e.g., *Monosiga*) or colonial (e.g., *Sphaeroeca*). They are important grazers of the smallest suspended bacteria, especially in the marine plankton (e.g., *Salpingoeca*, *Stephanoeca*, and *Parvicorbicula*).

Euglenids usually have two flagella which emerge from a small anterior pocket. This is a large group of planktonic and benthic protozoa, common in both fresh and marine waters. Species associated with surfaces typically keep one trailing flagellum in contact with the substrate, whereas the other pulls the cell forward. Many species are green and phototrophic (e.g., *Euglena*), but there is a great diversity of phagotrophs (e.g., *Astasia*, *Petalomonas*, *Entosiphon*, and *Notosolenus*). These are especially common in marine shallow-water sediments, in which they graze on bacteria. All bodonids are small flagellates with characteristic heterodynamic flagella (e.g., *Bodo* and *Rhynchomonas*). They are almost always found in organically polluted (even anoxic) waters or soils that are rich in bacteria, on which they feed. Many closely related (“kinetoplastid”) flagellates are parasites: *Ichthyobodo* causes “costiasis” in freshwater fish and the “trypanosomatids” are important parasites of man (*Trypanosoma brucei* causes sleeping sickness). In heterokont flagellates, each organism has two flagella—one hairy and one smooth, with the former creating the water current that brings mainly bacterial food to the (filter-feeding) flagellate. They include the bicosoecids that live in a lorica (*Bicosoeca*) or are naked (*Cafeteria* and *Pseudobodo*), that may be stalked

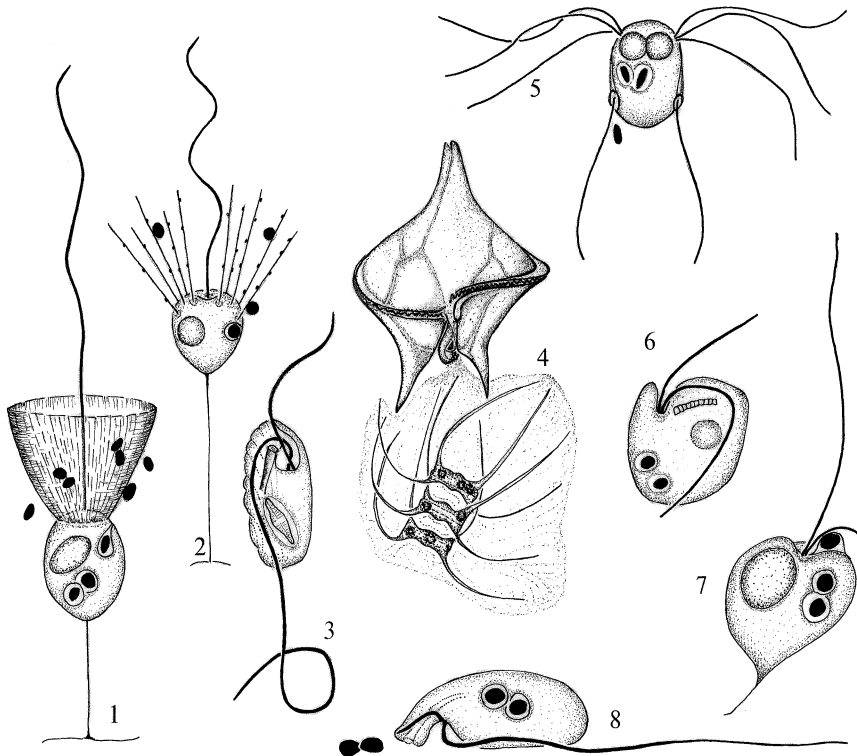


FIGURE 2 A selection from the variety of form and function in flagellated protozoa. 1, a choanoflagellate with bacterial food trapped on the external surface of the collar filter (*Monosiga*; body + collar  $\sim 0.015$  mm); 2, a bacteria-feeding helioflagellate (*Pteridomonas*; body 0.007 mm); 3, a phagotrophic euglenid (*Dinema*;  $\sim 0.03$  mm) with an ingested diatom; 4, a marine heterotrophic dinoflagellate (*Protooperidinium*;  $\sim 0.06$  mm) with a diatom trapped in its feeding veil; 5, a free-living, bacteria-feeding, anaerobic diplomonad (*Hexamita*; body  $\sim 0.008$  mm); 6, a small heterotrophic (bacteria-feeding) cryptomonad flagellate (*Goniomonas*;  $\sim 0.008$  mm); 7, a heterotrophic, typically bacteria-feeding, chrysomonad flagellate (*Spumella*; body  $\sim 0.008$  mm); 8, a bacteria-feeding bodinid (*Rhynchomonas*; body  $\sim 0.005$  mm).

or colonial, and which resemble the mixotrophic loricate chrysomonads (e.g., *Dinobryon*).

In the phagotrophic chrysomonads (e.g., *Paraphysomonas* and *Spumella*), the chloroplast is essentially vestigial (and the cell colorless). Functional types range from heterotrophs (e.g., *Paraphysomonas*) to mixotrophs (e.g., *Dinobryon* and *Uroglena*) and those that are predominantly phototrophs (e.g., *Synura* and *Mallomonas*). Chrysomonads are the most abundant heterotrophic flagellates in the freshwater plankton. The helioflagellates (also referred to as "pedinellid stramenopiles," e.g., *Actinomonas*, *Pteridomonas*, and *Ciliophrys*) are filter or diffusion feeders that superficially resemble heliozoa and/or choanoflagellates.

The haptomonads are typically photosynthetic, biflagellated organisms. They include the coccolithophorids and the prymnesiophytes, and some (e.g., *Emiliania*) are often abundant in the marine euphotic zone.

They have an additional flagellar appendage known as the "haptonema." In the marine *Chrysochromulina*, this is used in food uptake, but in *Prymnesium*, which has a very short haptonema, ingestion of food (dinoflagellates, chrysomonads, and green algae) is by means of a posterior pseudopodium. Prey organisms may be immobilized by toxins prior to ingestion (toxic bloom-forming species of *Prymnesium* are responsible for fish kills). The scale and importance of phagotrophy in freshwater haptomonads is unclear.

Most cryptomonads are yellow-brown and phototrophic, with two flagella of similar length (e.g., *Cryptomonas*). They are especially common in fresh waters. Heterotrophic species either lack plastids (*Goniomonas*) or have a plastid that is devoid of photosynthetic pigments (*Chilomonas*). Colorless species can be relatively abundant, especially when associated with benthic detritus.

The heteromitids and cyathobodonids represent a



varied assemblage of small, poorly studied flagellates (e.g., *Apusomonas*, *Heteromita*, *Cercomonas*, *Cyathobodo*, and *Kathablepharis*), some of which (e.g., *Cercomonas*) produce pseudopodia to ingest bacteria. Others secrete a stalk (*Cyathobodo*) to attach the cell to the substrate. Many are common in soil. Others feed mainly on bacteria in the freshwater plankton (e.g., *Kathablepharis*, which also feeds on small algae) and on sediment surfaces. The percolozoans are a diverse collection of bacteria-feeding organisms, including amoeboflagellates (e.g., *Naegleria*), the filter-feeding flagellate *Percolomonas*, and anaerobic flagellates with hydrogenosomes and endosymbiotic methanogens (*Psalteriomonas*).

Dinoflagellates are best known as a large group of photosynthetic flagellates and for the ability of some species to cause “toxic blooms” and “red tides” (*Protogonyaulax* causes paralytic shellfish poisoning). However, about half of all known marine species lack chloroplasts. Some of these (in the genera *Gymnodinium* and *Amphidinium*) feed on cryptomonads and sequester their chloroplasts, thus transforming themselves into mixotrophs. Moreover, many typically photosynthetic species can become mixotrophs through their ability to ingest particulate food. The non-photosynthetic and mixotrophic species are referred to collectively as heterotrophic dinoflagellates. These can be quantitatively important in marine food webs, especially as consumers of diatoms. Some species are benthic, but little is known about these. The diversity of heterotrophic species in fresh waters is much less, and it is known mainly from the genera *Katodinium*, *Peridinium*, *Gymnodinium*, and *Ceratium*. This may be due in part to the relative rarity in fresh waters of large, chain-forming diatoms—a typical food item of many marine dinoflagellates. These large food items are trapped and digested externally in a pseudopodial feeding veil or “pallium.” This explains why it has often been difficult to recognize ingested organisms within dinoflagellates. Other heterotrophic dinoflagellates use a feeding tube known as a peduncle. The common freshwater dinoflagellate *Peridiniopsis berolinensis* uses such a tube to ingest the fluid contents from injured and dying protists and small metazoans.

The number of marine dinoflagellate species with a recorded capacity for phagotrophy is increasing rapidly. These include thecate species (e.g., the peridinioids) previously considered to be incapable of phagotrophy. It is widely believed that most dinoflagellates—at least in marine environments—may be capable of phagotrophy. Heterotrophic dinoflagellates (e.g., *Oblea*, *Polykrikos*, *Heterocapsa*, and *Protoperidinium*) can dominate the protozoan biomass in coastal and oceanic waters, in which they feed on bacteria, flagellates, diatoms, other dinoflagellates, and ciliates. Their biomass can be

approximately the same as that of ciliates, and the two groups may compete with each other for food (although the size range of dinoflagellates is slightly greater—from 3 or 4  $\mu\text{m}$  in *Gymnodinium simplex* to 2 mm in *Noctiluca*, with the majority in the size range 20–200  $\mu\text{m}$ ). Also, most planktonic ciliates are incapable of ingesting the large diatoms that are consumed by dinoflagellates.

Almost all stony and stinging corals (Cnidaria) in shallow tropical waters harbor photosynthetic dinoflagellates (especially *Symbiodinium*) as symbionts. These consume carbon dioxide and bicarbonate. This promotes calcium deposition in the external skeletons of their hosts and may enhance the rate with which coral reefs are built. In the marine plankton, outbursts of parasitic dinoflagellates are sometimes associated with the collapse of zooplankton populations. The parasite *Ichthyodinium chabelardi* destroys the eggs of sardines, *Blastodinium* sp. in the copepod gut castrates its host, and *Gonyaulax catenella* (a dinoflagellate that can cause red tides) is parasitized by another dinoflagellate (*Amoebophrya ceratii*). The fossil record of dinoflagellate cysts extends over the past 220 million years, and they have a role as biostratigraphic tools.

Dinoflagellates in the marine plankton exhibit “latitudinal cosmopolitanism,” such that a morphospecies occurs in a circumglobal belt within fairly broad latitudinal limits which correspond to a specific temperature range in the surface waters. The recorded species richness is greatest in tropical marine waters.

### C. Ciliated Protozoa

This diverse and distinctive group uses cilia for locomotion and feeding (Fig. 3). They demonstrate a considerable adaptive radiation of feeding mechanisms and cell morphologies (e.g., the ribbon-shaped forms *Tracheloraphis* and *Geleia* adapted for life in the marine interstitial). The smallest species tend to feed on bacteria-sized particles and the larger species on unicellular algae, filamentous cyanobacteria, other protozoa, and even rotifers and other microzooplankton. The raptorial feeders (e.g., *Prorodon*) use a simple mouth to catch diatoms, dinoflagellates, and other large food items individually; some (e.g., *Lacrymaria*) kill motile prey, whereas others (e.g., *Chilodonella*) “hoover” diatoms and other elongate food particles from surfaces. The filter feeders (e.g., *Cyclidium*) use fine-mesh filters to sieve suspended bacteria, and some of these ciliates (e.g., *Paramecium* and *Tetrahymena*) thrive in habitats with very high bacterial concentrations. Other ciliates (e.g., *Pleuronema* and *Tintinnopsis*) use coarser filters to collect small algae. In many ciliates (e.g., *Oxytricha*,

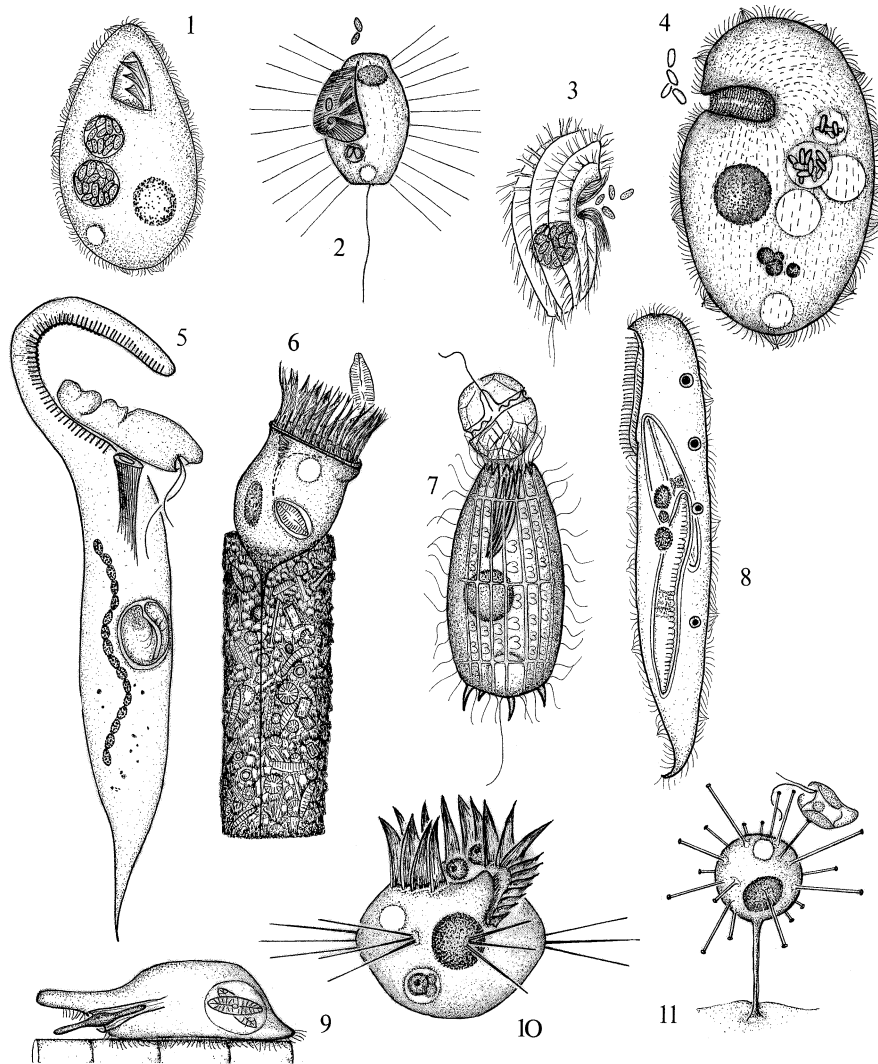


FIGURE 3 A selection from the variety of form and function in ciliated protozoa. 1, a bacteria-feeding tetrahymenid (*Tetrahymena*; length  $\sim 0.05$  mm); 2, a small scuticociliate (*Cyclidium*; length  $\sim 0.02$  mm) filter feeding on bacteria; 3, a bacteria-feeding colpodid ciliate (*Colpoda*; length  $\sim 0.02$  mm) from soil; 4, an anaerobic, typically benthic plagiopylid feeding on bacteria (*Plagiopyla*; length  $\sim 0.1$  mm); 5, a large haptorid (*Dileptus*; length  $\sim 0.4$  mm) with toxicysts used to kill motile prey (flagellates and other ciliates) prior to ingestion; 6, a planktonic, loricate, diatom-feeding tintinnid (*Tintinnopsis*; length  $\sim 0.1$  mm); 7, a gymnostome ciliate (*Coleps*; length  $\sim 0.07$  mm) about to ingest a dinoflagellate; 8, a diatom-feeding karyorelictid (*Remanella*; length  $\sim 0.12$  mm) typical of the marine interstitial; 9, a "hoover-feeding" cyrtophorid (*Chilodonella*; length  $\sim 0.04$  mm) ingesting a diatom; 10, a planktonic oligotrich (*Halteria*; diameter  $\sim 0.04$  mm) filter feeding on small algae; 11, a "diffusion-feeding" suctorian (*Podophrya*; diameter  $\sim 0.04$  mm) with a flagellate trapped on a feeding tentacle.

*Aspidisca*, and *Strombidium*), a row of membranelles generates a water current and acts as a relatively coarse feeding filter, and some of these ciliates (e.g., *Euplotes*) feed most efficiently if they are raised on cirri (fused cilia). Many ciliates are typically sessile and aligned perpendicular to the substrate (e.g., *Stentor* and *Vorti-*

*cella*). Diffusion feeders (e.g., *Podophrya* and other Suctorina) catch swimming prey (usually other protozoa) that collide with their sticky tentacles.

More is probably known about the biodiversity and ecology of free-living ciliates than any other protozoan group. There are several reasons for this, including the

distinctive and immediately recognizable ciliate morphology and swimming behavior and the relative ease with which many species can be cultured. They are a group of predominantly free-living protozoa. A few are parasites (e.g., *Ichthyophthirius* infects fish and there is one human endoparasite of minor importance—*Balantidium coli*), but most species are either free-living or harmless commensals of aquatic invertebrates (the ciliate epifauna of marine crustaceans is particularly diverse). Free-living species are known from all natural aquatic habitats in which temperatures are  $<45^{\circ}\text{C}$ , including oceanic sinking detritus, freshwater and marine sediments, anaerobic municipal landfill sites, and sewage treatment plants. They are abundant (up to  $\sim 10^6$  per milliliter) in activated sludge plants, in which they consume bacteria and also flocculate bacteria and other suspended particulate matter. These activities aid the clarification of the effluent and the formation of sludge. About a dozen ciliate species are frequently recorded in used-water treatment plants, although more than 200 ciliate species have been recorded. Ciliates also have a role as indicators of the level of organic pollution in river water.

Most ciliates are in the size range 0.02–2 mm, so they are generally larger than the heterotrophic flagellates and other nanoplankton (0.002–0.02 mm) on which many of them feed. Planktonic ciliates are relatively abundant (1–100 per milliliter) and important grazers of nanoplankton in marine and fresh waters. They are probably key grazers within the “microbial loop” which is responsible for the rapid remineralization of organic matter in the water column. The diet of planktonic metazoans includes ciliates, although the quantitative significance of this link is unclear. Benthic ciliates are often abundant ( $>1000$  per milliliter) and usually the most important grazers in freshwater sediments (especially in lakes) and marine sandy sediments of inshore waters.

Many ciliate species harbor prokaryotic and/or eukaryotic symbionts. At the oxic–anoxic boundary in the water column of lakes, most ciliates may harbor sufficient endosymbiotic algae (*Chlorella*) to render the consortia capable of net photosynthesis in dim light. The marine interstitial ciliate *Kentrophoros* carries ectosymbiotic chemolithotrophic sulfide-oxidizing bacteria, and most anaerobic ciliate species harbor methanogenic bacteria that act as a sink for (potentially inhibitory)  $\text{H}_2$  produced by the ciliate. The endosymbiotic nonsulfur purple bacteria in the anaerobic ciliate *Strombidium purpureum* use waste  $\text{H}_2$  from the ciliate as a reductant for photosynthesis. Most marine anaerobic ciliates carry ectosymbiotic sulfate-reducing bacteria.

Many ciliate species are microaerobic, and they seek out the (microbe-rich) oxic–anoxic boundary in sediment or the water column. Some of these species are facultative anaerobes (in the genera *Cyclidium*, *Euplotes*, *Strombidium*, and *Paranophrys*). There are approximately 70 known species of anaerobic free-living ciliate species, mainly in the genera *Metopus*, *Caenomorpha*, *Saprodinium*, *Epalxella*, *Trimyema*, and *Plagiopyla*. It is likely that all these contain hydrogenosomes.

Anaerobic ciliates also live as endocommensals in the enlarged forestomach (rumen) of ruminants and in the caecum of other mammals with postgastric fermentation. It is unlikely that any of these ciliates (e.g., in the genera *Dasytricha*, *Entodinium*, and *Polyplaston*) are capable of a free-living existence. Rumen ciliates are typically extremely abundant ( $10^5$ – $10^6$  per milliliter of rumen liquor). They consume bacteria and microscopic fragments of grass and other plant material. Some species are capable of degrading cellulose and other structural carbohydrates, and the endosymbiotic methanogenic bacteria in some species contribute to the methane emitted from ruminants. The economic significance of such large numbers of ciliates living in the rumen is unclear.

Ciliates also live in soil, in which all species are probably capable of producing desiccation-resistant cysts (e.g., *Colpoda*). The abundance of active ciliates in soil is extremely variable and reflects repeated cycles of cyst formation and excystment in response to fluctuating physical factors such as the level of soil moisture. Many species found in soil are frequently found in other (predominantly freshwater) habitats, although only about 100 ciliate species have been described from soil.

## D. Others

The following are often referred to as protozoa. Because they are endocommensals or parasites, or incapable of phagotrophy, they do not fall within the definition of protozoa used here (i.e., free-living unicellular phagotrophs). They are mentioned briefly for the sake of completeness but are excluded from Table I.

Chytridiomycetes, oomycetes, and hyphochytriomycetes have the typical fungal characteristics of saprotrophic (absorptive) nutrition and cell walls in the vegetative state. Some are economically important (the oomycete *Phytophthora infestans* causes potato blight). The plasmodiophorids are cell wall-free endoparasitic slime molds (*Plasmodiophora brassicae* causes club root disease in cabbages). The labyrinthulids (e.g., the “slime net” *Labyrinthula*) feed saprotrophically, especially on marine algae, and form complex branching colonies of

spindle-shape cells that move through slime channels. All five groups reproduce by means of unicellular flagellated zoospores (hence, they are also known as “zoosporic fungi”).

The opalinids superficially resemble ciliates, but all are mouthless endocommensals living in the hindgut of amphibians and some fish. The microsporidians are a large group (~800 species) of intracellular parasites of animals and other protozoa. They extrude a tubular filament through which the spore is injected into the host cytoplasm. They are believed to be highly derived fungi rather than early diverging eukaryotes. The haplosporidians are a small group of parasites of aquatic invertebrates, notably marine mollusks, and the parabasalans (anaerobic flagellates with hydrogenosomes, e.g., *Trichomonas*) are all parasites, with one or two possible exceptions (*Pseudotrichomonas*).

The organisms now referred to as apicomplexans used to be an important component of the “sporozoans.” All are intracellular parasites of animals, and many cause life-threatening diseases of man. They share the distinctive morphological feature of an “apical complex” at certain stages in the polymorphic life cycle. The degree of host specificity of many apicomplexans is not known. They include the “gregarines” that infect insects, marine polychaetes, and other invertebrates (e.g., *Monocystis* in earthworms), and also the “coccidian” parasites including *Toxoplasma gondii* (the domestic cat is the final host, but intermediate infection in the human can cross the placenta) and about 1050 species of *Eimeria* infecting mammals, chickens, and other farmed birds. *Cryptosporidium* is an apicomplexan, and 8 species are known that infect the digestive and respiratory systems of vertebrates. *Cryptosporidium parvum* infects humans and cattle, and waterborne transmission of oocysts is thought to complete the fecal–oral cycle. Infection can cause life-threatening diarrhea in immunocompromised humans. *Cryptosporidium* may be sufficiently different from other coccidians to warrant creation of the new higher taxon “cryptosporidia.” *Plasmodium*, the causative agent of malaria, is an apicomplexan, as is *Pneumocystis*, which is an enigmatic parasite that can persist in the human lung for long periods. *Pneumocystis carinii hominis* causes *Pneumocystis* pneumonia in AIDS patients. In the early 1970s, the organism was considered to be a fungus; in the early 1980s it was considered to be an apicomplexan, and by the late 1980s it was again considered to be a fungus, possibly an ascomycete. Confusingly, it can resemble *Cryptosporidium* when viewed with the light microscope. There are approximately 4000 nominal species of apicomplexans.

The Myxozoans are a large group (~1200 species) of parasites, of which the myxosporidians are important parasites of farmed fish (*Myxobolus cerebralis* causes “torsion disease” in trout). During their life cycle they show both protozoan and metazoan characters. Recent gene sequence data (16S-like rDNA and Hox) indicate that myxozoans are grossly modified and simplified metazoans.

## Acknowledgments

I am indebted to the following for the help and suggestions they provided: O. R. Anderson, S. Brown, D. Boltovskoy, J. O. Corliss, M. Cushion, G. F. Esteban, T. Fenchel, A. J. Gooday, P. J. Hansen, J. J. Lee, C. Nigrini, D. J. Patterson, and A. Rogerson.

## See Also the Following Articles

EUKARYOTES, ORIGIN OF • MICROBIAL  
BIODIVERSITY • MICROORGANISMS, ROLE OF • PLANKTON,  
STATUS AND ROLE OF • PREDATORS, ECOLOGICAL ROLE OF

## Bibliography

- Anderson, O. R. (1983). *Radiolaria*. Springer-Verlag, New York.
- Corliss, J. O. (1979). *The Ciliated Protozoa*, 2nd ed. Pergamon, New York.
- Corliss, J. O. (1999). Biodiversity and numbers of species of Protists. In *Nature and Human Society: The Quest for a Sustainable World* (P. Raven and T. Williams, Eds.), National Academy Press, Washington, D.C.
- Fenchel, T. (1987). *The Ecology of Protozoa*. Madison, WI: Science Tech.
- Fenchel, T. (1993). Are there more small than large species? *Oikos* 68, 375–378.
- Fenchel, T., and Finlay, B. J. (1995). *Ecology and Evolution in Anoxic Worlds*, Oxford Series in Ecology and Evolution. Oxford Univ. Press, Oxford.
- Finlay, B. J. (1998). The global diversity of protozoa and other small species. *Int. J. Parasitol.* 28, 29–48.
- Finlay, B. J., Maberly, S. C., and Cooper, J. I. (1997). Microbial diversity and ecosystem function. *Oikos* 80, 209–213.
- Gooday, A. J., Bett, B. J., Shires, R., and Lamshead, P. J. D. (1998). Deep-sea foraminiferal species diversity in the NE Atlantic and NW Arabian Sea: A synthesis. *Deep-Sea Res. Part II* 45, 165–201.
- Lee, J. J., and Anderson, O. R. (1991). *Biology of Foraminifera*. London, Academic Press.
- Lee, J. J., et al. (2000). *Illustrated Guide to the Protozoa*, 2nd ed. Society of Protozoologists/Allen Press. Lawrence, Kansas.
- Lee, J. J., and Patterson, D. J. (1998). Diversity and geographic distribution of free-living heterotrophic flagellates—Analysis by PRIMER. *Protist* 149, 229–244.
- Margulis, L., Corliss, J. O., Melkonian, M., and Chapman, D. J. (Eds.) (1990). *Handbook of Protozoa*. Jones & Bartlett, Boston.
- Page, F. C., and Siemensa, F. J. (1991). *Nachte Rhizopoda und Heliozoa*, Protozoenfauna Band 2. Fischer, Stuttgart.
- Patterson, D. J., and Larsen, J. (Eds.) (1991). *The Biology of Free-Living Heterotrophic Flagellates*, Systematics Association Special Volume No. 45. Clarendon Press, Oxford.





# PSYCHROPHILES, ORIGIN OF

Richard Y. Morita\* and Craig L. Moyer†

\*Oregon State University and †Western Washington University

- I. Definitions and Historical Background
- II. The Environment and Its Microflora
- III. Biodiversity of Psychrophiles
- IV. Evolution of Psychrophiles
- V. Physiology of Psychrophiles (Metabolic Activities)

focuses on their environment, biodiversity, and physiology.

## I. DEFINITIONS AND HISTORICAL BACKGROUND

### GLOSSARY

- barophile** Pressure-loving bacteria.
- cryobiosis** Anabiosis (latent life) due to freezing.
- endolithotrophic** Living inside rocks, usually sandstone.
- homeophasic adaption** The adaptation of the membrane to maintain the bilayer phase.
- meltwater** Ice or snow melted by radiant energy in the polar regions.
- permafrost** Ground (ice, bedrock, and soil) that remains frozen below 0°C for more than 2 years.
- thermocline** In the stratification of warm surface water over cold, deeper water, the transition zone of rapid temperature decline between the two layers.
- upwelling** Transport of water from the deep ocean to the surface, replacing the surface water that has moved offshore.

Psychrophilic bacteria are defined as cold-loving bacteria. Specifically, their cardinal temperatures are 20°C for maximal growth, 15°C or lower for optimal growth, and 0°C or lower for minimum growth (Morita, 1975), and this definition is accepted by most microbiologists. The old definition of psychrophiles applied to those organisms that produced a visible colony in 1 week at 0°C. However, it is important to recognize that there is a continuum of cardinal temperatures for the diverse microbes in nature. From an ecological standpoint, psychrotrophs and psychrophiles are both found in cold environments, but psychrophiles are not found at temperatures higher than 20°C. This temperature was selected as the maximum temperature for growth based on the fact that laboratory temperatures in the United States are approximately 21 or 22°C. In higher organisms, the cold-loving organisms are known as cryophiles.

Microorganisms capable of growing at 5°C or lower are psychrotrophs, regardless of the optimum temperature for growth. The psychrotrophs are cold-tolerant bacteria, but their maximal growth temperature ranges above 20°C and in many cases their optimal growth temperature is also above 20°C. A better term for these organisms that withstand cold temperatures is psychro-

**PSYCHROPHILES** are cold-loving bacteria, whereas psychrotrophs are cold-tolerant bacteria. This article

tolerant. However, because of common usage this term should be retained. For example, there are a few texts employing the term "psychrotrophs" in their title and this term is widely used by industry (mainly dairy and food), e.g., the normal souring of milk is due to psychrotrophs. Furthermore, precedence should be adhered to in keeping the name.

Bacteria capable of growing at 0°C were first reported by Foster in 1887 and 1892. The source material for these bacteria were from fish, natural waters, foods, wastes, rubbish, soil surface, and intestines of fish. The lowest temperature at which bacteria can grow remains to be determined definitely and -12°C is the lowest temperature reported. The term "psychrophile" was first used in 1902 by Schmidt-Nielsen. In 1903, Müller objected to this term because the organisms described actually grow well at higher temperatures. As a result, cold-tolerant bacteria were called cryophile, Glaciale Bakterien, rhigophile, psychrotolerant, psychrocartericus, psychrobe, thermophobic bacteria, facultative psychrophile, obligate psychrophile, and psychrotrophic. A review of the literature indicates that the organisms described were actually psychrotrophs (as defined previously) with the possibility of one psychrophilic exception. For many years, it was thought that there were no bacteria that could be termed psychrophiles, only yeasts and certain algae. The lack of refrigeration equipment in the early days added much to this confusion. Thus, the term psychrophile was considered a misnomer until true psychrophiles were isolated by three different laboratories in 1964. The confusion was ended when Morita (1975) defined the term psychrophile (as noted previously). For further details, Morita (1975) should be consulted.

## II. THE ENVIRONMENT AND ITS MICROFLORA

Eighty percent of the earth's biosphere is permanently cold and the average temperature of the earth is 15°C. Permafrost occupies 20% of the earth's surface: 80% of Alaska, 50% of Canada, 20% of China, and 50% of Russia are covered by permafrost. However, the amount of microbiological research done on psychrophiles is extremely low compared to that done on thermophilic bacteria.

The polar regions comprise about 14% of the earth's surface. Approximately 71% of the earth's surface is ocean, and more than 90% (by volume) of the oceans are 5°C or colder. Other cold environments include

caves, the tops of mountains, certain rivers and streams, the upper atmosphere (10°C or less at 1000 m, decreasing as the altitude increases), snow and ice, and the water below the thermocline of freshwater lakes; each has its own microbial flora (Baross and Morita, 1978).

### A. Upper Atmosphere

Air samples collected near the earth's surface up into the stratosphere exceeding 27,000 m have shown the presence of viable bacteria, viable fungi, pollen, and other microscopic particles. A high incidence of viable bacteria has been reported within the troposphere at approximately 10,000 m, where the temperature may be lower than -40°C. The highest incidence of bacteria appears to be at altitudes slightly higher than 10,000 m, where the temperature is lower than 10°C. Although this temperature is well within the cardinal temperature range of psychrophile and psychrotrophs, there are no reports of psychrophiles in the upper atmosphere. Furthermore, there is organic matter (cobalamin, biotin, and niacin) in the air but the concentration is very low. Thus, these organisms have the ability to resist starvation, drying, freezing, and radiation. These microbes can be the nucleation site for the formation of ice and snow.

However, in the lower atmosphere of the polar regions, nutrient agar plates were media exposed to the air and showed the presence of viable bacteria. Unfortunately, there are no reports of seawater-requiring psychrophilic bacteria from air samples. In all probability, the atmosphere, especially above the polar regions, does contain psychrophiles and psychrotrophs.

### B. Caves

There are many glaciated and subterranean caves in which the permanent temperature is about 10°C to below freezing. In addition to the low temperature, there is an absence of light, low levels of organic material, and relatively high moisture. Generally, there is an absence of psychrophiles in these caves, but psychrotrophs are found.

### C. Arctic

The similarities of the Arctic and Antarctic polar regions are the continuous sunlight during the summer and the total lack of it during the winter, the presence of sea ice during much or all of the year, and the cold temperature. On the other hand, they differ in that the Arctic has major riverine inputs, but the Antarctic does not.

They also differ in their landmasses, topographical features, large-scale water transport features, and the magnitude of nutrient supply. Relative microbial activities, as measured by the incorporation and respiration of  $^{14}\text{C}$ -labeled glucose and glutamic acid (called the heterotrophic potential method by marine microbiologists and substrate-induced respiration by soil microbiologists), are comparable in both the Arctic and Antarctic near-shore water. These measurements were made by the same laboratory working at different times in both regions.

#### D. Antarctic

About 98% of the surface of Antarctica is ice, leaving 2% of the continent ice free. The lowest temperatures on Earth occur on this continent and it has the lowest precipitation and relative humidity levels, making it the driest area on Earth. As a result, dry valleys and ice-free areas occur in addition to many other types of environments (e.g., temporarily and permanently ice-covered lakes, transient rivers and streams, oligotrophic lakes, and moss and peat bogs), with all sharing low temperatures. Freezing, like drying, will produce hypersaline lakes, both thallosaline (seawater derived) and athallosaline (brine derived from geological and geochemical sources). For a few weeks of the austral summer flowing liquid water can occur, producing ice melts around the margins of frozen lakes. Study of the microbiology of these meltwaters has just begun. In all probability, psychrophilic bacteria, which are also alkalophilic and/or halophilic, will be found. Even in the dry valleys of Antarctica, numerous psychrotrophs and mesophiles have been isolated. Because freezing is equivalent to drying, the indigenous microflora have the ability to tolerate low temperatures in addition to being xerotolerant (i.e., able to live in very dry environments). Within sandstones of the dry valleys, the presence of cryptoendolithotrophic bacteria has been shown along with that of algae (Friedmann, 1993). These bacteria, mainly cyanobacteria, only grow during the austral summer of the Antarctic when the temperature, due to solar radiation, rises sufficiently high to produce liquid water. It should also be mentioned that there is a greater predominance of cyanobacteria and psychrophiles in the Antarctic than in the Arctic.

#### E. Permafrost

The discovery of microorganisms in permafrost was initiated in the 1930s in the Trans-Baikal and North Ural regions, Central Yakutia, and Arctic islands (Gili-

chinsky, 1995). Permafrost cores yielded numerous microbes. Microbes have been reported in all the Arctic and Antarctic permafrost environments except in the lower strata of permafrost ice of Lake Vostok in Antarctica. This exception may be due to the techniques employed. Unfortunately, the early microbiologists did not recognize the abnormal thermolability of psychrophiles, and as a result psychrotrophs and mesophiles were mainly demonstrated. In the early literature, thermophiles (as high as 100 cells/g) were reported in Arctic permafrost (peat cores 10,525 years old). Viable microorganisms were reported from Holocene sedimentary deposits and Pliocene deposits (3–5 million years old). Viable fungi, spore-forming bacteria, and non-spore-forming bacteria were in the Antarctic ice sheet up to a depth of 300 m (11,600 years old) but not at 320 m (13,000 years old), whereas in a 107,000- to 200,000-year-old layer of glacier ice, viable organisms were found (Friedmann, 1993). In permafrost material millions of years old, thermophiles, mesophiles, and psychrophiles have been found (Vorobyova *et al.*, 1997). This ancient material does not have sufficient energy in it to keep the organism metabolizing for long periods of time; hence, one has to ask where the energy is coming from. In most cases, the energy supplied has been utilized since any ecological niche has a large number and consortium of microbes that utilize the energy source leaving the more recalcitrant material. This recalcitrant material, if used metabolically, will require exoenzymes to degrade it to its monomers so that it can be transported into the cell. In addition, many substrates that the cell utilizes require an active transport mechanism which also consumes energy. Because of this situation, microbes in ancient material, such as the deeper strata of permafrost, are in an anabiotic stage. Specially, the anabiotic stage is cryobiosis, brought about by the low temperature. However, the cells are also in a starvation-survival mode. Both processes generate stress proteins, some of which are shared. The Russian microbiologist Omelyanski (1911) attributed the existence of microbes in ancient cold environments to anabiosis, a term not employed by most microbiologists. Nevertheless, one always must ask “where is the energy?”

For further information concerning viable microorganisms in permafrost, Gilichinsky (1994) should be consulted.

#### F. Oceans

The area below the thermocline is approximately  $5^{\circ}\text{C}$  or colder. Thus, the deep sea is permanently cold. Grad-



ually, as the higher latitudes are approached, the depth of the thermocline decreases until it is at the surface of the ocean. In the deepest portions of the oceans (trenches and deeps), barophiles can be isolated that are also psychrotrophic or psychrophilic. At the edge of the ice where ice is melting or being formed, there is a sea ice environment with an associated microbial community. All types of microorganisms can be demonstrated in waters below the thermocline, mainly because there are many different types of habitats available.

### G. Sea Ice

Where the seawater is being frozen, a sea ice microbial community resides. The freezing process helps concentrate the dissolved nutrients, and when the ice melts the concentrated nutrients in the ice are released so that sufficient energy is present to permit the growth of this microbial community. Thus, this community is influenced mainly by the onset of spring and winter. Within this sea ice community a variety of bacteria are found displaying numerous morphological types. Seventy percent of the bacteria in the sea ice community are free living, and ice nucleation bacteria can also be found. Various pigmented and gas vacuolate bacteria are also present and 65% of the epiphytic bacteria are found associated with the diatom *Amphiprora*. The best growth response of the sea ice microbial community occurs between 5 and 10°C.

## III. BIODIVERSITY OF PSYCHROPHILES

The first known species of psychrophiles described taxonomically are *Vibrio* (*Moritella* gen. nov.) *marinus* MP-1 and *Vibrio* (*Colwellia* gen. nov.) *psychoerythrus*, both isolated in 1964. The biodiversity among psychrophiles in the various cold environments has yet to be studied extensively. Nevertheless, the various species within the genera *Achromobacteria*, *Alcaligenes*, *Altermonas*, *Aquaspirillum*, *Arthrobacter*, *Bacillus*, *Bacteroides*, *Brevibacterium*, *Clostridium*, *Colwellia*, *Cytophaga*, *Flavobacterium*, *Gelidibacter*, *Methanococcoides*, *Methanogenium*, *Methanosarcina*, *Microbacterium*, *Micrococcus*, *Moritella*, *Octadecabacter*, *Phormidium*, *Photobacterium*, *Polaribacter*, *Polaromonas*, *Pseudomonas*, *Psychroserpens*, *Shewanella*, and *Vibrio* have been found to be psychrophilic. Even a psychrophilic methanotroph (resembling *Methylococcus*) has been isolated from the tundra. Both the domains *Bacteria* and the *Archaea* are represented in the various genera listed. To date, only the genus

*Moritella* appears to be composed of psychrophiles only. However, it should be noted that many species of reported psychrophiles do not meet the definition given in this article. Because of this situation, the cardinal temperatures of many psychrophiles need to be determined.

All psychrophilic and barophilic bacteria that have been cultivated belong to the subdivision  $\gamma$ -Proteobacteria: *Shewanella*, *Photobacterium*, *Colwellia*, *Moritella*, and a new group designate containing strain CNPT3 and *Alteromonas haloplanktis*. The indication is that the combined barophilic and psychrophilic phenotype evolved independently in the different  $\gamma$ -Proteobacteria genera (DeLong *et al.*, 1997).

Several psychrophilic, gas vacuolate bacteria have been isolated from the sea ice and water from both the Antarctic and the Arctic. These include representatives belonging to the  $\alpha$ -,  $\beta$ -, and  $\gamma$ -Proteobacteria subdivisions as well as the Cytophaga-Flavobacterium-Bacteroides phylogenetic group (Gosink *et al.*, 1998). These results demonstrate a wide range of phylogenetic diversity capable of psychrophily across the domain Bacteria.

Incredibly, as much as 30% of the marine picoplankton (planktonic organisms with an average diameter of 0.2–2.0  $\mu\text{m}$ ) from both polar and temperate coastal waters are also Archaea (i.e., archaeoplankton) and the majority of these are associated with the Crenarchaeota. Despite this cosmopolitan distribution in the world's oceans, little is known about the physiological properties of these archaeoplankton, with the exception that they are hypothesized to potentially be psychrophilic based on the marine sponge symbiont *Cenarchaeum symbiosum* (Preston *et al.*, 1996). However, similar archaeal phylotypes have been located at a deep-sea hydrothermal vent system (Moyer *et al.*, 1998) in waters with an environmental temperature of 15–30°C. Many hyperthermophiles are members of the Archaea that can utilize H<sub>2</sub> as an energy source, and recently *Methanogenium frigidum*, a psychrophilic, slightly halophilic, H<sub>2</sub>-using methanogen, was isolated from the perennially cold, anoxic hypolimnion of Ace Lake, Antarctica (Franzmann *et al.*, 1997). Although it may be too early to state that this may be an indication of the evolutionary processes affecting both hyperthermophiles and psychrophiles, this does demonstrate that members of the domain Archaea are also capable of the psychrophilic lifestyle and that further research is needed. The main deterrent is the lack of isolation of psychrophiles, mainly because interest in this thermal group was lacking for many years. Yet to be isolated from the meltwater

in the Antarctic are psychrophilic bacteria that are also alkaliphiles. Each body of water in the Antarctic has its own cations and anions and the salinity may be very high due to freezing of the water.

In any cold environment in which microbes have been isolated, many of the isolates are psychrotrophs. When all things are equal, the psychrophiles will outgrow the psychrotrophs at low temperature.

#### IV. EVOLUTION OF PSYCHROPHILES

The common approach used in microbiology is to apply phylogenetic analysis to establish evolutionary relationships among organisms and to use this as a framework for making inferences about community structure, making inferences about genetic and thereby inferred organismal diversity, and (to a lesser degree) to infer physiological adaptation when applicable. This approach is possible due to the detailed theory of evolutionary relationships among the domains Bacteria, Archaea, and Eucarya that has emerged from comparisons of ribosomal RNA "signature" sequences. Most researchers currently infer that the earliest common ancestor to all three domains was a H<sub>2</sub>-metabolizing hyperthermophile based on the deepest branching lineages found in the domains Bacteria and Archaea (Pace, 1991). If we accept this premise, then the thermophiles must have evolved from the hyperthermophiles, followed by the mesophiles and finally the psychrophiles. This hypothesis has been challenged by criticism that the hot origin of life scenario is not compatible with the RNA world hypothesis (because of the instability of RNA at near water boiling temperatures) and/or by the argument that hyperthermophiles are not primitive but have instead successfully adapted to temperatures higher than the limits imposed on other life-forms (Forterre, 1996). In addition, dibiphytanyl ether lipids used in the formation of lipid "monolayer" cell membranes in hyperthermophilic Archaea were found in nonthermophilic members of the kingdom Crenarchaeota, which appear to be a major source of tetraether lipids in the marine planktonic environment (DeLong *et al.*, 1998). Recently, a nonhyperthermophilic earliest common ancestor was hypothesized based on the correlation between the G + C nucleotide content of ribosomal RNA sequences with the optimal growth rate temperature found in prokaryotes (Galtier *et al.*, 1999). If the nonhyperthermophilic hypothesis is presumed to be correct, it then implies that mesophiles were most likely the first to evolve with thermophily and psychrophily

adaptations followed thereafter. Psychrophily undoubtedly has evolved multiple times, thereby giving rise to polyphyletic origins (i.e., psychrophilic members are contained within lineages spanning the divergence of divisions and domains).

A phylogenetic tree generated exclusively from SUU rRNA sequences of cultivated obligate psychrophiles demonstrates five major lineages spanning the prokaryotic domains of Bacteria and Archaea (Fig. 1): the Crenarchaeota, the Euryarchaeota, the Flexibacter–Cytophaga–Bacterioides group, the gram-positive Bacteria group, and Proteobacteria. The majority of individual isolates are located in lineages contained either by the Flexibacter–Cytophaga–Bacterioides group or by the *Moritella*, *Colwellia*, and *Shewanella* subgroups of the  $\gamma$ -Proteobacteria.

#### V. PHYSIOLOGY OF PSYCHROPHILES (METABOLIC ACTIVITIES)

Psychrophiles were not isolated until 1964 mainly because of the fact that previous investigators did not recognize the abnormal sensitivity of this thermal class. Although it was known that most marine bacteria would not withstand the plating temperature of agar, microbiologists working in the Arctic and Antarctic did not take the precaution of keeping their sample material, media, pipettes, loops, and needles cold before use. Even for some psychrophiles, 10°C is too warm and the organisms will expire.

Many assume that growth at low temperature is slow. This assumption was shown to be incorrect when *Moritella marinus* was reported to have a doubling times of 80.7 and 226 min at 15 and 3°C, respectively. The lowest temperature recorded for growth of microorganisms is –12°C, but there are unconfirmed reports of growth at –18, –20, and –34°C for yeasts and fungi. Metabolic activity depends on the freezing properties of the cellular aqueous solutions and the liquid immediately adjacent to the cell. Aqueous solutions can be cooled to –10 to –20°C before freezing occurs. Franzmann *et al.* (1987) recorded liquid water at –14.1°C in Organic Lake, Antarctica. This liquid water naturally would be very saline since much of the water is freezing out. It is interesting to note that although *Halobacterium* and *Halomonas* are both halophilic as well as psychrophilic, no growth could be demonstrated in the liquid water from either Organic and Deep Lakes, Antarctica; hence, temperature was not the only lim-

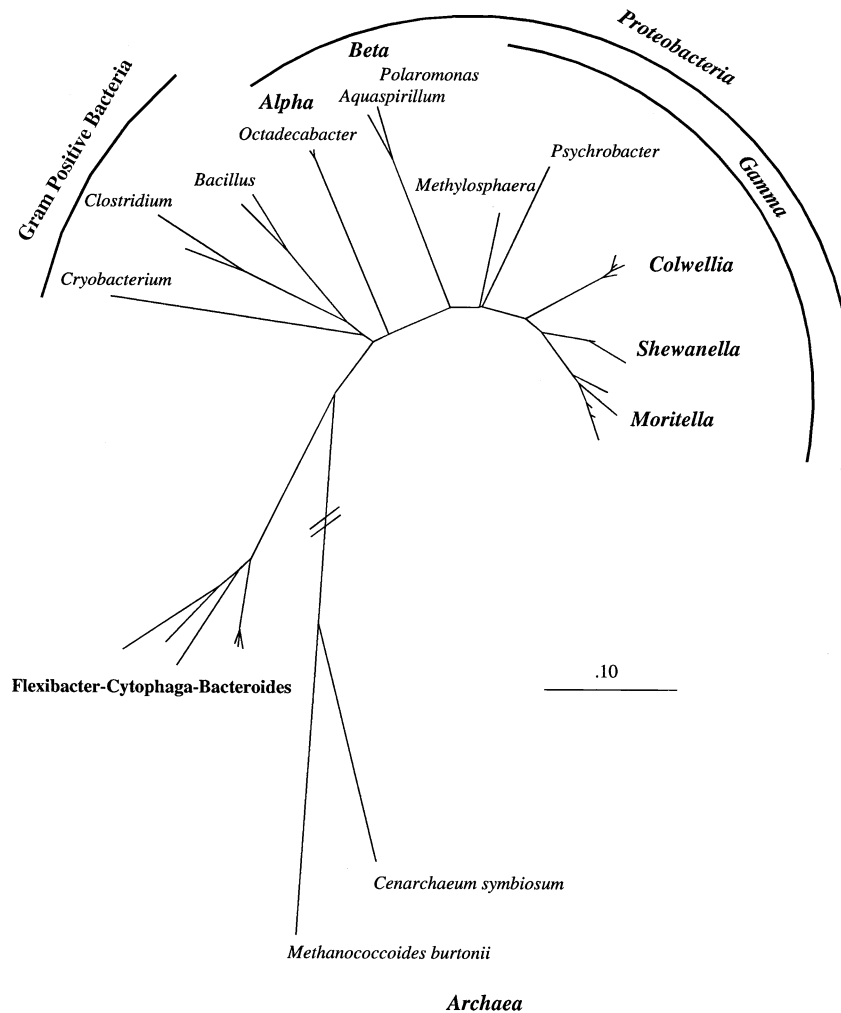


FIGURE 1 Radial phylogenetic tree using the neighbor-joining distance method demonstrating the evolutionary relationships among cultivated psychrophiles. The tree was constructed using complete SSU rRNA sequences contained in the Ribosomal Database Project from the 99/99 release of Version 7.1 (<http://www.cme.msu.edu/RDP/html/index.html>), with the additions of *Cenarchaeum symbiosum* and *Moritella* sp. ANT-300. The scale bar represents 0.10 fixed mutations per nucleotide position.

iting factor for growth. However, like most environments in the biosphere, the lack of other environmental or chemical conditions (mainly energy) usually prevents rapid growth of the microorganisms.

### A. Enzymes

Unlike thermophiles, research on the physiology of psychrophiles has been neglected over the years. The few studies dealing with enzymes (and cytoplasmic proteins) from psychrophiles indicate that they are more thermolabile compared to their counterparts in mesophilic bacteria—their maximum, minimum, and opti-

um temperatures for activity are lower. However, most of these enzymes will operate several degrees above the maximum growth temperature of the psychrophiles. Malic dehydrogenase, the first enzyme from a psychrophile to be studied, obtained from cells of *M. marinus* was found to be stable between 0 and 15°C and inactivation occurred between 15 and 20°C. This organism's optimum growth temperature is 15°C and its maximum is 20°C. If the enzyme is partially purified, inactivation at higher than 20°C becomes very pronounced. Phosphofructokinase/glyceraldehyde-3-phosphate dehydrogenase complex, lactic dehydrogenase, hexokinase, and aldolase from this same organism

lost activity when incubated for 1 h to 35–40°C. Surprisingly, glyceraldehyde-6-phosphate dehydrogenase was found to be stable at 26°C; however, at 36°C for 1 h, it lost 90% of its activity. Triosephosphate isomerase, a glycolytic enzyme, isolated from *Moritella* sp. strain ANT-300 (maximum growth temperature of 12°C) is extremely heat sensitive and has a half-life of heat inactivation of 520 s at 25°C. Other enzymes that have been examined from psychrophiles include  $\alpha$ -amylase, lipase, protease, alkaline phosphatase, and  $\beta$ -lactamase. As a general rule, the enzymes isolated from psychrophiles have not adapted entirely to correspond with the cardinal temperatures of the respective psychrophile cell. The abnormal thermoability of enzymes from psychrophiles is not the primary target of death from heat. Nevertheless, enzymes from psychrophiles are receiving more attention due to their possible industrial and biotechnological uses.

## B. Membranes

In order for the cell to maintain its bilayer structure of membrane lipids, homeophasic adaptation has been proposed. This adaptation permits the cell to function properly in terms of uptake, excretion, and regulation of intracellular ionic composition.

When psychrophiles are subject to temperatures above their maximum growth temperature, the cell leaks amino acids, DNA, and RNA, indicating that the membrane can no longer control its transport mechanisms. Furthermore, as the temperature reaches the maximum growth temperature or higher, the ability to respire is impaired. As a result, the cell expires.

Mesophiles and thermophiles cannot grow at low temperatures mainly due to the loss of membrane fluidity. This is why mesophiles (e.g., *Escherichia coli*) that have lost their membrane fluidity can no longer transport substrates into the cell. The fatty acid composition of the membrane changes in response to temperature. These homeophasic adaptations are often represented by changes in polyunsaturation, chain length, branching, cyclization, and often a combination thereof (Russell, 1992). Membranes of psychrophiles and psychrotrophs contain more polyunsaturated short chains and branch and/or cyclic fatty acids than do mesophiles and thermophiles. In general, the membrane lipids are composed of an increasing amount of unsaturated fatty acids as the temperature decreases. Membrane fluidity can also be changed by *cis/trans* isomerization of the double bonds of fatty acids. Depending on the species in question, psychrophiles are generally endowed with a higher proportion of unsatu-

rated fatty acids, especially hexadecenoic (16:1) and octadecenoic (18:1) acids, than mesophilic bacteria. The amounts of myristic acid (14:0) and palmitic acid (16:0) are higher in cells that are grown at higher temperatures. When grown at 0°C, there is an increase in docosahexaenoic acid (22:6) and/or eicosapentaenoic acid (20:5) in some strains of psychrophiles and barophiles.

In order to make the membrane lipids more unsaturated when the temperature is lowered, a desaturase enzyme is used which acts on the acyl chain of the membrane lipids (Russell, 1992). This desaturase activity is followed by temperature-dependent changes in fatty acid chain length and branching mediated by additional synthesis. These metabolic activities all serve to maintain the fluidity of the cell membrane so that it can function properly. The foregoing data were obtained by using a psychrotroph; therefore, the ability to alter the fatty acid composition may be restricted to psychrotrophs and not psychrophiles.

It should be noted that when the psychrophile *Moritella* sp. strain ANT-300 was starved, there was induced qualitative and quantitative changes in fatty acids (e.g., fatty acid 16:1 increased from 42 to 62.5%). When starved, cells of *Moritella* sp. strain ANT-300 (as with many other bacteria) underwent fragmentation (i.e., reductive division), resulting in ultramicrocells. The starvation-survival state is the normal state of most microbes since most of the biosphere is oligotrophic. The oligotrophic nature of the psychrophile's environment is predominant since organic matter is lacking for the growth of heterotrophic psychrophilic bacteria. Thus, over most (~80%) of the earth's biosphere and in both polar environments, most of the microbes are in a starvation-survival mode most of the time.

Thus, it appears that various species of psychrophiles evolved from their mesophilic component and they inhabit permanently cold environments. Their upper and lower temperature limits are 20 and -12°C, respectively.

## See Also the Following Articles

ANTARCTIC ECOSYSTEMS • ARCHAEA, ORIGIN OF • ARCTIC ECOSYSTEMS • BACTERIAL BIODIVERSITY • MICROBIAL BIODIVERSITY • THERMOPHILES, ORIGIN OF

## Bibliography

- Baross, J. A., and Morita, R. Y. (1978). Life at low temperatures: Ecological aspects. In *Microbial Life in Extreme Environments* (D. J. Kushner, Ed.). Academic Press, London.

- DeLong, E. F., Franks, D. G., and Yayanos, A. A. (1997). Evolutionary relationships of cultivated psychrophilic and barophilic deep-sea bacteria. *Appl. Environ. Microbiol.* **63**, 2105.
- DeLong, E. F., King, L. L., Massana, R., Cittone, H., Murray, A., Schleper, C., and Wakeman, S. G. (1998). Dibiphytanyl ether lipids in nonthermophilic crenarchaeotes. *Appl. Environ. Microbiol.* **64**, 1133.
- Forterre, P. (1996). A hot topic: The origin of hyperthermophiles. *Cell* **85**, 789.
- Franzmann, P. D., Deprez, P. P., Burton, H. R., and McMeekin, T. A. (1987). Limnology of Organic Lake, Antarctica, a meromictic lake that contains high concentrations of dimethyl sulfide. *Aust. J. Mar. Freshwater Res.* **38**, 409.
- Franzmann, P. D., Liu, Y., Balkwill, D. L., Aldrich, H. C., Conway de Macario, E., and Boone, D. R. (1997). *Methanogenium frigidum* sp. nov., a psychrophilic, H<sub>2</sub>-using methanogen from Ace Lake, Antarctica. *Int. J. Syst. Bacteriol.* **47**, 1068.
- Friedmann, E. I. (Ed.) (1993). *Antarctic Microbiology*. Wiley-Liss, New York.
- Galtier, N., Tourasse, N., and Gouy, M. (1999). A nonhyperthermophilic common ancestor to extant life forms. *Science* **283**, 220.
- Gilichinsky, D. (Ed.) (1994). *Viable Microorganisms in Permafrost*. Russian Academy of Sciences. Pushchino.
- Gilichinsky, D. (1995). Permafrost microbiology. *Permafrost Periglacial Processes* **6**, 281.
- Gosink, J. J., Woese, C. R., and Staley, J. T. (1998). *Polaribacter* gen. nov., with three new species, *P. irgensii* sp. nov., *P. franzmannii* sp. nov. and *P. filamentus* sp. nov., gas vacuolate polar marine bacteria of the Cytophaga-Flavobacterium-Bacteroides group and reclassification of "*Flectobacillus glomeratus*" as *Polaribacter glomeratus* comb. nov. *Int. J. Syst. Bacteriol.* **48**, 223.
- Morita, R. Y. (1975). Psychrophilic bacteria. *Bacteriol. Rev.* **39**, 144.
- Moyer, C. L., Tiedje, J. M., Dobbs, F. C., and Karl, D. M. (1998). Diversity of deep-sea hydrothermal vent *Archaea*. *Deep-Sea Res. II* **45**, 303.
- Omelyansky, V. L. (1911). Bakteriologicheskoe issledovanie Sanga mamonta Prilegayushchei pochvy. *Arkhiv Biologicheskikh Nauk* **16**, 335.
- Pace, N. R. (1991). Origin of life—Facing up to the physical setting. *Cell* **65**, 531.
- Preston, C. M., Wu, K. Y., Molinski, T. F., and DeLong, E. F. (1996). A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum* gen. nov., sp. nov. *Proc. Natl. Acad. Sci. USA* **93**, 6241.
- Russell, N. J. (1992). Physiology and molecular biology of psychrophilic microorganisms. In *Molecular Biology and Biotechnology of Extremophiles* (R. A. Herbert and R. J. Sharp, Eds.). Blackie, Glasgow.
- Vorobyova, E., Soina, V., Gorlenko, M., Minkovskaya, N., Zalinova, N., Mamukelashvili, A., Gilichinsky, D., Ravkina, E., and Vishnivetskaya, T. (1997). The deep cold biosphere: Facts and hypothesis. *FEMS Microbiol. Rev.* **20**, 277.