Nicholas G. Rambidi

Molecular Computing

Origins and Promises

Springer

Molecular Computing

Nicholas G. Rambidi

Molecular Computing

Origins and Promises

Springer

Nicholas G. Rambidi
Moscow
Russia

Revised and updated edition of the book in Russian:
Нанотехнологии и молекулярные компьютеры,
ISBN 978-5-9221-0869-0, 256 pp, 2007
Publisher: Fizmalit, Moscow

ISBN 978-3-211-99698-0 ISBN 978-3-211-99699-7 (eBook)
DOI 10.1007/978-3-211-99699-7
Springer Wien Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014952186

© Springer-Verlag Wien 2014
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

UDK 681.3
BBK 22.3.
R 21

The publication was supported by the
Russian Foundation for Basic
Research (project no. 07-08-07015d)

Rambidi N. G. Nanotechnologies and molecular computers. – M.: FIZMATLIT, 2007. – 256 pp. – ISBN 978-5-9221-0869-0

This book deals with one of the earliest and most actively developing directions of nanotechnology—the creation of computing devices in which separate molecules or their relatively small ensembles are used as the element basis.

The main principles of information processing on the molecular level, the complicated and controversial in many aspects history of the elaboration of the molecular computing devices, and the recent results, which make it possible to hope for a new breakthrough in the modern calculation techniques, have been presented. Digital molecular systems, the architecture of which is similar to that of the modern computers, and also devices mimicking the biological principles of information processing, which apparently may effectively solve the problems of the artificial brain, have been considered.

For students, postgraduates, and researchers who create new promising means for information processing and also for a wide range of readers interested in this problem.

© FIZMATLIT, 2007.
© N. G. Rambidi, 2007.

ISBN 978-5-9221-0869-0

Contents

Introduction

Nanotechnology occupies a special place among the modern directions of the human activity.

The history of its origin and formation is full of ingenious foresights and seemingly long-term negligence of their implementation. Half a century ago at an annual meeting of the American Physical Society, Richard Feynman, a multitalented physicist, reported in detail about the basics of what is called nanotechnology now. It was then when the main nanotechnological principles were formulated:

- Miniaturization of the devices up to the limit dimensions of the atomic–molecular level fundamentally improving their functional possibilities
- Control of the macroproperties of an object owing to the directed change of its structure on the nano- (molecular) level.

Nevertheless, the industrial production, i.e., to what the efforts of this field of activity should be directed, was not ready, as a matter of fact, to consciously percept it. Only in the end of the twentieth century the accumulated needs of the production led to the nanotechnological boom.

It should be noted that the idea of creating devices with the extreme miniature dimensions as it is was not alien to a human being before that. Suffice it to remember different miniature models, e.g., sailing boats confined in glass vessels, congratulations, and even pictures on rice grains as presents on outstanding events. At the same time, a nanotechnological principle—control of the macroscopic properties of a material due to the directed action on its microstructure—was widely used in the twentieth century in the industrial production. This principle was the basis of the industrial production of synthetic rubbers, synthetic fibers with desired properties, and many other industrial products. Nevertheless, among other directions of the human activity, the computer engineering played a specific role in the formation of nanotechnology.

The computer engineering steadily moved in the direction of the miniaturization of the electronic schemes during the second half of the previous century. Moreover, in the end of the twentieth century, it became one of the main touchstones on which nanotechnological ideas were and are perfected.

Historically the first intensive outbreak of the nanotechnological activity, which was not called nanotechnology then, occurred in the computer engineering. The source and the powerful driving force of it were the complexities of the industrial development of the planar semiconductor technology in microelectronics and the great interest of the defense departments in the fast improvement of the electronic computing devices.

The foundation was laid in the United States when Forrest Carter, a talented physicist and a member of the US Naval Research Laboratory, evoked the interest of the research community in the molecular electronics. Carter's activity was multifaceted. He developed several approaches to the building of the molecular element base of the computing devices on the basis of physical effects, which were not used before. He proposed to create switching schemes on the basis of the controlled passage of the electron through a system of energy barriers. At that time, the idea was implemented on the basis of the solid state semiconductor superlattices. Carter elaborated different models of switching devices with the use of molecular solitons. This phenomenon—the propagation of the excitation over the molecular chain—was earlier proposed and studied by a Soviet physicist A. S. Davydov. Carter's organization activity was an aspect of his activity of no less importance. In 1980s he organized three international conferences on molecular electronic devices. They attracted tens of scientists who worked in the fields close to molecular electronics and favored their unification. Unfortunately, the last conference took place in 1987 already after its organizer has untimely passed away.

Some sobering-up was achieved in the beginning of the 1990s after the period of the molecular boom in microelectronics. The understanding was strengthened that a molecule is a micro-object, the behavior of which is determined by the quantum-mechanical principles. The transitions between separate energy states of the molecule have a probabilistic character. Therefore even at the rather high probability of the molecular transition, one has to use not a separate molecule as switching elements but their ensemble, i.e., a small volume of the compound. The use of this approach led to the creation of the model molecular current rectifiers and simplest switching devices. The most promising elaboration was the creation of several molecular random access memory devices on the basis of the bacteriorhodopsin protein.

It seemed that molecular electronics entered the stage of peaceful development, accumulation of experimental facts, and search for new ways of creating molecular devices for information processing. But a breakthrough in this field took place again on the boundary between the past and present millennium. It was based on the successes of the modern synthetic organic chemistry. As a consequence, the elaboration of the molecular computing devices closely approached the stage of their industrial production.

In the Soviet Union, the investigations on the creation of the molecular devices on information processing were organized on the state level in the end of the 1970s. The head organization was the Scientific Research Institute of Physical Problems of the USSR Ministry of Electronic Industry that coordinated the activity of the research groups of the Academy of Sciences of the USSR and of some institutes of higher education. The work was performed under the aegis of the Commission of the Presidium of the Council of Ministers of the USSR that organized the Interdepartmental Council on the problem "molecular electronics" determining the main directions of the research. The irrepressible energy of B. A. Kiyasov, the deputy chair of the Scientific-Technical Council of the Commission of the Presidium of the Council of Ministers of the USSR, who was in fact the main organizer of the research in this field, played the extremely important role in its development. An efficient infrastructure was created in a decade in the Soviet Union, which made it possible to perform complicated theoretical and experimental studies. As an example, one can mention the elaboration of the memory devices on the basis of burning holes in optical spectra. These studies performed by the groups of R. I. Personov in Moscow and K. K. Rebane in Tartu not only did not yield to but rather surpassed the analogous studies abroad in the level of the research. In 1989, a perspective program of the research on creating molecular devices for the forthcoming 5 years was elaborated.

Unfortunately, the financing of the research in this field (the same as in other directions of the scientific research activity) was practically stopped starting from 1990. Of all elaborated program, only studies on the technology of the bacteriorhodopsin media and information processing by the distributed nonlinear media were slowly performed in the following years in Russia.

In this small book an attempt was made to tell about the formation and development of the research on molecular devices of information processing that took place in the atmosphere full of hopes, enthusiasm, successes, and disappointments. It seems to me that it is necessary to write about this now, when there appeared a real possibility to create fundamentally new computing and information-logical devices on the molecular level. They will hardly replace the existing semiconductor electronics. But these devices will make it possible to supplement it and drastically expand the possibilities of the modern industry of information.

Chapter 1
The Origins and the Making of Nanotechnologies

The emperor looked and saw that the tiniest sort of a speck really was lying upon the salver. The workmen say: 'Please spit on your finger, and take it in your palm.' 'But what am I to do with this speck?' 'It is not a speck,' they answer, 'but a nymfozoria.' 'Is it alive?' 'Not at all,' they reply; 'it is not alive, but it has been forged by us in the image of a flea out of pure English steel, and in the middle of it are works and a spring.' Please wind it up with the little key: it will immediately begin to dance.

NS Leskov *"The steel flea."*

1.1 A Few Words About the Roots of Nanotechnology in Human Consciousness

When talking about nanotechnology, the basis of not only the technical and technological revolution of recent years but also a source of fundamental shifts in the psychology of human society, one should not forget about its potential origins. Starting from ancient times, the idea of miniaturization seemed quite natural, i.e., the existence or creation of objects with dimensions much smaller than those usual for man that are still able to perform the functions inherent for their macro counterparts. This is manifested both in folk beliefs about various miniature creatures accompanying man, elves and dwarves and in literature.

Thus, over 300 years ago the great fantasist Jonathan Swift described the state of Lilliputians. Its inhabitants were less than one-tenth the size of normal human beings, but the organization and the laws of society differed little from those prevailing at that time in England. Swift brilliantly used his imagined model as a satirist. And, as a matter of fact, this practical significance is where one of the main goals of his nanotechnological approach, as we understand it today, manifested itself.

In the middle of the nineteenth century the famous Russian writer Nikolai Semenovich Leskov wrote his novel *The Steel Flea*. It narrates of the Emperor Alexander Pavlovich who, while traveling after the Congress of Vienna, received in England a "nymfozoria" (footnote: infusoria)—a steel flea—as a gift. It was visible

© Springer-Verlag Wien 2014

N.G. Rambidi, *Molecular Computing*, DOI 10.1007/978-3-211-99699-7_1

only in "melkoskop" (footnote: the old Russian word for a microscope), and when wound up, it moved its feelers and danced. Upon the emperor's return this miracle was shown to the Tula artisans for their edification. The most skillful gunsmith (the left-handed man) stayed unimpressed and promised to modernize the flea. As a result the flea could still move its feelers, but was not able to dance anymore. The left-handed man explained to the angry emperor that the artisans shod the flea with horseshoes. Stamped upon each shoe was the maker's name, and it was the left-handed man who forged the tiny nails with which the shoes were fastened on.

More recently, in 1963–1964, the Polish writer Stanislaw Lem published a science fiction novel *The Invincible*. In this novel, he designed a strange mechanical civilization living on a distant planet—a complex system built from a huge number of minuscule particles. Each of them housed a primitive sensor, a logic device, an appliance for communicating with other particles, and a microscopic engine enabling the particles to move freely. The logical capabilities, minimal for a single particle, increased dramatically upon their association. The assembled conglomerate, the swarm as Lem calls it, was a distributed self-organized system with decentralized control exercised by the collective intelligence of the system, the level of which is determined by its size. In the novel the system particles initially reside individually on the surface of the planet. If a foreign intruder appears a number of proximate particles rise in the air. Dependent on the extent of the threat more and more particles join the swarm, resulting in the growth of both the system's intelligence and its operational capabilities. Particle sensors respond to biological entities made of protein. The main objective of the population of particles is life competition with biological life-forms. It suppresses biological organisms by surrounding them by a cloud of particles and erasing their memory by a directed electromagnetic pulse.

1.1.1 A Little Detail: Systems of Microparticles and Their "Emerging" Properties

It is surprising that in his novel Stanislaw Lem anticipated the nanotechnological principle of system design known today as "bottom-up." The modern industrial practice is dominated by the opposite "top-down" approach when based on preestablished instructions and a selected technology the initial metal bar is, for example, successively turned, milled, and reamed, to obtain the final product—the desired component part (Fig. 1.1). In contrast, the "bottom-up" principle in its pure form implies that the product is successively created from its basic parts by means of self-organization processes. The starting material for manufacturing the product represents a system consisting of a large number of elements (atoms, molecules, or their assemblies), each of which can perform certain physical actions. A crucial factor determining the "technological" capabilities of the system is the nature of the interaction between these elements. Under certain conditions the system acquires properties lacking in its separate elements. More details on the occurrence of such "emerging" properties will be given below.

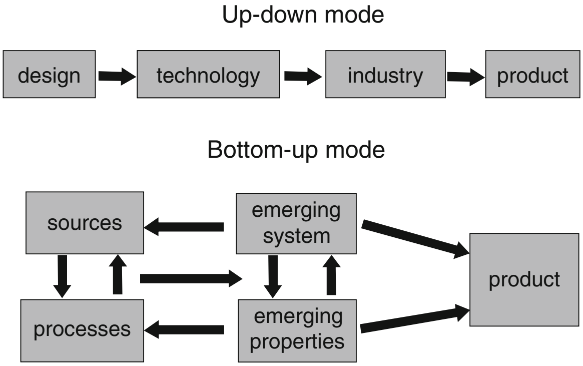


Fig. 1.1 Scheme of technological "top-down" and "bottom-up" principles

Further literature examples could be given where nanotechnological ideas would become apparent in one form or another. However, those already described are sufficient to emphasize the basic principles of nanotechnological approaches they convey:

– Maximum micro(nano)-miniaturization of macro-objects, which leads to their micro-analogues with properties modified in the intended direction
– Directed alteration of the structure of the macro-objects at the micro (nano)-level aimed at achieving the intended change in the properties of the macro-object
– The technological principle "bottom-up"

It should be noted that even before the nanotechnology boom researchers in various fields of human activity were using these principles, similar to the well-known hero of the play by Molière who, unbeknownst to himself, had been speaking prose.

1.1.2 A Little Detail: Just One Example from Chemistry

Caoutchouc played an important role in the engineering of the first half of the past century, with a variety of industrial products produced on its basis. It was shipped to the industrial countries from Southeast Asia. It was known that natural caoutchouc was a polymer of an organic compound—isoprene. Nevertheless, all attempts to obtain synthetic caoutchouc on the isoprene basis in the prewar years failed (Fig. 1.2). Polymerization took place, but the resulting product was so different from the natural one that its use as a substitute was impossible. Upon the advent of the Second World War the sources of natural rubber were cut off due to the rapid capture of Southeast Asia by Japan. By that time the major industrialized countries had mastered the production of caoutchouc substitutes on another synthetic base. In the Soviet Union polybutadiene was created after the fundamental studies by I. L. Kondakov and S. V. Lebedev. In Germany the production of butadiene–styrene

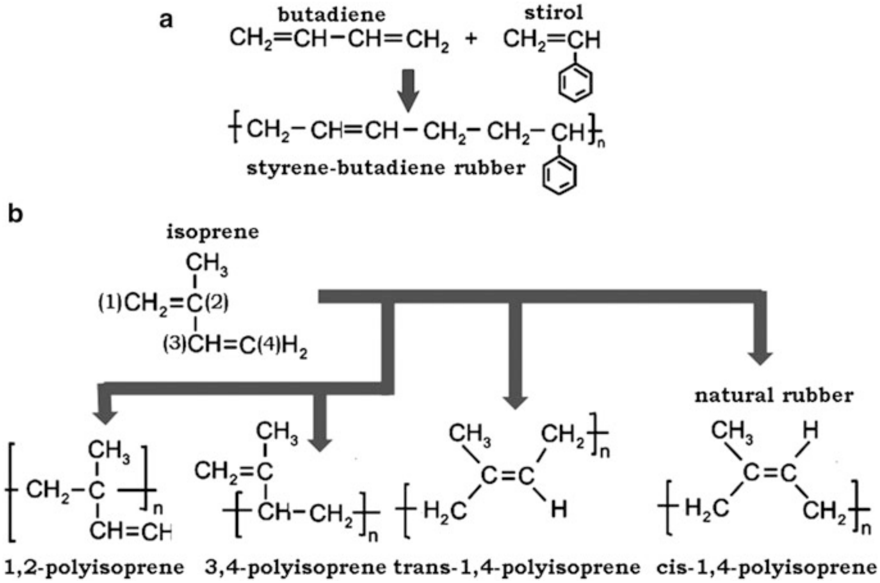


Fig. 1.2 Synthesis of synthetic (**a**) and natural (**b**) caoutchouc

rubber was organized. In the United States, a group of researchers led by the famous chemist Carothers synthesized neoprene—a product of polymerization of chlorinated butadiene. But despite the large-scale production of these synthetic rubbers, the synthesis of natural rubber remained a pressing task because its properties were far superior to the properties of substitutes. This problem was solved only in the postwar years when the molecular structure of natural rubber was determined. It turned out that only one of the four possible isomers occurring in the process of isoprene polymerization corresponded to the structure of the natural rubber. Only after that, in the end of the 1950s, did it become possible to develop methods for stereospecific synthesis—a complex process at the molecular level allowing to selectively obtain the required polyisoprene isomer during polymerization—which then became the basis for its bulk production.

1.2 Nanotechnology, Created in the Pages of Science Fiction, Becomes the Basis of a New Industrial Revolution

Starting from the middle of the last century the possibility of using natural micro-objects or creating artificial ones capable of performing certain macroscopic actions is being actively discussed among scientists. One of the pioneers was apparently Erwin Schrödinger, the famous twentieth-century physicist, who in February 1943 delivered a lecture at Trinity College (Dublin) entitled “What Is Life?” In this lecture, later published in book form and reprinted in many countries, he first proposed the idea of an aperiodic crystal—a microsystem of large information capacity. In fact, he anticipated the principle of storing genetic information in the DNA structure, noting that:

> A small molecule might be called “the germ of a solid,” Starting from such a small solid germ, there seem to be two different ways of building up larger and larger associations. One is the comparatively dull way of repeating the same structure in three directions again and again. That is the way followed in a growing crystal. Once the periodicity is established, there is no definite limit to the size of the aggregate. The other way is that of building up a more and more extended aggregate without the dull service of repetition. That is the case of the more and more complicated organic molecule in which every atom, and every group of atoms, plays an individual role, not entirely equivalent to that of many others (as is the case in a periodic structure). We might quite properly call that an aperiodic crystal or solid and express our hypothesis by saying: “We believe a gene—or perhaps the whole chromosome fibre—to be an aperiodic solid.”

The emergence of nanotechnology as a separate field of science and technology is usually associated with the outstanding physicist of the twentieth century, Richard Feynman. On December 29, 1959, he made a presentation at the meeting of the American Physical Society entitled “Plenty of Room at the Bottom.” In this lecture he drew attention of the physics community to the fact that among the various areas of physics research, there is one area that has not yet come into the view of physicists although it offers a wide variety of scientific and technical applications. This area is the detailed investigation of micro- and nano-sized objects. As a first example Feynman considered the problem of compact information storage—*Why cannot we write the entire 24 volumes of the Encyclopedia Brittanica on the head of a pin*? Answering this question, Feynman pointed out that if one magnified the head of a pin (having the diameter of 1/16 of an inch) by 25,000 diameters, the area of the head of the pin would then be equal to the area of all the pages of the Encyclopedia Britannica. The smallest element of the text, a dot, would contain in its area 1,000 atoms. Feynman went on to address a number of opportunities that constructing and using super miniaturized devices would offer. Those include ultradense recording and storing of information, the development of miniaturized computers, and the creation of autonomous tools that can perform surgery directly in the human body. According to Feynman “it would be interesting in surgery if you could swallow the surgeon. You put the mechanical surgeon inside the blood vessel and it goes into the heart and ‘looks’ around (Fig. 1.3) ... Other small machines might be permanently incorporated in the body to assist some inadequately-functioning organ.” In his lecture Feynman mentioned the possibility of synthesizing chemical substances directly from atoms, by adding them sequentially to the structure being created. In general, thinking about the many Feynman predictions that have actually been implemented today, one perceives in a new light the final words of his lecture: “In the year 2000 [people] will wonder why it was not until the year 1960 that anybody began seriously to move in this direction.”

The term “nanotechnology” was first coined in 1974 by Norio Taniguchi of the University of Tokyo. In its original meaning this term referred to precision

Fig. 1.3 Mechanical surgeon inside the blood system

manufacturing of parts in the process of industrial production with ultralow tolerances for dimensional error. In his article, “On the Basic Concept of ‘Nano-Technology,” Taniguchi wrote: “Nanotechnology is the production technology to get the extra high accuracy and ultra fine dimensions. … The smallest bit size of stock removal, accretion, or flow of materials is probably of one atom or one molecule, namely 0.1–0.2 nm in length. … Accordingly, nano-technology mainly consists of the processing of separation, consolidation, and deformation of materials by one atom or one molecule.”

This concept was expanded and, in fact, modified by the American scientist Eric Drexler. In 1981, while at the Massachusetts Institute of Technology, and inspired by the ideas of Richard Feynman, he published an article entitled “Molecular engineering: An approach to the development of general capabilities for molecular manipulation.” Based on the technique to design complex protein macromolecules from simple molecular fragments available at that time, he considered the possibility of establishing various devices at the molecular level. These included actuators, motors, pumps, and even wires and bearings. Several years later Drexler released his main book *Engines of Creation: The Coming Era of Nanotechnology*. The best and most concise description of what this book conveys apparently belongs to one of the leading scholars in the field of informatics of the last century, Marvin Minsky, who wrote the foreword to this book.

> Engines of Creation begins with the insight that what we can do depends on what we can build. This leads to a careful analysis of possible ways to stack atoms. Then Drexler asks, "What could we build with those atom-stacking mechanisms?." For one thing, we could manufacture assembly machines much smaller even than living cells, and make materials stronger and lighter than any available today. Hence, better spacecraft. Hence, tiny devices that can travel along capillaries to enter and repair living cells. Hence, the ability to heal disease, reverse the ravages of age, or make our bodies speedier or stronger than before. And we could make machines down to the size of viruses, machines that would work at speeds which none of us can yet appreciate. And then, once we learned how to do it, we would have the option of assembling these myriads of tiny parts into intelligent machines, perhaps based on the use of trillions of nanoscopic parallel-processing devices which make descriptions, compare them to recorded patterns, and then exploit the memories of all their previous experiments. Thus those new technologies could change not merely the materials and means we use to shape our physical environment, but also the activities we would then be able to pursue inside whichever kind of world we make.
>
> ... It seems to me, in spite of all we hear about modern technological revolutions, they really haven't made such large differences in our lives over the past half century. Did television really change our world? Surely less than radio did, and even less than the telephone did. What about airplanes? They merely reduced travel times from days to hours—whereas the railroad and automobile had already made a larger change by shortening those travel times from weeks to days! But Engines of Creation sets us on the threshold of genuinely significant changes; nanotechnology could have more effect on our material existence than those last two great inventions in that domain—the replacement of sticks and stones by metals and cements and the harnessing of electricity. Similarly, we can compare the possible effects of artificial intelligence on how we think—and on how we might come to think about ourselves—with only two earlier inventions: those of language and of writing.
>
> We'll soon have to face some of these prospects and options. How should we proceed to deal with them? Engines of Creation explains how these new alternatives could be directed toward many of our most vital human concerns: toward wealth or poverty, health or sickness, peace or war. And Drexler offers no mere neutral catalog of possibilities, but a multitude of ideas and proposals for how one might start to evaluate them. Engines of Creation is the best attempt so far to prepare us to think of what we might become, should we persist in making new technologies.

Eric Drexler widely advocated the ideas of creating molecular machines in a large number of journal articles and books. One of the most popular among them is the book entitled *Nanosystems: Molecular Machinery, Manufacturing, and Computation*. In order to give a more vivid picture of the opportunities offered by the implementation of Drexler's ideas, let us consider some examples. Figure 1.4 shows simple variants of a molecular transmission gear and some more complex similar devices. A molecular implementation of the bearing is also depicted.

In fact, despite the detailed discussion of the possibilities for creating molecular devices, everything that Eric Drexler considered is reminiscent of Jules Verne-type science fiction rather than technically sound specification for the development. Marvin Minsky wrote about it in the foreword to the book *Engines of Creation*, drawing an analogy between Drexler's suggestions and the ideas of Jules Verne, HG Wells, Frederick Paul, Robert Heinlein, Isaac Asimov, and Arthur C. Clarke. Drexler conceptually designed analogues of macroscopic devices, using as "building material" nanoscale components, including single atoms and molecules. But at the same time, being bound at the level of ideas of the 1960s and 1970s, he could

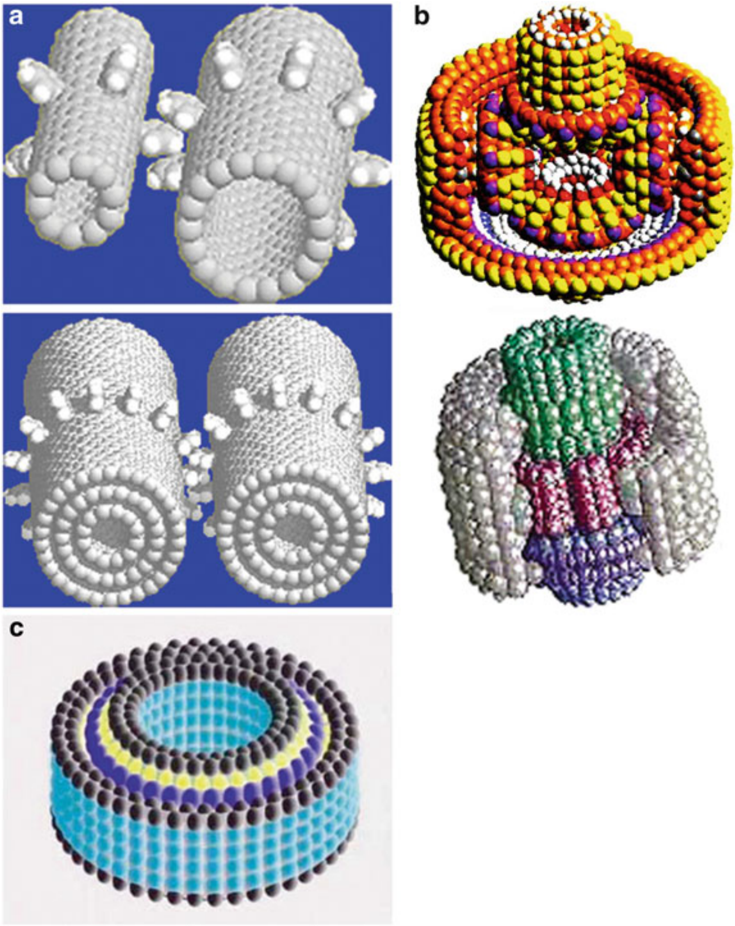


Fig. 1.4 Drexler's molecular devices: "transmission gear" (**a**), more complex "transmission gear" (**b**), bearing (**c**)

not envision a number of fundamental limitations and identify concrete ways to build devices. The problem of micro–macro interfaces, i.e., the problem of interaction between the human and the micro device, was also practically not discussed.

At the same time, rapid progress in several areas of human activity, and, first of all, semiconductor electronics, naturally led to the need to use nanotechnological principles. In several technological areas, nanotechnological base was being built up. The most important development was arguably the creation in the 1980s of the scanning tunneling microscopy and then atomic force microscopy. This provided the opportunity to not only see objects at atomic resolution but also to manipulate atoms and molecules.

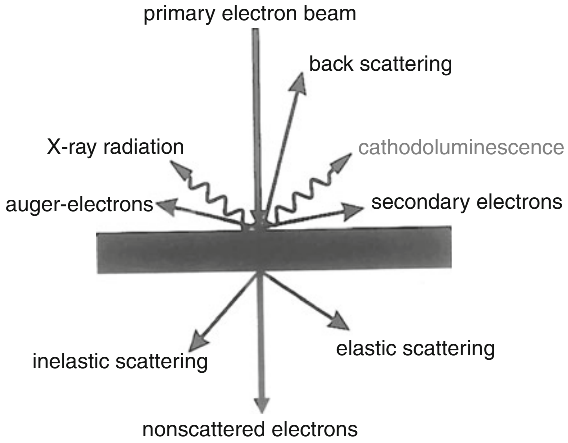


Fig. 1.5 Mechanisms of interaction of the electron beam with matter

1.2.1 More Details: Before Building Something at the Nano-level, One Must Learn to See What One Does

In order to create micro- and nano-sized objects, it is of course necessary to be able to determine their characteristics—shape and structure, the gross composition, and the composition of their local areas. The capabilities of the common optical microscopy are limited by its low resolution which is determined by the wavelength of visible light. Therefore, the maximum resolution is ~0.0002 mm, and the achievable magnification does not exceed 500–1,000 times.

In contrast to the light radiation, electron beam proved to be an effective means of studying the structure of matter at the micro- and nano-level. Depending on the electron energy its corresponding wavelength may amount to 10^{-2}–10^{-3} nm, with not very high acceleration voltage of electrons (tens of thousands of volts). The processes of interaction of electrons with matter are characterized by significant variety. Let the beam of accelerated electrons fall on a thin layer of substance (Fig. 1.5). Some of them are elastically scattered. Besides it, the interaction of electrons with the atoms of the substance leads to luminescence in the visible spectrum, X-ray radiation, and reflected secondary electrons knocked out from the object's atoms. A part of the secondary electrons, the so-called Auger electrons, provide an opportunity to determine the composition of surface layers and even to identify the distribution of a particular chemical element on the surface of the studied object. A significant number of electrons falling on the object pass through it without scattering and without losing energy, while some of them undergo both elastic and inelastic (i.e., with a loss of energy) scattering.

In the transmission electron microscope, the electron-optical system creates a monochromatized and focused electron beam which passes through the studied specimen (Fig. 1.6). The distribution of electrons obtained as a result of the interaction between the electrons and the specimen is focused by the projection system on a fluorescent screen or on a photographic plate. Sometimes it is said that

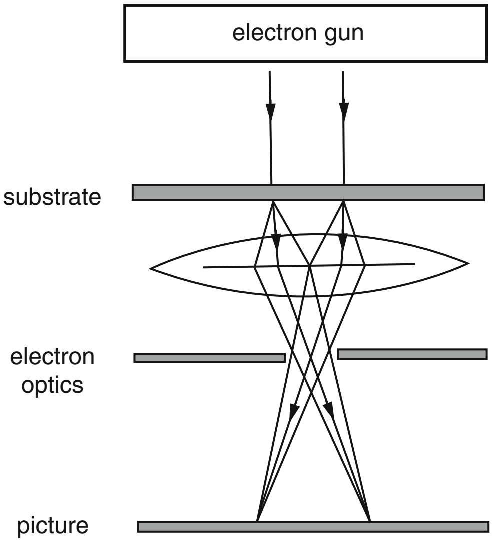


Fig. 1.6 Scheme of the transmission electron microscope

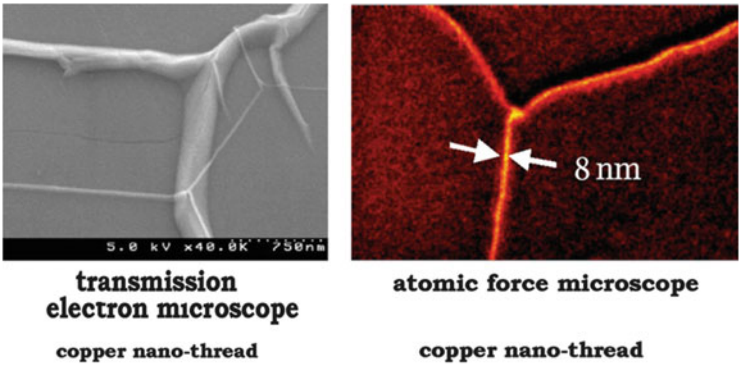


Fig. 1.7 Image of the copper nanothread recorded by a transmission electron microscope and atomic force microscope

the transmission microscope acts as a slide projector, which, instead of the light beam and a slide, has an electron beam and the studied specimen.

The resolving power of the transmission microscope can be brought to a few tenths of a nanometer. As an example, in Fig. 1.7, the image obtained by transmission electron microscopy is compared with the image recorded by an atomic force microscope (this and several subsequent figures exemplifying new physical methods make use of the images that are easily found in various video galleries that abound on the Internet; see, e.g., references 7–9 for Chapter 1). However, the higher the resolution power of the transmission the microscope gets, the smaller

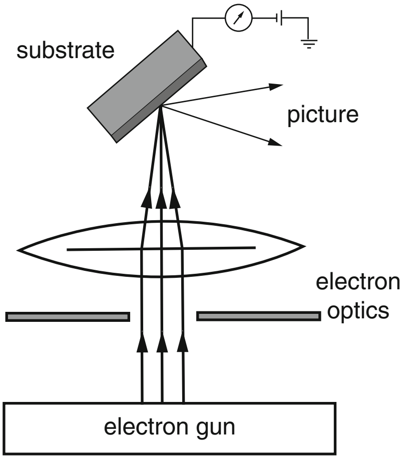


Fig. 1.8 Scheme of the scanning electron microscope

should be the wavelength of electrons, and this necessitates an increase of their energy. In its turn, higher energy of the electrons incident on the specimen results in a weakened interaction with matter, primarily with light atoms. Therefore, the image loses contrast. Various techniques have been developed to improve the image obtained by high-energy electrons. One possibility is to spray a thin layer of heavy atoms, such as tungsten, on the studied object. However, this generally causes the distortion of the information received. To a major degree, the way out of this situation was the creation of the scanning electron microscope (Fig. 1.8).

In it the electron beam is reflected from the surface of the investigated specimen, and the resulting image is recorded with the electron-optical system. Scanning electron microscope allows to obtain contrasting images of objects. Several typical examples shown in Fig. 1.9 include:

- The head of mosquito with magnification of 200 (A) and 1000 (B) times
- Bird blood
- Crystalline silver dendrites

However the resolution power of scanning electron microscopes is much lower compared to transmission microscopes. The minimum size of structural features that can be registered amounts to merely a few nanometers.

In the 1970s–1980s of the last century, different versions of electron microscopes were refined to become convenient commercially available tools. Nevertheless, they could not meet the needs of nanotechnological research. A dramatic turning point occurred in the early 1980s, when the scanning tunneling microscope was created.

In 1981 Gerd Binnig and Heinrich Rohrer at the IBM research laboratories in Zurich created the first scanning tunneling microscope. The significance of this work is apparent from the fact that just 5 years later, in 1986, they won the Nobel

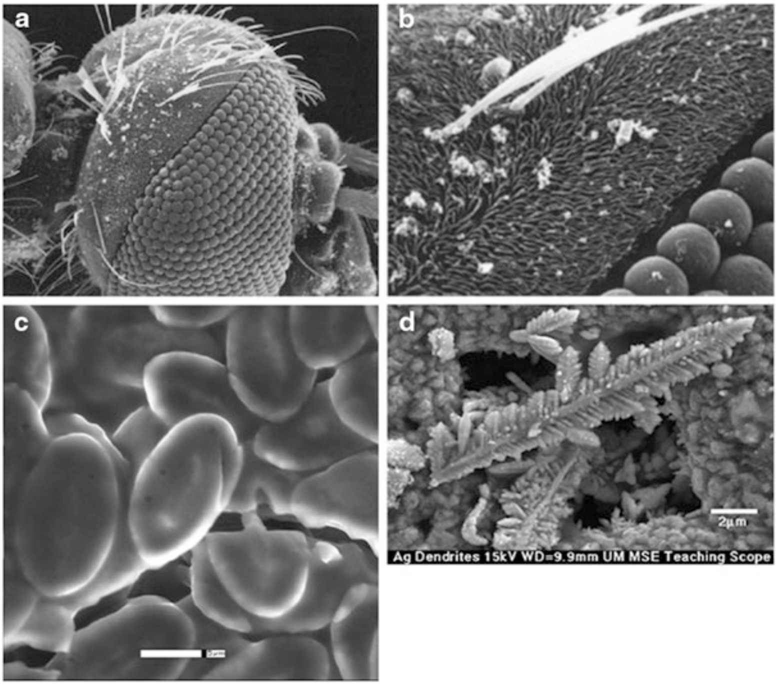


Fig. 1.9 Images obtained on a scanning electron microscope (see text for explanations)

Prize for Physics, jointly (and that is remarkable), with Ernst Ruska, who was awarded "for his fundamental work in electron optics and for the design of the first electron microscope."

The design of Binnig and Rohrer is based on the principle of quantum-mechanical tunneling of electrons through a nonconductive barrier.

The quantum-mechanical tunneling effect is a process whereupon particles leak through a potential barrier and penetrate the areas that in classical mechanics would have been inaccessible to them. Suppose there is a particle held in a potential well by a barrier of finite height and width. Suppose that the energy of the particle is such that, based on the laws of classical mechanics, it is not sufficient for the particle to escape the well, passing over the potential barrier. A quantum-mechanical treatment of this problem shows that there is some chance of the particle tunneling through the barrier and exiting the well. The possibility of tunneling arises from the requirement of continuity of the wave function on the walls of the well. If the amplitude of the wave function is not equal to zero at the inner edge of the barrier (which is permissible, provided that the potential at this point does not become infinite), it cannot just disappear inside the barrier. Instead, it begins to approach zero more or less quickly (Fig. 1.10). If the drop of the amplitude happens not too fast, it may not reach zero at the outer edge of the barrier. At this point, the wave function should make a smooth transition to the function characteristic for free

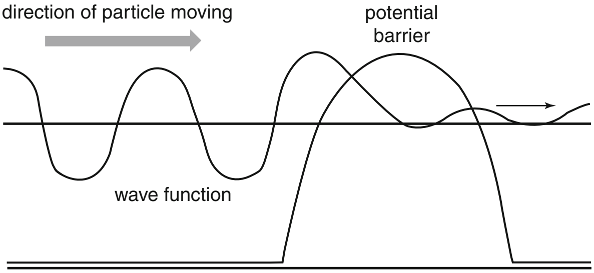


Fig. 1.10 Scheme of tunneling of electrons through a potential barrier

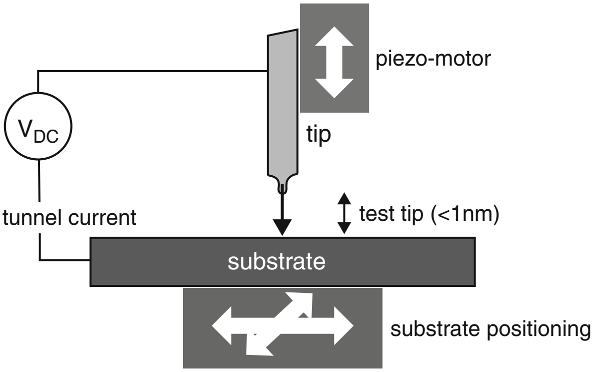


Fig. 1.11 Scheme of the scanning tunneling microscope

particles outside of the barrier, and, from that point on, the wave ceases to decay. But since the wave function has not disappeared in the area outside of the barrier, there exists a nonzero probability of finding the particle in this area, i.e., the particle performs tunnel passage out of the potential well.

The scanning tunneling microscope is a system with an extremely thin needle tip and the sample studied, to which constant voltage is applied (Fig. 1.11). The distance between the needle tip and the specimen is controlled by a piezo element. If this distance becomes ~1 nm, tunnel current starts to flow between the tip and the specimen which dramatically depends on the distance between the needle and the specimen. Therefore, by moving the sample relative to the needle tip, the current corresponding to each point of the sample surface can be measured. This relationship corresponds to the contour of the surface.

The disadvantage of this microscope design is that the tunnel current strongly depends on the distance between the needle tip and the specimen which complicates the measurement of the current. For this reason another method to record the relief of the surface is employed in tunneling microscopes. In this case a constant tunnel current is maintained by the system which controls the piezo element by adjusting the distance between the tip and the surface of the sample (Fig. 1.12).

The creation of the scanning tunneling microscope was a revolutionary event, which allowed to study the surface structure of solids and entities on such surfaces while being able to discern individual atoms. One can get an idea about the

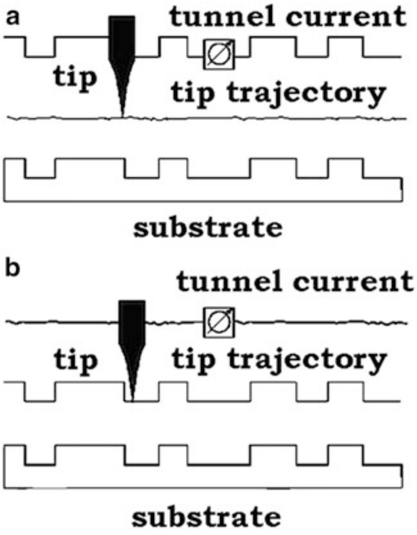


Fig. 1.12 Scheme of recording the tunnel current: the constant tip–sample distance (**a**), constant current value (**b**)

possibilities of the method looking at the images of the platinum surface with observed vacancies (Fig. 1.13) or annealed decanethiol films on a golden substrate. The method of scanning tunneling microscopy allows obtaining unique information about the processes occurring on a solid surface. In particular, the diffusion of atoms and molecules on surfaces has been studied and new stages of catalytic processes identified. At the same time the capabilities of tunneling microscopy are much broader than just the study of the structural characteristics of surface phenomena. This technique is being successfully utilized to navigate the surface of individual atoms and molecules and to create on this basis complex atomic and molecular structures.

Let an atom be adsorbed on the surface of a crystal, with the needle tip of the tunneling microscope being close by. In the field of the tip, a force arises which attracts the atom to the tip. The normal and the tangential components of this force with respect to the surface will vary depending on the relative position of the atom and the needle (Fig. 1.14). Therefore, if the needle tip moves along the surface, the atom starts to move along the surface in “jumps,” from one energetic minimum of the crystal surface to another one. This effect was used in 1990 by Eigler and Schweitzer of the IBM research center in San Jose (California). They dragged adsorbed xenon atoms along the surface of a single nickel-35 crystal so that they formed on the surface the three letters IBM (Fig. 1.14). Since then a variety of structures (the Brandenburg Gate, a cattle coral, triangles, etc.) have been crafted on solid surfaces with the help of the tunneling microscope in different laboratories. In 1991 Eigler, Lutz, and Rudge also proposed a surface construction of a bistable element in which switching was achieved by transferring a xenon atom on the surface of a nickel monocrystal from one steady state to another. Unfortunately, practically usable nanoelements of this type have not been developed so far.

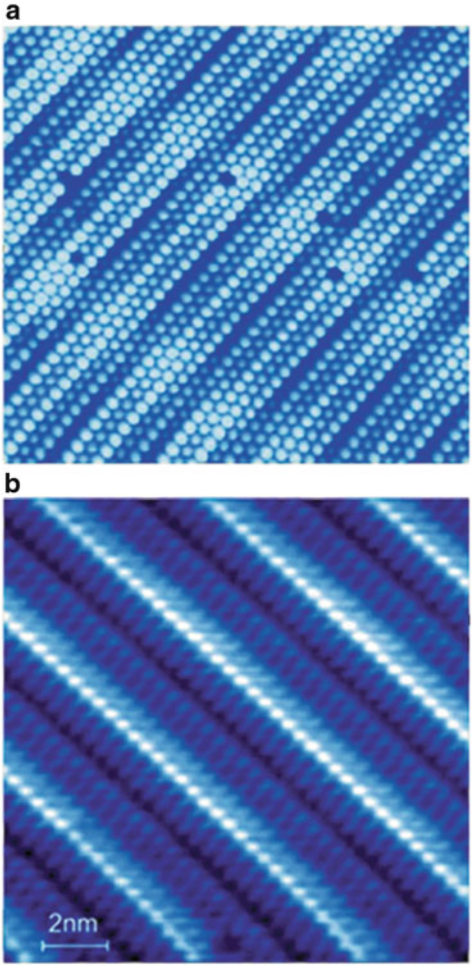


Fig. 1.13 Structure of the platinum single-crystal surface (**a**) and annealed decanethiol film (**b**) on a golden substrate recorded on a scanning electron microscope

The possibilities of reorganizing atomic and molecular surface structures are not limited to spatial rearrangements. The influence of the spectrometer's needle can lead to the dissociation of molecules into atoms or molecular fragments.

The scanning tunneling microscopy plays an important role in the studies of the surface structure of solids. But at the same time, it has a major shortcoming—the objects of investigation can only be materials that conduct electric current. Therefore, only a few years after the invention of the tunneling microscope, in 1985, IBM's Binnig and Gerber in Zurich and Quate, professor at the Stanford University in California, developed an atomic resolution scanning microscope that allows to study nonconducting objects. In this microscope, the image is created due to the interaction of the needle tip with the surface atoms of the investigated sample. However, the quantum effect of the interaction between atoms at short distances used in this case, the so-called dispersion force, differs from tunneling. This

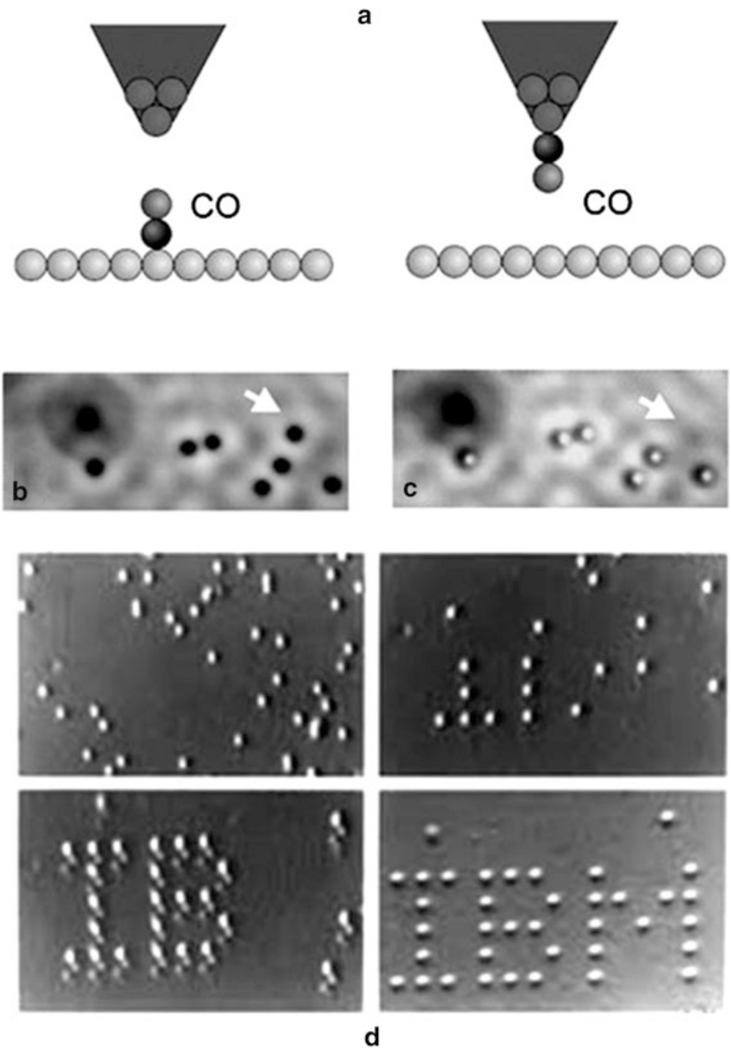


Fig. 1.14 Transfer of molecular fragments by a scanning electron microscope: (**a**) scheme, (**b** and **c**) transfer of separate atoms, (**d**) the first sign formed by tungsten atoms in the IBM laboratory

self-excited force arises due to the correlation of fluctuations of the electron density of atoms or molecules. A fluctuation in the atom (molecule) may induce instantaneous dipole moment. This dipole polarizes the electron shell of the neighboring atom and, at small distances, gives rise to a local dipole moment at the neighboring atom. The interaction of induced moments leads to attraction of the atoms, which is suppressed as the atoms come closer to each other due to electron shell repulsion. The mechanism of the emergence of such self-excited force was investigated in the

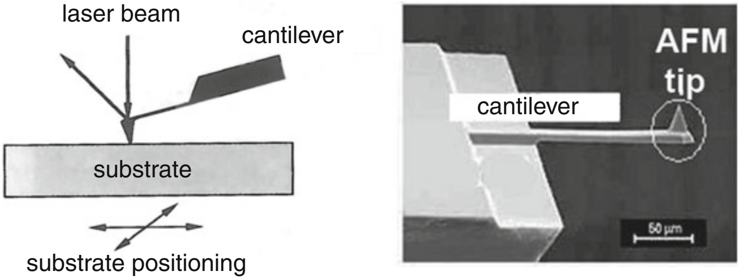


Fig. 1.15 Scheme of the atomic force microscope

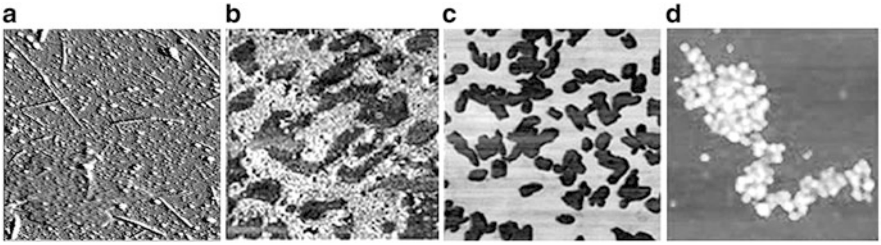


Fig. 1.16 Images obtained on an atomic force microscope (see text for explanations)

beginning of the 1930s by the English physicist London and was dubbed dispersion interaction. It is also known as van der Waals interaction.

Atomic force microscope uses dispersion interaction between the atoms of the needle tip and the surface of the investigated sample (Fig. 1.15). The needle is fixed to an elastic support—cantilever. The displacement of the tip due to dispersion interaction is described by Hooke's law, i.e., it is determined by the strength of the dispersion force and the elasticity of the cantilever. It is usually measured by registering the laser beam reflected by the cantilever with a photodiode matrix.

Today atomic force microscopy is a powerful method for studying the structure of diverse objects, including also the biological ones, and is widely used in research practice. Some of the typical examples are shown in Fig. 1.16: (a) carbon nanotubes (2 × 2 μm), (b) natural rubber (20 × 20 μm), (c) MoO_3 crystallites on a MoS_2 substrate (8 × 8 μm), and (d) chromatin, a complex consisting of proteins and DNA which is part of the chromosomes (400 × 400 nm).

1.3 Nanotechnology at the State Level

A turning point in the advent of nanotechnology as a distinct area of engineering and technology was the development in the United States of a document called "National Nanotechnology Initiative. Leading to the next industrial revolution." Formally this document was a report of the Interagency Working Group on

Nanoscience, Engineering, and Technology and the National Science and Technology Council and was seen as the justification of the 2001 budget requested from the Congress by the President of the United States. The Working Group was composed of the representatives of leading US agencies, including the National Science Foundation; Departments of Defense, Commerce, Energy, and Transportation; National Aeronautics and Space Administration; National Institute of Health; and several others. In terms of its significance, this section of the budget was considered a top priority for the development of science and technology. In a note accompanying the budget the US President Bill Clinton wrote:

> My budget supports a major new National Nanotechnology Initiative, worth $500 million. ... the idea of nanotechnology [is] the ability to manipulate matter at the atomic and molecular level. ... Imagine the possibilities: materials with ten times the strength of steel and only a small fraction of the weight, shrinking all the information housed at the Library of Congress into a device the size of a sugar cube, detecting cancerous tumors when they are only a few cells in size. Some of our research goals may take 20 or more years to achieve, but that is precisely why there is an important role for the federal government.

Almost immediately upon the release of this document it became a concept of industrial development in the United States due not only to a significant increase in funding but, not in the last turn, to the comprehensive coverage of the problem as a whole.

The document defines virtually all major aspects of the work in the field of nanotechnology, introducing a number of crucially important provisions. The authors of the document propose a detailed definition of the nature and fundamental goals of nanotechnology:

> The essence of nanotechnology is the ability to work at the molecular level, atom by atom, to create large structures with fundamentally new molecular organization. Compared to the behavior of isolated molecules of about 1 nm (10^{-9} m) or of bulk materials, behavior of structural features in the range of about 10^{-9}–10^{-7} m (1 to 100 nm—a typical dimension of 10 nm is 1,000 times smaller than the diameter of a human hair) exhibit important changes. Nanotechnology is concerned with materials and systems whose structures and components exhibit novel and significantly improved physical, chemical, and biological properties, phenomena, and processes due to their nanoscale size. The aim is to exploit these properties by gaining control of structures and devices at atomic, molecular, and supramolecular levels and to learn to efficiently manufacture and use these devices. Maintaining the stability of interfaces, and the integration of these "nanostructures" at the micron-length scale and macroscopic scale is another objective.

The document defines long-term objectives (grand challenges) for the future:

- Expansion of mass storage electronics to multi-terabit memory capacity.
- Making materials and products from bottom-up, that is, by building them up from atoms and molecules. Bottom-up manufacturing should require less material and pollute less.
- Developing materials that are 10 times stronger than steel, but a fraction of the weight for making all kinds of land, sea, air, and space vehicles lighter and more fuel efficient.

– Improving the computer speed and efficiency of minuscule transistors and memory chips by factors of millions making today's Pentium IIIs seem slow.
– Using gene and drug delivery to detect cancerous cells by nanoengineered MRI contrast agents or target organs in the human body.
– Removing the finest contaminants from water and air and to promote a cleaner environment and potable water.
– Doubling the energy efficiency of solar cells.

The document describing the National Nanotechnology Initiative included two Appendices. The first of them established the basic principles of funding the fundamental investigations of the infrastructure required in 2001. While determining the directions of basic research, the authors of the paper argued:

> Nanoscience is still in its infancy, and only rudimentary nanostructures can be created with some control. The science of atoms and simple molecules, on one end, and the science of matter from microstructures to larger scales, on the other end, are generally established. The remaining size-related challenge is at the nanoscale, roughly between 1 and 100 molecular diameters, where the fundamental properties of materials are determined and can be engineered. A revolution has been occurring in science and technology, based on the recently developed ability to measure, organize, and manipulate matter on a scale of 1–100 nm (10^{-9}–10^{-7} m) and on the importance of controlling matter at nanoscale on almost all human-made products. Recently discovered organized structures of matter (such as carbon nanotubes, molecular motors, DNA-based assemblies, quantum dots and molecular switches) and new phenomena (such as magnetoresistance and size confinement) are scientific breakthroughs that merely indicate future potential developments. Nanotechnology creates and utilizes functional materials, devices, and systems by controlling matter on this scale.

Based on these concepts Appendix 1 details specific areas of nanotechnological research of the highest priority (grand challenges): the creation of new materials; development of electronics and optoelectronics; improvement of healthcare and environmental protection; development of tools for energy conversion and storage; industrialization of space research; development of bio-nanodevices for detection and mitigation of threats to humans, maintaining defense superiority; and use of nanotechnological ideas in economical and safe transportation.

The nanotechnology initiative proposed by the US administration nearly doubled the funding for these activities in 2001 compared with 2000. Agencies participating in the initiative included:

- The National Science Foundation (NSF)
- The Department of Defense (DOD)
- The Department of Energy (DOE)
- National Aeronautics and Space Administration (NASA)
- Department of Commerce's (DOC) National Institute of Standards and Technology (NIST)
- National Institutes of Health (NIH)

Table 1.1 Funding of the US agencies under the program "National Nanotechnology Initiative"

Agency	2001	2002	2003	2006	2007
National Science Foundation	150	199	221	344	373
Department of Defense	123	180	201	436	345
Department of Energy	88	91	139	207	258
Department of Justice	1.4	1.4	1.4	1	1
Department of Transportation	0	2	2		
Environmental Protection Agency	5	5	5	5	9
NASA	22	46	51	50	25
National Institutes of Health	40	41	43	175	173
National Institute of Standards and Technology	33	38	44	76	86
Department of Agriculture	1.5	1.5	2.5	2	2
Total	464	604	710	1301	1277

The allocation of funding for these agencies is shown in Table 1.1. In comparison with 2000 ($270 million), in 2001 it has more than doubled, with further significant growth in subsequent years.

A major achievement of the developed program was the plan to establish an infrastructure for nanotechnological research, discussed in detail in Appendix 1. The plan provided for:

- Creating networks of shared research facilities equipped with state-of-the-art instruments in order to strengthen the capacity of various research teams and greatly facilitate their work
- Support for infrastructure development, including metrology, instrumentation, modeling, and simulation
- Addressing ethical, legal, and social implication and promoting workforce education and training efforts in order to attract skilled workers to this new and promising area

Table 1.2 gives an overview of the funding allocated in 2001 for various sections of Annex 1 of the Nanotechnology Initiative.

Appendix 2 to the "National Nanotechnology Initiative" lists nanotechnological devices developed over the past several years. The list includes magnetoresistance materials for high-density information storage, nanostructured catalysts, systems for drug delivery via the human blood circulatory system, metal–polymer composites, bio-detection of anthrax, water desalination, and random access storage based on rotaxane molecules.

The "National Nanotechnology Initiative" was approved by the US Congress in 2001, triggering a significant growth in spending on nanotechnological research and engineering (see Table 1.2). In 2006 it more than doubled compared to 2000. Moreover, the program stimulated similar developments in all industrialized countries. The volume of research carried out in 2002 and its growth in 2003 are shown in Fig. 1.17. It is worth noting that the growth of research in the field of nanotechnology is correlated with the pace of technological development in respective countries.

Table 1.2 Funding of different sections of the National Nanotechnology Initiative

Section	2000	2001
Fundamental research	87	170
Grand challenges	71	140
Centers and networks of excellence	47	77
Research infrastructure	50	80
Ethical, legal, and social implications	15	28
Total	270	495

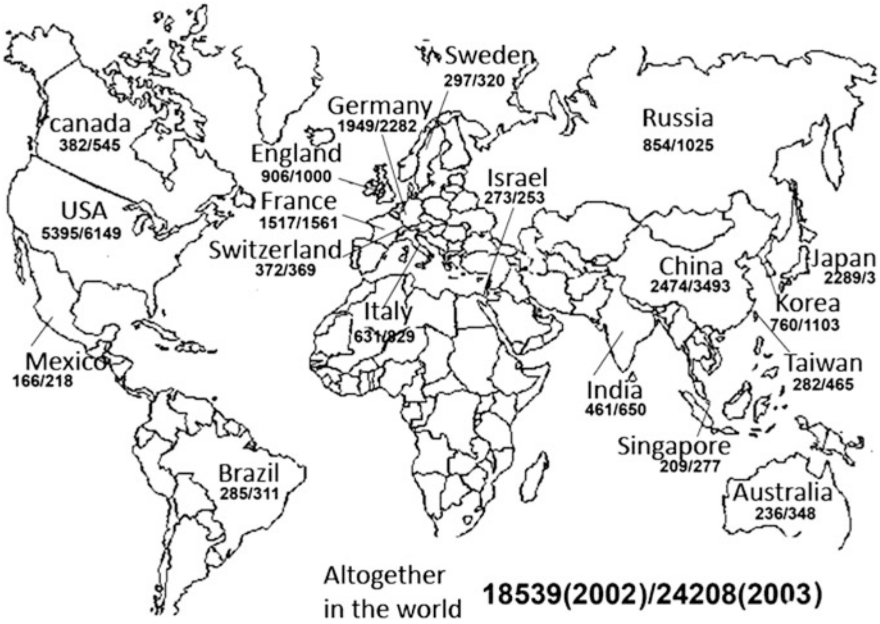


Fig. 1.17 Distribution of the volume of research in the field of nanotechnology performed in different countries on 2002/2003

Thus, research efforts in the People's Republic of China in 2002–2003 trailed only those in the United States.

1.4 Nanotechnology Today: Basic Principles and Their Unexpected Incarnations

In recent years the development of concrete nanoscale systems confirmed the significance and attractiveness of this field of activity, as predicted in the end of the last century. As a result, the current state of affairs can be justifiably called a nanotechnological boom, which has spread to most of the areas of human activity

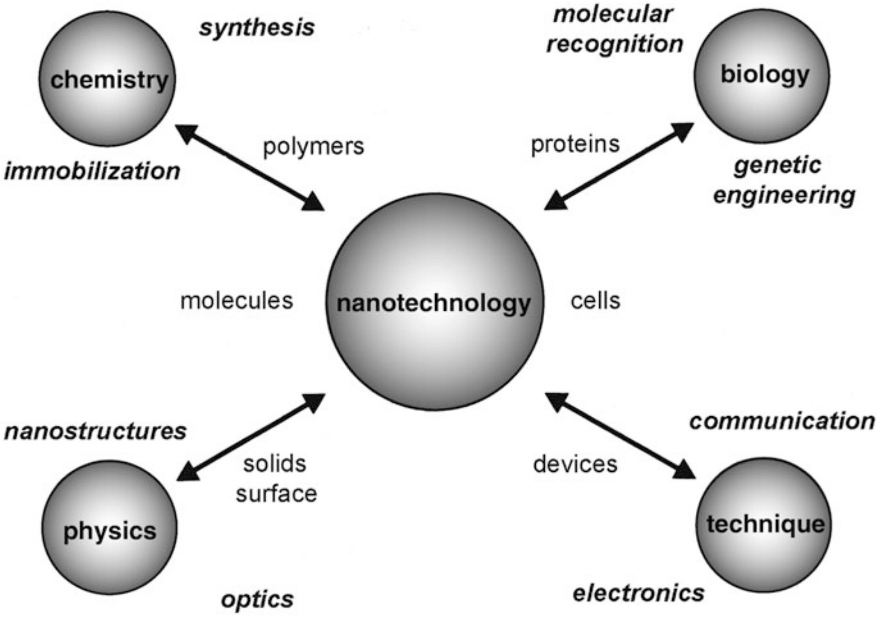


Fig. 1.18 Structure of nanotechnological studies

(Fig. 1.18). New concepts such as "nanochemistry," "nanobiology," and "nanomedicine" came into existence. And, in fact, nanotechnology has acquired the status of an interdisciplinary approach, with its main areas of application being physics, chemistry, biology, and engineering. One of the attempts to bring together the basic tools of this approach and the specific systems to which it can be applied is illustrated in Fig. 1.18. However, due to significant differences in specific features of different application areas, the notion of "nanotechnology" is undergoing change. First of all it is manifested by the attempts to separate the technology and research components of this concept. Thus, for example, in 2002 Koelling of the Materials Science Division of the Argonne National Laboratory (USA) attempted to clarify the content of these components. According to Koelling:

> The nanoscale is not just another step towards miniaturization. It is a qualitatively new scale where materials properties depend on size and shape, as well as composition, and differ significantly from the same properties in the bulk. 'Nanoscience' seeks to understand these new properties. 'Nanotechnology' seeks to develop materials and structures that exhibit novel and significantly improved physical, chemical, and tribiological properties and functions due to their nanoscale size. The goals of nanoscience and nanotechnology are:
>
> - To understand and predict the properties of materials at the nanoscale
> - To "manufacture" nanoscale components from the bottom up
> - To integrate nanoscale components into macroscopic scale objects and devices for real-world uses

Koelling's approach is by no means the only attempt to determine the essence and scope of nanotechnology. But, in fact, all of them incorporate the same basic principles, including:

(a) Extreme product miniaturization leading to the acquisition by the micro-objects of new properties that are distinct from the properties of their macro-analogues.
(b) Control of the macro-object properties by means of directed change in its structure at the microlevel. Such control is based on the processes of self-assembly and self-organization.

An essential role in the value system of nanotechnology is played by the economic potential of products. One of the many examples is constituted by metal–polymer composites. The introduction of metal nanoparticles into polymers leads to strength comparable with that of metals, while at the same time, these materials are much lighter and cheaper, which makes them attractive for the automotive industry. In this case, besides the reduction of the car's cost due to the use of metal composites as a building material, reducing the weight of the car leads to fuel economy, resulting in additional economic benefits.

It is easy to see that the definition of the basic nanotechnological principles given above does not explicitly use the word "nano," and this is no accident. These principles are so multifaceted that they now apply to the areas, which, strictly speaking, cannot easily be attributed to nanotechnology.

The most striking example is the US project to establish "pico-" and "nanosatellites." These systems promise a serious breakthrough in space exploration. In recent years, the possibilities of spacecraft miniaturization have dramatically increased. Significant progress in the semiconductor planar technology, rapid development in the area of MEMS (microelectromechanical systems), and the emergence of new construction materials have led to the emergence of spacecraft in a wide weight range.

In this connection, a new satellite classification system based on weight has been established. Satellites heavier that 1,000 kg are classified as standard, while those with the weight in the ranges of 100–1,000 kg, 10–100 kg, 1–10 kg, and below 1 kg are called small, micro, nano, and pico, respectively. Pico- and nanosatellites have turned out to be the most attractive ones because of their potential to serve as a basis for developing promising, methodologically novel space research programs. One example of such programs is the joint effort of the US National Aeronautics and Space Administration (NASA) and the Goddard Space Flight Center (GSFC) aimed at detailed investigation of the Earth's magnetosphere. Its centerpiece is the creation of a "constellation" of ~100 identical picosatellites that are simultaneously launched into different orbits with the same perigee of ~3 Earth radii. The apogees of the orbits must be in the range of 12–43 Earth radii with the difference of 3 radii. Each satellite is a simple specialized system (Fig. 1.19) with an engine and stock of fuel for orientation and orbit correction. They are launched into space from an intermediate platform which distributes them in different orbits (Fig. 1.20). The NASA/GSFC program provides that the data measured by the orbiting satellites are transmitted to the Earth station in the perigee. Satellites are designed to carry out

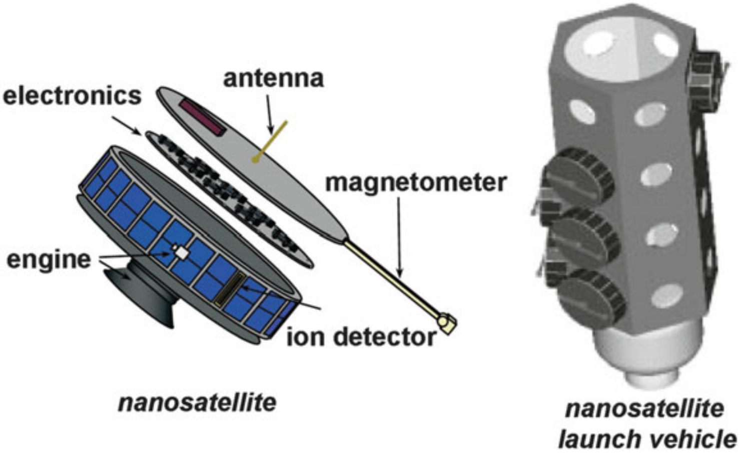
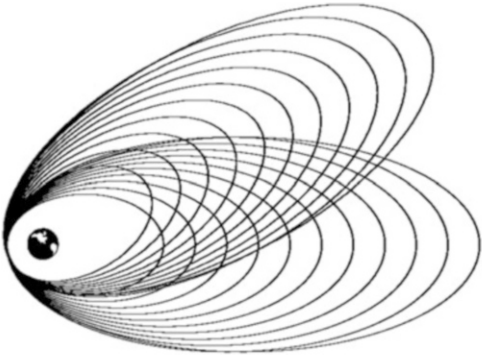


Fig. 1.19 Nanosatellite and its carrier

Fig. 1.20 Orbits of nanosatellites in the US National Aeronautics and Space Administration and the Goddard Space Flight Center program



both local (in the current point of the orbit) and remote measurements. To this end their orientation with respect to Earth is stabilized by rotation or in three spatial axes.

Thus, the dimensions of the individual elements of the picosatellite system are far from nanoscale. Nevertheless the basic principles of the system design comply with the nanotechnological ones. Simple microelements form a macrosystem performing much more complex functions than its constituent individual parts. The significant advantages of such a system compared to separate standard multifunctional satellites are launch weight reduction, the possibility of using a less powerful spacecraft carrier, and, consequently, a significant reduction in the cost of the program.

The NASA/GSFC system is one of the simplest cases discussed in the literature in terms of structure and functions. Other system designs are possible in which dedicated satellites collect and process in the orbit the information received from other satellites, analyze it, and, if necessary, initiate repeated measurements. An even more attractive option is a system of picosatellites built on the principles of self-organization and capable of reorganizing its structure and functions depending on data received by each satellite.

Thus, the nanosatellite system represents in a sense a macro model of the system operating based on nanotechnological principles. Indeed, an object whose macroscopic properties change due to directed structural changes at the microlevel represents one of the varieties of distributed dynamic systems. A distributed dynamic system is defined as spatial ensemble of the components, elementary for the given system, that perform specific functions and interact with each other. Atoms and molecules in a crystal or in an amorphous body, microorganisms in a biological culture, individual molecular units in block copolymers—all these are basic components of distributed systems at different levels of structural organization. The dynamics of a distributed system is determined not only by the processes taking place in each of its elements (i.e., at each point of the system) but also by the interactions between these elements. Therefore it turns out to be fundamentally different from the processes in its basic components and is much more complex. In the theory of distributed systems, the term "emerging" properties refers to the properties that are inherent to the system as a whole and cannot be directly derived from the properties of its elementary constituents.

In a general sense, nanosystems are distributed environments with complex mechanisms of interaction at the nano-level. It is these mechanisms that determine the processes of self-assembly or self-organization at the structural level and lead to the appearance of new, emerging properties of the system at the macro-level. Because of the similarity of construction and operational principles of distributed systems at different levels of structural organization as well as the processes taking place in them and their new properties, some analogies between the nanoscale systems built and the macroscopic distributed systems become apparent.

Today, at the height of the nanotechnological boom, the principles and methods of nanotechnology are gaining new positions in various fields of human activity. But among all these areas, there is one which was like a "testing ground" for nanotechnology, passing successive stage of miniaturization of the devices created. Since the 1940s of the last century, computer technology has made an amazing journey from vacuum tubes to very large integrated circuits (VLSI) and persistently seeks today to use the molecular component base. In this book an attempt will be made to describe the main principles and ways of establishing the molecular component base and the possible role of molecular information processing devices in information technology.

Chapter 2
Computer Engineering and Nanotechnology

In those days computers were distributed by government order. Korolev and Mishin personally, wherever they could, attempted to attain the delivery of computers to Special Design Bureau number 1
Boris E. Chertok "Rockets and People. Race to the moon"

The most powerful experimental supercomputers in 1998, composed of thousands or tens of thousands of the fastest microprocessors and costing tens of millions of dollars, can do a few million MIPS. They are within striking distance of being powerful enough to match human brainpower, but are unlikely to be applied to that end. Why tie up a rare twenty-million-dollar asset to develop one ersatz-human, when millions of inexpensive original-model humans are available? Such machines are needed for high-value scientific calculations, mostly physical simulations, having no cheaper substitutes. AI research must wait for the power to become more affordable.
Hans Moravec "When will computer hardware match the human brain?"

2.1 A Brief History of Computing

When people began to understand themselves as reasonable beings, they felt the need to describe the world around them—to count everything their eye caught. The choice of the number system was quite natural. Ten fingers on the human hands, ten toes on the feet—such was the decisive argument in favor of the intuitively chosen decimal system. It has remained that way up to now. The symbols denoting numbers changed from time to time, but the system itself, most psychologically acceptable for us, remained the same practically everywhere.

It seems that very soon fingers and toes ceased to be sufficient, and the question of tools to facilitate calculations came up. The progress of computer technology is a century-long process which involved ingenious representatives of the human society (see the excellent reviews by B. N. Malinowski and B. A. Gladkikh). We know

© Springer-Verlag Wien 2014

N.G. Rambidi, *Molecular Computing*, DOI 10.1007/978-3-211-99699-7_2

that the great Leonardo da Vinci drew up plans for an adding machine which was reproduced in metal in our days and proved to be quite operational. This machine operated with 13 digit-registering wheels.

2.1.1 A Little Detail: The Mechanical Calculators

The basic idea of such a device is shown in Fig. 2.1. Suppose that each position of an arbitrary decimal number corresponds to a shaft on which two gear wheels with different numbers of operative teeth are mounted. Suppose that each shaft is set to its initial position corresponding to the number "0." Other numbers correspond to subsequent rotations of the shaft by 36°. With a gear ratio between the gear wheels at adjacent shafts equal to 1:10, the shaft corresponding to units must spin 10 times by 36° to cause the shaft representing tens to move one position. This design allows dialing any number limited only by the number of shafts and to add to or subtract from it any number. Mechanical systems based on gear transmission were being perfected and used almost until the middle of the last century. Back in the 1950s the Soviet Army had in operational service the PUAZO antiaircraft artillery director which determined correction for velocity, humidity, temperature, etc., while controlling antiaircraft fire. PUAZO was a transportable cube with edges roughly one meter in length stuffed with drive gears, worm gears, electric motors, etc.

Over the centuries that followed the Leonardo da Vinci era, outstanding scholars from different countries—Blaise Pascal, Gottfried Leibniz, and Charles Babbage—occupied themselves with the development of computer technology. However, by the time the complex, cumbersome computations turned out to be vital, the level of computing technology was inadequate for addressing the most pressing challenges.

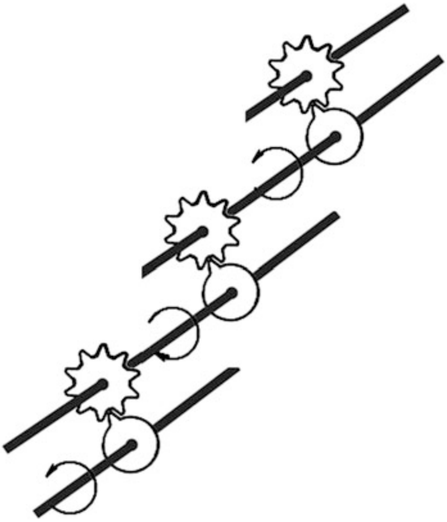


Fig. 2.1 Scheme of a mechanical computing device

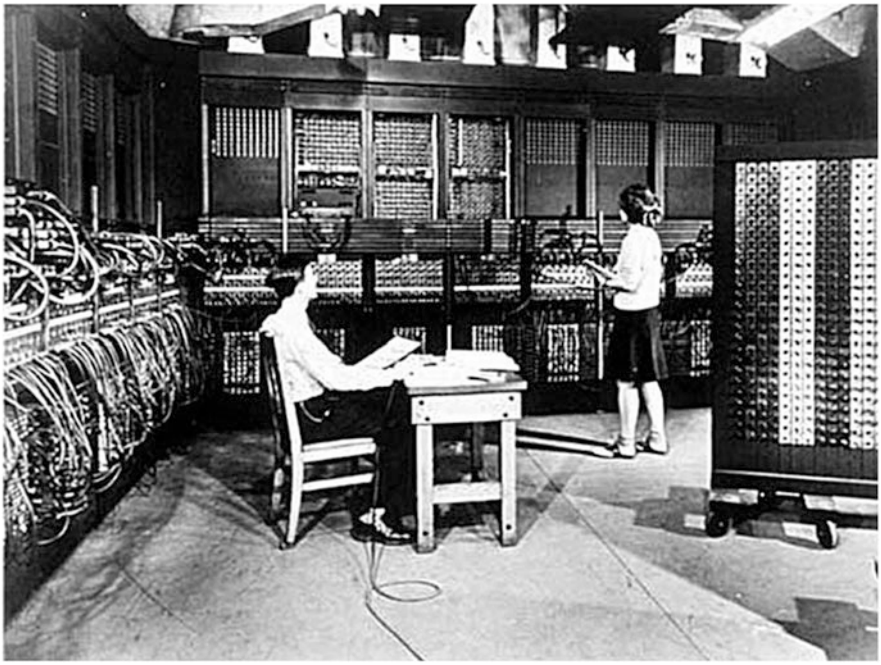


Fig. 2.2 Electronic computing device ENIAC—input of initial data

During the Second World War, improvement and development of new types of weapons generated overwhelming requirements to computing.

In 1941, the personnel of the Aberdeen Ballistic Research Laboratory in the United States approached the University of Pennsylvania's School of Electrical Engineering located nearby for help in creating firing tables. They suggested using the available Bush differential analyzer—a bulky mechanical analogous computing device. However, John Mauchly, a physicist at the School, proposed to create for this purpose a then powerful computer based on electronic valves. In April 1943 a contract was signed between the Ballistic Research Laboratory and the University of Pennsylvania to develop a computing machine, called the Electronic Numerical Integrator and Computer (ENIAC) with a budget of $400,000. About 200 people took part in this work, including several dozen mathematicians and engineers. The project was placed under the supervision of J. Mauchly and a talented electrical engineer Presper Eckert. The hard work was completed in late 1945 when ENIAC was successfully tested. In early 1946, the machine was first applied to solving real problems. Its dimensions were impressive: it measured 26 m in length, 6 m in height, and weighed 35 tons. It used the decimal numeral system and could hold in memory 20 ten-digit decimal numbers. ENIAC was programmed by being rewired via plugs and a patch panel which caused inconvenience since it could take many hours and even days to reprogram (Fig. 2.2). Therefore in 1945, while still completing the ENIAC project, its creators were already developing a new electronic

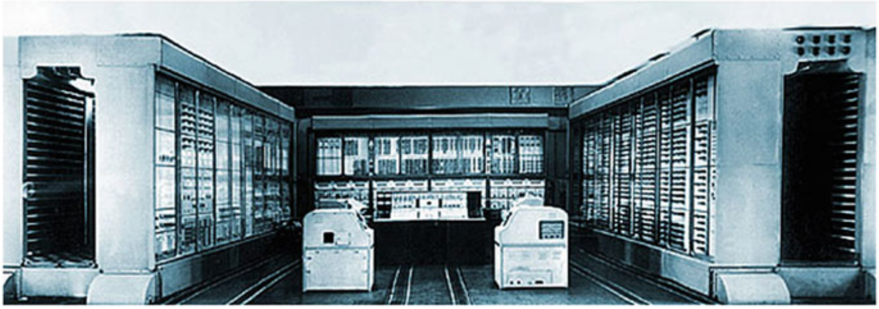


Fig. 2.3 First Soviet series-produced mainframe "Strela"

digital computer EDVAC. They suggested to store programs directly in the computer's random access memory, thus eliminating the main drawback of ENIAC—configuration using physical switches and plugs. At this time the world-renowned mathematician John von Neumann, member of the Manhattan project to develop the atomic bomb, joined the team. He immediately appreciated the perspectives of the new technology and took active part in creating EDVAC. The part of the report on the machine he wrote contained a general description of EDVAC and the fundamental principles of its design. In spite of the fact that a number of engineers were involved in the development of these principles, they later were called "von Neumann principles" (von Neumann paradigm). In any case, the establishment of these principles was a revolutionary event in computer technology that determined its further development.

Without going into detail of the history of the development of digital von Neumann computers, the most significant milestones are as follows.

In 1953 IBM began the production of a general-purpose computer. The series-produced IBM-701 had a random access memory with the capacity of 2,000 words of 36 bits each and was capable of ~10,000 operations per second.

Just 6 years later IBM-7030 installed at the Los Alamos Scientific Laboratory reached one million operations per second. Its random access memory capacity was 256,000 64-bit words.

The pace of progress in computer technology over the second half of the last century is illustrated by the characteristics of one of the most powerful supercomputers of our time, Top-500 Earth Simulator. It consists of 640 modules, each containing 8 processors. The system has 10 terabytes of RAM and is theoretically capable of 40 teraflops (Flop stands for the number of floating-point operations per second).

In the Soviet Union the first electronic computer was created in 1952 under the supervision of academician S. A. Lebedev. His first full-fledged computer, BESM, could perform 8,000 operations per second and store 1,000 39-bit words.

The first series-produced mainframe "Strela" was built in 1953 under the supervision of Yu. Ya. Bazilevsky (Fig. 2.3). Its speed was 2,000 operations per second, and its random access memory was based on cathode-ray storage tubes

(43-bit words). The machine occupied the area of ~300 m^2 without air conditioning which required almost as much space.

Very popular among computer specialists was Lebedev's BESM-6 which went into production in 1968. Its random access memory contained between 32,000 and 128,000 48-bit words, and its speed reached one million operations per second. Until 1987 355 such machines were manufactured.

A significant event in the history of computing was the introduction of personal computers, an entirely novel direction distinctly different from the general line of development in the 1940s–1970s. This has been facilitated by at least two factors. First, an urgent need was felt in different fields of human activity for computers that would be sufficiently powerful and at the same time simple in operation. Second, the electronic industry was essentially prepared for this need. A number of manufacturing companies brought to market electronic components, including microprocessors which were comparatively powerful for that time.

The first attempt to combine components into a single unit was made in 1975 by a company called MITS which released for sale a kit dubbed "Altair," essentially a set of parts and a housing. Buyers were supposed to solder together and test the assembled units and to create computer programs in machine language.

Subsequently a number of other versions of simple electronic computing devices appeared on the market. But the real beginning of the history of personal computers is associated with the names of American electronic engineers Steve Jobs and Steve Wozniak.

In the early 1970s Jobs worked at Atari where he met the senior developer of the company, Ron Wayne. Along with Wozniak, who at that time worked at Hewlett-Packard, they spent nights in a garage working in a personal computer which they called Apple-1. Soon after, on April 1, 1976, they created a company called Apple Computer. While establishing the company ran into significant difficulties. Constant need for funding, a cautious attitude of consumers to this new direction, and a number of other factors led to Ron Wayne's departure. Nevertheless, after the first model Apple-1 which received lukewarm response, Jobs and Wozniak created the Apple-2 computer which was a tremendous success. As a consequence, large companies started to produce personal computers. In the early 1980s, IBM has released its 5150 model, conspicuously called IBM PC, thus introducing the term "personal computer" widely used today.

It is now hard to find an area of human activity where these amazing computing devices would not be used for a broad range of tasks—from powerful computing to relaxing entertainment.

During the second half of the last century the information capacity of computer technology was steadily growing. A characteristic feature of this process was the preservation of the basic principles of computing devices and the continuous refinement of the components implementing these principles. In turn, progress in the component base was accompanied both by a change in the physical principles underlying the mechanisms of function of the components and by their consistent miniaturization, resulting in the need to change the production technology.

2.1.2 Some Details: von Neumann Paradigm and Its Implementation

The von Neumann paradigm encompasses six main principles:

1. Computers built on electronic elements should work in the binary rather than in the decimal system.
2. The program must be located in a storage device with sufficient capacity and appropriate read/write speed for commands of the program.
3. The program and the numbers with which the machine operates are represented in binary code. Thus, in terms of representation, commands and numbers are of the same type. This leads to the following important consequences:
    - Intermediate results of calculations, constants, and other numbers may be stored in the same storage device as the program.
    - Numerical notation of the program code allows the machine to perform operations on values that serve to encode program instructions.
4. The challenge of physical implementation of a mass storage device with a speed of operation corresponding to the performance of logic circuits requires a hierarchical organization of the memory.
5. Arithmetic units of the machine are constructed on the basis of circuits that perform the operation of addition. Implementing specialized devices for other operations is impractical.
6. The machine uses the principle of parallel operation on words, executed simultaneously over all bits. At the same time execution of program instructions (computer operations) is performed sequentially, one after another.

Today's computer technology is very diverse (Fig. 2.4). It includes digital computers in which program instructions, input data for solving problems, and computational results are recorded in memory as sets of binary characters. It also includes analog devices, processing continuous sequences of values of some physical quantity. Furthermore, specialized devices for mass solutions of a single task or a group of similar tasks also exist. However, the overwhelming majority of com-

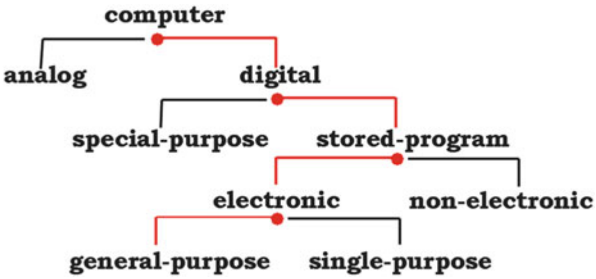


Fig. 2.4 Main directions of computer technology

putational devices that surround us in our daily life are universal digital computers. Let us consider the implementation of von Neumann principles on the example of a typical representative of this category of devices—a personal computer.

Computers of this type are built on the modular principle and represent a set of separate units. They communicate with each other using a special information highway—the bus. To create a flat bus multicore cables are usually used. The set of bus wires is divided into separate groups to transfer the address code of the operation to be performed, data, and control signals.

The principal computer components include (Fig. 2.5) a storage device, an arithmetic logic unit, and a control device.

The storage device, or memory, is a collection of cells designed to store information. Each cell is assigned a unique number called the address. Information stored in the cell may be both machine instructions and data. A machine instruction

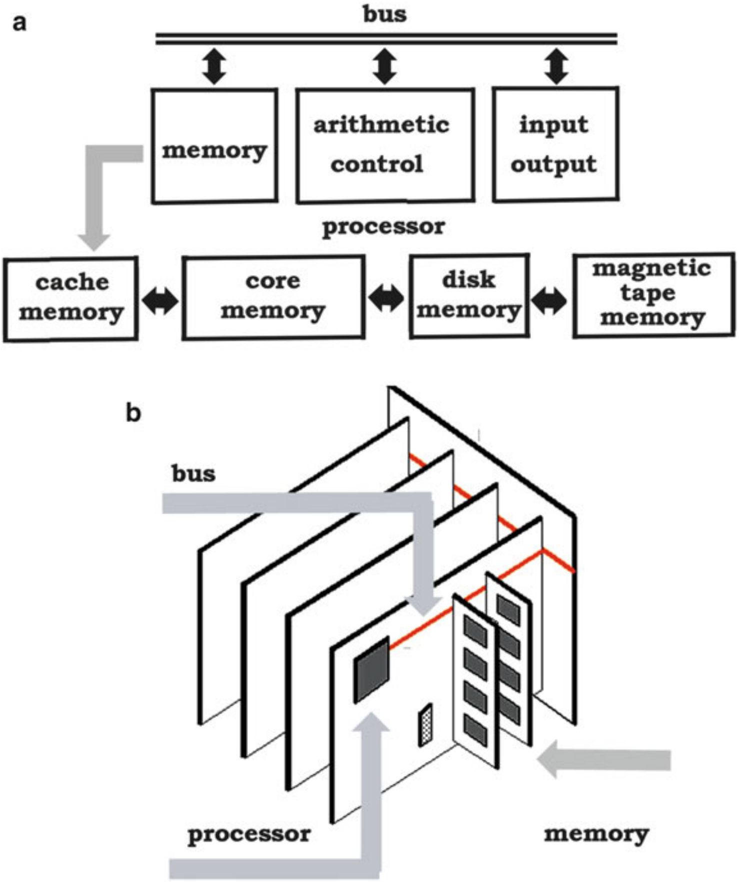


Fig. 2.5 Structure of a modern personal computer: (**a**) block scheme, (**b**) spatial location of devices

is a binary code that defines the operation to be carried out, the addresses of its operands (i.e., the codes of the numbers on which the operation is to be performed), and the address of the cell in which the result of the operation will be recorded. Storage devices of modern computers have a hierarchical structure (Fig. 2.5). The main memory is a solid state random access storage device with the read/write speed comparable with the processing speed of the arithmetic logic unit. In order to neutralize the difference in these speeds, an additional, small capacity high-speed memory can be used known as cache memory. In Pentium processors it contains 8,000 cells for code and another 8,000 cells for data. For technical and economic reasons the storage capacity of memory devices is limited. Today it reaches several gigabytes. Storage capacity can be extended by using slower storage on magnetic disks (up to hundreds of terabytes) and magnetic tapes with practically unlimited capacity.

All operations in a computer are controlled by the signals generated by the control unit. The control unit generates the address of the next command to be executed and sends a control signal for the contents of an appropriate memory cell to be read. The command readout is transmitted to the control unit. According to the information contained in the address fields of the command, the control unit generates the addresses of operands and control signals for reading the operands from the storage and transmitting them into the arithmetic logic unit. Subsequently the control unit sends signals for executing the operation to the arithmetic logic unit. The result is stored in the machine memory. Result attributes (sign, overflow flag, zero flag, etc.) are delivered to the controller where they are written in a status register. This information can be used while carrying out subsequent commands, e.g., conditional jump instructions.

2.2 Semiconductor Devices: A Revolution in Electronics

Tremendous progress made by computer technology is due to the development of the component base of computing devices. Over the past half century it underwent revolutionary changes that led to modern means of information processing utilized in virtually all areas of human life and activity.

The first major shift in digital computers was the transition from mechanical and relay systems to vacuum tubes. It is now even difficult to remember what vacuum tubes looked like and how radios, amplifiers, and control devices on their basis functioned. In the simplest case three electrodes—cathode, anode, and an intermediate electrode called the grid—were sealed into a glass vacuum tube. The cathode was heated by electric current and emitted electrons that were accelerated toward the anode by voltage passed through the grid. An electron tube, even in this simplest three-electrode implementation, is a natural embodiment of a switching element required for logical circuits. Depending on the potential on the grid, it either lets the current flow through or not, thus constituting an element with two stable states. Therefore, starting with ENIAC tube-based computing systems were created in different countries. However, thousands of electron tubes in a single device consumed a lot of energy and required labor-intensive maintenance by skilled personnel. Electron tube-based computing systems were capricious in operation. In order to achieve stable operation, days had to be spent on debugging the system. The temperature in the computer room was sometimes 10–15 °C. Therefore, the appearance of semiconductors dramatically increased the reliability of computers and paved the way for their further improvement.

It is not by accident that semiconductors won a strong position in computer engineering. As is known, in contrast to the quantum objects like atoms and molecules, solids have electronic band structure.

In order to understand the underlying reasons for the formation of the band structure, let us consider a simple model—a one-dimensional chain of atoms (Fig. 2.6). If there were two atoms, the electronic levels of such a system would be split into two components—a bonding and an antibonding orbital (see Chap. 3).

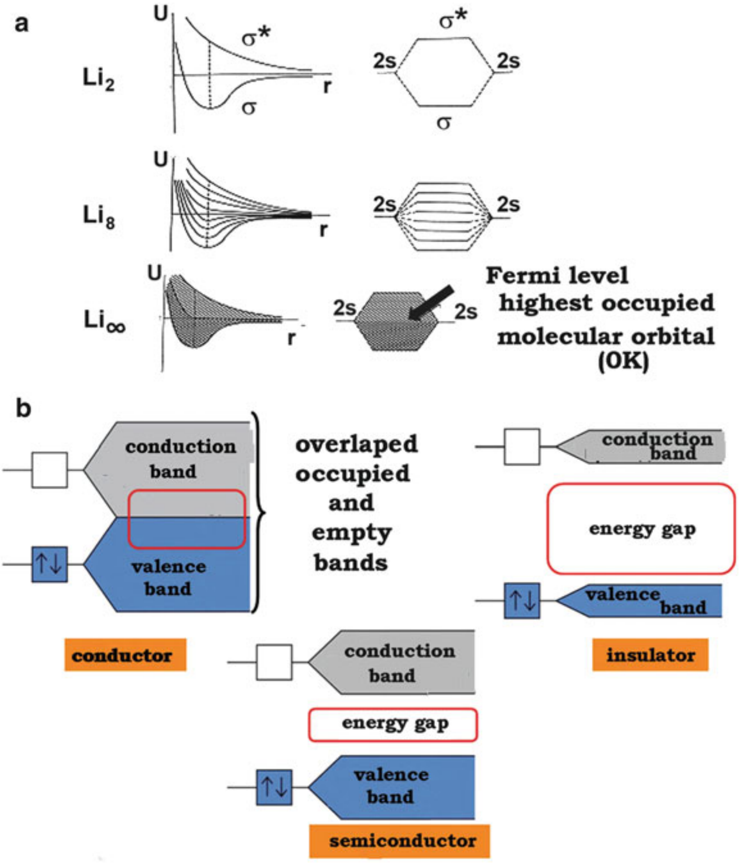


Fig. 2.6 Scheme of the formation of bands in a chain of Li atoms (**a**) and band structure of a conductor insulator and semiconductor (**b**)

As the number of atoms in the chain increases, so does the degree of level splitting and ultimately two energy bands arise which are called the valence band and the conduction band. They are continuous but contain a finite number of electronic states. For a number of elements these zones overlap (Fig. 2.6). The electrons of the atoms occupy the lower levels of the valence band. The other levels remain vacant and the electrons can move from the valence to the conduction band. Electronic conductivity arises in the system. Such electronic structure corresponds to conductors of electric current. In general the properties of solids are determined by the distance between the valence and conduction bands, i.e., by the band gap between them, and by the degree to which the valence band is filled by electrons. If the band gap is wide and the valence band is completely filled, the solid is an insulator. In the most interesting case, corresponding to a semiconductor, the valence band is nearly completely filled and the band gap is relatively narrow.

Semiconductors that practically do not contain dopants are called intrinsic, or undoped. In this case, when an excited electron is promoted from the valence to the conduction band, a positively charged vacancy is produced in the valence band. Of course the neighboring electrons can neutralize this vacancy, but while doing so they will form a new vacancy elsewhere. Thus, a positively charged moving entity appears in the semiconductor that is called a hole. In intrinsic semiconductors charge carriers must appear in pairs (electron–hole pair). The situation changes significantly if a certain amount of dopants—alloying additions—is introduced into the semiconductor. We will consider silicon whose electronic structure and properties correspond to a semiconductor. Tetravalent silicon forms four covalent bonds with neighboring atoms. If trivalent boron is introduced into the structure of the silicon crystal, one of the bonds remains unfilled (Fig. 2.7). It can be filled by an electron of any other neighboring silicon atom, leading to the formation of a hole. Dopants of this kind are called acceptors, and the resulting holes are situated just above the valence band. Such semiconductors are called p-type semiconductors. A different situation arises when a pentavalent atom of a dopant (phosphorus or antimony) is introduced into a silicon semiconductor. These atoms have five valence electrons, one more than silicon. The fifth electron is easily detached from the atom containing it. As a result a static ionic charge as well as an energy state corresponding to the fifth electron, situated slightly below the conduction band, arise. Such dopants are called donor dopants and the semiconductors n-type semiconductors.

Remarkable properties arise upon contact of p- and n-type semiconductors. In this case, due to large difference of concentrations—a whole sea of holes on one side and a sea of electrons on another side—strong diffusion currents of holes and electrons arise. As a result minority carriers appear in the system—electrons in p-type semiconductors and holes in n-type semiconductors. At the same time ions of the electron–hole pairs will be approaching the junction causing electric field around it (Fig. 2.8). In turn, this field will cause drift currents of electrons toward the n-type material and of holes toward the p-type material. If the external voltage is absent, the diffusion and drift currents will be equal in magnitude and opposite in direction. Therefore, the total current will be zero. Suppose that electric field is

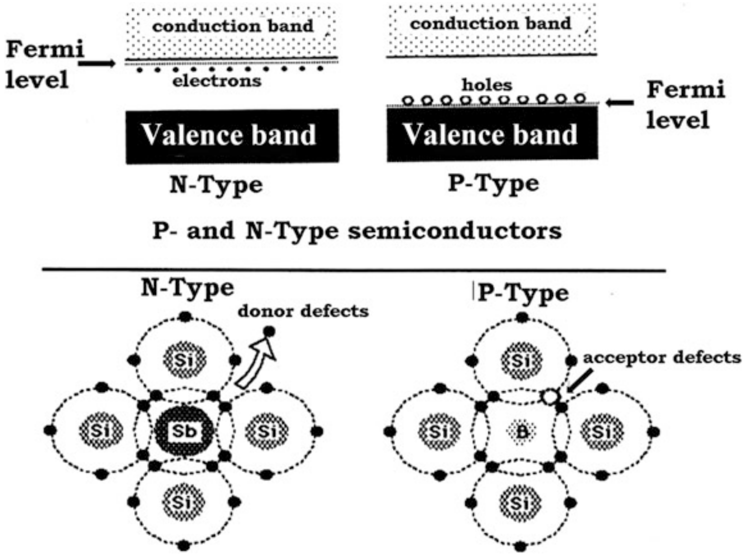


Fig. 2.7 Semiconductors of n- and p-type

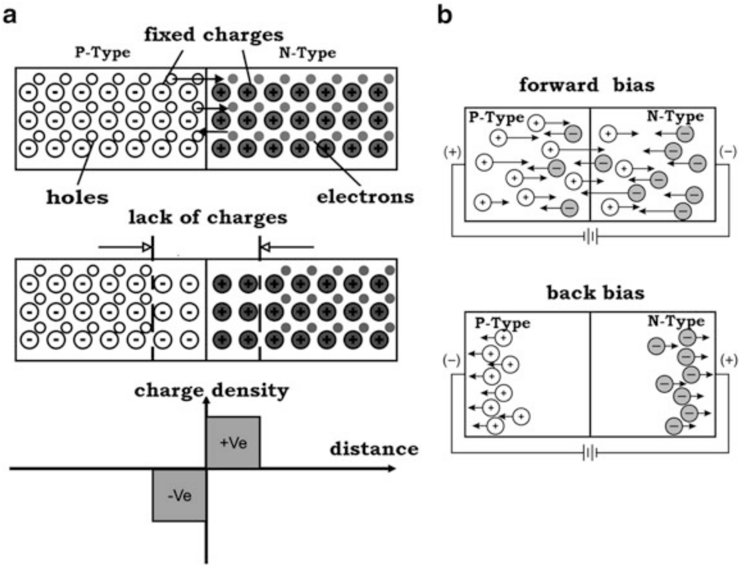


Fig. 2.8 p–n transition (**a**) and semiconductor diode (**b**)

applied to the junction (direct shift—a positive potential is applied to the n-type semiconductor and a negative one to the p-type semiconductor). This voltage will increase the concentration of minority carriers which will penetrate deeper into the material and recombine with majority carriers. Those minority carriers that

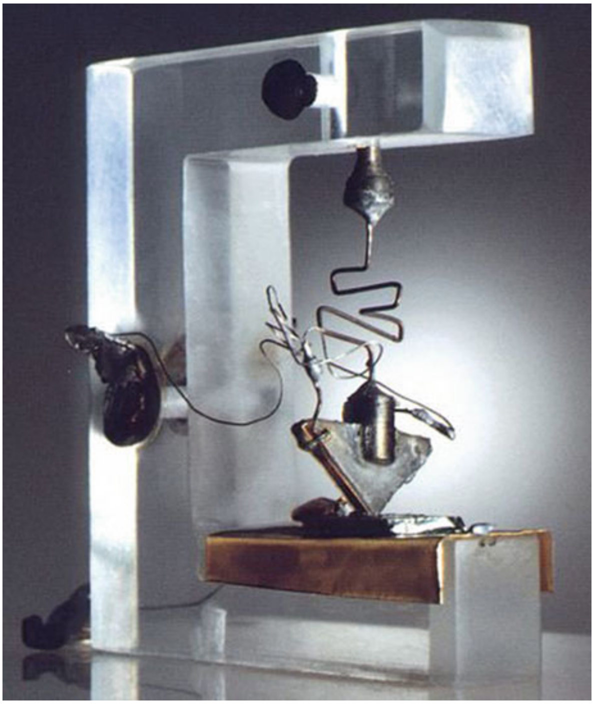


Fig. 2.9 First Bardeen and Brattain transistor

disappeared as a result of recombination will be replaced by new ones due to diffusion through the junction, leading to constant direct current. Similarly, if the field is applied in the reverse direction (reverse bias), the current through the junction will be vanishingly small (Fig. 2.9). Thus, the p–n transition behaves as a semiconductor diode.

All these properties of semiconductor materials and p–n and n–p junctions were used to create a unique semiconductor device—the transistor—which changed the face of computing devices. Today this name applies to a large group of semiconductor switching devices with two stable states.

In 1946 William Shockley, Walter Brattain, and John Bardeen at Bell Laboratories in New Jersey (AT&T Bell Labs) started to work on a semiconductor device—the transistor. In 1947 Bardeen and Brattain demonstrated the first implementation of a transistor on the basis of a germanium crystal with p- and n-zones, with metallic wires connected to the junction (Fig. 2.9, reference 6 in Chapter 2). Based on their work Shockley analyzed the physics of the device and a few months later proposed a fully planar semiconductor transistor. In 1956 Shockley, Bardeen, and Brattain received a Nobel Prize for this work. In his Nobel lecture John Bardeen said: “I knew the transistor was important, but I never foresaw the revolution in electronics it would bring.”

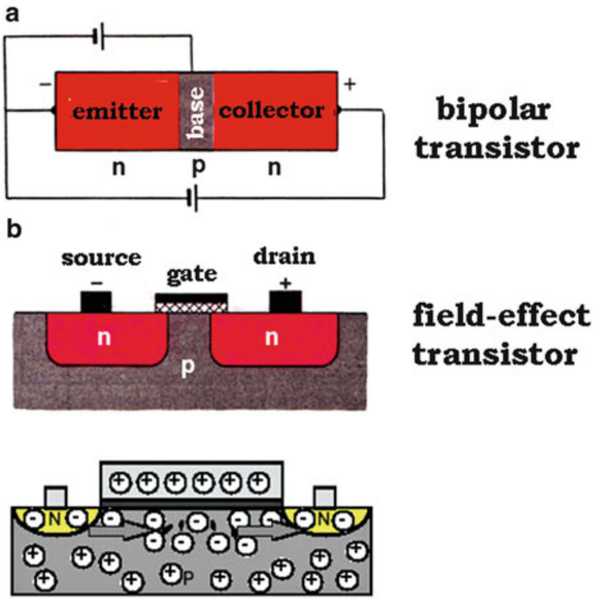


Fig. 2.10 Bipolar (**a**) and field-effect (**b**) transistor

As a result of continued rapid development of the theory and semiconductor technology, various versions of two basic types of transistors—bipolar and field-effect transistors—were created (Fig. 2.10). They employ a combination of p–n and n–p semiconductor junctions.

A bipolar transistor comprises two semiconductor areas of the same type (emitter and collector), separated by a thin layer of semiconductor of the other type (base). In simple terms it can be thought of as two p–n junctions joined back-to-back. If voltage is applied only between the emitter and the collector, then for any polarity of the voltage current will not flow. One of the two p–n junctions will be closed. But when the voltage is applied between the emitter and the base, current flows in the chain emitter–collector, and the strength of this current can be controlled by much weaker emitter–base current.

The field-effect transistor introduced somewhat later is based on the idea expressed as early as 1925 by the American researcher Julius Lilienfeld. He proposed to control the resistance of a semiconductor layer in a system which is essentially a capacitor, with one plate made of metal and the second one from doped semiconductor, using voltage applied between the metal and the semiconductor. If negative potential is applied to the metallic plate, the field will displace the electrons from the surface layer of the semiconductor, leading to a lack of current carriers and an increase of resistance. When the polarity is reversed the number of carriers in this area will go up and the resistance will increase. The mechanism of action of the field-effect transistor is shown schematically in Figure.

A very important factor in creating the field-effect transistor was the availability of suitable materials: silicon (semiconductor) and silicon dioxide (insulator).

The latter can be easily grown on the silicon surface by oxidation. In this fashion the foundations of the modern MIS (metal–insulator–conductor) technology, also called MOS (metal–oxide–semiconductor) technology, were laid.

A revolutionary step in the development of the semiconductor technology was the transition to integrated circuits in which all the elements of the transistor are formed on the surface of a silicon crystal or some other semiconductor media. Their appearance was due to an acute need to improve the reliability of equipment and to automate manufacturing and assembly of electronic circuits. Assembling equipment at that time was mostly manual—a very laborious and time-consuming process poorly amenable to automation. As the number of switching devices in the digital equipment, especially in computers, increased manifold, the reliability and the mean time between failures dropped sharply. For example, the CD1604 computer released in 1960 by the US firm Control Data Corp. contained about 100,000 diodes and 25,000 transistors and could work without failure for not more than 2–3 h.

The world's first integrated circuits were designed and built in 1959 independently by Jack Kilby at Texas Instruments and Robert Noyce at the Fairchild Semiconductor Company.

In 1958 Kilby began to work on integrated circuits in which electronic components were supposed to be located on the same substrate. By this time semiconductor materials could be used to produce resistors, capacitors, and transistors. Resistors were produced using the ohmic properties of the semiconductor "body," while capacitors were built based on reverse biased p–n junctions.

In 1959 Kilby demonstrated the design of a flip-flop on one monolithic piece of germanium. For its production, photoengraving techniques patented by Texas Instruments were used. This "solid circuit" was introduced in 1960 at the exhibition organized by the American Institute of Radio Engineers.

Many of the shortcomings of "solid circuits" were later addressed by Robert Noyce, working at Fairchild. He developed technological processes that anticipated the modern semiconductor planar technology. A patent application was filed, and developers of components started to work on bringing diffusion resistors and transistors together on silicon wafers.

In 1960, a group of researchers at Fairchild Semiconductor headed by Jay Last produced the first integrated circuit containing four transistors (Fig. 2.11).

"You and I agree that while the world loves a hero, semiconductor progress depended on the efforts and ideas of a large number of people and that moving forward depended on contributions going back a few decades in some cases. Also, as is the case with most inventions, a number of people with access to the same pool of common knowledge were working independently at the same time to put it altogether and to make the necessary extensions to the existing technology and who realized that the time was right for society to accept the new concepts."—Jay Last later wrote in a letter to one of his friends.

The development of integrated circuits began to progress at a feverish pace. This was the beginning of a new era. To obtain a rough estimate of the rate of development,

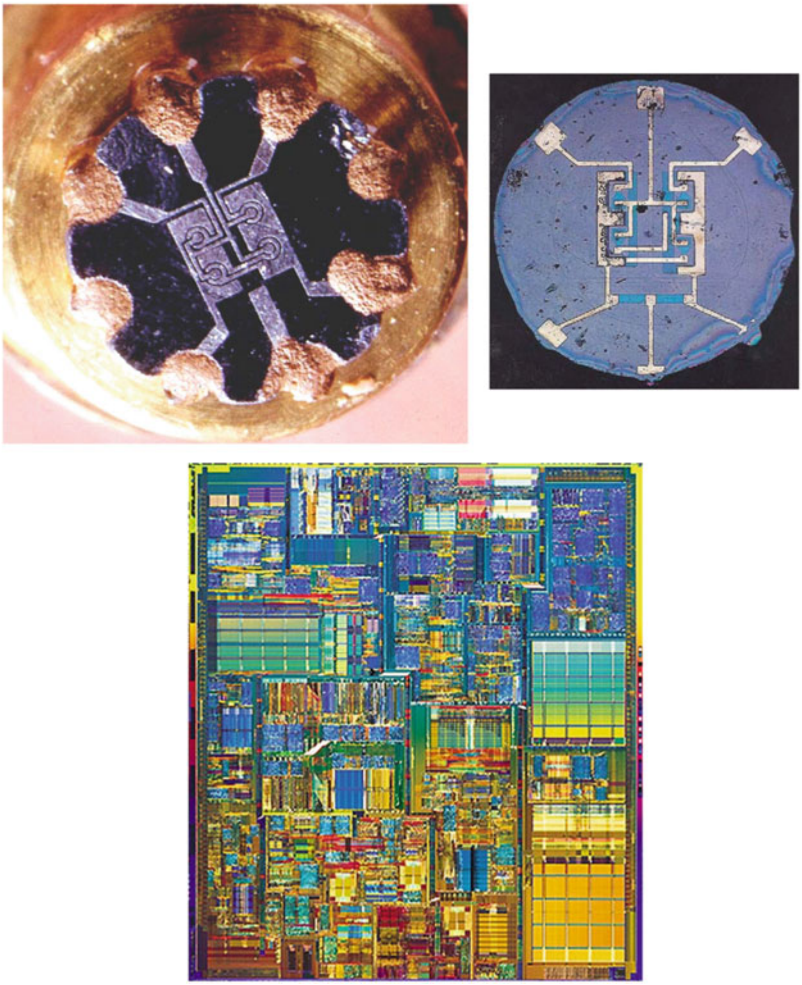


Fig. 2.11 First planar integrated circuit and one of modern variants of IC

it suffices to compare the first integrated circuit with the Pentium 4 processor, released in 2000, which harbors 4.2 × 10^7 transistors on the surface of 224 mm^2.

The final important step toward the creation of the modern semiconductor planar technology was the introduction of batch fabrication of integrated circuits, when a large number of identical circuits are made on the same substrate. This step was in fact quite natural given the large size and high quality of silicon wafers produced at that time (Fig. 2.12). Anticipating a little bit, we note that the size of available wafers grew rapidly—from wafers of ~25 mm in diameter in 1960 to 200 mm and more in the 1990s of the last century. At the same time the wafer area per one circuit went up from ~1 mm^2 in 1960 to ~100 mm^2 in the 1990s, with up to three million elements in each circuit.

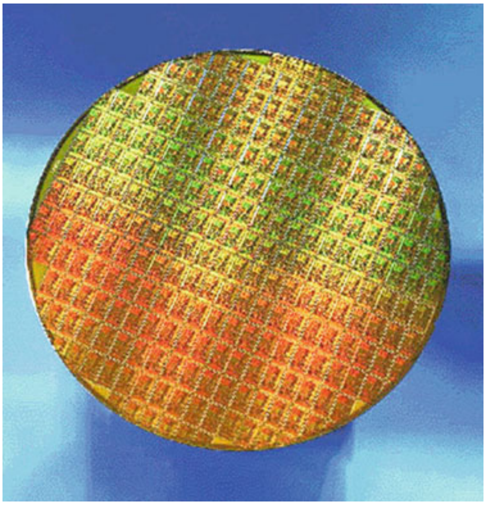


Fig. 2.12 Silicon wafer at the stage of producing integrated circuits

Today the semiconductor planar technology plays a central role in the production of electronic circuits (chips) that are used in a large number of devices—computers, control, and communication devices. It relies on a wide variety of specific processes that differ in their physicochemical nature and the instrumentation used. Typical for semiconductor technology are the extreme requirements imposed on the purity of raw materials, handling medium (water, auxiliary materials), and the atmosphere of the production facilities. In chemical practice, both in research and production, a substance is considered pure if the concentration of impurities does not exceed 0.001 %. The number of atoms in 1 cm^3 of the semiconductor is 10^{22}. Doping a semiconductor usually involves introducing 10^{16}–10^{19} dopant atoms per 1 cm^3, i.e., 0.0001–0.1 %. This means that the concentration of harmful impurities in silicon, which may affect its semiconducting properties, must be below 0.00001 %.

The planar technology is characterized by a number of other important features. However, since the main focus of this book is on the interaction and mutual influence of the computer technology and nanotechnology, we will confine consideration to the most important problem in this context—the lithographic process and the limits it imposes on miniaturization of electronic circuits.

2.3 Planar Semiconductor Technology: Universal Acceptance and Limitations

Planar technology (see Fig. 2.13) involves successive application on the surface of the silicon substrate of thin layers of material which serves to form the individual elements of the scheme, with subsequent processing of this layer.

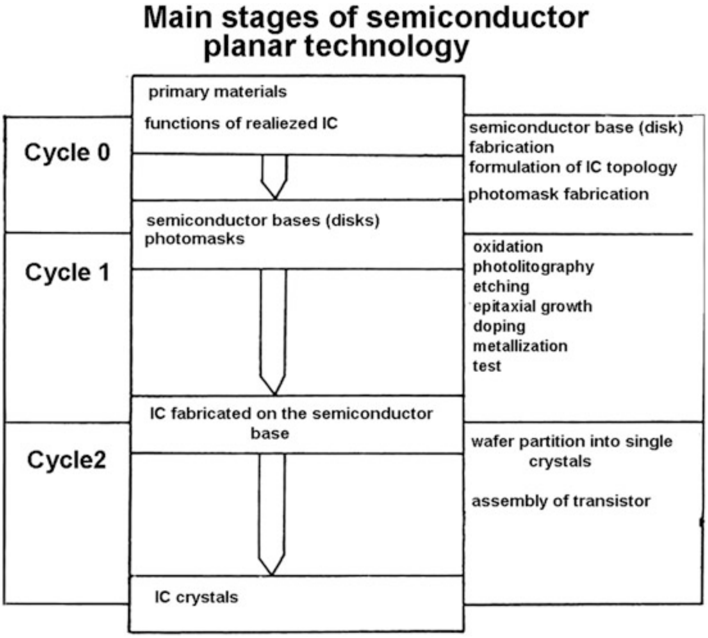

Main stages of semiconductor planar technology
Cycle 0
Cycle 1
Cycle2
primary materials
functions of realiezed IC
semiconductor bases (disks)
photomasks
IC fabricated on the semiconductor base
IC crystals
semiconductor base (disk) fabrication
formulation of IC topology
photomask fabrication
oxidation
photolitography
etching
epitaxial growth
doping
metallization
test
wafer partition into single crystals
assembly of transistor

Fig. 2.13 Main stages of semiconductor planar technology

It is thus an approach utilizing the top-down principle. The essence of this principle, characteristic, for example, for the production of macroscopic parts of mechanical devices, is easy to illustrate using the manufacturing process of complex metal parts as an example. The process starts with the part blank which is successively subjected to turning, machining, boring required holes, cutting thread, etc. As a result of these successive operations, the part blank is turned into the final detail specified in the drawing. The planar technology uses various operations to create a layer of the required material. Insulating layers of silicon oxide are grown by controlled oxidation of the silicon substrate surface. Alternatively spraying, deposition from the liquid phase, etc., can be used to form the films as well. The main tool for creating individual elements of the chip is photolithography (Fig. 2.14). During the photolithographic process, photoresist is applied to a film of the material from which the elements of the integrated circuit are formed. Photoresist is a photosensitive compound which either decomposes upon light exposure (positive photoresist) or polymerizes, forming a solid film (negative photoresist). Photoresist is exposed through a photomask whose black and white pattern determines the shape and the location of details to be formed on the surface of the circuit. The photoresist film is subsequently "developed," that is, treated with solvent, which removes film areas unconverted by light. As a result, the photoresist film is transformed into a stable mask. Through its windows the material can be affected, e.g., by oxidation, doping, etc.

Fig. 2.14 Scheme of the photolithographic process

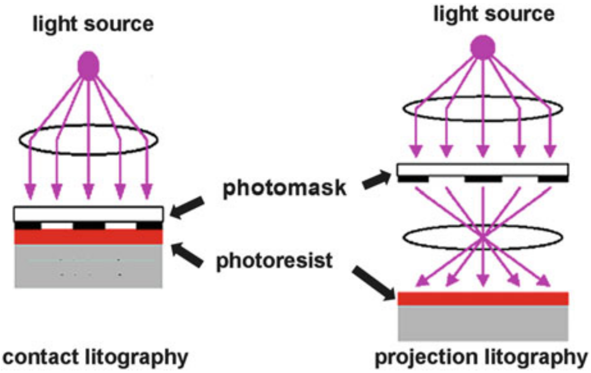
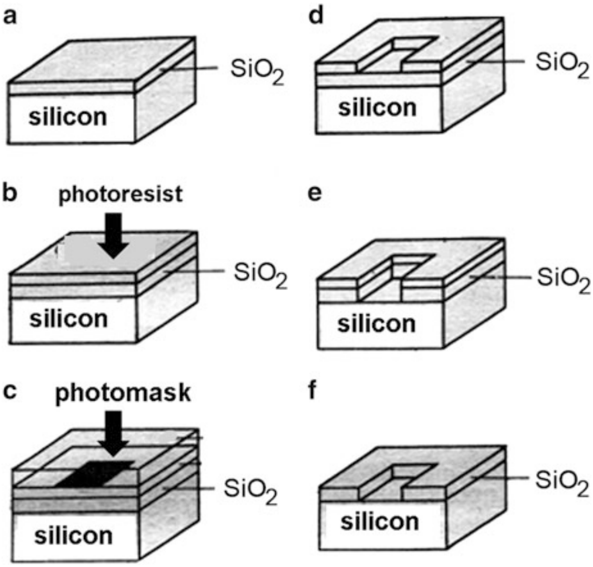


Fig. 2.15 Successive stages of the positive lithographic process (a–f)



As an example, consider the formation in a semiconductor of n-type zones corresponding to the source and drain of a planar transistor (Fig. 2.15). A negative photoresist is applied to the layer of silicon oxide grown on a silicon substrate. Non-transparent areas of the photomask used for illumination correspond to the n-type areas to be formed. As a result of illumination of the photoresist its entire surface polymerizes except for those areas being created, from which unreacted resist is cleaned up with a suitable solvent (benzene, toluene). Then, through the windows in the polymerized film of the resist, silicon oxide is removed with hydrofluoric acid, after which the polymerized resist is removed by a new solvent. In this fashion windows are formed in a layer of silicon oxide on the surface of the integrated circuit that are used for doping silicon by phosphorus or antimony. The process of doping involves diffusion: alloying additive is applied to the surface and

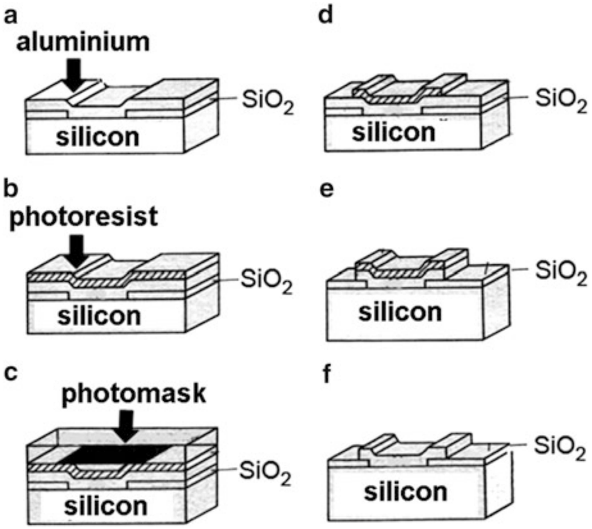


Fig. 2.16 Successive stages of the negative lithographic process (a–f)

then the chip is heated. Alternatively ion implantation by direct action of dopant ions can be employed. Figure 2.16 illustrates the positive lithographic process in which a layer of aluminum deposited on the surface of the chip is removed from all its elements except for the gate area of the field-effect transistor.

Various resists are known which differ in terms of their light sensitivity and resolution. Often a positive photoresist PMMA (polymethylmethacrylate, Fig. 2.17) is used. As a negative resist, COP (a polymer of glycidyl methacrylate and ethyl acrylate) is used.

Photolithographic technology allows for the creation of complex semiconductor circuits, but at the same time it represents the limiting factor of the planar semiconductor technology. Due to its wave nature light is diffracted by the photomask elements (Fig. 2.18). The maximal achievable resolution of the exposed pattern is of the same order as the wavelength of the light used for illumination. As is known, the wavelength of visible light is in the range from 0.38 μm (violet region of the spectrum) to 0.76 μm (red area). This determines the minimum line width in a semiconductor structure which can be obtained by optical lithography.

Today's scientific and engineering projects are aimed at improving the key step in the production of integrated circuits—lithography—which will determine the physical limits of semiconductor technology in the foreseeable future. Experts note that because of these restrictions lithography may exhaust its possibilities already in the beginning of our century.

The development of the lithographic technology since its inception in the early 1970s was directed at reducing the wavelength of the light used. This allowed to reduce the size of the elements of the integrated circuit. Since the mid-1980s ultraviolet laser radiation is used in photolithography. In order to apply the pattern of the circuit to the plate, computer-controlled machines (steppers) are employed. Configuration of "window" is determined by appropriate masks, and the resulting

Fig. 2.17 Positive (**a**) and negative (**b**) photoresists

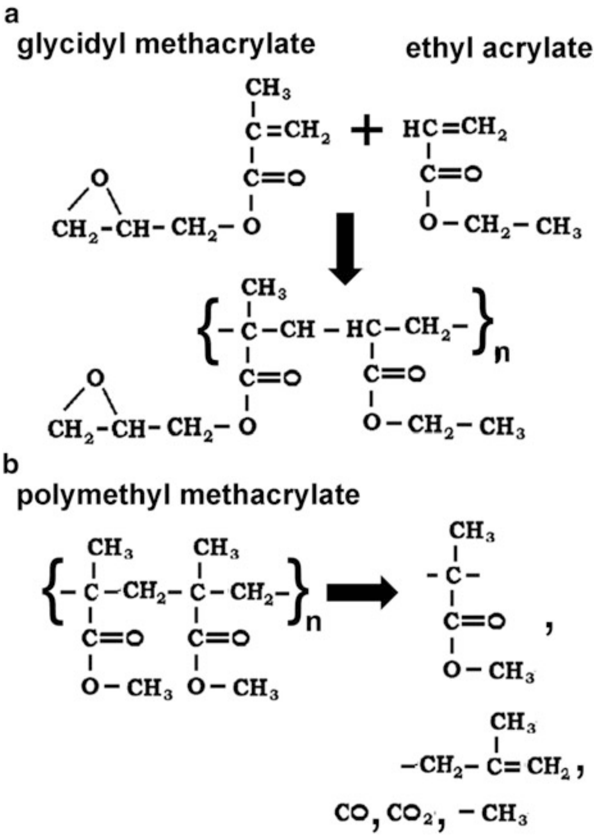


Fig. 2.18 Limitations of photolithography

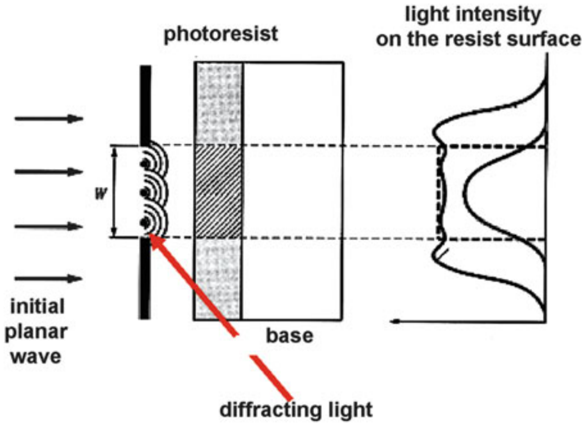


image is focused by a special lens system which reduces the given mask pattern to a microscopic size. A modern photolithography machine handles several tens of 8-in. semiconductor wafers per hour.

Currently, most chips are made using ultraviolet radiation with a wavelength of 0.248 μm. For some circuits, a lithographic technology with the wavelength of 0.193 μm has been developed. However, beyond 0.2 μm, serious problems put under question further progress in photolithography. For example, at a wavelength of less than 0.2 μm too much light is absorbed by the photosensitive layer, complicating and slowing down the process of transferring the pattern of the circuit template. Such problems motivate investigators and manufacturers to seek alternatives to conventional lithographic technology. For example, the possibility of replacing ultraviolet rays by X-rays has been under investigation in US scientific laboratories for more than two decades.

One technology, called EUV (extreme ultraviolet) and supported by several well-known companies, aims to improve the process of lithography in chip manufacturing.

As already noted, modern equipment for printing circuits on silicon substrates based on deep ultraviolet radiation (deep ultraviolet, DUV) uses light sources with a wavelength of 248 nm. It is assumed that the wavelength of EUV radiation can be as short as 13 nm, i.e., approximately 20 times shorter. The transition from the DUV to the EUV lithography provides for more than tenfold decrease of the wavelength, making it comparable to the size of just a few tens of atoms.

Nevertheless, in addition to purely physical problems, there are other factors in the manufacturing process of circuits limiting miniaturization and the degree of integration of transistors. Generally speaking, the properties of the devices created on the same silicon wafer, as well as on different wafers, are not identical. Deviations can occur at each stage of production. The nature of the possible differences between the produced circuits and the frequency of occurrence of completely defective devices may hamper further miniaturization of integrated circuit elements. Note that miniaturization affects not only the length and the width of the circuit but also the thickness of the crystal on which transistors and connections are implemented through a series of levels. In modern chips, there may be four or five such levels. Reducing the size of transistors and increasing their density on the crystal brings about an increase in the number of levels. However, the more layers exist in the circuit, the more thorough control of the production process is required, since each of the levels will be affected by the levels underneath it. The cost of improving control and creating connections between multiple layers may deter the increase in the number of layers.

Among other things, the increasing complexity of integrated circuits necessitates further improvement of production conditions, to which unprecedented requirements are already posed. A more precise mechanical control over the positioning of the original silicon wafer is necessary. Sterile rooms (so-called clean room) in which chips are manufactured should become even cleaner to exclude the penetration of tiny dust particles that can destroy a complex circuit.

Taken together, all this not only demonstrates the need to improve the planar semiconductor technology but also motivates the quest for fundamentally new approaches to building computing devices. One of the most promising ones is the transition from semiconductor to molecular components.

Chapter 3
Molecular Elements of Computers

> *Who knows how many words the God had tried, before He found the one that could create the world.*
>
> Stanisław Jerzy Lec, *Aphorisms*

3.1 Preliminary Remarks

For many decades molecules and molecular complexes, with their discrete energy levels and the ability to switch the molecular system from one state to another, have been considered ideal elements for computing devices. However, the technological possibilities available in the first half of the last century and foreseeable at that time did not allow even to think about the practical use of the molecular primitives. Moreover, beginning from the 1960s the planar semiconductor technology, with its promise of increasing the capabilities of computing device manifold, was advancing at a rapid pace.

Nevertheless, the difficulties in establishing the planar semiconductor technology in the 1970s and 1980s, and at the same time, its successful advent as an entirely new field of technology have revived the idea of building computing devices based on molecular components. Credible estimates based on the principles of the theory of molecular structure, developed in the postwar years and confirmed by experimental studies on a large number of molecules, indicated that, compared to semiconductor electronics, molecular components can provide:

- A higher degree of integration
- Significantly lower switching energies
- Enhanced stability of circuits with respect to radiation, especially for circuits with a high degree of integration

At the same time it became clear that the molecular component base could make possible fundamentally new features, such as:

- Complete identity of the molecular elements with characteristics not subject to scatter due to unavoidable technological errors

© Springer-Verlag Wien 2014

N.G. Rambidi, *Molecular Computing*, DOI 10.1007/978-3-211-99699-7_3

- Noise-free single-electron processes
- Specific molecular mechanisms of signal transmission, which may allow to create logically more complex primitives

The beginning of the 1970s of the last century coincided with the molecular boom in electronics. In 1974, the IBM scientists Ari Aviram and Mark Ratner, who worked at Northwestern University in Illinois, published a work entitled "Molecular Rectifier." For the first time, it was attempted to use a rigorous physical approach for evaluating the possibility to construct a molecular system with unidirectional electron conductivity: a molecule consisting of two molecular fragments with different electron affinities and placed between two electrodes. If electric potential is applied to electrodes in one direction, then the closest molecular fragment captures an electron from the electrode and transfers it to another fragment which, in its turn, passes it to the second electrode. When the electric potential is applied in the opposite direction, no electron transfer is observed. Aviram and Ratner were the first to propose a specific molecule which they believed could be used as one of the basic elements of electronic circuits.

Also in the beginning of 1974 Michael Conrad at the Wayne University in Detroit conceived a molecular information processing system based on the principle of a specialized neural network (see next chapter for details).

But the greatest progress in the development of this new field of research, which people started to call "molecular electronics," was due to Forrest Carter, a scientist at the US Naval Research Laboratory. Over the course of many years, he considered the physical principles of molecular systems that could be used as primitives for information processing devices. Moreover, it was Forrest Carter who tried to unite the scientists who were seeking ways to use molecular objects for creating electronic devices. In 1982 and 1987 he organized two international conferences on molecular electronic devices. Unfortunately, the third conference he had organized was held in July 1988 after his death.

As a result, activity and expectations in this new field of research were high in the 1980s and early 1990s. For example, supporting the widespread belief in the importance of molecular primitives for the further development of information processing, Alan Berman, director of research at the US Naval Research Laboratory, declared at the Workshop on Molecular Electronic Devices in 1981 [25]:

"I think the possible advantages of a computer based on molecular level of electronics are fairly obvious. When one evaluates whether one should invest one's time, assets or career in this field, one must consider the following points:

In going from a modern day two-dimensional (2-D) computer to three-dimensional (3-D) molecular configuration, wiring costs must be significantly decreased and fabrication more fully automated, or the device will not be economically feasible.

By reducing the switching elements to molecular size the memory density could be increased by several orders of magnitude and power input decreased very significantly.

Three-dimensional construction plus switching elements of molecular size could enhance computer speed by several orders of magnitude. To use this speed, faster

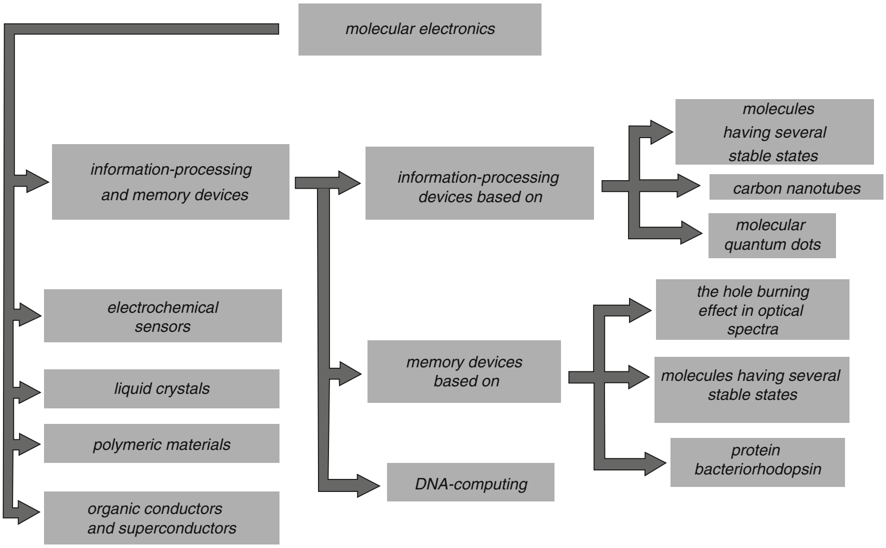

molecular electronics
information-processing and memory devices
electrochemical sensors
liquid crystals
polymeric materials
organic conductors and superconductors
information-processing devices based on
memory devices based on
DNA-computing
molecules having several stable states
carbon nanotubes
molecular quantum dots
the hole burning effect in optical spectra
molecules having several stable states
protein bacteriorhodopsin

Fig. 3.1 Main directions of molecular electronics

data pumps will need to be developed. These possible advantages are all very significant. However, the technical barriers which now appear to limit the utility of such devices must be overcome. The decade of the 1980s should be the time where chemists, physicist, and engineers will begin to become sensitive to the possibility of useful switching devices at the molecular level."

Interest in the practical use of molecular systems affected not only the field of computer technology. Over the last decades of the past century, a new area of research called "molecular electronics" emerged. Figure 3.1 gives an overview of its main directions, and there are review articles and books dedicated to it. Therefore in the following only utilization of molecular systems in the computing and information-logic devices will be considered in detail.

In the end of the past century a number of theoretical studies analyzed the general principles and suggested ways to design molecular systems suitable for designing electronic circuits. Several experimental studies confirmed the feasibility of creating molecular primitives. Let us consider in more depth some of the major publications of this period, starting with a brief summary of the main theoretical concepts about the structure of atoms and molecules.

3.1.1 Some Details: Molecular Structure—Electronic Levels of Atoms and Molecules

The theoretical basis for understanding the structure of molecules and calculation of their characteristics is constituted by the Schrödinger equation:

$$\left\{-\frac{\hbar^2}{2}\sum_\alpha \frac{1}{M_\alpha}\left[\frac{\partial^2}{\partial X_\alpha^2}+\frac{\partial^2}{\partial Y_\alpha^2}+\frac{\partial^2}{\partial Z_\alpha^2}\right]-\frac{\hbar^2}{2m}\sum_i\left[\frac{\partial^2}{\partial x_i^2}+\frac{\partial^2}{\partial y_i^2}+\frac{\partial^2}{\partial z_i^2}\right]\right.$$

$$-\sum_{i\alpha}\frac{Z_\alpha e^2}{r_{i\alpha}}+\sum_{\alpha\beta}\frac{z_\alpha Z_\beta e^2}{r_{\alpha\beta}}+\sum_{ij}\frac{e^2}{r_{ij}}\}\Psi_n=E_n\Psi_n.$$

Here X_α, Y_α, Z_α and x_i, y_i, z_i denote spatial coordinates of the nuclei (α) and of the electrons i) of the molecule.

M_α, m—masses of the atomic nuclei and electrons.

$r_{\alpha i}$, $r_{\alpha\beta}$, r_{ij}—distances between the nucleus and the electrons, between two nuclei, and between two electrons, respectively.

$\hbar$—Planck's constant.

The physical meaning of this equation is simple. The first and the second terms on the right side correspond to the kinetic energy of the nuclei and electrons of the molecular system. Subsequent terms describe electrostatic interactions between nuclei and electrons as well as interactions of nuclei and electrons with each other. A quantum system is characterized by discrete energy levels E_n and wave functions Ψ_n or each energy state. The square modulus of this function reflects the probability of finding nuclei and electrons in the corresponding point in space.

Let us consider molecules with discrete energy levels E_n, with their electronic states described by the wave functions Ψ_n. The square modulus of the wave function determines the probability of finding electrons in space.

The Schrödinger equation was derived under some very serious assumptions. First of all, it does not take into account the electron spin, which has to be introduced based on physical considerations.

The Schrödinger equation is a partial differential equation for which exact solutions are in general unknown. Nevertheless, experience shows that approximate ab initio numerical methods (i.e., methods not making use of additional experimental information) developed so far allow for calculating molecular characteristics with an error not exceeding the experimental error.

For some fairly simple systems the Schrödinger equation does have an exact solution. One of them is the problem of the hydrogen atom, which plays a fundamental role in describing atomic and molecular structure. The Schrödinger equation for the hydrogen atom (one nucleus and one electron) is written in the laboratory coordinate system (see Fig. 3.2) as

$$\left\{-\frac{\hbar^2}{2M}\left[\frac{\partial^2}{\partial X_l^2}+\frac{\partial^2}{\partial Y_l^2}+\frac{\partial^2}{\partial Z_l^2}\right]-\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial x_l^2}+\frac{\partial^2}{\partial y_l^2}+\frac{\partial^2}{\partial z_l^2}\right]-\frac{Ze^2}{r}\right\}\Psi_{\mathrm{ln}}=E_{\mathrm{ln}}\Psi_{\mathrm{ln}},$$

where the index l indicates that these distances are measured from the origin of coordinates (see Fig. 3.2) and the distance r from the nucleus of an atom to the electron.

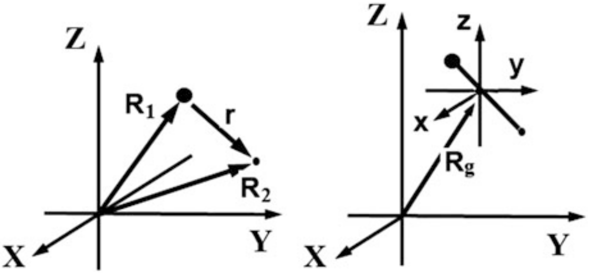


Fig. 3.2 Coordinate system for the quantum-chemical calculation of the hydrogen atom

Upon moving the origin of coordinates to the center of mass, this equation splits into two ordinary differential equations. The first one describes the motion of a particle with the mass $M+m$, while the second one describes electronic characteristics of the hydrogen atom:

$$\left\{-\frac{\hbar^2}{2\mu}\left[\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right]-\frac{Ze^2}{r}\right\}\Psi_n=E_n\Psi_n.$$

Here the coordinates x, y, z are measured from the center of mass, and μ is the reduced mass of the hydrogen atom:

$$\frac{1}{\mu}=\frac{1}{M}+\frac{1}{m}.$$

A remarkable property of the hydrogen atom is its high spherical symmetry. By going over to spherical coordinates with the origin at the nucleus of the hydrogen atom, an exact solution of the Schrödinger equation can be obtained. It is defined by three quantum numbers n, l, m, with electronic energy levels depending only on the principal quantum number n:

$$E_n=\frac{A}{n^2},\quad A=-\frac{Z^2}{2\hbar}me^4,\qquad\qquad \Psi_{nlm}=\text{const}\,R_n(r)Y_l^m(\vartheta,\varphi).$$

Thus, the energy levels of the hydrogen atom are degenerate. Their degree of degeneracy is determined by the quantum numbers l, m. In accordance with the solution of the Schrödinger equation, restrictions are imposed on the quantum numbers:

$l \leq n-1$, an integer which can take values 0, 1, 2, … $n-1$, if $l=0$, the electronic state is called s-state, with $l=1$—p-state, with $l=2$—d-state, etc.

A quantum number "m" may take integer values from $-l$ to $+l$. It determines the degree of degeneracy of each l-sublevel, with $l=0$ corresponding to no degeneracy, $l=1$ to threefold degeneracy, and $l=2$ to fivefold degeneracy.

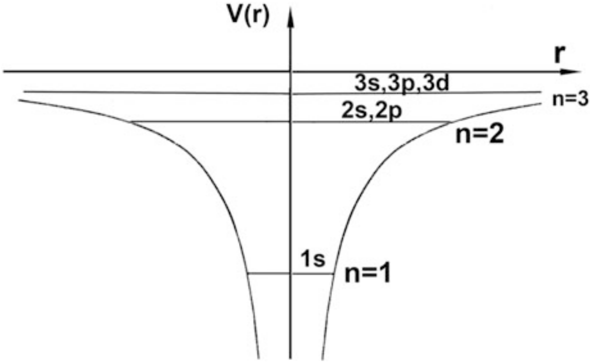


Fig. 3.3 Energy levels of the hydrogen atom

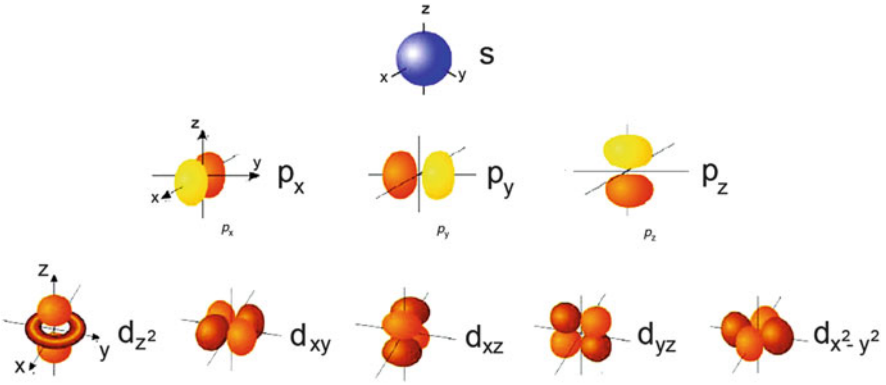


Fig. 3.4 Angular parts of the wave functions of the hydrogen atom

Electronic energy levels of the hydrogen atom (the energy levels of the electron in the Coulomb potential) are shown in Fig. 3.3. The form of the basic angular functions for the "*s*" and "*p*" states of the hydrogen atom is shown in Fig. 3.4.

The hydrogen atom problem forms the basis for describing molecular systems. Schrödinger equations for multi-electron atoms and molecules (even for the simplest two-electron helium atom) do not have an exact solution. In this case, strictly speaking, the symmetry of the atoms with more than one electron does not satisfy the spherical symmetry group, since the Schrödinger equation contains a term corresponding to the interaction of electrons. Nevertheless, it is assumed that at least for the first half of the atoms of the periodic system, the spherical symmetry is approximately preserved. Therefore, as in the case of the hydrogen atom, each electron is assigned four quantum numbers *n*, *l*, *m*, and a *s*-spin quantum number, which takes only two values ±½. The overall structure of energy levels is also preserved. In addition, the state filling principle is introduced, according to which in the ground electronic state, the electrons occupy successive levels, starting from the

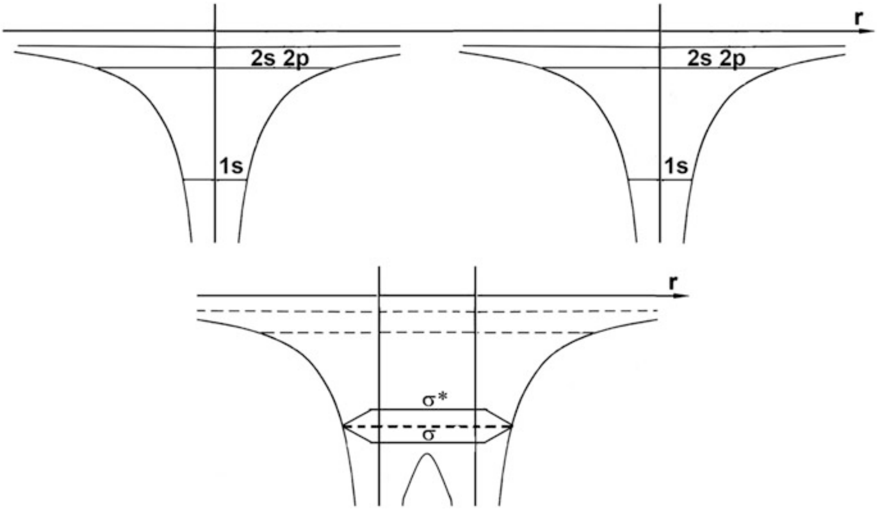


Fig. 3.5 Scheme of the formation of the hydrogen atom

lowest one, with two electrons with opposite spins at each level. In the case of a degenerate state, a single electron first occupies each component, which is subsequently filled by electrons with opposite spin.

This simple model of the electronic structure of atoms can also be used for qualitative description of more complex systems. Consider two identical hydrogen atoms that are far enough from each other, such that any interaction between them can be neglected (Fig. 3.5). Suppose that the atoms move toward each other. At small distances between the nuclei of these atoms, their Coulomb potentials will start to overlap, forming a single system. Simultaneously electron levels will be changing as well. Let us estimate qualitatively the nature of these changes.

We introduce the notation for the hydrogen molecule (two nuclei at a distance R_{ab} from each other), as shown in Fig. 3.6.

In this notation, the Schrödinger equation can be written as

$$H(1,2)\Psi(1,2) = E\Psi(1,2), \quad H(1,2) = H(1) + H(2) + \frac{e^2}{r_{12}},$$

$$H(i) = -\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2}\right] - \frac{Ze^2}{r_{ai}} - \frac{Ze^2}{r_{bi}}, \quad i = 1,2.$$

It consists of three parts. $H(1)$ and $H(2)$ depend only on the coordinates of the first and the second electrons, respectively, while the interaction between the electrons is represented by a separate term. This interaction makes a significant contribution to the calculated energy of the hydrogen molecule and cannot therefore be neglected. At the same time, since the exact solution of the Schrödinger equation is not known, necessitating the use of approximate methods, an analytical form of

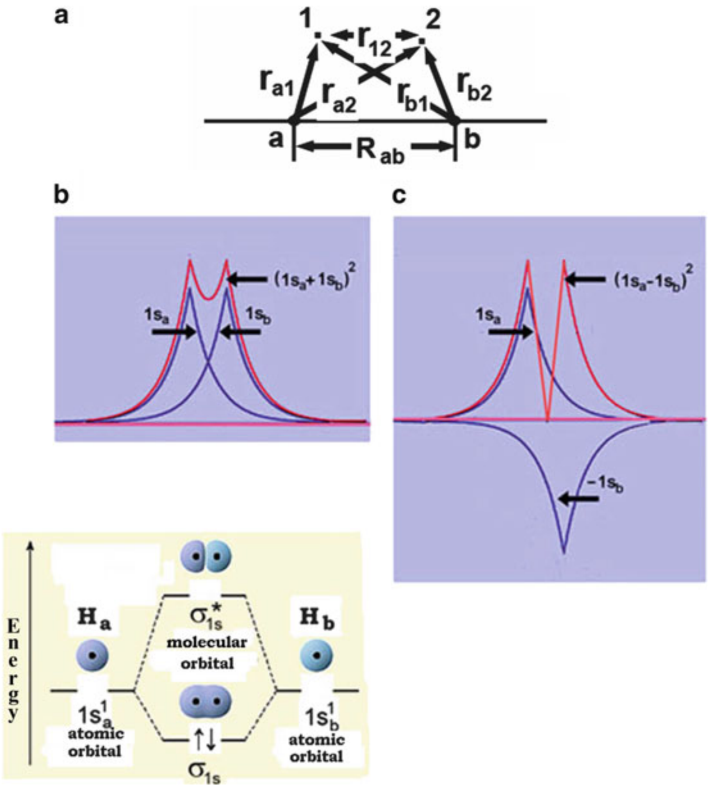


Fig. 3.6 Bonding and antibonding orbitals of the hydrogen molecule (**a**), notations of the distances between nuclei and electrons (**b**), and scheme of molecular orbitals (**c**) in the H_2 molecule

the wave function must be chosen. Practice shows that this can be done based on a simplified Schrödinger equation, where the interaction between electrons is neglected:

$$H(1,2) \rightarrow H^{(0)}(1,2) = H(1) + H(2),$$
$$H^{(0)}(1,2)\Psi^{(0)}(1,2) = E^{(0)}\Psi^{(0)}(1,2).$$

It can be shown that this simplified equation can be partitioned into two ordinary differential equations:

$$H(1)\varphi_1(1) = \varepsilon_1\varphi_1(1),$$
$$H(2)\varphi_2(2) = \varepsilon_2\varphi_2(2),$$
$$\Psi^{(0)}(1,2) = \varphi_1(1)\varphi_2(2), \quad E = \varepsilon_1 + \varepsilon_2.$$

Consider solutions $H(1)$ when the distance between the first electron and the nucleus "a" is small and therefore $r_{a1} \ll r_{b1}$. In this case in $H(1)$ the term Ze^2/r_{b1} can be neglected. Then $H(1)$ coincides with the corresponding expression for the hydrogen atom with index a, and the wave function $\varphi_1(1)$ should coincide with the wave function of the hydrogen atom at the point "a." Similarly, if $r_{b1} \ll Re$, the wave function $\varphi_1(1)$ should coincide with the wave function of the hydrogen atom at the point "b":

$$r_{a1} \ll R_e, \quad H(1) \to -\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial y_1^2} + \frac{\partial^2}{\partial z_1^2}\right] - \frac{Ze^2}{r_{a1}}, \varphi_1(1) \to \varphi_a(1); r_{b1}$$

$$\ll R_e, \quad H(1) \to -\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial y_1^2} + \frac{\partial^2}{\partial z_1^2}\right] - \frac{Ze^2}{r_{b1}}, \varphi_1(1) \to \varphi_b(1).$$

Thus one can choose the wave function $\varphi_1(1)$ as $\varphi_1(1) = \mathrm{const}_{(1)}\varphi_a(1) + \mathrm{const}_{(2)}\varphi_b(1)$.

Note that the hydrogen molecule possesses a rather high degree of symmetry, with the rotation axis of infinite order passing through the two hydrogen atoms. Furthermore, among the symmetry elements there exists a plane that passes through the middle of the distance between the nuclei of atoms and perpendicular to it. Therefore, the wave function $\varphi_1(1)$ can be written as

$$\begin{aligned} \varphi_1^{(s)}(1) &= N_s[\varphi_a(1) + \varphi_b(1)], \\ \varphi_1^{(as)}(1) &= N_{as}[\varphi_a(1) - \varphi_b(1)]. \end{aligned}$$

Upon reflection in the symmetry plane the atoms exchange positions (Fig. 3.6). However the square modulus of the wave function (i.e., the electron density distribution) remains unchanged. As shown in the figure the function $\varphi_1^s(1)$ must correspond to the bond between the two hydrogen atoms, since the probability density of finding electrons in the space between the nuclei increases. According to this criterion the function $\varphi_1^{as}(1)$ does not contribute to that bond. Therefore it is conventional to call $\varphi_1^s(1)$ and $\varphi_1^{as}(1)$ bonding and antibonding orbitals. Numerical solution confirms these qualitative considerations. The function $\varphi_1^s(1)$ leads to the potential energy of the nuclei of the molecule with a minimum corresponding to the bonding of two hydrogen atoms (figure), whereas the function $\varphi_1^{as}(1)$ leads to their repulsion.

The simple model we have considered, which qualitatively describes the order and relative positions of electronic levels in the potential well, is often applied to more complex molecular fragments. It is especially useful when analyzing electron transitions in a system consisting of fragments which can be regarded as independent.

3.2 Pioneering Ideas of Molecular Electronics

In 1974 Ari Aviram and Mark Ratner considered a hypothetical molecule, which they called “molecular rectifier” (Fig. 3.7). It contains two fragments, one being the electron acceptor (tetracyanoquinodimethane, TCNQ) and the other the electronic donor (tetrathiafulvalene, TTF). Electron acceptor is the molecule possessing an electronic structure with the lowest molecular orbital unoccupied. BF_3 can serve as the simplest example of such a molecule. The boron atom has the following electronic structure: $1s^2 2s^2 2p_x^1 2p_y 2p_z$. Upon formation of the BF_3 molecule this electron distribution is changed to $1s^2 2s^1 2p_x^1 2p_y^1 2p_z$. Three electrons form three bonds with fluorine atoms, with one orbital remaining free. In the donor molecule electrons are situated at the highest occupied molecular orbital. The electronic structure of the nitrogen atom forming the NH_3 molecule is $1s^2 2s^2 2p_x^1 2p_y^1 2p_z^1$. Three electrons form three nitrogen–hydrogen bonds, but two more electrons are still available in the outer electron shell. Such electronic structure of NH_3 and BF_3 leads to the formation of a stable molecular complex BF_3NH_3 due to the so-called

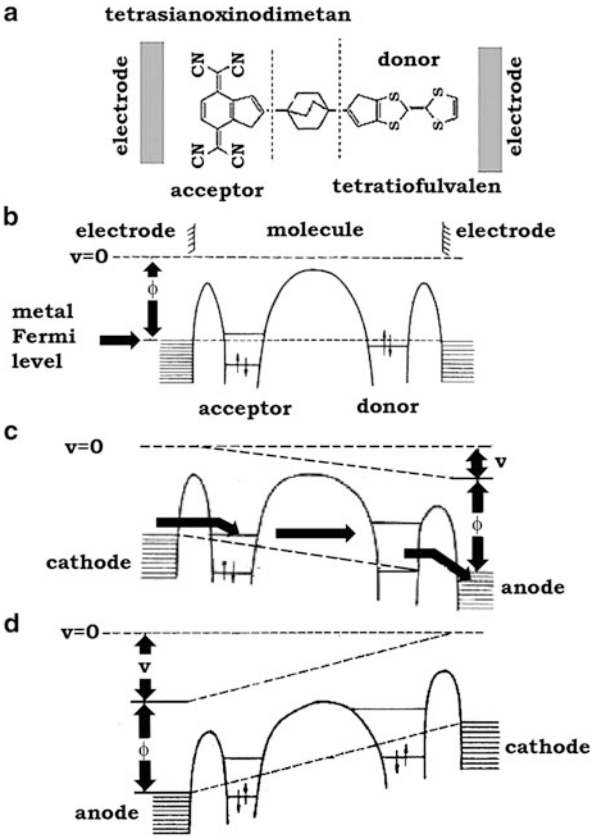


Fig. 3.7 Aviram and Ratner molecular rectifier: chemical structure (**a**), scheme of the electronic levels (**b**), shift of the electronic level under the application of the direct (**c**) and reverse (**d**) voltage

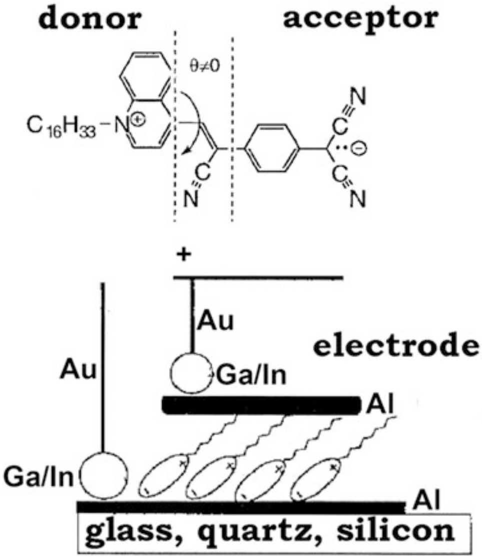


Fig. 3.8 Experimental model of a molecular rectifier

coordination (donor–acceptor) bond. In this complex a pair of electrons of the nitrogen atom essentially fills the free orbital of the boron atom.

In the Aviram and Ratner model acceptor and donor groups are separated by a hydrocarbon fragment that does not conduct electrons, and the whole system is placed between two electrodes.

Let us investigate what causes unidirectional flow of electrons in the Aviram and Ratner model. Initially it is assumed that in the absence of the electric field, the highest filled level of the acceptor is located above the Fermi level of the nearest electrode, while the lowest filled level of the donor below the Fermi level of another electrode. Suppose that voltage is applied to the electrodes in the direction acceptor–donor. If the voltage at the cathode is sufficiently large, the upper level of the acceptor is lowered, facilitating electron capture by the acceptor from the cathode. Likewise, a possibility arises for the donor to transfer the electron to the anode. If the structure of the barrier between the acceptor and donor is such that tunneling through the barrier between them is possible, the electrons will move from the cathode to the anode. The situation changes radically if the voltage between the electrodes is applied in the reverse direction. The arrangement of levels then becomes such that the probability of electron transition through the molecule becomes much smaller than in the previous case.

The work of Aviram and Ratner naturally stimulated experimentalist to test the proposed model. In 1997, 23 years after the work of Aviram and Ratner was published, an American physicist Metzger with his team discovered the effect of electrical current rectification by a molecular film consisting of chemical compounds of the donor–acceptor type. The structure of the molecule and the experimental scheme are shown in Fig. 3.8. The peculiarity of this work is that the object of the study was not the conductivity of a single molecule, but rather of their

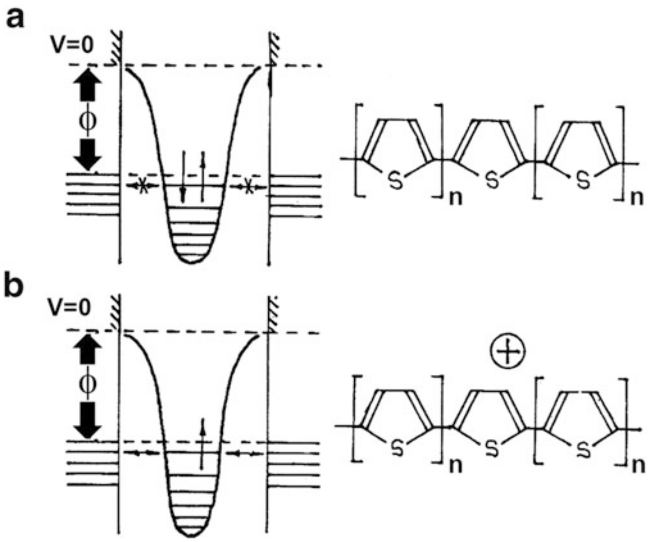
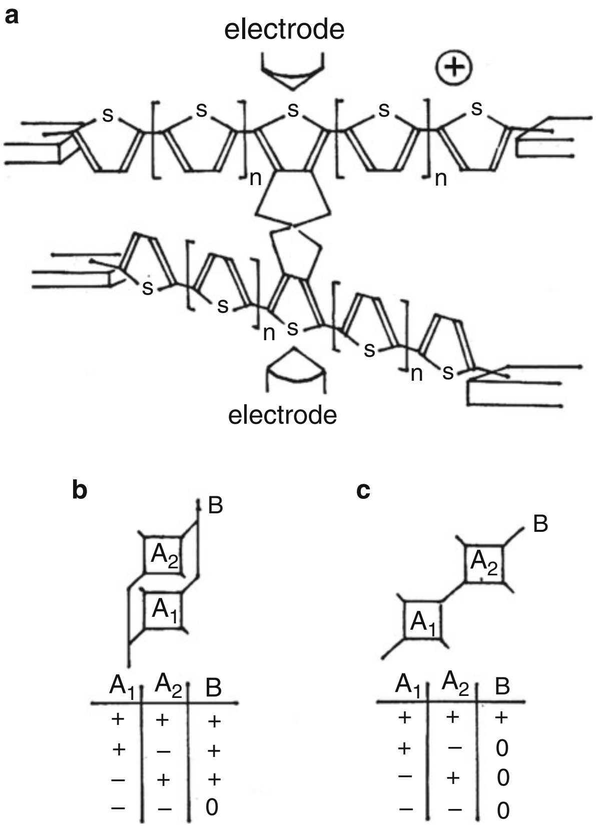


Fig. 3.9 Structure of the electronic levels and the chemical formula of the Aviram molecular switch in the nonconducting (**a**) and conducting (**b**) form

ensemble—an ordered molecular film. This is no accident. In essence, rather than describing the passage of electric current through a molecule, the model of Aviram and Ratner describes single-electron transfer between the fragments of the molecule. This immediately casts doubt on the suitability of systems of this type for the creation of molecular switching elements. Because the molecule of Aviram and Ratner is a quantum object, there is only a certain probability that the molecule will switch to the conducting state under the influence of the voltage applied to the electrodes. In contrast, a switching element should always respond to the applied stimulus.

Therefore, in 1988 Aviram proposed another version of the molecular complex designed, as he defined it, for "memory, logic, and amplification." The structure of this complex was based on extensive molecular fragments (at least ~50 Å long). It was known that the electronic structure of these fragments is similar to the band structure and that they can be in two states—conducting and nonconducting, moving from one state to another during reductive–oxidative processes. The highest filled molecular orbital of the nonconducting form is situated below the Fermi level of the metal (Fig. 3.9). Therefore, if such a molecule is placed between two electrodes, the applied voltage does not cause the passage of electrons through the molecular fragment. However, the current will pass if one electron is removed from the top orbital of the fragment, turning it into a positively charged ion. The molecular switch proposed by Aviram is shown in Fig. 3.10. In this complex conductive and nonconductive parts are perpendicular to each other and are connected by a nonconducting hydrocarbon fragment.



A_1	A_2	B
+	+	+
+	−	+
−	+	+
−	−	0

A_1	A_2	B
+	+	+
+	−	0
−	+	0
−	−	0

Fig. 3.10 Structure of the Aviram molecular switch (**a**) and logic elements built on the basis of this device (**b** and **c**)

The efficiency of such an element is defined by three fundamental properties.

The fragments of this element must be long enough to ensure their electronic band structure. In other words, they should be similar to conductive polymers.

The structure of the molecule as a whole and of the nonconducting bridge between the two fragments (Figs. 3.9 and 3.10) must ensure the possibility of electron tunneling at a fairly low potential, which converts the conducting form to the nonconducting one and vice versa.

The switching of the fragments from one state to another must occur under the influence of the potential applied to the microelectrodes located on the central axis of the molecular complex.

Thus, the molecular component proposed by Aviram is a switching device with two stable states, controlled by electric field. Aviram considered a number of logic circuits which can be constructed on the basis of the proposed molecular element. Two of them are shown in Fig. 3.10.

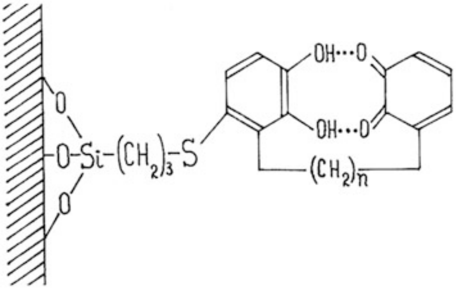


Fig. 3.11 Experimental model of a molecular switch

In order to test his ideas experimentally Aviram made two attempts to create molecular switches. In particular, Aviram, Seiden, and Ratner used chemical reactions to bind semiquinone molecules to the surface of metallic aluminum. In semiquinones hydrogen atoms can be in one of two positions and can move from one position to another (Fig. 3.11). It was shown that they are situated perpendicular to the surface and that their absorption spectra differ at different positions of hydrogen atoms relative to the skeleton of the molecule (Fig. 3.11). This allowed for integral (i.e., for the entire set of molecules on the surface) observation of the transitions of these atoms under the influence of an electric field. Later Aviram, Joachim, and Pomerantz synthesized a similar system and demonstrated the feasibility of controlling molecular states using a scanning tunneling microscope.

The development of molecular electronics in the 1980s was strongly influenced by the theoretical studies of Forrest Carter who treated in detail various molecular mechanisms that could be used for creating molecular electronic devices.

Forrest Carter's two main areas of research were:

- The use of the tunnel mechanism of electron conductivity in a system with consecutive potential barriers and the control of this mechanism by shifting the levels in one of the potential wells between the two barriers (the concept of control groups)
- The use of the soliton mechanism (see below) of signal transmission to change the electronic structure of molecular systems and thereby to transfer molecules from one stable state to another (the concept of switching molecules)

Electronic conductivity of extended molecular systems, i.e., the process of electron transfer from an electron-donating to an electron-accepting group along the chain of atoms, attracted the attention of many researchers in the postwar years. Of particular interest are conjugated (polyene, polyacene, etc.) systems in which electrons form extended molecular orbitals upon overlap. These molecules represent quasi one-dimensional systems whose properties are substantially different from the conventional three-dimensional systems.

Without going into details of the theory of electronic conductivity of such molecules, let us only consider the mechanism of electron passage through a

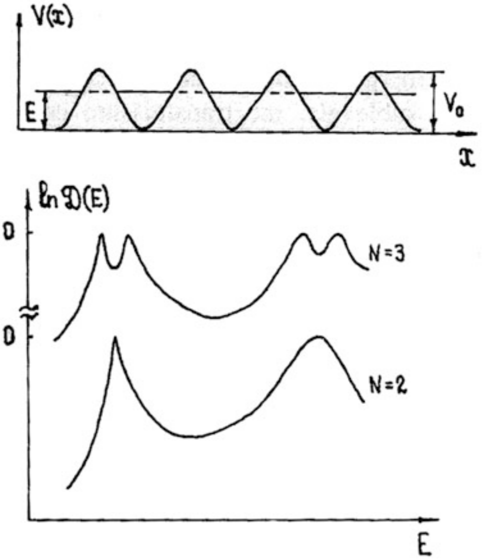


Fig. 3.12 Tunneling of electrons in the system of potential barriers

sequence of N-dimensional barriers $V(x)$ of arbitrary shape, separated $(N - 1)$ potential well, with $x \rightarrow \pm\infty$, $V(x) \rightarrow 0$, as considered by Carter.

The qualitative properties of this process were calculated analytically for the barriers of the type δ functions. It turned out that they did not depend on the particular kind of potential energy. In 1962, the Soviet physicist Pshenichnov showed that in quasi classical approximation for barriers of arbitrary shape under certain conditions the passage may be abnormally large (Fig. 3.12).

Pshenichnov calculated for a system of identical barriers the transmission coefficient for a particle with the mass μ and energy with energy E lower than the height of the barrier V_0.

The study showed that when $0 < E < V_0$ stationary states arise in each of the minima whose spectrum is, roughly speaking, the spectrum of one of the individual minima, with all levels split into $(N - 1)$ sublevels.

Moreover, if the energy of the incident particle equals the energy E of one of the sublevels, the transmission coefficient shows a non-monotonic dependency on E and has $(N - 1)$ maxima in which the values of the transmission coefficient are equal to unity. Thus, the possibility of free electron passage through a succession of barriers arises. At the same time, a change in the form of at least one of the barriers causes the change in the position of the levels (relative to the energy of the passing particle), and the transmission coefficient drops sharply. Therefore, in such a system both the mechanism of signal transmission over long distances and effective interruption of the signal are possible.

It should be noted that Pshenichnov's calculations were not confirmed experimentally in a particular molecular system. However, in 1973 Chu and Esaki performed similar calculations for multilayer semiconductor heterostructures using one-dimensional approximations. Their results fully reproduced the findings

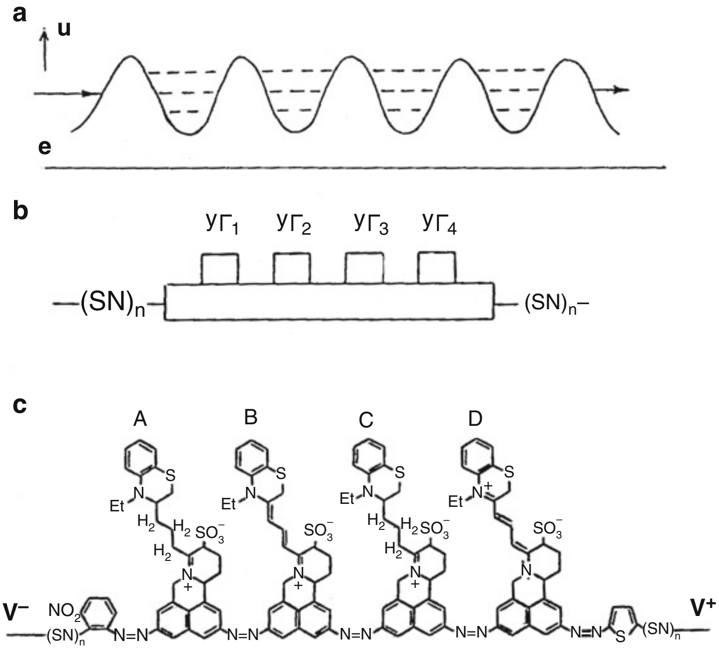


Fig. 3.13 NOT–AND molecular element proposed by Carter

of Pshenichnov. Thus, the dependence of the transmission coefficient on the energy of the tunneling electron demonstrated a sharp increase in transmission when the energy of the electron was equal to one of the levels in the system of rectangular barriers, as well as the splitting of this peak. Later, the work of Chu and Esaki was convincingly experimentally confirmed, in particular for gallium–arsenic heterostructures.

We also note that in practice the situation is complicated by the application of an external electric field. This leads to a change in the form of barriers, they become unequal and the coefficient of the electron transfer is substantially smaller than unity. Therefore, in order to use the tunneling, the potential has to be chosen such that all coefficients of electron transfer through barriers are equal upon the application of external electric field.

The capacity to control molecular devices based on tunneling mechanisms can be illustrated by the NOT–AND and NOR molecular elements proposed by Carter. It is important that even one of these elements is sufficient to construct any logic circuit.

Figure 3.13 shows a hypothetical molecular structure of a NOT–AND element in which positively charged aromatic heterocyclic groups represent potential wells for the electron, and the diazo groups connecting them—potential barriers. If the electron energy coincides with the energy level of the well, the transmission coefficient becomes equal to unity, and the electron passes freely along the chain of potential wells, as shown in Fig. 3.13. Electron passage can be interrupted by either changing the height of the potential barrier or the depth of the potential well, which alter the level position. In Fig. 3.13 the depth of the potential well is determined by the charge of the nitrogen atom N^+. Out of four control groups in two—the first and the third, the charge on N^+ does not change. At the same time for the second and the fourth control groups, a restructuring accompanied by a change of the charge on the nitrogen atom is possible which may be caused, for example, by the absorption of light by one of the molecular groups at the entry point, as shown in the figure (compare entry points *B* and *D*). Since neutralization of the aromatic group leads to substantial change in the shape and location of the well as well as in the position of energy levels within it, free passage of the electron will be impossible and the transmission coefficient of the system will decrease sharply. Using molecular fragments of different structures as control groups, one can ensure that structure rearrangement of the device will occur at different photon energies at different entry points. Such molecular element operates using optical information input. There may be other ways of influencing control groups, such that an electron enters the input, or a structure rearrangement caused by the proliferation of charged solitons takes place.

The molecular NOR element proposed by Carter, and acting on the same principles, is shown in Fig. 3.14. It is a stack of molecules which are gallium

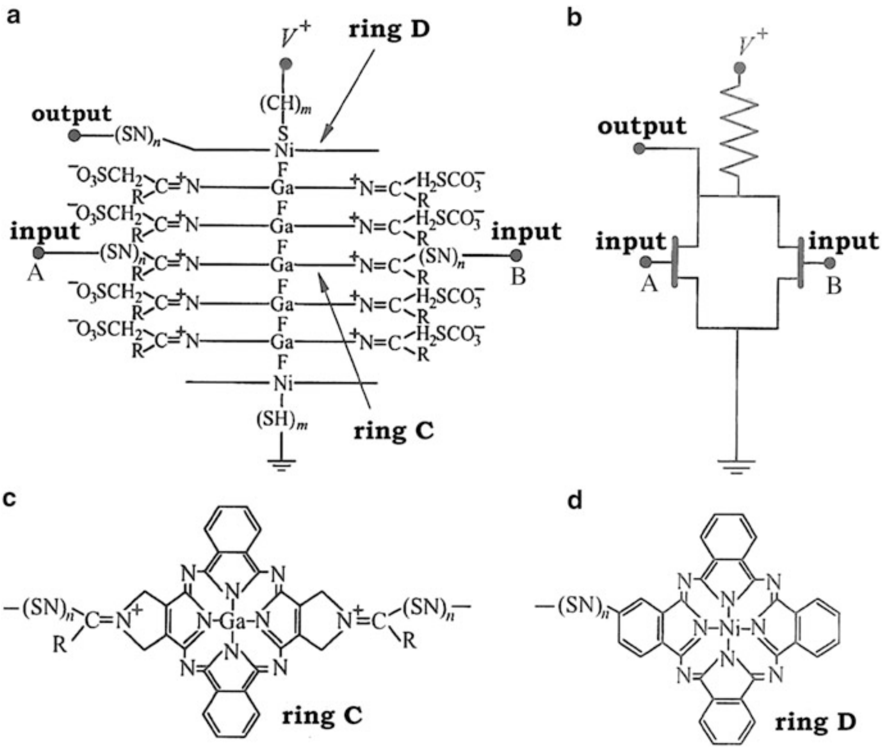


Fig. 3.14 NOR molecular element proposed by Carter

phthalocyanine derivatives (ring C). Potential barriers between the wells are due to intermediate fluorine atoms. As input signal the chain $(SN)_n$ is utilized along which the electron comes to neutralize the charge of the $C{=}N^+$ group. This leads to a change in the shape of the potential well and cessation of electron tunneling along the stack of molecules. Rings of type *D* at the element edges ensure signal exit through the $(SN)_n$ chain and connection to the voltage difference across the Ni–S link.

The basis of these two examples is the concept of control groups, i.e., molecular fragments with the electronic structure which is rearranged under the influence of some factor, thereby interrupting electron motion in the main chain. Other types of control groups also exist.

Thus, in addition to the mechanism of charge neutralization caused by electron transfer along the molecular chain at the nitrogen atom, as discussed above, Carter proposed to use the electronic structure of control groups through:

- Tautomeric rearrangement driven by the change in the external electric field (Fig. 3.15a)
- Intramolecular charge photo-transfer (Fig. 3.15b)

Figure 3.15d shows an example of a more complex molecular control group, which can be manipulated both by an external electric field and light radiation.

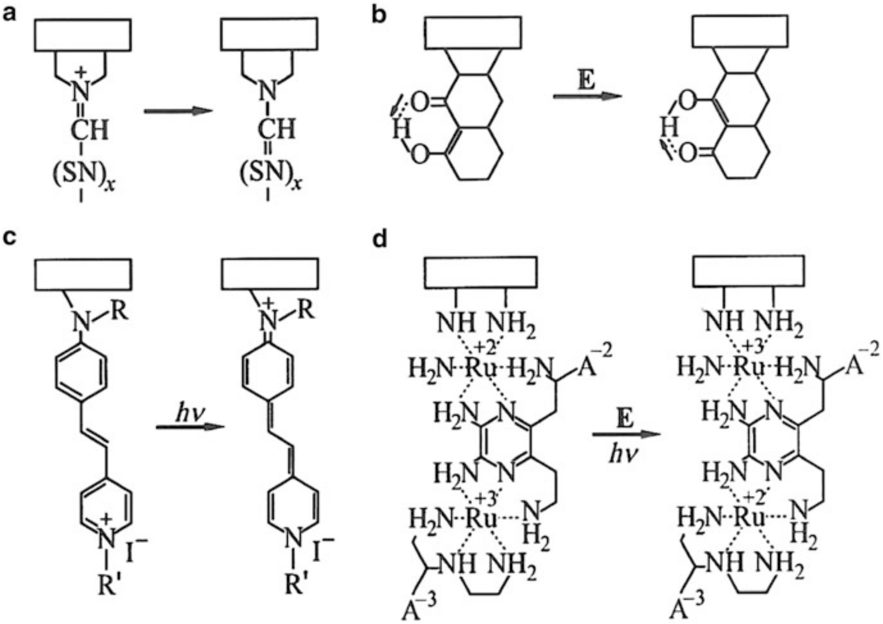


Fig. 3.15 Molecular control groups proposed by Carter

Among the processes of excitation energy transfer in quasi one-dimensional molecular chains, collective processes, in particular, soliton mechanisms, are of special importance. A soliton (also called solitary wave) is described as energy excitation of the medium, propagating along the medium at long distances. A fundamental role in studying these mechanisms in molecular systems and in using them to explain a number of biological phenomena was played by the prominent Soviet physicist A. S. Davydov and his school.

The concept of soliton switching, developed by Carter, is based on the rearrangement of electronic structure initiated by collective processes of this type. Below follows a brief description of these interesting molecular entities.

3.2.1 More Details: Molecular Structure—Spatial Configuration of Molecular Nuclei

We use the term molecular structure to refer to the set of constants and functional characteristics that describe the relative position and relative motion of the nuclei and electrons of the molecule. Based on this premise, three main groups of molecular characteristics are usually considered:

- Parameters of the geometric configuration of the nuclei of the molecule (a full set of internuclear distances and bond angles is often called the structure of the molecule in a narrow sense)
- Dynamic characteristics that determine the relative motion of the nuclei of the molecule (energy levels, frequencies of oscillation, average displacements of nuclei from their equilibrium positions, and so on)
- Electronic characteristics (energy levels of the electrons of the molecule, the electron density distribution and its further characteristics, such as electric dipole and quadrupole moments of the molecule, etc.)

It is conventional to characterize the relative position of the nuclei of the molecule by their equilibrium configuration corresponding to the minimum potential energy of the nuclei.

Among the vast number of molecules known today organic carbon compounds display the largest variety. A unique feature of carbon is that its atoms are combined into chains of repeating units of varying length. In this manner molecules are formed that belong to various classes of organic compounds—saturated and unsaturated hydrocarbons, aromatic compounds, etc. The carbon atoms forming organic compounds can exist in three structural states (Fig. 3.16). The first one is the tetrahedral state in which the carbon atom is at the center of a tetrahedron, and its bonds with other atoms are directed to the vertices of the tetrahedron. In the second state, which is called trigonal, the bonds of the carbon atom are directed from the center of a plane triangle to its vertices. Finally, in the linear state, all three atoms—

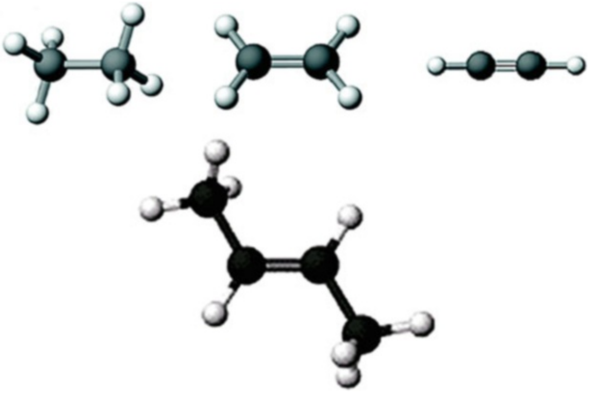


Fig. 3.16 Main structural forms of carbon compounds

the carbon itself and the two atoms associated with it—lie on a straight line. Since in the vast majority of organic compounds the valence of carbon is equal to four, trigonal states appear in the molecules with double carbon–carbon bonds, while the linear states in the molecules with triple bonds. Moreover, different structural elements can be combined within the same molecule.

These structural features of the molecules of carbon compounds can be explained based on the properties of their electronic structure. Today, the geometric configuration of the nuclei of relatively large organic molecules can be determined with experimental accuracy by solving the Schrödinger equation, but this is a tedious and expensive approach. Therefore, already in the 1930s of the last century, when computational quantum-mechanical methods were still in their infancy, the famous chemist Linus Pauling proposed a theoretical semiempirical method. He suggested an explanation of the experimentally determined variants of bond positions of the carbon atom based on the principle of hybridization of atomic orbitals.

As is known, the valence electrons of the carbon atom are located at the 2s and 2p orbitals. The 2s orbitals have spherical symmetry, and the 2p orbitals are located in the space perpendicular to each other. Since linear combinations of the solutions of the Schrödinger equation also represent its solutions, Pauling introduced the concept of hybrid orbitals—linear combinations of s and p orbitals of the carbon atom. If one imposes symmetry constraints on these combinations, it turns out that they correctly describe the spatial arrangement of carbon atom bonds (Figs. 3.17, 3.18, and 3.19). In the case of tetrahedral symmetry, for example, upon turning the hybrid orbitals by 120° with respect to one of them, they will coincide with each other. Taking into account the symmetry principles it could be shown that in the case of tetrahedral symmetry all s and p atomic orbitals of carbon are present in all four hybrid orbitals. This kind of hybridization is called sp_3 hybridization. At the same time, in accordance with symmetry requirements the trigonal (sp_2) and linear (sp) hybrid orbitals must consist of an s and two p and an s and one p orbitals, respectively. Consequently, there remain orbitals that are not part of hybrid orbitals—one p orbital in the case of sp_2 hybridization and two in the case of sp

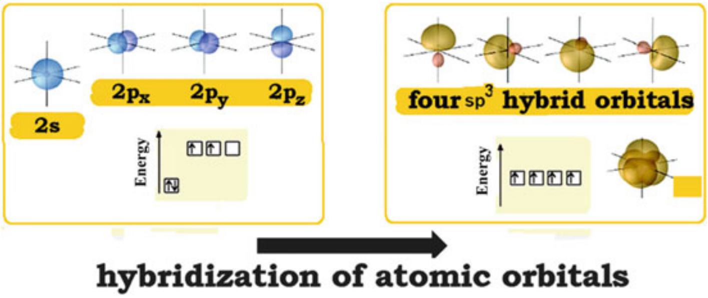


Fig. 3.17 *sp*^{3} hybridization of atomic orbitals

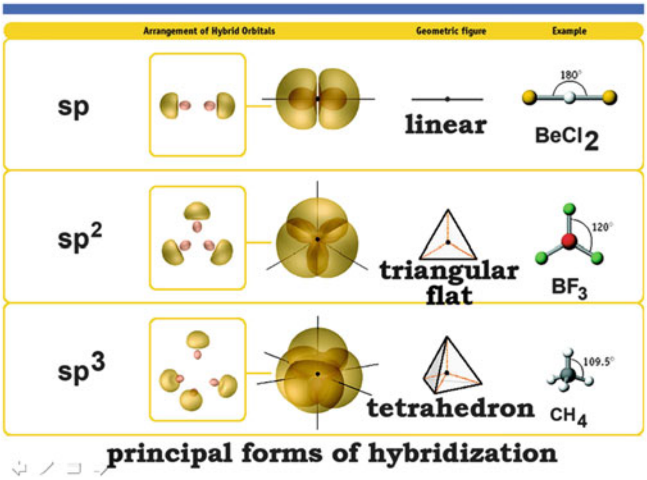


Fig. 3.18 Different variants of hybridization of atomic orbitals

hybridization. Pauling explained their overlap by the formation of double and triple carbon–carbon bonds.

Later, Gillespie and Nyholm offered another semiempirical method for determining the structure of simple molecules, which applies not only to compounds of carbon but also to molecules involving all atoms of the periodic system. They suggested that the relative position of atoms in a molecule is determined by the electrostatic repulsion between bonds and lone pairs of electrons. Despite its exceptional simplicity, this assumption describes the structure of simple molecules with surprising accuracy (Fig. 3.20).

A characteristic feature of large molecules is that with an increase in the number of nuclei in the molecule the probability of the existence of isomers—compounds with the same stoichiometry (i.e., the atomic composition of the molecule), but with different equilibrium configurations of the nuclei—also increases. Accordingly, several minima corresponding to each isomer appear on the potential energy

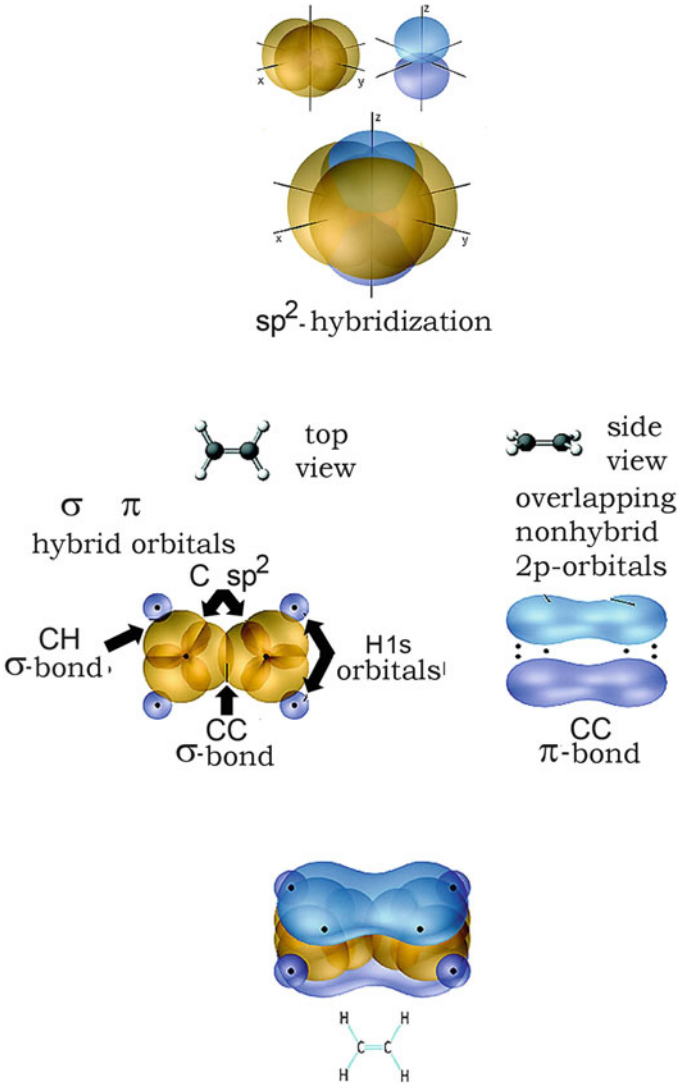


Fig. 3.19 Scheme of the formation of the ethylene molecule according to Pauling

surface of nuclei of a polyatomic molecule. In other words, in this case the molecule represents a system with several states such that the transitions between them (isomerization) may become possible under the influence of physical factors. In the literature such states are usually called conformers. A good example is the molecule of cyclohexane (Fig. 3.21). Its lowest energy state is called "armchair." It is situated lower than another conformation called "bathtub," at ~22 kJ/mol. Transitions "chair"–"bath" are made through the intermediate conformations of the type "elbow." In general, the transition from one conformation to another is

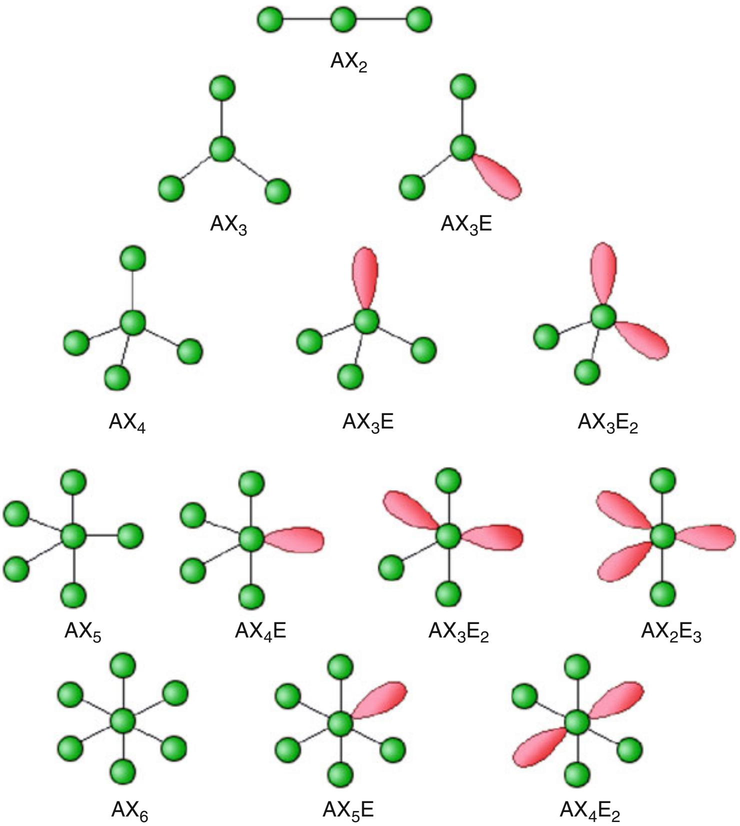


Fig. 3.20 Structure of simple molecules according to Gillespie and Nyholm

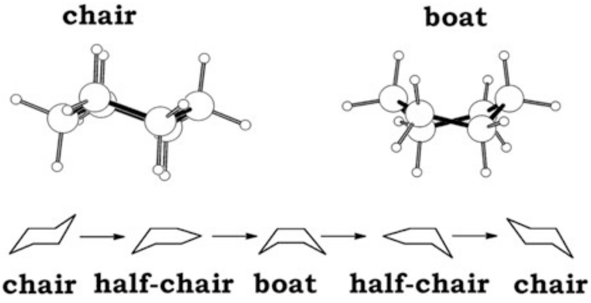


Fig. 3.21 Conformations of cyclohexane

accompanied by a rearrangement of the electron structure of the molecule. Therefore, significant energy is spent on these transitions, which leads to the need to use macroscopic effects (rather than individual quanta of radiation) and to take into account the fact that a part of the impact energy is ultimately dissipated as heat.

As the barriers between the minima corresponding to individual conformations are sufficiently large, these conformations correspond to long lifetimes. Therefore in recent years, electron-conformational transitions are widely used to create molecular switching devices.

At least two types of nonlinear systems in which solitons can propagate are known. The first one is constituted by molecular chains built from interacting fragments, which are additionally linked by weak interactions (e.g., hydrogen bonds), i.e., electronic and vibrational properties of the fragments in the chain do not substantially differ from the properties of individual molecules. An example of such a system is a chain of peptide groups connected by hydrogen bonds in a helical protein fragment. In the peptide group intramolecular vibrational excitation with relatively large dipole moment of transition is possible. It provides a strong interaction between neighboring molecules, which leads to collective, rather quickly decaying, excitation, called exciton. Another collective excitation occurs in these molecular chains as a result of the interaction of intramolecular excitations (vibrations) and the nonlinearity, caused by the coupling of these vibrations with the local displacements of equilibrium positions of the molecules. Solitons are stable entities that propagate over large distances along the chain without energy dissipation. The second type of systems that manifest collective effects is conjugated polymers, such as trans-polyacetylene, in which the possibility of soliton excitation is caused by the degeneration of the ground electronic state.

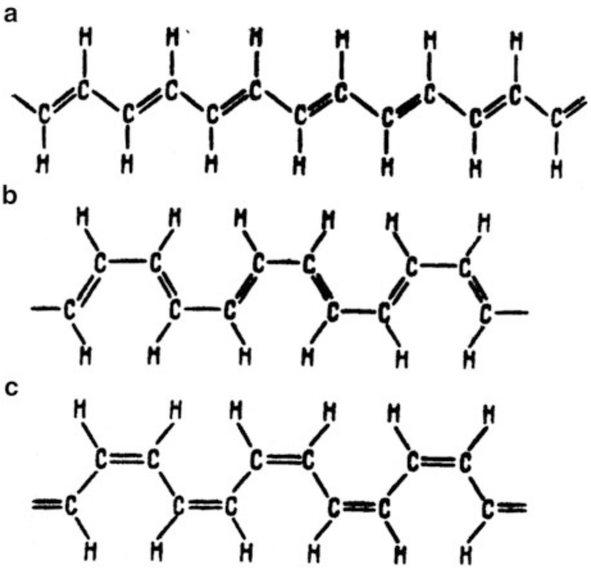
Polyacetylene (Fig. 3.22) is a conjugated hydrocarbon $(CH)_n$ with the simplest structure. It is well known that the carbon atom has four valence electrons. In accordance with the generally accepted semiempirical notions in molecular fragments such as ethylene (the main structural fragment of the polyacetylene chain), three of them are located on four hybrid sp^2 orbitals and form single bonds with the neighboring hydrogen and carbon atoms situated in the same plane. These connections correspond to completely filled zones deep in the atom. The fourth valence electron of the carbon atom (π-electron) corresponds to the wave function $2pz$ which is perpendicular to the plane of the molecule. The functions of neighboring atoms of this type overlap, which leads to the formation of π-orbitals. It is precisely these delocalized electrons that determine the characteristic properties of conjugated molecules, while σ-electrons form a rigid skeleton of the molecule and the field in which the mobile π-electrons move. Therefore, the usual approximation for conjugated systems is the splitting of energy into two parts:

$$E = E_\sigma + E_\pi,$$

where E_σ is the energy of the "σ-skeleton" of the molecule in the absence of interaction with π-electrons and E_π is the energy of π-electrons in the effective field of the σ-skeleton.

Polyacetylene exists as two isomers: trans-$(CH)_n$ and cis-$(CH)_n$ (Fig. 3.22). While performing quantum-mechanical calculations it is usually assumed that equal distances between CH groups correspond to the trans-$(CH)_n$ skeleton, whereas alternation of σ-bonds takes place in the cis-$(CH)_n$. This follows primarily

Fig. 3.22 Conformations of polyacetylene



from general considerations—the symmetrical arrangement of the neighboring hydrogen atoms with respect to carbon atoms in the trans-$(CH)_n$ and the asymmetrical arrangement in cis-$(CH)_n$, which, due to steric interactions, H…H leads to a series of shortened and elongated distances between the CH groups. A similar conclusion may be reached on the basis of ab initio quantum-mechanical calculations.

Soliton solutions (more precisely, topological defects) arise upon detailed consideration of the electronic structure of such a system of conjugated bonds of sufficient length. First of all, based on obvious geometric considerations, it is easy to see that the chain of conjugated bonds of the trans-$(CH)_n$ type can exist in two degenerate energy forms. Their interaction leads to the possibility of occurrence of a topological defect in the electronic structure (Fig. 3.17), which is one of the elementary excitations of this conjugated system.

In 1962 Pople and Walmsley investigated this type of a defect making the simple assumptions that the defect is localized on one CH group and that only two single bonds surround this defect (as shown in Fig. 3.23) and that there are only two types of bonds: single and double. As a result they showed that in addition to the valence band formed by bonding orbitals and to the conduction band, formed by the antibonding orbitals, two localized nonbonding orbitals arise that correspond to the two defects being considered. They are situated in the middle of the gap between the valence and the conduction band (Fig. 3.23).

A soliton in trans-polyacetylene is a topological defect which connects two geometrically inequivalent configurations (A and B, Fig. 3.23) of the degenerate ground state and is entirely due to this degeneration. In the long conjugated chains different mechanisms may be responsible for the formation of defects of the soliton

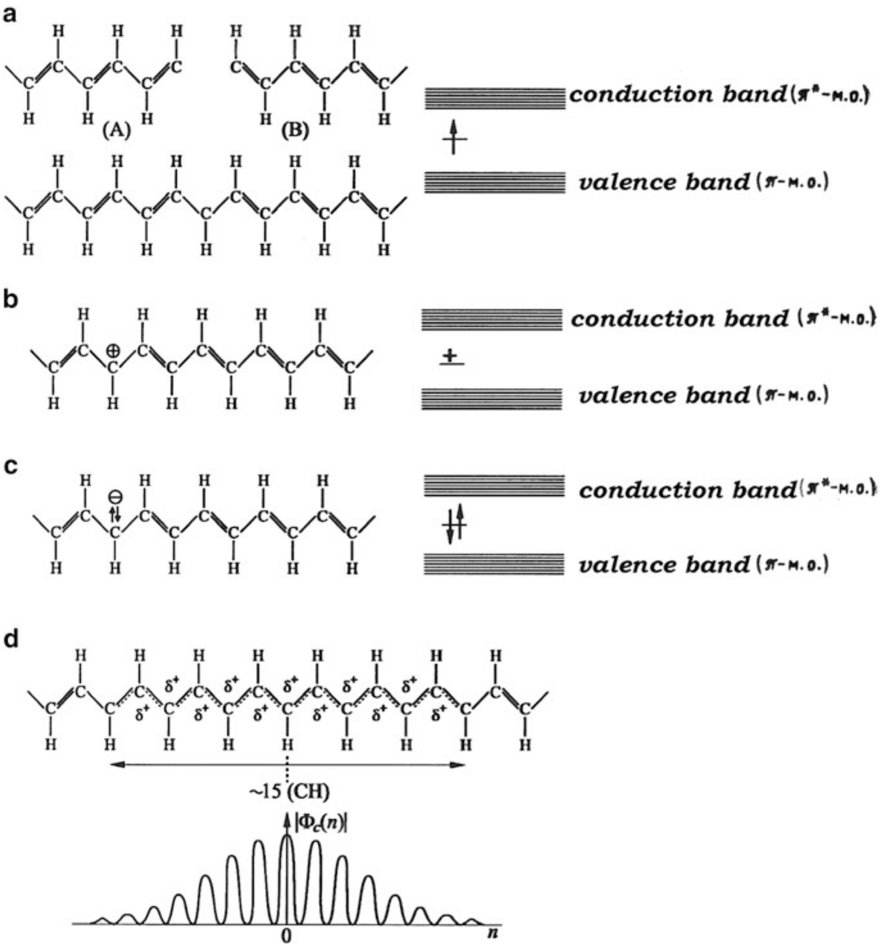


Fig. 3.23 Scheme of the formation of a soliton in the polyacetylene chain (**a**); different types of the excitation solitons (**a**–**c**); spatial character of the soliton excitation (**d**)

type. In particular, absorption of a photon by the polyacetylene molecule may lead to pair defects. It is easy to see (Fig. 3.24) that the defect arising on the odd (even) atom of the polyacetylene chain can only propagate via the odd (even) sites, and the defect is called either soliton or antisoliton, respectively. Upon breaking one of the double bonds in polyacetylene a diverging soliton–antisoliton pair is created. At the same time, two defects moving toward one another, the soliton and the antisoliton, are capable of recombination.

The concept of soliton switching, developed by Carter, is based on the rearrangement of the electronic structure of the molecular system caused by collective processes of this type. Propagation of a soliton in a quasi one-dimensional conjugated system is due to the replacement of double bonds by

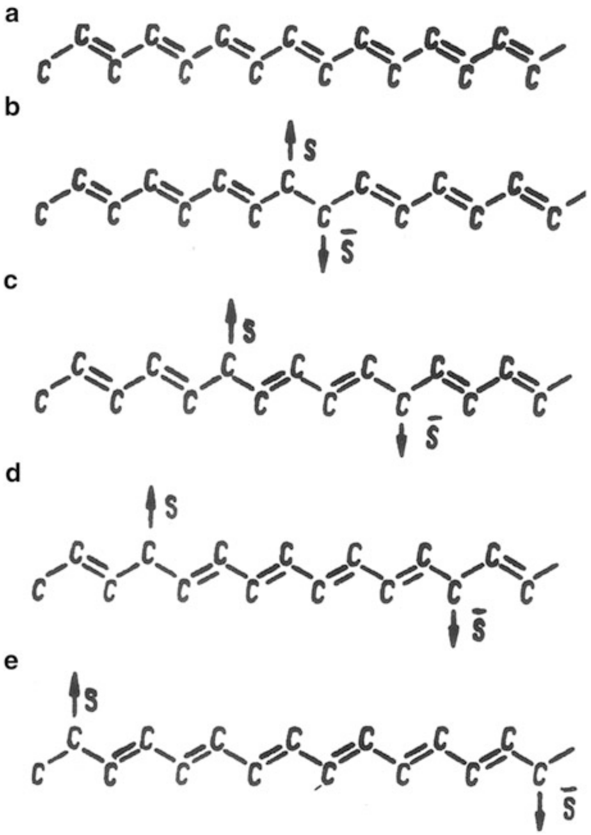


Fig. 3.24 Excitation of a soliton in the polyacetylene chain

single ones and vice versa. This can be used to create switches by incorporating a switchable molecular fragment capable of rearranging the conjugation chain into the original conjugated system. One possible fragment of this type is the molecule of 1,1-N,N-dimethyl-2-nitro-ethanamine, which, being embedded in the trans-polyacetylene chain, can undergo phototransformations upon light absorption and rearrange the conjugation chain (Fig. 3.25). Herewith:

- After the passage of a soliton in the main chain and the substitution of double bonds by single ones and vice versa, no photoisomerization will occur, which can serve as an indicator of the soliton transmission.
- Isomerization reaction, initiated before the passage of the soliton in the main chain, changes the electronic structure of not only the switching fragment but also of the adjacent region of the chain (breaks the conjugation chain) and thus makes subsequent passage of the soliton impossible.

The concept of soliton switching can be generalized to multiple conjugated quasi one-dimensional chains, along which solitons can propagate. One such example is

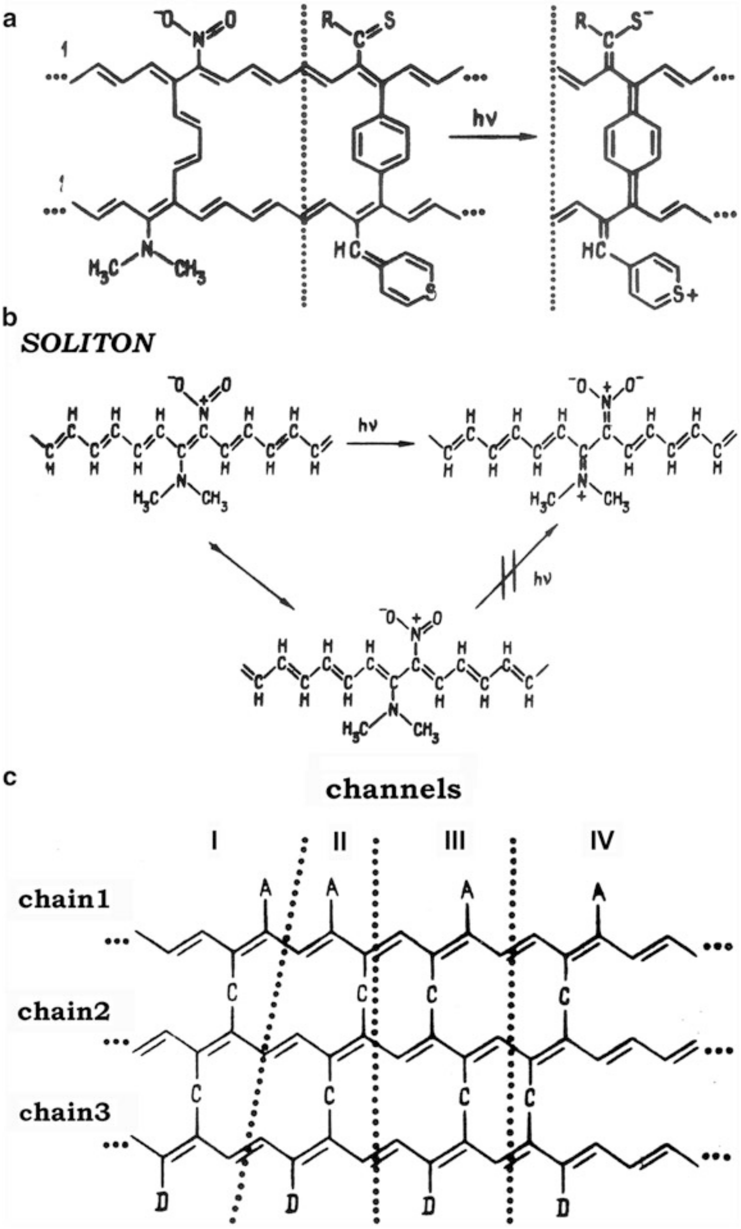


Fig. 3.25 Different variants of switches on the basis of solitons

Table 3.1 State of the logic scheme with several inputs and outputs controlled by soliton switching

Channel	1	2	3	4
Chain 1	1	0	1	1
Chain 2	1	1	0	1
Chain 3	1	1	1	0

presented in Fig. 3.25 which shows two conjugated chains and two different chromophores. The configuration of the chain 1 is such that the first chromophore does not absorb light, while the second one may be photoexcited. Passage of a soliton along the chain 1 will switch on the first chromophore and switch off the second one. A soliton traveling along the chain 2 will switch off both chromophores.

More complex switching systems with a sufficiently large number of inputs and outputs can be built based on the same principles. Figure 3.25 shows an example of such a system built from several chains $A\ldots C\ldots C\ldots D$ joined into a common system by strings of conjugated bonds. Here A, C, and D are the electron acceptor, the intermediate group, and the electron donor, respectively. In this system, as in the case of the NAND element considered above, Carter proposed to use the tunneling transition of the electron. Soliton channel switching (I–IV) can be carried out through three channels of polyene chains (horizontal directions in the figure). Note that each group (A, C, D) separated from each other by a dotted line can be attributed to different states, depending on the distribution of single and double bonds between them. Variants of the system states are shown in Table 3.1. Here unity denotes the presence of a double carbon–carbon bond between A and C, C and C, and C and D in the corresponding channel. It is easy to see that the passage of the electron along the vertical chain $A\ldots D$ in Fig. 3.25 is only possible with all unities in the corresponding table columns. For example, the fragment C is disconnected from in the channel II by the chain 1, while the fragment C is disconnected from the second fragment C in the channel III by the second chain. The system state changes significantly during the passage of a soliton along the corresponding chain.

The soliton switching mechanism can serve as the basis for implementing more complex logic functions. Thus, the systems shown in Fig. 3.26 can perform quite complex operations. The structure depicted in this figure consists of a carbon atom linking three semi-infinite trans-polyacetylene chains. Assuming that the presence of two successive single bonds blocks further propagation of the soliton, one can see that the soliton propagation along one of the chains will shift and redistribute the single and double bonds, as if rotating the system around the central carbon atom. In such a case, each soliton propagation corresponds to the group operation—rotation by 120°—blocking the propagation of the soliton to the other chain (figure, soliton propagation between the left and the bottom chain is impossible). A cyclic system combining three fragments of type *a* (Fig. 3.26) ensures the execution of group operations corresponding to the relevant point group D_2. Similar operations are performed by the system shown in Fig. 3.26c.

While the simple system (a) has three different states, the cyclic configuration of three such systems (b, c) has four different states, and the configuration shown in

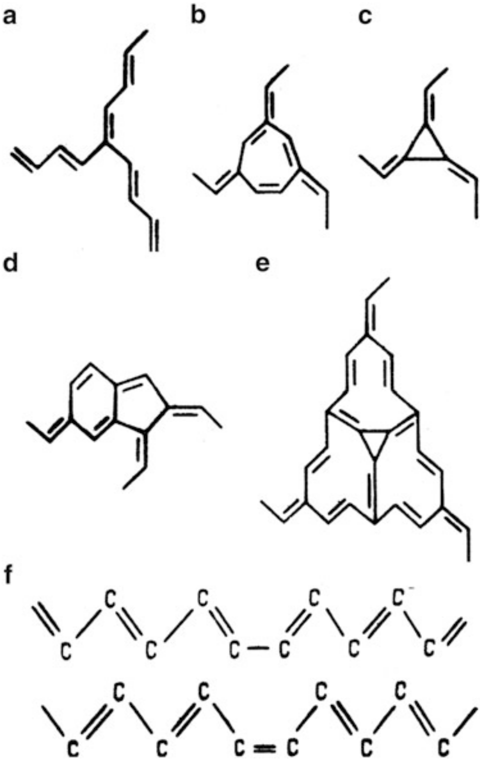


Fig. 3.26 Cyclic elements on the basis of solitons

figure has five states; therefore they can be used in the machines operating with nonbinary number systems. The configuration shown in figure has ten states and may be of interest for the systems that perform calculations in the decimal system. Thus, these elements could be used in nonbinary machines and perform complex group operations.

In principle, such soliton structures, combined in a traditional fashion, can be used to construct the elements that perform the functions of Boolean logic and serve as memory elements.

Pioneering research of Aviram, Ratner, and Carter provoked the study of various aspects of information processing at the molecular level. Over the next few years a number of theoretical works were published, suggesting possible ways of using molecular components. Several interesting experimental studies were performed. One must admit, however, that they merely demonstrated fairly trivial facts. Basically it was shown that molecules with a certain structure can switch from one stable state to another and that these transitions can be recorded. But at the same time these studies led to a real understanding of the peculiarities of molecular electronics.

First of all, it was realized that due to their quantum nature small molecules consisting of a relatively small number of atoms could not be expected to act as deterministic switches analogous to semiconductor devices.

The theoretical feasibility of information-logic devices at the molecular level is only helpful in determining general criteria rather than specific rules for the selection of molecular elements for building devices with desired functions.

In order to explain specific difficulties in solving this problem, let us consider one possible option when individual molecules are regarded as logic elements, with control of molecules and transmission of information carried out by monochromatic light and optical communication.

It is known that an information processing system of arbitrary complexity can be constructed using logic elements of only three types: NOT, AND, and OR.

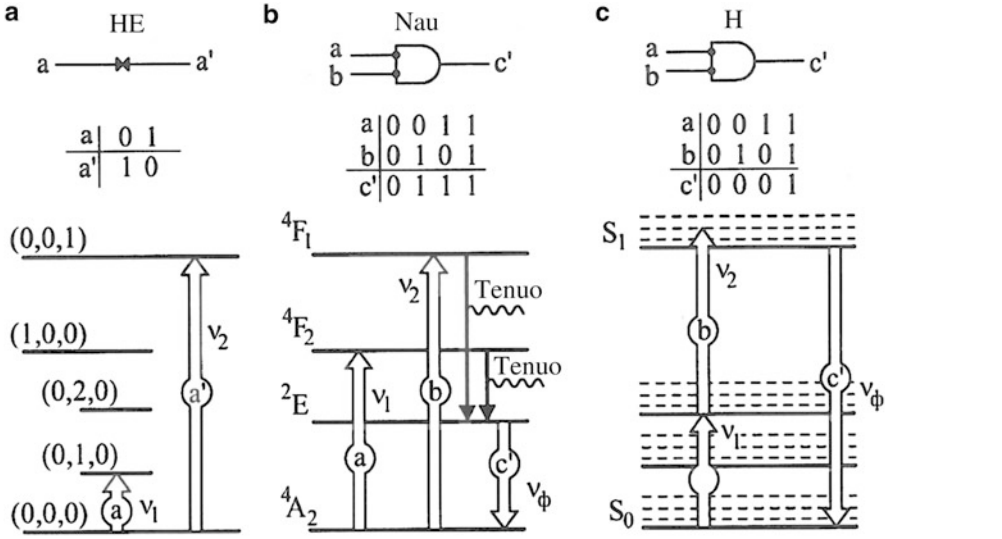
At first glance, many atomic or molecular systems with their discrete energy levels are ideally suited as the basic logic elements. Let us consider, for example, the processes of selective excitation of electronic or vibrational states of the atom (ion) or polyatomic molecule by monochromatic radiation. Let us assume that initially only the ground state is populated in the atom or the molecule and that the excitation of a particular level by a quantum of radiation occurs with a high (almost complete) probability. Naturally, we assume that the transitions between these states are allowed.

Figure 3.27 shows primitive examples of possible implementations of basic logic functions at the molecular level.

It is easy to see (Fig. 3.27) that the possibilities of excitation of the CO_2 molecule, which is situated in the ground vibrational state (0, 0, 0), into one of the excited states (0, 1, 0) or (0, 0, 1) are mutually exclusive over the lifetime of the excited state. This process can be compared with the NOT operation. Excitation of the chromium ion in a ruby crystal by radiation with a frequency ν_1 or ν_2 leads to the emission of a photon with the frequency ν_φ (the OR element). Selective two-step excitation of the ground singlet electronic state S_0 of the rhodamine molecule by joint action of two ultrashort pulses with the frequencies ν_1 and ν_2 simulates the logical operation AND.

Nevertheless, despite the fact that the possibilities of using individual atoms and molecules as basic elements in computational devices seem obvious, fairly simple considerations reveal a number of discrepancies between the characteristics of atomic and molecular systems and the requirements imposed on the basic elements. Thus:

– Basic elements should have high reliability (probability of response) upon control activation.
– The average power of element’s reaction transmitted to another element should not be much smaller than the average power of the input stimulus applied to the element. In our case this means that the number of quanta per unit of time at the input of the element should not substantially exceed their number at the output. Otherwise, the probability of element response after combining them in a chain will decrease with the distance from the beginning of the chain.



a	0	1
a'	1	0

a	0	0	1	1
b	0	1	0	1
c'	0	1	1	1

a	0	0	1	1
b	0	1	0	1
c'	0	0	0	1

Fig. 3.27 Logic possibilities of molecular systems

- Conversion efficiency of the signal by an individual element should be close to unity. This means that the reaction of the molecule upon excitation has to be unambiguous and that the excited molecule must continue to respond unambiguously without losing excitation energy through intramolecular dissipation.
- It must be possible to switch over an element by a control stimulus to any required state.
- During element's transition from one state to another, it must remain in it long enough for the next control stimulus to be guaranteed to transfer the element to a new state.
- The state of the element should be readable, i.e., it should be possible to unambiguously determine its state.

Detailed analysis of these requirements shows that in the case of electronic states they are not fulfilled, at least in part. In the following it will be shown that when the configuration of the nuclear core changes the conformational transitions turn out to be suitable for creating molecular switching elements.

At the same time during the 1980s and 1990s, it was shown that ensembles of molecules with certain characteristics are promising targets for the development of devices for handling and storing information, i.e., that chemical media can be used in computing or storage devices. In fact this was a shift from the ideology of molecular devices (i.e., systems in which individual molecules serve as basic elements) to the devices built on the basis of chemical environments, i.e., macro-objects, preserving at the same time certain advantages of molecular objects. These two major trends (using molecular media and synthesis of large switching molecules) resulted in the second half of the 1990s in the establishment of molecular information processing devices.

3.3 Molecular Memory Based on the Protein Bacteriorhodopsin

Absorption spectra of various compounds exhibit bands corresponding to the transition of their molecules in one of the excited states. If the lifetime of the excited state is large enough and if it can be transferred back to the ground state of the molecule by some physical effect, such chemical compounds can be used to record and store information.

Technically, the easiest approach is to use for recording quasi two-dimensional films made of suitable material. This can be done, for example, by a laser with a frequency equal to the frequency of the molecular transition. In this case, the laser beam scanning the surface of the film and recording information can be focused to the spot with a diameter of ~1 μm. However such focusing requires complex and expensive equipment. Therefore systems focusing the laser beam to a diameter an order of magnitude greater are commonly used. Thus the recording density of information on the film will be 10^6–10^8 bits/cm^2.

Work in this field led to the creation of a number of functional memory devices. Apparently, the most important among them and the closest to being practically useful were different versions of storage for computers designed by Robert Birge on the basis of a unique protein, bacteriorhodopsin.

Bacteriorhodopsin is a photosensitive protein contained in the purple membranes of halobacteria *Halobacterium halobium*. Halobacteria live in salt water or in springs where salt is mined. The latter appear reddish due to halobacteria. This shade of yellow is conferred to the cells by carotenoids and the purple bacteriorhodopsin. Bacteriorhodopsin molecules form the photosynthetic center of halobacteria. By absorbing a quantum of light, it acts as a proton pump, contributing to ATP synthesis. The structure of the bacteriorhodopsin molecule was determined in detail (Fig. 3.28). It is a cyclic combination of seven polypeptide helices, inside of which a light-sensitive fragment—chromophore—is situated. When light is absorbed, structural rearrangements of the molecule take place.

Bacteriorhodopsin possesses stability which is unique for proteins. It is capable of keeping its properties intact for many years both as a dry sample and in polymer films as a monolayer with the thickness varying from 5 nm to several tens of microns.

The fundamental property of the bacteriorhodopsin molecule—the photochemical cycle—is a sequence of excited states (intermediates) which the molecule passes after being excited by light radiation. The spectra of the intermediates differ significantly from each other (Fig. 3.29). During the photocycle, optical characteristics of the protein—absorption and refraction—also undergo change. Thus, natural bacteriorhodopsin at room temperature behaves as a photochromic medium with a short time of information storage. The ground state of the bacteriorhodopsin molecule is bR(570). After excitation by light into the K(610) state, the bacteriorhodopsin molecule spontaneously passes through a number of intermediates and returns to its original form. However in the course of this process, an additional

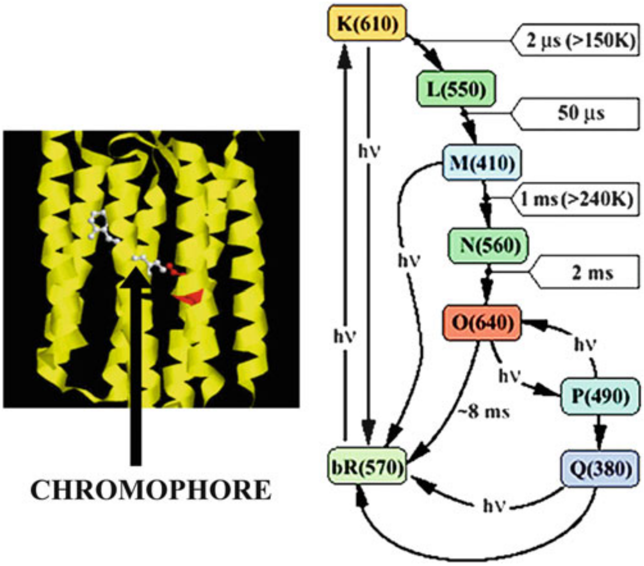


Fig. 3.28 Structure and photocycle of the bacteriorhodopsin molecule

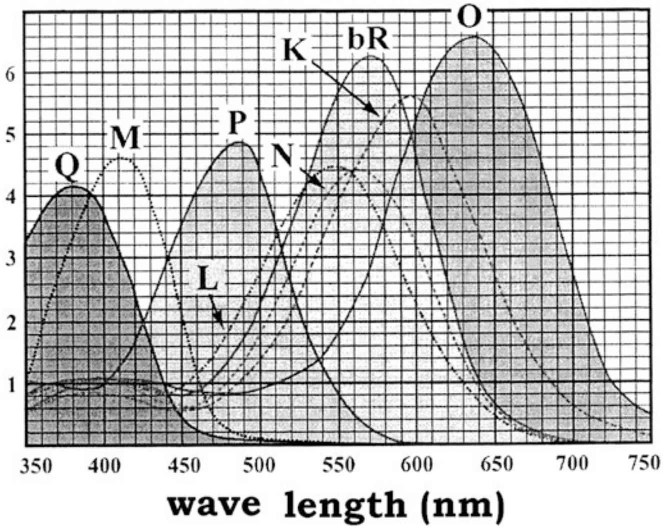


Fig. 3.29 Spectra of bacteriorhodopsin conformers

opportunity arises to convert the molecule from the state O(640) to another long-lived state, Q(380), by selective radiation. Mechanisms of transition between excited states of the molecule may vary and have different temperature dependencies. At 77 K the photocycle of the natural bacteriorhodopsin gets disrupted, and the

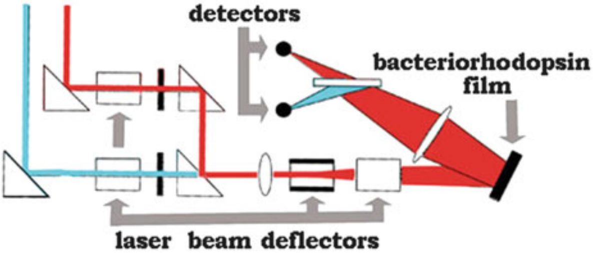


Fig. 3.30 Scheme of a planar molecular memory on the basis of the bacteriorhodopsin protein

molecule behaves as a system with two stable states, bR(570) and K(610), with transitions between them initiated by light in the visible spectral range.

In the early 1990s, based on this property of bacteriorhodopsin, Robert Birge developed a cryogenic optical random access memory for digital computers. In this device information was recorded on a film made of bacteriorhodopsin-containing polymer and read from it by laser beams with various frequencies of radiation (Fig. 3.30). The memory with 25 MB of storage capacity and access time of 10–100 ns was operated at 77 K. At the same time, 25 MB cache storage was developed, operating at temperatures close to room temperature, in which a different excited state of bacteriorhodopsin was used.

During the 1990s and in the beginning of this century, the Birge group has also developed a holographic associative memory capable of reading out information with only a part of it being available (e.g., reconstructing an image based on a fragment). But the main efforts of the group were directed at the development of ultrahigh-capacity memory.

Unlike earlier versions of storage devices in which information was recorded on a flat bacteriorhodopsin film, in the volumetric memory, as its name implies, the entire volume of the medium is used to store information.

It should be said that the bacteriorhodopsin-based volumetric memory is a technically sophisticated electron-optical device. It includes several lasers emitting in different spectral regions (Fig. 3.31).

In a memory device, the main bR-state and the Q-state of the bacteriorhodopsin molecule correspond to the digit values of “0” and “1,” respectively. These long-lived states can remain unchanged for several years. The storage medium is a cuvette 1 × 1 × 2 in. in size filled with polyacrylamide gel with embedded bacteriorhodopsin molecules. Around the cuvette lasers and devices that convert and record light information are situated.

Consider a simplified process for recording and reading information in the volumetric memory (Fig. 3.32). The process of writing begins with the selection and activation of a thin layer (page) within the volume of the storage medium. The thickness of the page can vary from 15 to 100 μm. From the information point of view, the page is a quasi-flat storage media with 4,096 × 4,096 bits capacity. The system is activated by green lasers. They excite all bacteriorhodopsin molecules located in the medium layer constituting the page. After about 2 ms, the concentration of molecules existing in the O-state on the page reaches a maximum. At this

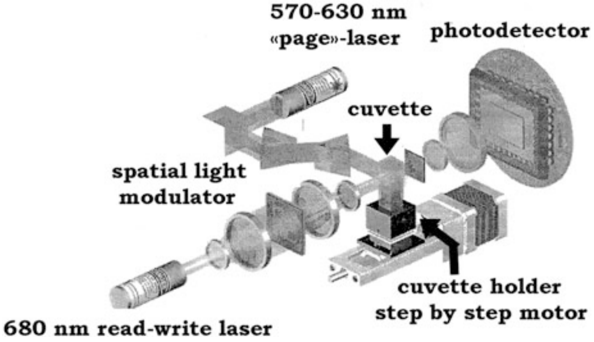


Fig. 3.31 Scheme of a volumetric molecular memory on the basis of bacteriorhodopsin

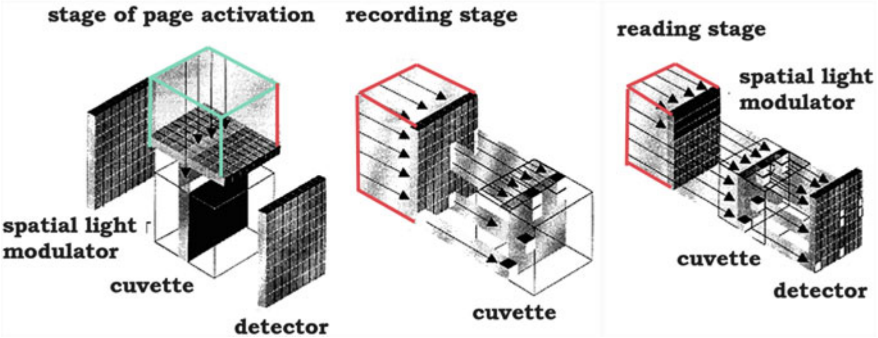


Fig. 3.32 Recording and reading information in a volumetric molecular memory on the basis of bacteriorhodopsin

point, the system of red lasers is switched on with radiation direction which is perpendicular to the plane of the page. It illuminates the page via an LCD display such that only selected elements of the page are switched to the Q-state corresponding to the digit value "1." The rest of the bacteriorhodopsin molecule are switched to the initial state, thus preserving the digit values of "0."

In order to read out the data the page is again activated by a system of green lasers. In this case the bacteriorhodopsin molecules that remained in the ground state ("0") switch to the O-state within ~2 ms. After that the page is illuminated by red lasers whose radiation is absorbed by the molecules in the O-state and not absorbed by the molecules in the Q-state. The positions of the medium elements containing molecules in the Q-state can be registered by measuring the red radiation passing through the cuvette by a CCD (charge-coupled device) detector.

The memory can be erased by switching all bacteriorhodopsin molecules into the ground state using a blue laser.

Small lifetimes of bacteriorhodopsin intermediates and parallel simultaneous information processing on the page allow for reducing the read/write times to several tens of gigabits per second. Estimates also indicate that in the volume of 3 cm^3, a huge amount of information, hundreds of gigabytes, can be stored.

It should be emphasized that the creation of such a system is a major scientific and technical challenge, with the memory capacity of currently existing prototypes not exceeding tens of Kbits.

The development of bacteriorhodopsin-based devices and some other studies in the 1990s brought nearer the industrial use of molecular media in computer engineering.

3.4 Storage Devices Based on the Effect of “Hole Burning” in the Absorption Spectra

Along with the volumetric molecular memory, another way to create ultrahigh storage capacity is of importance. It is based on the principle of frequency-selective recording.

The mechanism of optical frequency-selective recording exploits the phenomenon of hole burning in broad inhomogeneously broadened absorption bands of organic molecules in solid matrices. The essence of this phenomenon, discovered and investigated by the school of K. K. Rebane, the R. I. Personov group, and some other authors, is that a narrow band laser creates a narrow dip in the absorption spectrum of the material which will persist for a rather long time in darkness and at low (helium) temperatures. For practical implementation of optical frequency-selective recording of information, it is necessary to have answers to two questions: how fast is hole burning and what is its life span.

Extensive experimental data indicate that many organic molecules have no structural optical spectrum. To explain this fact it is usually suggested that molecules in an amorphous body or in a solid solution are arranged in arbitrary orientations, which leads to interactions of different strength with the environment. As a result, the frequencies of vibronic transitions are shifted somewhat in comparison with those molecules not interacting with the environment. These transitions correspond to narrow homogeneously broadened bands Γ_0. A single inhomogeneously broadened band Γ_H, which is the superposition of bands of individual molecules, corresponds to the entire ensemble of molecules. 10^2–10^3 components contribute to each inhomogeneously broadened band. Their use for recording information allows to achieve recording density of 10^9–10^{11} bits/cm^2 (Fig. 3.33).

Each vibronic state will correspond to the band which can be represented as a sum of two terms:

$$I(\omega) = J(\omega) + \Phi(\omega),$$

where the function $J(\omega)$, having the form of a sharp peak, describes the zero-phonon line directly corresponding to the vibrations of the atoms of the molecule.

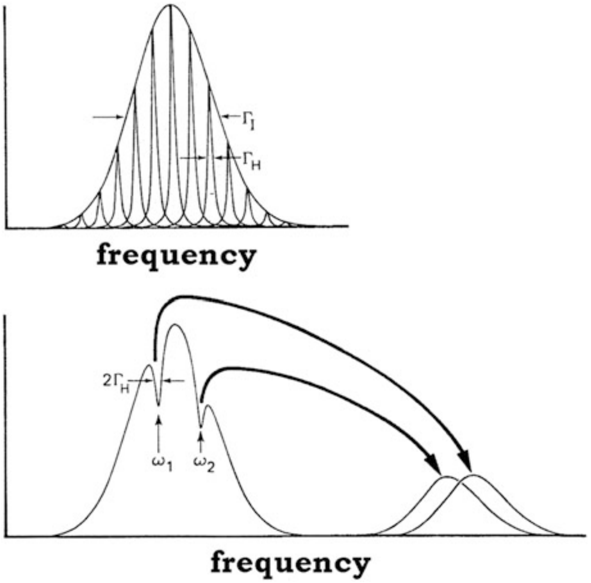


Fig. 3.33 Recording and reading information in a volumetric molecular memory on the basis of bacteriorhodopsin

The function $\Phi(\omega)$ describes the phonon wing which accompanies the zero-phonon line. It arises as a result of collective molecular vibrations in a solid. Suppose that the broad band in the spectrum of a chemical compound is the result of superposing zero-phonon lines corresponding to different local conditions for individual molecules in the matrix. This is possible when the molecule has a spectrum consisting of a zero-phonon line and a weak phonon wing.

It is in this case that a hole can be burnt in the absorption band by monochromatic laser radiation. This phenomenon occurs due to high intensity of the zero-phonon line if, upon light absorption, there exists nonzero probability that something will happen to the molecule and it will not return to its exact original state. Regardless of the specific nature of phototransformation, it is essential that the energy transition either disappears altogether or has a shift in the spectrum of the photoproduct exceeding the uniform width of the initial transition. An additional requirement for observing the effect of hole burning is for phototransformation to be irreversible on the time scale required for the registration of the spectrum.

Experimental evidence shows that for a number of investigated molecules the lifetimes of holes are in the order of at least hours (in the dark, at helium temperature). As for the duration of hole burning, in currently known experimental studies, it is in the range of seconds to tens of seconds. Naturally, this excludes the possibility of recording sufficiently large volumes of information (albeit at relatively low laser intensities used in experimental studies). However, it was suggested that the recording time of information can be reduced to 30 ns/bit.

In the 1980s of the last century at the IBM Research Laboratory in San Jose (California, USA), experimental studies as well as materials science and

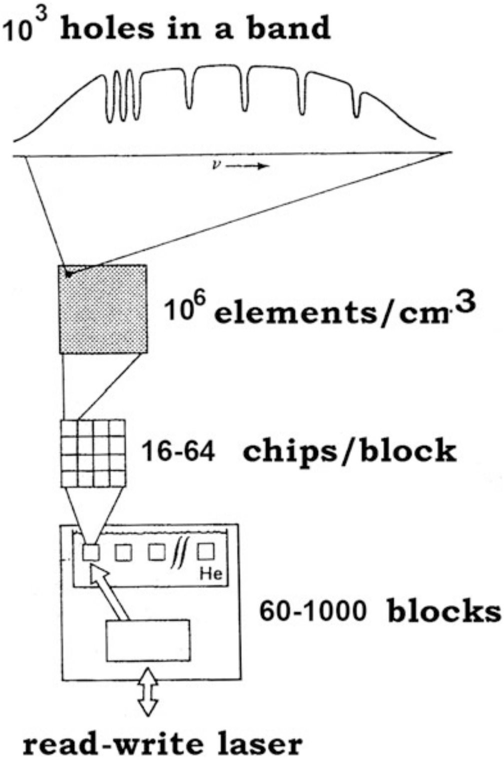


Fig. 3.34 Memory based on the "hole burning" effect

engineering assessment of the feasibility of creating extra high-capacity memory based on the "hole burning" effect were conducted (Fig. 3.34). Consideration was given to the option in which chips of 1 cm^2 with the record density of 106 bits/cm^2 were grouped into subsystems containing 16–64 chips, for a total capacity of 2.8 GB. 50–1,000 of such subsystems, in their turn, formed a memory with up to 8 TB of capacity.

It was suggested to use a sodium fluoride crystal with color centers, in which holes of 50 MHz in width could be burnt in the band with a maximum of 575 nm and a width of 100 GHz (~0.1 nm). Later it was reported that a new group of materials for frequency-selective recording of information was discovered—the crystals of fluorides of alkaline elements in which ions of rare earth elements and special glasses based on one of the boric acids were introduced, with embedded carbazole molecules. This storage medium had to be kept at the temperature of liquid helium. The studies demonstrated that there were no technical constraints to the creation of such a system with record characteristics for that time. However, operational characteristics of this device were not satisfactory. One problem was the helium temperature required for storing information. Furthermore reading and writing speed was inadequate even for that time. With a maximum speed of recording of one hole of 30 ns, it took 30 s to record 1 GB.

Nevertheless, the idea of frequency-selective recording of information has not lost its appeal until now. Active search is ongoing for novel memory architectures and new materials which will be capable of acting as memory devices of giant capacity.

3.5 Molecular Memory: Developer Dream or Reality?

Let us return to the molecules in which the switching from one stable state to another occurs due to electron-conformational transition. The understanding that they may be promising elements of electronic circuits dates back to the 1980–1990s of the last century as a result of the rapid development of theoretical foundations of molecular electronics. The last step that remained to be made was to find the specific molecules of this type which would meet the expectations of the theoreticians.

The capabilities of the synthetic methods of organic chemistry increased dramatically in the second half of the last century. The development of the theory of chemical structure on the basis of quantum-chemical concepts dramatically increased technical capabilities both of synthesizing new compounds and determining their structure—all of this together allowed for a synthesis of a number of molecules, of which synthetic chemists had dreamed for decades. Among them were the molecules constructed from cycles threaded through each other, i.e., chained fragments of cyclic compounds (Fig. 3.35). Their name—catenanes—is derived from the Latin word catena (chain). Rotaxanes are structurally similar to catenanes. In these chemical compounds, a linear chain of molecular fragments penetrates one or several cyclic groups. Three-dimensional molecular groups at the ends of the chain preclude the collapse of the molecule due to the elimination of a cyclic fragment from the chain. Similar structures involving purely mechanical connection of fragments are called molecules "without a chemical bond."

Synthesizing such molecules is clearly a very challenging task. Only in the early 1980s of the last century, a group of French synthetic chemists proposed fundamentally new ways of synthesis. Without going into details, we note that the synthesis was based on the idea of self-organization of the molecule's structure and on the analogy with the biological principles of the synthesis of large molecules.

It is these molecules of catenanes and rotaxanes that turned out to be the basis of a breakthrough in the creation of molecular electronic circuits in the end of the past century and in the beginning of the current century. It was initiated by the well-known American company Hewlett-Packard and the University of California, Los Angeles.

Over the past decades, Hewlett-Packard took interest in developments in the area of molecular technologies. This interest dramatically increased in the late 1990s, particularly after Stan Williams, a physicist at the company's research lab,

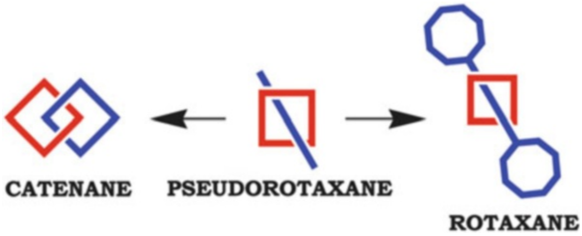


Fig. 3.35 Principles of catenane and rotaxane structures

organized a group of 15 people focusing exclusively on molecular switching elements. Williams said in an interview:

“I would say let someone else working in this field, say ‘Let’s make diodes and transistors’. Instead of re-inventing the silicon electronics at the level of individual elements, we’re going to attack at a higher level. What we want should look as a fully integrated device.” Hewlett-Packard started to work in the field of molecular components in close collaboration with chemists from the University of California, Los Angeles, Fraser Stoddart and James Heath. At this time, they studied the possibility of using molecular catenanes to create molecular switches. They chose the molecule which was a combination of the cyclic cyclophane molecule (Fig. 3.36) containing two bipyridine groups connected to a complex crown ether molecule. Crown ethers are well-known nonplanar cyclic molecules, in the central cavity of which other comparatively small atomic fragments can be built in. In particular, crown ethers retain well metal ions, which is widely used in practice. The crown ether used for the formation of the catenane molecule has a tetrafulvalene group and dioxynaphthalene fragment built in on opposite sides of the crown ring. Thus, the catenane molecule involved linked cyclic fragments, one of which contained an electron donor, tetrathiafulvalene, and the second one—a cyclophane ring with an electron acceptor (bipyridine). Electron transition from donor to acceptor leads to Coulomb repulsion of the ring fragments, and the molecule undergoes a conformational change resulting in increased distance between the charged groups.

The process of conformational rearrangement is illustrated in Fig. 3.36 which shows a three-dimensional structural model of the catenane molecule. Its compact size and the flexibility of the crown ether ring should greatly facilitate this process.

Based on the catenane molecule considered above, an electric current switch was created. The molecule was placed between a polycrystalline layer of *n*-type silicon and a metal electrode. The switch opened when the potential of 2 V was applied, it closed at −2 V, and its state could be read at 0.1 V (Fig. 3.37).

At the end of the 1990s, the group involving Hewlett-Packard and the University of California set a goal to develop a working model of a molecular mass storage device. The rotaxane R molecule (Fig. 3.38) was chosen as a molecular memory element.

A typical rotaxane molecule is a polyester chain on which a cyclic molecular fragment is mounted like a bead on a string. This fragment with embedded electron acceptor group can move along the chain. In order for the cyclic fragment not to slip

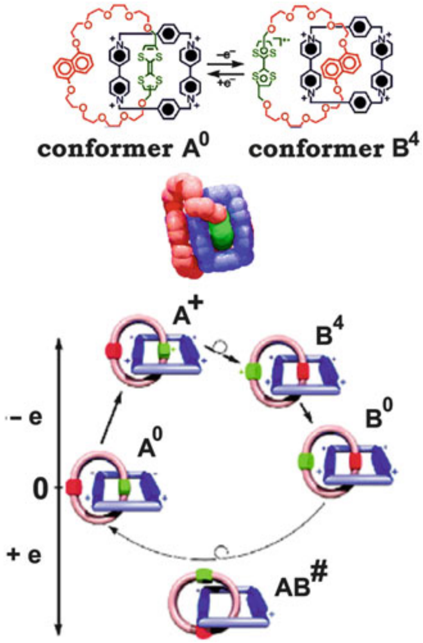


Fig. 3.36 Main states of catenane molecules

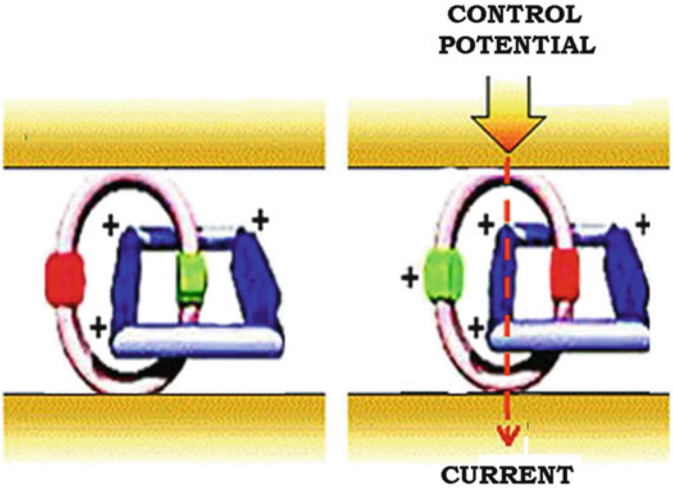


Fig. 3.37 Switch based on the catenane molecule

off the chain as a result of external forces, three-dimensional molecular stopper groups are placed at its ends. The characteristic property of the chain is that it contains electron donor groups. Through their interaction with the electron acceptor

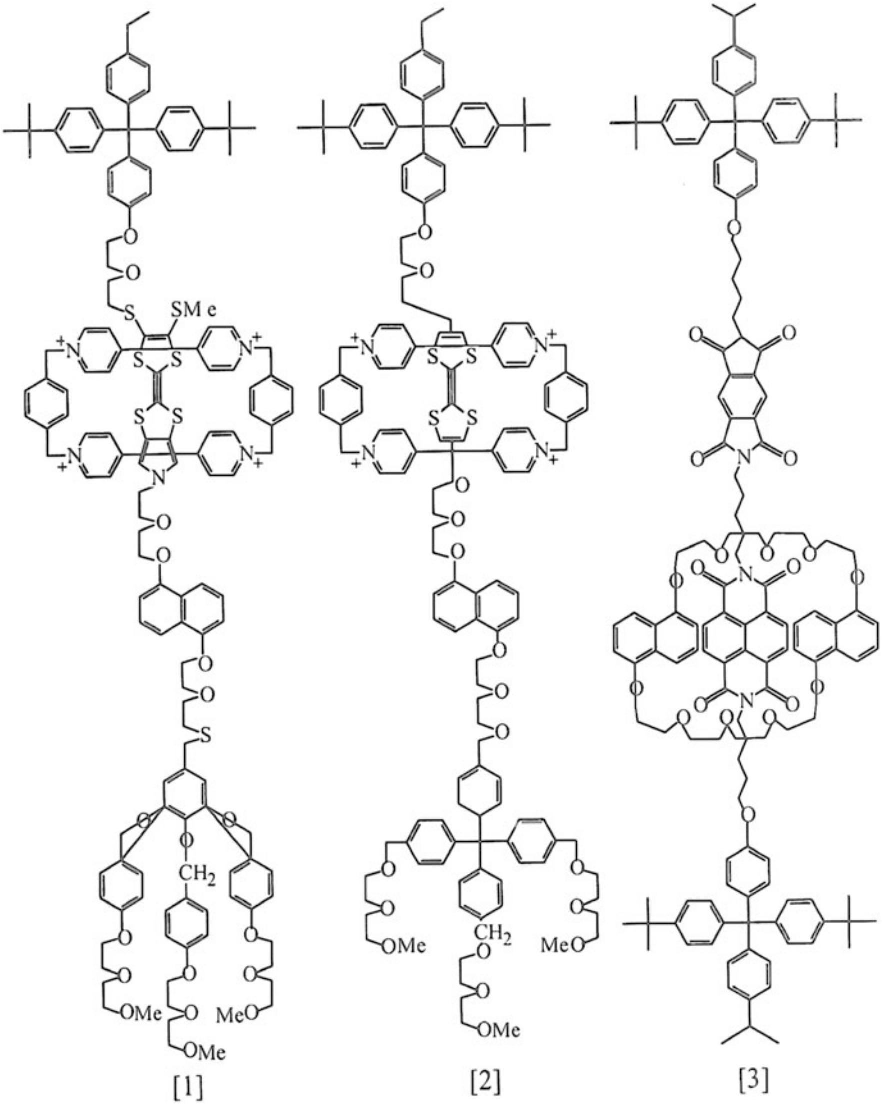


Fig. 3.38 Rotaxane molecules

groups of the cyclic fragment, dominant conformations arise in the rotaxane molecule. This is the mutual arrangement of the cycle and chain corresponding to the minimum potential energy of the system. A remarkable property of rotaxane molecules is that the transitions from one conformation to another can be controlled by both chemical (protonation of groups) and electrical (electron transfer) stimuli.

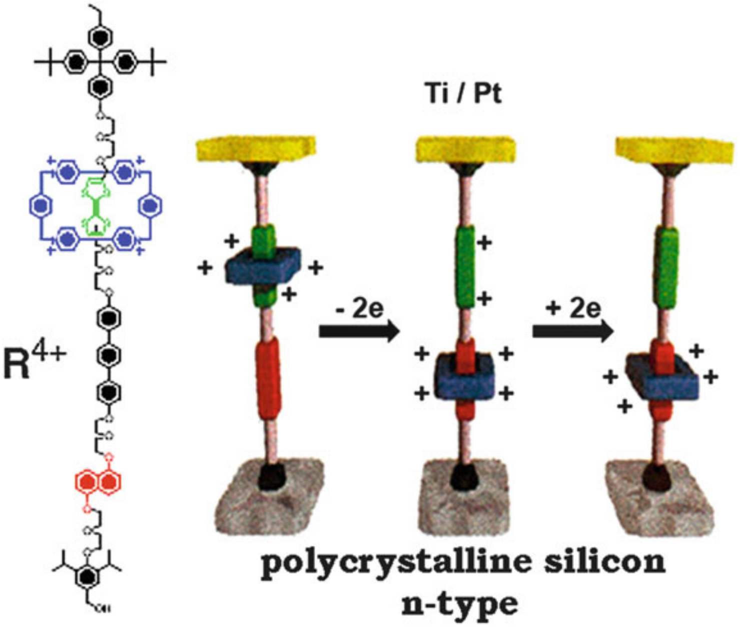


Fig. 3.39 Switch based on the rotaxane molecule

In the molecule of [2] rotaxane R, selected by the Hewlett-Packard, University of California group, an electron donor group—tetrathiafulvalene—is built into the main chain, and the cycle moving along the chain is constituted by a tetracation nitric group, an electron acceptor. A conformational rearrangement of the molecule occurs (Fig. 3.39) when the donor loses an electron. As a result, an electrostatic repulsion of the cycle from the donor group occurs with subsequent rearrangement of the molecular structure of the rotaxane. The peculiarity of this molecule is its amphiphilic character.

This term implies that the molecule is a long molecular chain with pronounced hydrophobic properties. A hydrophilic group is located at one end of this chain. This specific structure of amphiphilic molecules underlies the Langmuir–Blodgett method for forming structurally ordered monomolecular films.

This method involves pouring the solution of amphiphilic compounds on water surface using the device called Langmuir–Blodgett trough (Fig. 3.40). Due to hydrophobicity of the chains, molecules position themselves on the water surface in an arbitrary manner, although the terminal hydrophilic groups actively interact with water. A barrier swept across the surface compresses the film, with hydrophilic groups remaining in the water and the hydrophobic tails lining up perpendicular to the surface. If a solid substrate is introduced into the film being compressed and is gradually extracted from the trough upon film contraction, the monomolecular layer of the amphiphilic substance gets transferred to the substrate. Repeating this process many times, one can obtain multimolecular layers.

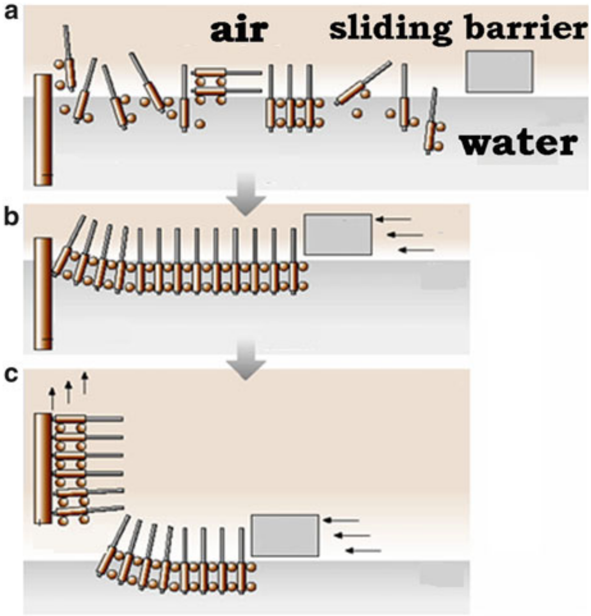


Fig. 3.40 Scheme of the formation of molecular layers using the Langmuir–Blodgett method: (**a**) amphiphilic molecules on the water surface, (**b**) compression of the film of amphiphilic molecules, (**c**) transfer of the film to the substrate

The electrical characteristics of monomolecular films of [2] rotaxane R, measured in the course of experiments, showed that the chosen molecule switches from one state to another due to a voltage in the range between −2 V and +2 V applied to the film (Fig. 3.41). To fabricate a memory device based on the molecules of [2] rotaxane R, a unique technique was used employing the technological principle "bottom-up." The main characteristic features of this technique are:

- The so-called crossbar architecture of the memory device
- Imprint lithography which allows to form nanometer-size elements on the substrate surface
- The Langmuir–Blodgett method for fabricating monomolecular films of oriented [2] rotaxane R molecules

The crossbar architecture (Fig. 3.42) comprises two sets of linear electrodes placed perpendicular to each other, with rotaxane molecules situated between them. Thus, the *i*,*j*th memory element is the entire population of molecules between the *i*th and the *j*th electrodes. A set of electrodes is transferred onto the substrate surface by imprint lithography (Fig. 3.43). To this end a thin polymer layer, with a system of grooves for the electrodes imprinted by a solid stamp, is applied to the substrate. After removing the stamp and polymer residues, metal is sprayed into the grooves to fill them. In a final step the polymer layer is removed from the substrate.

To assemble a storage device monomolecular rotaxane film was applied to the substrate with the generated set of electrodes, and on top of it a second substrate was placed with the electrodes perpendicular to the electrodes of the first electrode set (Fig. 3.44).

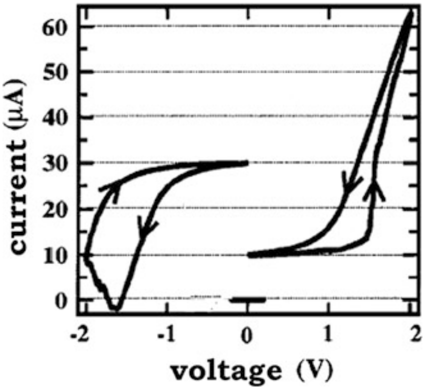


Fig. 3.41 Switch of the rotaxane molecule by the electric potential

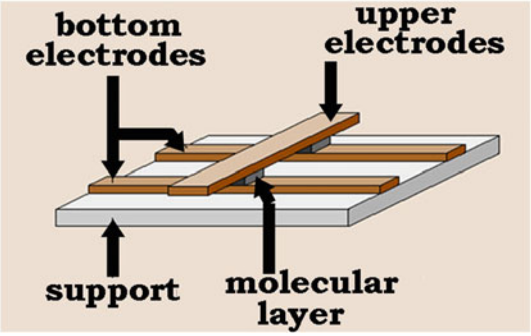


Fig. 3.42 "Crossbar" architecture of the memory device

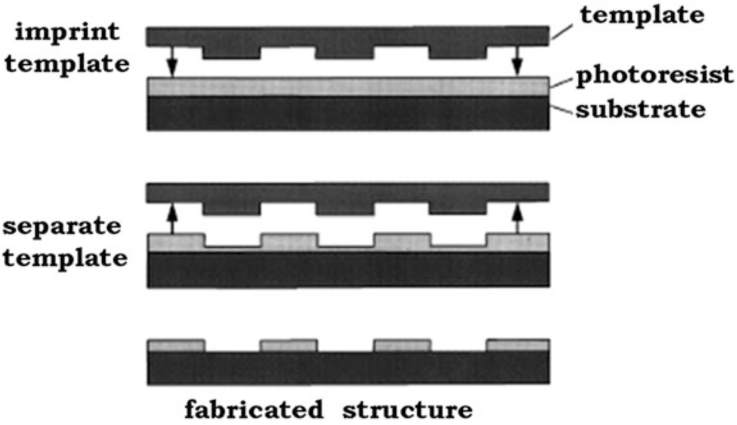


Fig. 3.43 Scheme of imprint lithography

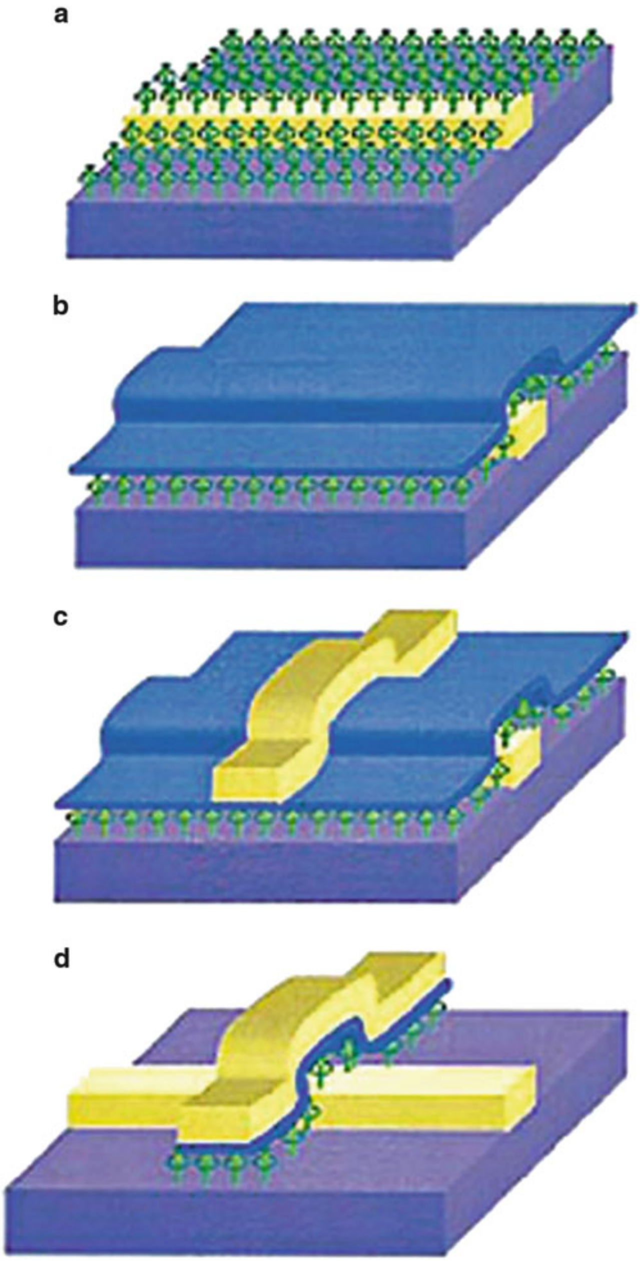


Fig. 3.44 Scheme of the formation of a molecular device based on rotaxane: (**a**)–(**d**) successive stages of the formation of a switch

Overall, the results received by the Hewlett-Packard, University of California group, were aptly summarized by one of the project participants, Yong Chan: "... a base element of the device – Pt/molecules/Ti is formed, which acts as a switching device. Cross 8×8 memory modules were formed on the basis of 64 switches."

Of course, this work (reference 8 in this chapter) is considered by the authors only as the beginning of complex scientific and technological research which will ultimately lead to industrial production of innovative digital computing devices. Hence, of natural interest is the developers' vision of the future of these devices.

In 1996, in an interview following the publication of the first results, James Heath, a team leader at the University of California, explained the newly discovered characteristics of molecular circuits. He drew particular attention to the fact that molecular circuits perform the same functions as the silicon ones, but many times better. First of all, they are much smaller, faster, and cheaper. Moreover, technological defects posing a significant problem in the manufacturing of semiconductor integrated circuits do not occur in molecular circuits. In particular, Heath pointed out that the power consumption of molecular circuits is orders of magnitude smaller than that of semiconductor circuits.

Although there still remain a number of complex problems, such as the optimal architecture of huge arrays of molecular elements, Heath believes that the first prototypes of molecular logic and storage circuits will appear in a few years. In his opinion, in the end of the first decade, the first hybrid computer with molecular logic and memory elements will be created.

So the wait will not be long.

3.6 One More Attempt: Chiropticene

But can a single molecule be used as a switching element of integrated circuits after all?

In March 1997 the California Molecular Electronics Corporation (CALMEC) was founded by leading scientists in the field of switching and conducting molecules: Michael P. Cava, Robert M. Metzger, Joseph Michl, Chad Mirkin, Mark A Ratner, Robert R. Schumaker, and Fred Wudl. James J. Marek, an expert possessing extensive experience in engineering and organizational work in large electronic companies, became the president of the corporation.

As stated by Marek in his interview to "Nano Magazine," the main mission of the corporation is to explore the concepts of constructing molecular information processing, prototyping devices with subsequent licensing, and commercialization by other companies. Furthermore, the company sought to organize, either internally or with selected partners, manufacturing of molecular objects, which would be the signature product of the corporation, and to stay their leading distributor.

The central know-how of the corporation was the creation of chiropticene—a switchable molecule whose structure has absorbed the invaluable experience of the company's founders and possesses the classical switching properties. The structure

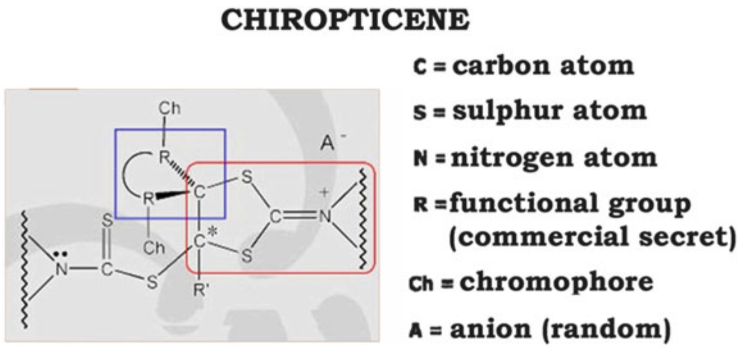


Fig. 3.45 Chiropticene molecule

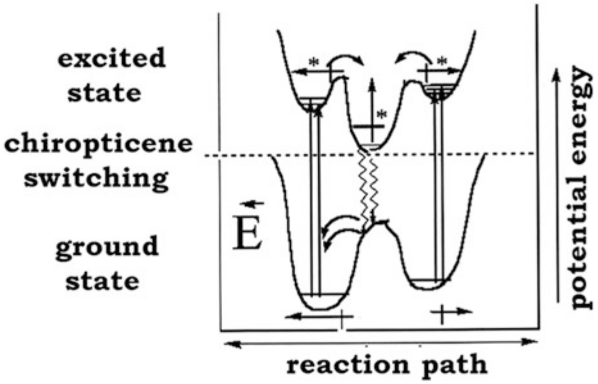


Fig. 3.46 Scheme of the intramolecular transitions in the chiropticene molecule

of this molecule is schematically shown in Fig. 3.45 (the exact structure is a commercial secret of the corporation). This optically active molecule was not discovered, but rather invented. Its two optical isomers—two stable states of the molecule—have oppositely directed dipole moments, making it possible to use electrical signals for switching. The barrier between the conformations is large enough to eliminate accidental switching caused by random fluctuations, impeding at the same time the switching by an electrical signal. For this reason, a photosensitive group is introduced into the structure of the molecule, and the characteristics of the molecular system are chosen such that the energy barrier between the conformations of the molecule in its excited state would be significantly smaller than in the ground state (Fig. 3.46). Thus, the molecule gets switched by a simultaneous effect of light radiation and electric field. The absorption band of the photosensitive group, whose structure is a commercial secret, coincides with the wavelength of a common laser. Under this scheme the switching time is in the femtosecond range, making chiropticene a fast switching element.

The Corporation (or more accurately, Robert Shoemaker, supported by the corporation) obtained two patents, of which one defines the structure of the chiropticene molecule and the other one the technology for creating three-dimensional systems based on these molecules. The technology utilizes modern chemical methods for forming molecular environments, including Langmuir–Blodgett films.

James Marek determines the prospects of chiropticene as an opportunity to create memory devices with a capacity of about 16 terabits per cubic inch. This is approximately two trillion data units. According to Marek, the molecular memory developed stores in the same volume 34 times more information than the modern solid-state memory. The design of the memory device allows for parallel writing/reading of data. It is capable of reading a million bits of information in a single read operation, with 2,000 read operations per second, thus achieving the read/write speed of about 2 gigabits per second.

Remarkable advances of the developers of chiropticene and of rotaxane-based memory represent the cutting edge of modern technology. But it will take more time and effort to make these devices commercially available.

Chapter 4
Seemingly Incompatible: Chemical Reaction–Diffusion Media and Artificial Intelligence

The variety of phenomena that may be described by this sort of reaction-diffusion equation is quite amazing
I. Prigogine. "*From Being to Becoming*"

4.1 Information Requirements of the Postindustrial Society and the von Neumann Paradigm

In the early 1940s of the last century humanity faced grand technical and technological challenges associated with enormous demand for computing. The John von Neumann paradigm proposed at that time, which laid the foundation for designing digital computers, made it possible to create effective computing devices to address the most pressing engineering problems and to continue improving them up to the present day. The success of the von Neumann paradigm rested on its suitability for solving massive computational tasks which were at the core of the most important engineering problems. Human information processing style discussed in the 1940s—the neural network approach of McCulloch and Pitts—turned out to be premature, since the main trends in the development of information-logical devices in those years were determined by engineering and technological computing tasks that could be effectively solved with digital computers.

Development priorities and needs of the human society were changing during the second half of the last century. And, since the 1980s–1990s, along with engineering problems, the understanding of the dynamics of large dynamic systems and the development of methods to control them was becoming increasingly relevant. This applies to a wide range of objects, from biological communities (a group of apes, a wolf pack, an anthill) to physical and chemical environments with complex nonlinear interactions between components. The processes in such systems play an important role in nature and in human society. They become apparent in a variety of physical, chemical, and biological objects, in large transportation systems, as well as in human communities, in particular in the form of various economic and sociological phenomena. Their study allows us to understand and prevent epidemic diseases, to create

© Springer-Verlag Wien 2014
N.G. Rambidi, *Molecular Computing*, DOI 10.1007/978-3-211-99699-7_4

autonomous devices capable of replacing the man in critical conditions, to reliably predict weather, and to regulate the economy and social problems.

In general investigation of the mechanisms of functioning of large dynamic systems and their external manifestations involves a series of information tasks, such as:

- Recognition of images, scenes, and situations which may be somewhat arbitrarily reduced to:
  - Classification of objects based on a given set of features
  - Object segmentation, i.e., subdividing it into a set of simpler fragments
  - Context-based recognition of fragments with subsequent construction of the source object (important in such areas as medicine, material science, etc.)
- Study of evolution of systems with complex dynamics of behaviors (e.g., predator–prey problems, evolution of populations of biological cells, etc.)
- Selection of the optimal (in some predefined sense) structure or behavior of a complex multifactor system with a complex decision tree ("the traveling salesman problem," strategic and tactical decision games)
- Control problems, which include:
  - Continuous recognition of situations
  - A continuous selection of the optimal strategy (navigation of autonomous robots operating in a complex changing environment and other similar devices)

A fundamental feature of large dynamic systems is their distributed nature. This means that the system represents a huge collection of simple (for this system) elements which interact with each other. At the level of the human society such an element is a single person in the crowd, in physicochemical reaction media—individual molecular components of the reactions taking place in the system. But regardless of the level of complexity of behavior of a separate element, distributed systems are characterized by common structural and behavioral characteristics.

Processes in a distributed system (environment) occur simultaneously at its each point. In other words, it is characterized by high (or rather gigantic, compared, e.g., with modern parallel digital computers) parallelism of the system's actions.

In a large number of distributed systems nonlinear interactions between individual elements of the system occur. This leads to much more complex behavior of the system than that of its individual elements. A simple but impressive example, frequently used in recent years, is the activity of ant communities. One of the main functions of the community is the search for and delivery of food to the anthill. Elementary actions of an individual person are simple: random search, exudation of specific chemicals (pheromones) once food has been found and is being transported to the anthill, and following the pheromone trail. These simple mechanisms help memorize the way to accidentally found food, involve other individuals in this process, optimize delivery routes, etc. A characteristic feature of distributed systems with nonlinear interactions of the elements is that the behavior of the system as a whole is much more complex than that of its individual elements. Moreover, it cannot be derived from the behavior of the elements. For example, there is a

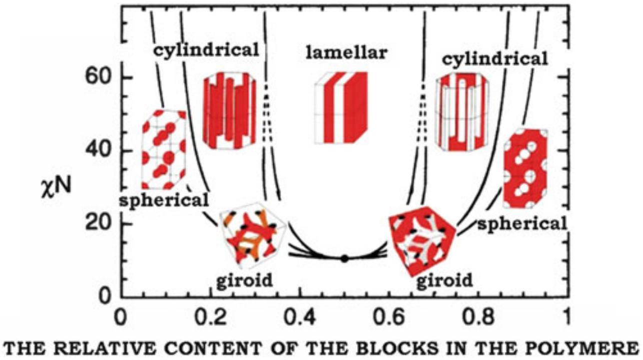


Fig. 4.1 "Traveling salesman problem"

complex implicit dependence between the optimization of food delivery routes to the anthill complex and the mechanisms of behavior of individual ants. This effect, called emerging mechanisms, will be discussed in more detail below.

Distributed systems are characterized by multilevel operation. In the ant community there is a strict differentiation of persons in terms of actions they perform—searching for food, protection from enemies, etc. This differentiation significantly increases the efficiency of the system in terms of the main goal—survival in a complex, dangerous, and changing environment.

In general information tasks required to understand the processes occurring in large dynamic systems and to control them proved to be a stumbling block for digital computers with von Neumann architecture due to their high computational complexity. As the size of the problem (the required accuracy of calculations, dimensionality, etc.) grows, resources needed for its solution, i.e., the required number of computer cycles or the number of memory cells, grow sharply and in some cases exponentially. An impressive example is the famous "traveling salesman problem" (Fig. 4.1).

Suppose we have a number of cities to be visited. The traveling salesman problem involves finding the shortest possible tour that visits each city exactly once. Let us denote the start and end points of the route by $N(0)$ and $N(\mathrm{max})$. Then the route can be written as $N(0) - 1 - 2 - \ldots - N - N(\mathrm{max})$. It is easy to see that the number of all possible routes will be $N!$ Suppose that the problem is solved by brute force, i.e., by comparing the lengths of all paths. For $N = 4$ we need to compare the 24 routes, but if the number of cities increases to just 10, the number of routes will have a huge value of 3,628,800.

4.1.1 Some Details: Structural, Behavioral, and Computational Complexity

The world around us is complex. This is a reality that defines our understanding of the phenomena encountered every day. They become apparent in various fields of human activity, from isolated, seemingly simple engineering tasks to sophisticated economic and social problems.

Stafford Beer, a former president of the Operations Research Society of America, viewed complexity as an inherent property of our world. He pointed out:

"There exist the real world, which can be observed and measured, and a reality, in which a dead man is a dead human being rather than an object of statistics. There also exists a reality that we encounter every day. It is not arranged in orderly pigeonholes and is not provided with labels for the corresponding bureaucrats who make decisions. The essence of such reality is complexity."

Among the many aspects of the complexity of the world around us, there are three basic concepts that are fundamental to the concept of complexity. By defining them, the famous American mathematician John Casti believed that:

"Static complexity describes the formation of the system from its component subsystems, dynamic complexity is based on computing the length which is required to describe the behavior of the system, the complexity of governance is a measure of computational resources required for detailed calculation of the dynamics of the system."

Preserving the essence of these definitions, in what follows we will introduce a somewhat different terminology more widely adopted today to describe physical, chemical, and biological systems. We will use the terms:

- Structural complexity of the system (static)
- Behavioral complexity (dynamic), which determines the spatiotemporal evolution of the system which carries out information processing operations
- Computational complexity of the algorithm (control complexity), which describes the information processing operations being performed

A great contribution to the concept of complexity was made by Academician A. N. Kolmogorov, who created the foundations for its quantitative description. His concept of algorithmic complexity is arguably the most adequate description of the evolution of dynamic systems.

Let there be a system that converts information at system's input $x \in X$ into information at its $y \in Y$ output according to a certain program p. Here X and Y are sets of values x and y. The structure of the system is defined as a description of its hardware characteristics, i.e., some function $\varphi(p/x) = y$ which can be used by the program p. Importantly, this description does not depend on the choice of input data.

The algorithmic complexity of the process which takes place in the system with structure $\varphi(p/x) = y$, i.e., behavioral complexity of the system, is defined as the minimum length of the program $l(p)$ out of many programs S which adequately describe the process:

$$K_{\varphi}(y/x) = \begin{cases} \min l(p): & \varphi(p/x) = y, \\ \infty : \forall p \in S & \varphi(p/x) \neq y. \end{cases}$$

Based on the definition of the system's structure, one can introduce the definition of its structural complexity $K(x/y)$, as complexity with fixed values $x = x_0$:

$$K_{\varphi}(y_0/x_0) = \begin{cases} \min l(p): & \varphi(p/x_0) = y_0, \\ \infty : \forall p \in S & \varphi(p/x_0) \neq y_0. \end{cases}$$

Computational complexity of the algorithm describing the behavior of the system (the complexity of the problem) characterizes the practical feasibility of understanding the behavior of the system in detail. It can be reduced to the dependence of computational capacity (resources of the computer system) required to model the behavior of the system, from the specific characteristics of the problem—the size of the problem.

Different versions of the functions that characterize the computational complexity (so-called signaling functions) are used for its quantitative assessment. The most common among them are:

Time function:

$$t_A(n) = \max_{|x|=n} t_A(x).$$

Here $t_A(x)$ is the number of steps of the program that implements the algorithm A to solve the individual problem x, whose size is determined by word length $|x|$ equal to n.

Spatial function:

$$S_A(n) = \max_{|x|=n} S_A(x),$$

where $S_A(x)$ is the number of memory cells required to solve the problem A.

The algorithm for solving the mass problem is called polynomial algorithm if the signaling function depends polynomially on the size of the problem. If, however, the signaling function depends on the size of the problem exponentially, it is defined as intractable. At present time a detailed classification of hard problems is utilized (NP, NP-complete, etc.).

Speaking about the structural and behavioral complexity of an information processing system and about computational complexity of the tasks being solved by the system, particular properties of the concept being used should be emphasized. Unfortunately, so far no sufficiently strong correlation between these characteristics has been established.

It is popularly believed (mainly in connection with the practice of creating wireless devices) that increasing complexity of the device leads to more complex behavior. Nevertheless, today we know a large number of objects in which simple structure coexists with very complex behavior. The most well-known example is the Hénon problem describing the dynamics of two harmonic oscillators with nonlinear coupling. This seemingly simple system demonstrates a great variety of different modes, ranging from trivial harmonic vibrations to the state of chaos.

Even more uncertain is the relationship between computational complexity of problems and the structure of the devices used to solve them, which is particularly important for building computational tools that optimally solve problems of high computational complexity.

Apparently, one of the most promising opportunities to find constructive principles for such devices is constituted by heuristic approaches. In particular, it can be assumed that the ability to effectively carry out operations of high computational complexity is inherent to systems with highly complex behavior. Therefore, the problem of designing effective devices for solving problems of high computational complexity becomes a problem of finding the design principles of systems showing high behavioral complexity. Given the main factors determining the functioning of distributed systems, the following principles can be chosen:

- Distributed nature of the system, resulting in high parallelism of processes in the system
- The complexity of system's dynamics, which leads to logically complex local behavior
- Layered system architecture

In the following, an attempt will be made to show that consistent use of these principles in designing information processing devices leads to a new paradigm which is fundamentally different from the von Neumann paradigm and to optimal solutions for problems of high computational complexity.

4.2 Computer Engineering and the Problem of Artificial Intelligence

Modern digital computers, the principles of information processing, and the information technology have radically changed the world around us. As a consequence, in the second half of the last century the dominant paradigm of von Neumann for constructing information processing devices seemed most optimal, if not the only possible one. Explicitly or implicitly, it was assumed that the progress of semiconductor technology, combined with new physical ideas, would lead to a further increase in productivity of digital (in fact, von Neumann) devices that will be able to solve all or almost all pressing information tasks. Nevertheless, the ever-growing significance of information for large dynamic systems has substantially undermined this widely accepted view.

In the 1980s of the last century, the famous American cybernetician Michael Arbib expressed his belief that further development of computing concepts will take the path of imitating the style of information processing by the human brain. "The human brain is a metaphor of the next (sixth)-generation computers"—he wrote—"Its style is determined by the interaction between systems, many of which correspond to the joint operation of the spatio-temporal structures in layered neuron structures."

The basis for understanding of the importance of this approach, as noted Arbib, is that:

"The human brain is a calculator oriented on action. This implies that such systems (of a man, animal or remote controlled robot) must correlate the action itself and its result so that to build an internal «model» of a phenomenon.

-The human brain has a hierarchical layered organization. It is extremely important that no single-level model is able to reproduce its functions.

-The human brain is not a system of the information processing with the consequent execution of operations."

In the late 1990s of the last century, Hans Moravec published a forecast for the development of computer technology, which he called: *When the existing computers will be equal to the human brain?* Based on the number of neurons in the brain and the number of synapses (10^{11} and 10^{14}, respectively) and upon having analyzed Man vs. Machine chess game, he concluded that the performance of a digital computer (~10^8 MIPS) will be sufficient to simulate the functions of the human brain. According to his estimates such computer systems should appear in the first half of the twenty-first century, not earlier than 2020.

In this article, attention was drawn to three features of modern information processing devices that are important for understanding the problem being considered.

First of all, virtually any discussion of the perspectives of development of the digital technology is related to solving intellectual problems (problems of artificial intelligence). Indeed, this direction has acquired critical importance during the last decades due to the need to understand and effectively control large dynamic systems. As noted above, economic and social problems, traffic management, global communication, weather prediction, assessment of environmental pollution, and a host of other equally important aspects of modern society are among the most pressing tasks. Virtually all of these tasks can be better handled by man than by modern digital computer and can be attributed to problems of artificial intelligence.

Secondly, Moravec notes that some users of supercomputers, including G. Kasparov, "feel" the rudiments of intelligence in computers. But this can rather be attributed to self-deceptiveness. Even the designers of Deep Blue—the computer which played with Kasparov—deny the presence of intelligence in the machine, believing that it is "only a big database on openings and endgames, incorporated and advanced functions, which are proposed by chess grandmasters, and, especially, the high efficiency, making it possible to examine the current situation on 14 moves onward."

The third important observation is that a digital computer with the capabilities approaching the human brain will represent a super complex system with tremendous cost of creation and operation. Therefore, computers being created are fundamentally inferior to man in terms of efficiency, simplicity, and cost of solving intellectual problems.

In general, solution of problems of high computational complexity, which basically come down to the problems of artificial intelligence, leads to a manifold increase in computational resources required to address them and forces developers of computational devices to work on further increasing storage capacity and speed of execution of elementary operations. Nevertheless, addressing these challenges with modern digital computers turns out to be ineffective and often completely impossible. An alternative to this race "computational complexity of the problem—computer performance" emerged in the end of the last century due to further development of the ideas of McCulloch and Pitts.

4.3 Biologically Inspired Information Processing Devices: Neural Networks and Neurocomputers

In 1943 McCulloch and Pitts proposed the neural network approach to processing information based on the current knowledge of the structure of the cerebral cortex.

By the time it was already known that the cortex is a complex system of interconnected nerve cells—neural network (Fig. 4.2). Each neuron has branches—dendrites—through which a neuron receives signals from other neurons. They are summed up algebraically (i.e., taking into account the sign of the incoming signal), and if the sum exceeds a certain threshold, the neuron transmits via the output branch (axon) the signal which, in principle, may reach all other network nodes. The signals at each neuron's input are controlled by groups of cells—synapses—that determine the structure of the neural network, i.e., the particular wiring of neurons. This, in turn, determines what specific tasks are solved in the cerebral cortex. Synapses play a fundamental role in learning processes. Their number and distribution varies significantly in the course of infant development.

The neural network model of McCulloch and Pitts gives a simplified description of the structure of the cerebral cortex. A neural network is a system of elementary processors—formal neurons (Fig. 4.2). Each of them receives either positive or negative signals from all neurons of the network, weighted to simulate synaptic connections. The neuron sums up these signals algebraically and, if their sum exceeds a specified threshold, generates a pulse which propagates through the network. The initial state of the network is given by modifiable weights which determine the structure of the problem to be solved by the network. Once the initial state of the network is defined, the network structure evolves over time, and its final state represents the solution of the specified task. A remarkable feature of the neural network is that information processing is carried out simultaneously by all its neurons, i.e., with tremendous parallelism unrivaled even by modern semiconductor multiprocessor computers. In contrast to the von Neumann computer, the particular task solved by the network is determined not by an input program, but rather by the initial states of neurons and the network structure—the system of weights with which neuron signals are transmitted over the network.

A major step in the development of neural network concepts was made in 1962 by the American neuroscientist Frank Rosenblatt. He suggested an arbitrary neural network structure, called the perceptron, which was based on three types of neurons

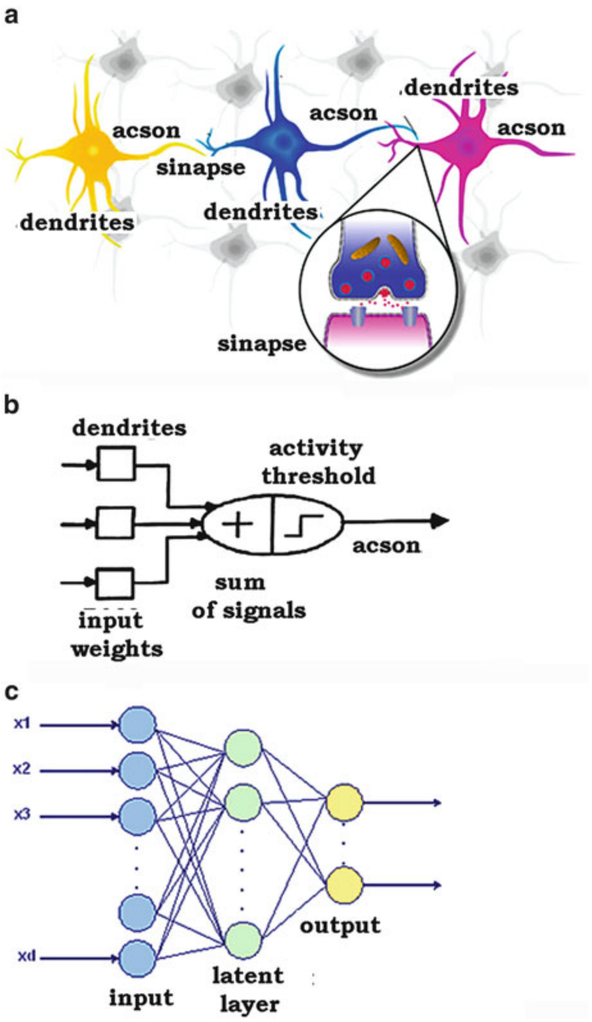


Fig. 4.2 Natural (**a**) and "formal" (**b**, **c**) neural networks

(Fig. 4.2). Sensory neurons were considered sensitive elements which generate and send a signal to the network under the influence of some external stimulus (such as electric signals, light, sound, etc.). Associative neurons were defined as logic elements that issue an output signal when the algebraic sum of input signals exceeds a certain threshold value. Reactive neurons are elements that receive input from sensory neurons and form control stimuli in the external environment. Rosenblatt developed an operating device—single-layer perceptron—capable of classifying input signals into one of two classes.

Rosenblatt's approach was used in the early 1960s to explain various psychological and physiological phenomena. But, in general, interest in them soon dropped significantly because of the absence of demand for neural network ideas

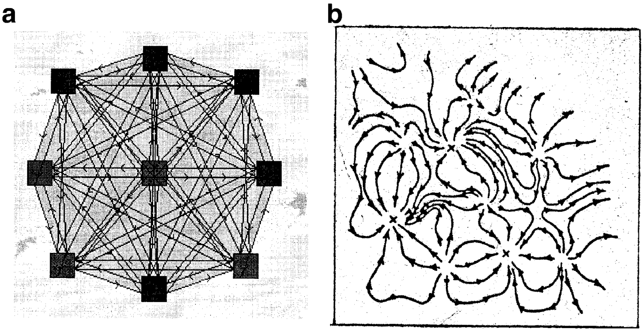


Fig. 4.3 Hopfield neural network: (**a**) network structure; (**b**) surface of the network potential

in those years. In fact, the work of Rosenblatt was subjected to criticism in 1969 by Marvin Minsky—the founder of a scientific approach to artificial intelligence—in his book *Perceptrons*, written jointly with Seymour Papert. Only about 20 years later, in 1982, American physicist Hopfield again sparked interest in neural networks, establishing a neural network model named after him.

The Hopfield neural network is a two-dimensional array of formal neurons, connected with each other in a pairwise fashion (Fig. 4.3). Each neuron is considered as an element with two possible states described by the binary variable s. One of the states corresponds to the "excited" neuron ($s = 1$), the other state to the ground state ($s = 0$). In general, the neuron is characterized by the function $f(s)$ determining its dynamics.

The status of a network of neurons N at the time t is defined as the configuration of all variables at this moment of time. The evolution of this state in phase space s_i is determined by the interaction of neurons. Neurons are connected with each other by synaptic connections. The strength of the connection between the ith and the jth neurons is characterized by the value T_{ij}. This matrix is called weight matrix (also known as the matrix of long-term memory). The condition of neurons at the initial moment of time depends on the image presented to the network.

The image is converted by the network in accordance with certain rules. Among the rules governing the evolution of different network models one property remains invariant: the change of neuron's state is determined by the total excitation reaching the neuron from all other neurons in accordance with the synaptic weights.

In the Hopfield network all neurons are mutually connected. The memory matrix T is symmetrical ($T_{ij} = T_{ji}$) and has zero diagonal elements. The function $f(s)$ is binary.

Nonlinear transformation of the original image $A = (S_1, \ldots, S_N)$ follows the rule:

$$s_i^* = f\left(\sum_{j=1}^{N} T_{ij}s_j\right).$$

Successive repetition of this transformation leads to subsequent changes in the original image.

This transformation can be performed in different ways. A neuron can be selected at random and its condition recalculated, then another neuron gets randomly selected and subjected to the same procedure, etc. Such process of network transformation is called asynchronous. If the network transformation is performed in “cycles” simultaneously for all neurons, then such a process is called synchronous.

Dynamical evolution of the Hopfield model, for which $T_{ij} = T_{ji}$ and $T_{ii} = 0$, and the transformation of the image happens synchronously, corresponds to reducing the function usually called “energy”

$$E = -\frac{1}{2}\sum_{ij} T_{ij}s_i s_j.$$

The steady state of the network corresponds to the local minimum of this function. Sequential transformation of the original image involves the motion along the phase surface to the point (more precisely, to one of the points) of a local minimum corresponding to the solution of the given problem (Fig. 4.3).

Particular properties as well as computing, information, and logical capabilities of the Hopfield model were studied in detail in the 1980s of the last century. Moreover, this model stimulated the development of a wide range of variants of both single-layer and multilayer neural networks. But the most important consequence of the rapid development of the theory of neural networks was apparently engineering development of commercial information processing devices—neurocomputers.

Let us try to understand what advantages of neural networks were considered particularly promising by the developers of neurocomputers.

The basis of the neural network approach is constituted by the biological principles of information processing and, above all, the general principles of functioning of the cerebral cortex. Therefore, one could expect that the devices that mimic biological neurons will be able, at least partially, to reproduce their function.

Over a long period of biological evolution the human brain developed properties inaccessible to modern digital computers with von Neumann architecture. These include:

- Distributed representation of information and parallel computing
- Ability to learn and generalize
- Adaptability
- Tolerance to faults and errors in the structure
- Low power consumption

Devices built on biological principles of information processing must have these capabilities, which will be of great importance for the industry of information processing.

Without attempting to outline in detail the history of neurocomputers, extensively treated in [], let us focus only on the main factors characterizing this important and rapidly developing field.

Today there are three active development tracks of neurocomputers:

First, emulators, i.e., systems based on digital von Neumann computers, implementing typical neural network operations at the software level.

Second, neuroaccelerators, neural network systems implemented on the basis of universal digital computers in the form of expansion cards. They may be both "virtual" (compatible with expansion slots of standard PCs) and "external," connecting to the host computer via a specific interface or bus.

Finally, neurocomputers, employing specialized neural chips which execute all operations in the neural network logical basis.

Without touching on neuroemulators and neuroaccelerators, we will mention briefly the basic principles of construction and functional characteristics of neurocomputers.

A special circuitry was developed in neuroinformatics for describing algorithms and designing devices, in which elementary devices are combined into networks designed to solve specific problems.

The following basic elementary devices are used (Fig. 4.4):

- Adaptive adder which computes the scalar product of the input vector x (i.e., information coming from all neurons) with the parameter vector
- Nonlinear signal converter, which receives a scalar signal x and converts it into a given function $f(x)$
- Branch point, which is used to send the incoming signal to multiple addresses

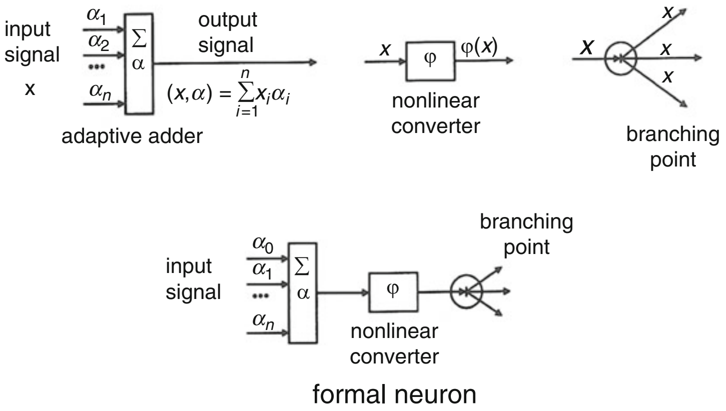


Fig. 4.4 Circuitry of neurocomputing

- Standard formal neuron, which is a combination of the input adder, the nonlinear transducer, and the branch point at the output

On the basis of these elements neural chips are produced from which the required neural network can be assembled.

Nowadays industry of various countries produces dozens of digital, analog, and hybrid neural chips, including neural chips with a rigid neuron structure (hard-wired) and neural chips with a custom structure (reprogrammable). Typically 32–256 neurons with the word length of variables 16–8 bits are formed on the chips. They are manufactured based on silicon planar semiconductor technology.

As an example, let us consider neural chips with programmable weight coefficients developed by Bell Laboratories in 1987.

In order to allow adjustment of weight coefficients of the neural network, a specialized circuit employing a combined technology was designed such that both the neural network and the memory for controlling reprogramming of weight coefficients were placed on a single chip. Digital input and output signals are used by the circuit. To calculate the output of the neural network analog circuits for multiplication and summation were used, while digital signals served to control the functioning of the circuit.

This VLSI was manufactured using the 2.5 μm CMOS technology and contained approximately 75,000 transistors. The 6.7 × 7 mm crystal accommodated 54 operating amplifiers performing the functions of neurons and 5 kilobytes of static RAM for reprogramming weight coefficients on the crystal. The circuit allowed to implement a neural network with full connectivity. The 54 × 54 matrix of connection coefficients occupied about 90 % of the crystal surface and allowed for connecting the output of any neuron to the input of any other neuron.

4.4 The Future of Neurocomputing: The Dubious Legacy of Digital Computers

The intense revival of neural representations and the development of neurocomputers became a reality in recent years.

Nevertheless, despite the fact that the neural network paradigm is rooted in the 1940s of the last century, approaches to its optimal practical implementation in hardware were hardly considered for many years. Therefore, the establishment of neurocomputers took a conventional path—the use of semiconductor circuitry and planar technology, well proven in a variety of microelectronic devices.

It must be admitted that so far, in fact, there is no alternative to semiconductor circuitry and the technology being used to simulate neurons and neural networks. This is easily explained because, due to enormous advantages of the planar technology, it all but outcompeted almost all previously proposed technological implementations of computing devices. Moreover, discrete circuitry was created to implement the von Neumann architecture. In this sense, the von Neumann architecture is optimal, and, just as important, it is far from exhausting its possibilities.

The situation changes radically when it comes to the development of neural information-logical devices.

The basic principles of the neural network paradigm are fundamentally different from the principles of the von Neumann paradigm. At the core of the von Neumann paradigm is the notion of a program with unalterable structure. Changing the program, even a single operator, leads to its disintegration and renders it unable to perform its functions.

By contrast, the ideology of neural networks, even the initial approach of McCulloch and Pitts, is fundamentally different in that it provides for the possibility of small changes in the structure of the network. The concept of variable weights of neurons allows to vary them in a certain range of values, without causing qualitative changes in the operating mode of the network.

It is precisely this feature where the discrete semiconductor implementation of neural networks fundamentally contradicts to their nature.

Removing or adding even a single transistor to a planar circuit, in general, leads to loss of its function (information redundancy is not considered here as it does not change the overall conclusions). Therefore, gradual adaptation of the circuit toward a more efficient solution of a specific problem encounters major difficulties when using semiconductor planar technology.

At the same time, known biological systems with neural network architecture are built of basic molecular fragments that are qualitatively different from semiconductor components (transistors). One of the main and probably most important features of such systems is the structural redundancy of molecular objects with respect to their functions. Thus, biopolymer molecules of protein enzymes play an important role in the functioning of biological systems. Their structure is a combination of functional groups, defining the function of the enzyme, and an extended (polypeptide) "tail." A remarkable feature of this structure is that the removal of even relatively large fragments from the tail leads to only marginal change of enzyme function.

Functional redundancy of biological systems also manifests itself when the change of dynamic characteristics of the system in a fairly wide range does not lead to qualitative changes of dynamics, i.e., to transition to another regime. This can be defined as a dynamic redundancy of the system (see below).

In general, structural and (or) dynamic redundancy of elements (molecular fragments) used to construct logic information systems constitutes the basis of their variability. This, in turn, should be the basis of evolutionary selection. Therefore, such source elements may be, in principle, utilized to build information-logical devices, capable of gradually learning the most efficient solution of the problem in the course of the decision process.

4.5 Reaction–Diffusion Information Processing Devices

Distributed (continuous and discrete) systems represent one of the most important types of natural objects, possessing high complexity of behavior and capable of focused action. It is these systems that have attracted attention in recent years as a promising basis for the establishment of efficient biologically motivated information processing tools. In this case, processing of information occurs at every point of the physical environment, which leads to a high degree of parallelism not comparable with the possibilities of parallel computing by digital discrete processors.

The dynamics of distributed environments is described by systems of differential equations giving the local changes in concentrations of the medium components (u_i) during the evolution of the system in space and time:

$$\frac{\partial u_i}{\partial_t} = F_i(u_1, u_2, u_3 \ldots u_N) + \sum_j D_{ij} \Delta u_j.$$

Here, the functions $F_i(u_1, u_2, u_3 \ldots u_N)$ describe the local dynamics of the interaction between the components of the medium, and the second term on the right corresponds to the diffusion of its components.

In the state space corresponding to a reaction–diffusion system, basins of attractors appear, i.e., certain dynamic regimes in which the system spontaneously comes over into a stationary state. However, this movement within the basin does not lead to qualitative changes of dynamics, i.e., to transition to another regime (another attractor's basin). This can be defined as a dynamic redundancy in the system. Therefore, there exists a fundamental possibility to create on the basis of reaction–diffusion media devices that are capable, within certain limits, of changing their functions under the influence of external factors, i.e., possessing the ability to learn.

The equations become considerably simpler in the case of continuous complete intermixing of the components of the system. Due to the uniform distribution of the components of the environment, diffusion mechanisms do not play any role, and the dynamics is determined solely by the mechanisms of interaction between the components:

$$\frac{\partial u_i}{\partial t} = F_i(u_1, u_2, u_3 \ldots u_N).$$

The media described by such systems of equations are called the reaction–diffusion media. Modes of operation of reaction–diffusion media are determined by the function $F_i(u_1, u_2, u_3 \ldots u_N)$. From the standpoint of information processing, the most interesting media are those with nonlinear mechanisms of interaction between components. They demonstrate a high complexity of behavior and carry out basic operations of information processing that are logically complex actions,

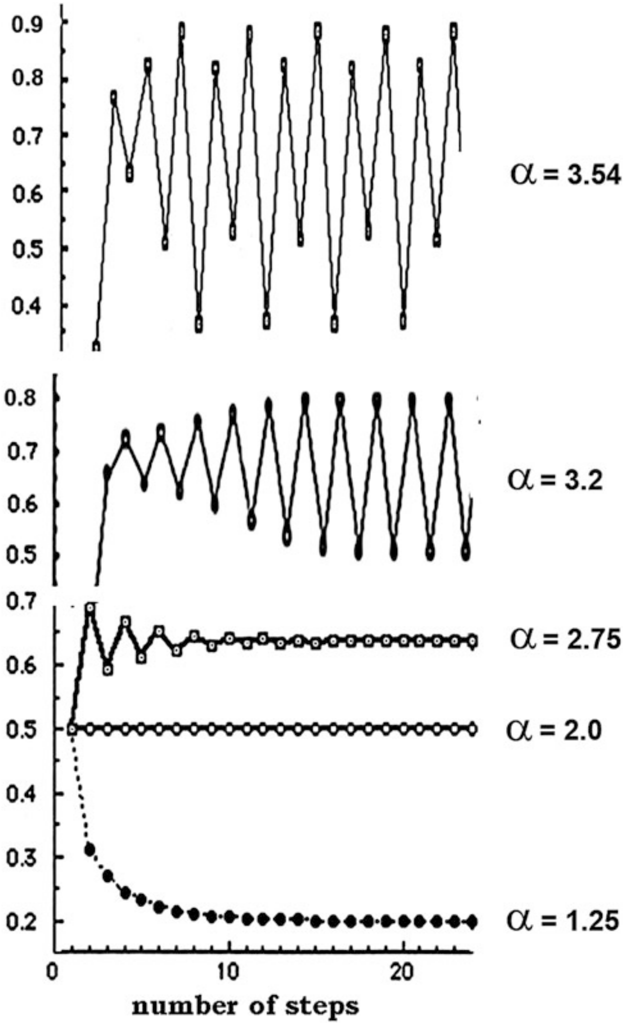


Fig. 4.5 Collective processes in complex dynamic systems

equivalent to tens or sometimes hundreds of elementary binary operations of a digital computer.

In nature, reaction–diffusion media are found at different levels of structural organization (Fig. 4.5).

Thus, complex behavior is manifested at the level of colonies of unicellular organisms during their growth, leading to nontrivial spatial structures (concentric, circular, or spiral formations).

At the level of body tissues, an obvious example is the function of the cerebral cortex. At the same time, reaction–diffusion phenomena are characteristic for other organs. For example, cardiac arrhythmias and the phenomenon of sudden death occur as a result of pathological wave modes in the myocardium.

Widely known are the modes of concentration fluctuations in chemical and biochemical systems, in biological membranes and cells, i.e., on the supramolecular level.

Finally, nonlinear dynamics can lead to collective excitations, the so-called solitons which are waves propagating over long distances along the molecular core, i.e., at the molecular level.

A remarkable feature of these environments is that regardless of the physical implementation, they all show the same macroscopic behavior:

- Local or global (throughout the entire volume of the medium) fluctuations in the concentrations of medium components
- Local areas of high concentration (concentration pulses) propagating in the bulk of the medium
- Trigger modes propagating in the medium switching from one state to another one
- Formation of stable and persistent over time dissipative structures with a nonuniform distribution of concentrations of medium components

Chemical reaction–diffusion systems and, above all, those media in which Belousov–Zhabotinsky reactions take place apparently show the best promise for creating neuro-like information processing tools. Trigger mode and the "leading center" mode of these media are shown in Fig. 4.6.

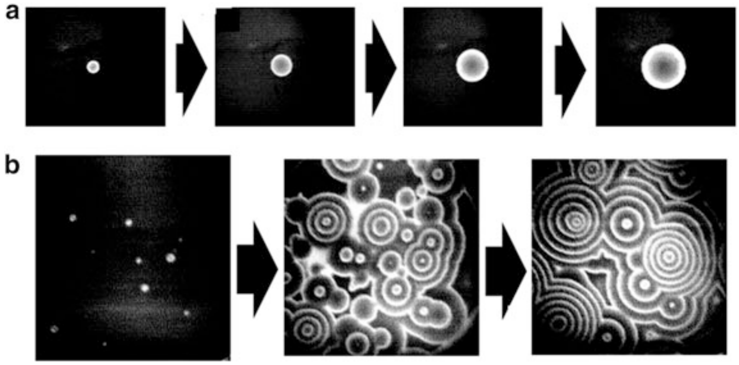


Fig. 4.6 Trigger mode (**a**) and the "leading center" mode (**b**) in the Belousov–Zhabotinsky medium

4.6 Media of the Belousov–Zhabotinsky Type: A Neural Network Architecture

The most famous among the chemical reaction–diffusion media used to create information processing devices are the media of the Belousov–Zhabotinsky type. Dynamics of this environment is based on the oxidation of an organic compound (malonic acid $C_3H_4O_4$) by an inorganic oxidant (sodium or potassium bromate), catalyzed by transition metal ions (mostly iron). Schematically, the reaction can be described by the following equation:

$$C_3H_4O_4 + NaBrO_3 + H^+ \overset{Fe}{\rightarrow} C_3H_3BrO_4 + H_2O + CO_2.$$

In fact, the Belousov–Zhabotinsky process is a set of intermediate reactions, the exact number of which has not yet been fully established.

Let us consider the simplest version of the chemical reaction–diffusion medium—a flat quasi two-dimensional layer (Fig. 4.7). For a clearer understanding of the information properties of such a system, it is convenient to subdivide it into microvolumes with linear dimensions smaller than the diffusion length of the medium. This value

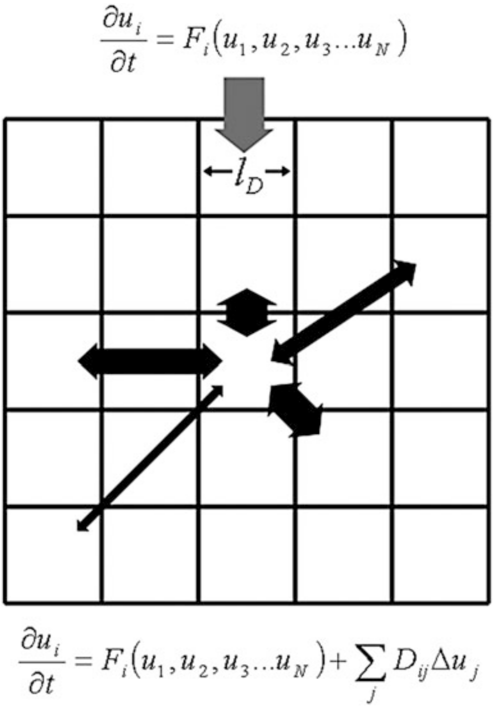

$\frac{\partial u_i}{\partial t} = F_i(u_1, u_2, u_3 \ldots u_N)$
l_D
$\frac{\partial u_i}{\partial t} = F_i(u_1, u_2, u_3 \ldots u_N) + \sum_j D_{ij} \Delta u_j$

Fig. 4.7 Neural network architecture of the Belousov–Zhabotinsky medium

$$l_D = (D_\tau)^{\frac{1}{2}}$$

is determined by the average diffusion coefficients of the medium components D and by the period τ of the processes occurring in the environment. The state of the medium inside such elementary volume corresponds to full intermixing. The dynamics of the environment in such volumes is relatively simple. At the same time, their interaction due to diffusion leads to complex spatiotemporal regimes.

4.6.1 Some Details

Various approaches are known to describe the Belousov–Zhabotinsky reaction. The conventional model of the process (the Field–Körös–Noyes (FKN) approximation) is based on 11 stages, whose dynamics is described by two kinetic equations for the reaction inhibitor u ($HBrO_2$) and activator v (the highest valence of a metal ion—Fe^{3+}):

$$\varepsilon \frac{\partial u}{\partial t} = \frac{(\mu - u)}{(\mu + u)}[qv + \varphi] + u(1 - u) + D_u \Delta u, \qquad \frac{\partial v}{\partial t} = u - \lambda v + D_u \Delta v.$$

Here, ε, q, and μ are the parameters determining the initial concentrations of the components of the reaction and the rate constants of the intermediate reactions occurring in the system. Unfortunately, these constants are not known today with an accuracy sufficient for practical applications. However, over a number of years, the composition of the media to suit the specific dynamic regimes has been defined. Thus empirical connections between the values of the parameters ε, q, and μ and the relative content of molecular components in the environment were established. The value φ is introduced into kinetic equations only when photosensitive catalyst of the reaction is employed. This value takes into account the influence of light radiation on the dynamics of the environment (see next chapter).

For systems with complete intermixing, the term in the kinetic equations responsible for the diffusion of medium components is discarded. In this case, it is easy to qualitatively describe the main dynamic regimes of the medium. Suppose that the initial state of the Belousov–Zhabotinsky medium corresponds to arbitrary concentrations of its components. Then the reactions occurring in the medium should lead it to a steady state or states, which correspond to the point of intersection of the curves

$$\frac{\partial u}{\partial t} = 0, \quad \frac{\partial v}{\partial t} = 0.$$

These curves, called zero isoclines, are a convenient tool for the qualitative description of dynamic regimes of the Belousov–Zhabotinsky media.

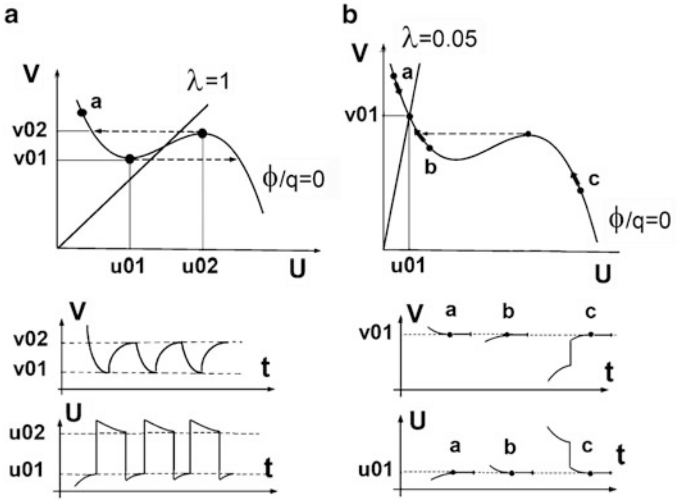


Fig. 4.8 Zero isoclines of the Belousov–Zhabotinsky reaction for the vibrational (**a**) and the excitable (**b**) modes: the time variations of the activator (v) and inhibitor (u) concentrations are shown below

Based on the kinetic equations and equating their right sides to zero, it is easy to obtain two equations for zero isoclines:

$$v = \frac{u(u-1)(u+\mu)}{(u-\mu)q} - \frac{\varphi}{q}, v = \frac{u}{\lambda}.$$

The first of these equations is an S-shaped curve, while the second one is a linear dependence of the activator of the reaction on the inhibitor concentration (Fig. 4.8).

A theoretical analysis shows that the points of intersection of zero isoclines can correspond to either stable or unstable states. In the first case, the derivative at the crossing point must be negative; in the second case it is positive.

Consider the case of the stable point of intersection (Fig. 4.8c). The figure shows three possible options that correspond to the gradual approaching of the medium to a steady state, based on the arbitrary concentrations of the molecular components of the medium. After that the medium remains in this steady state until a perturbation occurs. A detailed examination shows that when the diffusion of medium components is taken into account, this variant of the intersection of isoclines, called excitable regime, corresponds to concentration pulses propagating in the medium. Moving concentration pulses in the reaction–diffusion medium are autowave structures, whose properties differ from those typical for conventional physical wave phenomena. They are not reflected, but rather fade at an impermeable boundary. When concentration pulses encounter each other, they annihilate. Figure 4.9 shows how concentration pulses bend around an obstacle and pass through small holes.

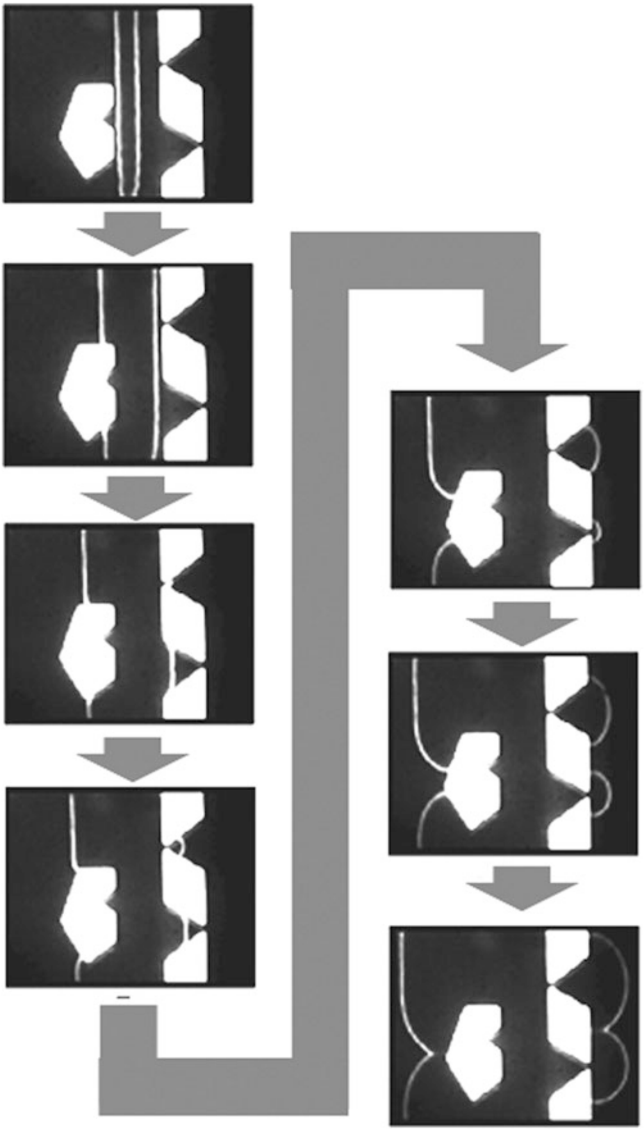


Fig. 4.9 Excitable mode of the Belousov–Zhabotinsky medium. The concentration pulse bends around an obstacle and passes through holes

The dynamic mode is much more complex and may become vibrational if the intersection of the zero isoclines corresponds to an unstable state. In this case, the evolution of the medium initially located at the point *a* (Fig. 4.8a) reaches along the stable branch of the S-shaped isocline the point (u01, v01). Here, stability is lost, and the medium spontaneously jumps to the opposite stable branch and again tends

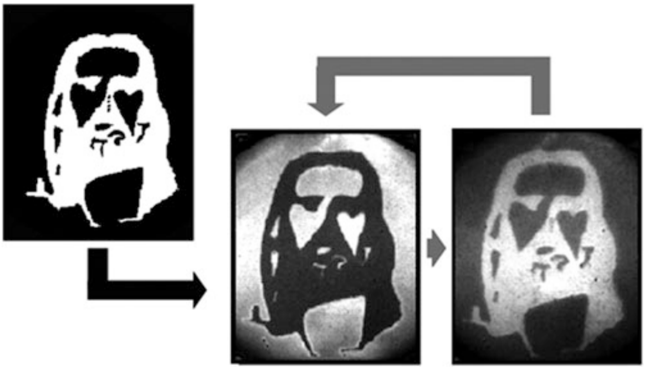


Fig. 4.10 Vibrational mode of the Belousov–Zhabotinsky medium

to the point of intersection of the isoclines. During this process, the concentrations of the activator and of the inhibitor of the reaction change periodically. When taking into account the diffusion, the oscillatory nature of the regime is kept, but the evolution of the medium becomes more complex (see next chapter). A relatively simple situation where one can observe the vibrational mode of the Belousov–Zhabotinsky medium, into which an arbitrary image is introduced, is shown in Fig. 4.10. The original image in the medium consequently passes through the stages "negative–positive–negative . . ." etc.

An important feature of an information medium of the Belousov–Zhabotinsky type is its neural network architecture.

Qualitatively, the Belousov–Zhabotinsky medium can be regarded as a realization of a neural network (Fig. 4.7), where:

- Each elementary volume of the medium can be thought of as a simple processor.
- The dynamics of the processor and the operations it executes are defined only by nonlinear kinetics of reactions occurring in the microvolume.
- Processes occurring in microvolumes are related via short-range (diffusion) interactions. Specifically, each microscopic volume is linked by diffusion with any other microvolume of the environment. But, because of the low rate of diffusion, interaction between volumes occurs with a delay and with attenuation proportional to the distance between them.
- Depending on the state of the medium (concentration of the reaction components and temperature) and external excitation, the system described by reaction–diffusion equations taking diffusion into account can operate in different dynamic regimes.

It should be noted that distributed neural networks are more rigorously described by a system of integro-differential equations that cannot be, in general, reduced to a reaction–diffusion equations. Nevertheless, under certain rather relaxed assumptions, these two models are adequate.

An interesting confirmation of the neural network architecture of Belousov–Zhabotinsky media was obtained on the basis of neural networks describing the characteristics of human vision.

Vision in mammals and, in particular, in human represents a complicated photobiological process. The image of the environment is projected by the optical self-regulating system of the eye, converted and compressed by a set of horizontal amacrine and ganglion cells, and then transferred by the optic nerve. Optic nerves of two eyes overlap, sharing some information, and pass it through the lateral geniculate body to the visual cortex, where the image is integrated into a unified whole. According to the generally accepted concepts, the cortex preprocesses information: it enhances the contours of image fragments, the lines of the orientation, the boundaries between individual blocks, etc. Further interpretation of the external information is a complex psychophysiological process, carried out by the cerebral cortex.

A distinctive feature of human vision is visual fields in which information can be amplified in the center of the field and suppressed at the periphery ("on center—off surround") or, conversely, suppressed in the center and amplified on the periphery ("off center—on surround"). As a result, the retina does not perceive uniform diffuse illumination, but it does capture point and ring structures, the boundaries of dark and light.

Among the two-dimensional neural networks with lateral (side) interaction of special importance are those that allow for simulating specific functions of the human brain and, in particular, features of human vision.

In the late 1960s, Pozin and colleagues carried out a detailed study of the neural networks described by kinetic equation:

$$\frac{\partial s_i}{\partial t} = -as_i + F(p_i) + I_i.$$

Here, s_i is the state (potential) of the ith neuron, $F(p_i)$ is a step function describing the state of the neuron depending on the sum of signals from all other neurons:

$$p_i = \sum_j T_{ij} s_i,$$

and I_i is external stimulus on the ith neuron.

This model (Fig. 4.11) represents a neural network with excitatory and inhibitory inputs. More precisely, the distribution of excitatory and inhibitory signals is described by a coupling function $g(x)$ which depends on the distance between neutrons on the surface of the network. Extended one- and two-dimensional spatial effects on the neural network were considered. The dimensions of the input signal features were significantly greater than interneuron distances. Therefore, a neural network can be viewed as a continuous, homogeneous medium.

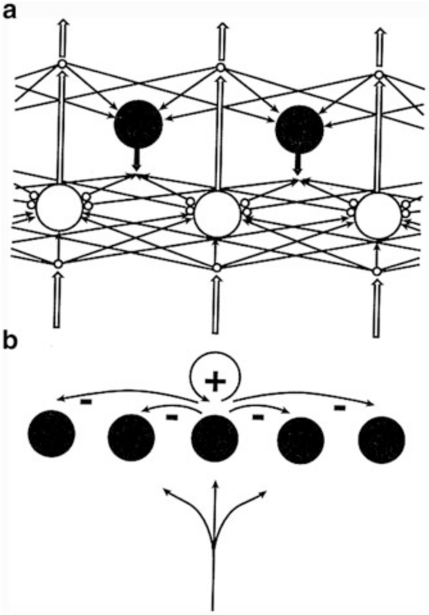


Fig. 4.11 Pozin (**a**) and Grossberg (**b**) neural networks

The main result obtained in these investigations was that a rectangular pulse could be sharpened or broadened, or a contour of the signal could be enhanced depending on the shape of the coupling function $F(p_i)$. Numerical simulation of these effects performed using the technique of Pozin is shown in Fig. 4.12. Convolution integrals of the rectangular distribution with the coupling function were computed, with the latter approximated by the expression:

$$g(x) = A1\exp\left(-x^2/B1\right) - A2\exp\left(-x^2/B2\right).$$

Constants $A1$, $A2$, $B1$, $B2$ are shown in Fig. 4.12.

In the end of the 1960s of the last century, American mathematician Stephen Grossberg began a series of studies on neural networks essentially similar to Pozin networks.

In order to explain the mechanisms of information processing by visual cortex, Stephen Grossberg proposed the concept of specialized neural networks, incorporating particular features of visual fields.

Based on psychobiological and neurobiological data, Grossberg concluded that neural networks with a central activation and lateral inhibition (Fig. 4.11) can be used to interpret a variety of phenomena of human vision, including optical illusions. Neural networks of this type are described by kinetic equations:

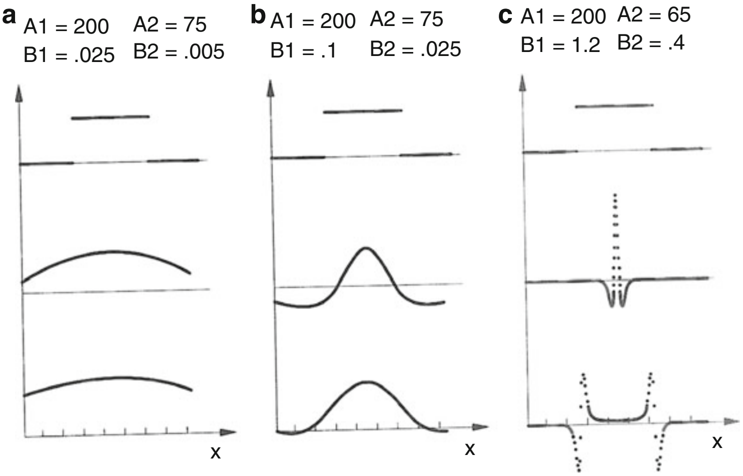


Fig. 4.12 Processing of information by the Pozin neural network

$$\frac{\partial s_i}{\partial t} = -as_i(B - s_i)[I_i + f(s_i)] - s_i\left[J_i + \sum_{j \neq i} f(s_j)\right].$$

Here, basic notation coincides with the notation adopted for Pozin equations, I_i and J_i are external stimuli on the excitable and inhibitory neurons.

In this equation, the first term on the right side describes the rate of decay of the excitation of the *i*th neuron, the second term limits the value of excitation to a certain value, and the third term corresponds to the impact on the *i*th neuron by its inhibitory environment.

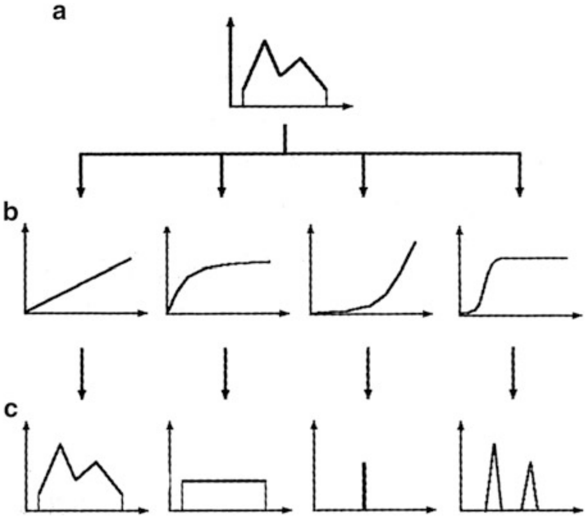
Grossberg showed that neural networks of this type have a short-term memory. In addition, depending on the form of the function, this network sharpens or broadens the spatial signal acting on it or enhances its contour (Fig. 4.13).

Grossberg noted the analogy between the dynamics of the proposed neural networks and reaction–diffusion systems used by Gierer and Meinhardt to explain the biological pattern formation.

Later, this analogy was demonstrated for the chemical reaction–diffusion systems of the Belousov–Zhabotinsky type. It is easy to see that this analogy has a real physical–chemical basis. Chemical nonlinear systems are characterized by autocatalytic mechanisms (activation in the elementary microvolume) and by significantly higher values of diffusion coefficients of the inhibitor compared with the activator.

Apparently, the neural network architecture of the Belousov–Zhabotinsky media explains the specifics of elementary operations of image processing by the environment (image contour enhancement, amplifying or quenching of its features), coinciding with the elementary operations of Grossberg networks.

Fig. 4.13 Processing of information by the Grossberg neural network: (**a**) initial signal, (**b**) the function *f*(*s*), and (**c**) result of the transformation



4.7 "Emerging" Information Mechanisms

The possibility of information processing by reaction–diffusion media can be understood on the basis of the concept of "emerging" information mechanisms, which is being actively developed in recent years. A reaction–diffusion medium may be considered a multilevel system.

The initial level of consideration is the kinetics of chemical reactions between the components of the medium. The change in concentration of the medium components over time is determined mainly by their initial values, medium's temperature, and the rate constants of the reactions occurring in the medium. This represents so to say the lower level of the dynamics of the medium. At the same time, considering the medium as a whole and solving reaction–diffusion equations, one can proceed to the next level of the dynamics of the reaction–diffusion medium—the ensemble of its steady-state regimes. Characteristic for systems with complete intermixing are the bistable regime (the system can move from one stationary state to another), the sleep regime (stable stationary state), and the regime of concentration fluctuations. If diffusion of components can exist in the medium, then these regimes will also include trigger switching from one stable state to another, moving concentration pulses, spiral structures, leading centers, and many other dynamic states. All of these regimes form a basis for the next level of dynamics—the interaction of stationary structures. The principles of this interaction are specific and not similar to the principles of seemingly analogous physical objects. Thus, concentration waves, unlike physical wave processes, while bending around obstacles, do not reflect from them, but rather annihilate. Annihilation also occurs during mutual collision of concentration waves. This dynamic occurs on the basis of preceding dynamic levels and, at the same time, cannot be reduced to them.

This "emerging" dynamic is a characteristic feature of biological nonlinear distributed systems arising from self-organization processes occurring in such systems.

Thus, in a reaction–diffusion system, three levels of dynamics can be distinguished:

- The level of interaction between the elements of the medium, i.e., nature of their interconnections (microlevel)
- The level at which the system can be in a stationary state (mesolevel)
- The level of interaction between the states of the system and its environment (macro-level)

It is the dynamics of the macro-level which is responsible for information processing by the reaction–diffusion media.

4.8 Principles of Information Processing by Reaction–Diffusion Devices

Nonlinear reaction–diffusion media represent a means of information processing fundamentally different from digital von Neumann computers. The distributed nature of the medium leads to a high parallelism of information processing multiple times greater than the capabilities of multiprocessor digital systems. Nonlinear mechanisms of the dynamics of reaction–diffusion media cause high logical complexity of the elementary operations executed by them. Therefore, the performance of a reaction–diffusion device is determined not by an increase of the performance of the elements (microminiaturization of circuits) but rather by more complex dynamics of the device, which leads to increased logical complexity of elementary operations. A natural extension of this approach should be the creation a multilevel media with high complexity of behavior.

Chapter 5
Reaction–Diffusion Processor: Possibilities and Limitations

> *The future of analog calculations is boundless. Being a daydreamer, I see how they eventually take the place of digital ones, especially first for solving the differential equations in partial derivatives and for simulation in neuroscience. Several decades are need for this to occur. In the meantime, I believe that this is a very fruitful and challenging field of research although (but maybe therefore) it is not popular today.*
>
> Lee A. Rubel, *From a letter to a friend*

5.1 Reaction–Diffusion Processor: Basic Principles

At the heart of information processing by reaction–diffusion systems is the concept of reaction–diffusion processor. A schematic diagram of this device, which is based on a chemical or a biochemical system, is shown in Fig. 5.1. Chemical reaction–diffusion media of the Belousov–Zhabotinsky type represent a convenient starting material for the creation of information processing devices. They are stable and not toxic.

The temperature range and the temporal scales of processes in them are suitable for recording the characteristics of these media by available physical methods. Chemical components required for the formation of these media are readily available, and their cost is low. At the same time, those dynamic regimes that are most important for information processing appear when the environment is operating in a steady state (far from equilibrium). Therefore, the initial components of the reaction must be continuously fed to the environment, maintaining their initial concentrations constant, and the products of the reactions occurring in the medium must be carried off. Dynamical regimes of the Belousov–Zhabotinsky media are sensitive to temperature changes. Therefore, the medium processing the information must be thermostatted.

Information processing by a processor involves several levels characteristic of this system.

© Springer-Verlag Wien 2014

N.G. Rambidi, *Molecular Computing*, DOI 10.1007/978-3-211-99699-7_5

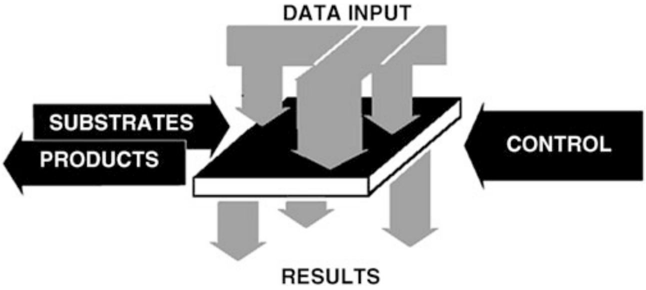


Fig. 5.1 Scheme of a reaction–diffusion processor

5.1.1 Input of Information

For entering data into a reaction–diffusion medium, a "macro–micro" interface is required. It converts the input information (a spatial distribution of physical stimuli) into the corresponding spatial distribution of molecular components of the medium. Particularly convenient for entering information are photosensitive media. In this case, the information being entered represents an image (generally, an arbitrary distribution of light intensity), projected by an optical system on the surface of the layer or into the volume of the medium. The medium itself contains a photosensitive catalyst of the Belousov–Zhabotinsky reaction initiating a sequence of photochemical reactions that lead to a change in the content of its basic components:

$$\mathrm{Ru}^{+2} + \mathrm{h}\nu \rightarrow \ * \,\mathrm{Ru}^{+2},$$

$$*\mathrm{Ru}^{+2} + \mathrm{C_3H_3BrO_4} \rightarrow \mathrm{Ru}^{+3} + \mathrm{Br}^{-} + \text{organic products},$$

$$\mathrm{Ru}^{+3} + \mathrm{C_3H_3BrO_4} \rightarrow \mathrm{Ru}^{+2} + \mathrm{Br}^{-} + \text{organic products}.$$

As a result, at each point of the medium changes of the concentrations of its components take place that depend on the intensity of light emission at this point. In other words, a chemical implementation of the input image appears in the medium.

Before the source data for solving the selected task are input into the medium, it should be reset to some initial form not containing significant information. For this purpose it is convenient to illuminate the medium by intense light emission. Let us consider this process in more detail and turn again to the zero isoclines of the photosensitive Belousov–Zhabotinsky medium (Fig. 5.2). The analytical expression for an S-shaped isocline includes a parameter describing the effect of light radiation on the dynamics of the medium (see previous chapter). It is easy to see that this stimulus leads to a shift of the S-shaped isocline along the axis of ordinates.

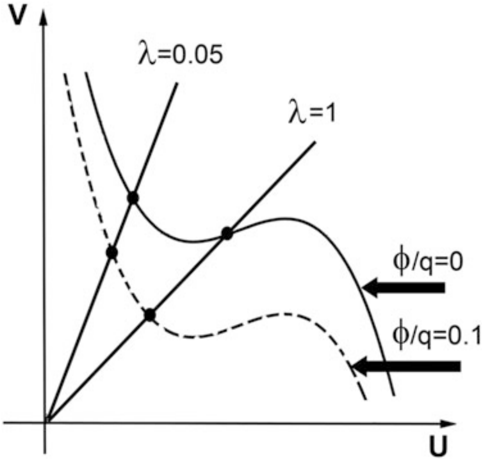


Fig. 5.2 Zero isoclines of the photosensitive Belousov–Zhabotinsky reaction in the dark and under light irradiation. *v* and *u*—activator and inhibitor

In this case, light emission does not affect the second linear isocline (Fig. 5.2). As a result, intersection points of the isoclines shift along the S-shaped curve to the area of higher concentrations of the activator of the reaction and extremely low concentrations of the inhibitor.

Let us consider qualitatively the process of entering information. Suppose that a black-and-white image is projected on the medium, whose composition corresponds to the vibrational mode. In this case, the spatial regions of the medium affected by intense radiation are transferred to a state with a high content of the activator of the reaction. At the same time, spatial areas on which no light is falling do not change their status. When the exposure of the light emission ends, the evolution of the input image begins in the media driven by the reactions in the medium. Those areas of the medium on which no light emission fell immediately enter the oscillatory process, while illuminated areas will need some time until chemical reactions transfer them into an oscillatory regime.

Orange-red and blue colors of the reagent correspond to high content of activator and inhibitor in the medium, respectively. In practice, when monitoring the evolution of the image, blue light filters are used in order to increase its contrast. As a result, the registration of the evolution of the image by a black-and-white video camera makes the regions of the medium with a high content of activator appear dark and those with a high content of inhibitor appear bright.

Thus, during the evolution of a black-and-white image introduced into an oscillating medium, bright fragments (which are dark on the input image) appear on a dark background corresponding to the illuminated fragments (which are bright on the original input image). In other words, at the initial stage of the evolution, a negative form of the input picture appears. This process is shown in Fig. 4.10 of the previous chapter.

The possibilities of the optical information input option can be significantly enhanced by using a set of molecular components of the medium selectively sensitive to radiation in different spectral regions. Other physical stimuli, such as electric fields, electrochemical processes, local changes in the temperature of the environment, etc., can also be used to enter information in a reaction–diffusion medium.

5.1.2 Information Processing

In order to accomplish a specific selected operation of information processing, a corresponding dynamic mode of the spatiotemporal evolution of the medium needs to be specified. The medium converts its original distribution of reactant concentrations into some final state, which is regarded as a solution to the problem. In turn, the dynamic mode is determined by the state of the medium, i.e., its composition (relative concentrations of its molecular components) and temperature.

As shown below, the properties of the input information also play an important role. Since optical input of information is particularly convenient, the source data represent some images. An image can exist in two forms, which are equivalent in terms of information it contains—positive and negative. However, these variants behave differently in the course of their evolution in reaction–diffusion media.

There is some arbitrariness in determining the positive and negative forms of the image. For definiteness, let us call positive an image corresponding to its natural perception by the human vision. However, in some cases it is difficult to determine what perception should be called natural. In this case we call a black image, carrying information, on a white background a positive image. In any case the input picture may be reduced to a set of values—the optical densities D_i ($D_0 < D_i < D_\infty$, where D_0 and D_∞ are the minimum and maximum values of optical density). A negative image we will define is a corresponding set of inverted optical densities $DN_i = D_\infty - D_i$.

Optical input of information determines one more characteristic property of the functioning of the reaction–diffusion processor. The solution of the problem is based on continuous recording of the evolution of the image in the medium, e.g., by a video camera. This, however, is only possible if the process of evolution can be observed, i.e., the medium is continuously illuminated by light. This additional "technological" illumination of the light-sensitive medium can naturally influence its dynamic regime (see zero isoclines, Fig. 5.2). Let the state of the medium correspond to the vibrational mode. If after image input it gets illuminated in the process of registering by a minimum light intensity that can be registered, a typical oscillatory process occurs in the medium (Fig. 5.3). But if the same medium at the same exposures of the input image is illuminated during the recording by sufficiently intense light emission, a process corresponding to the excitable regime emerges and develops in the medium.

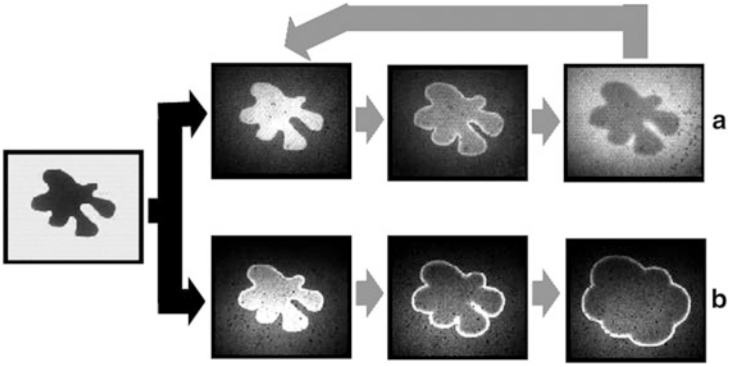


Fig. 5.3 Change of the vibrational mode of the photosensitive Belousov–Zhabotinsky medium (**a**) under light irradiation (**b**)

At the same time, the evolution of the initial distribution of concentrations can be altered by affecting the environment by different physical and chemical control stimuli and (or) designing media with complex structure. All this leads to a manifold increase in the information capabilities of reaction–diffusion media.

5.1.3 Visualizing (Deriving) the Results of the Problem Solution

In order to derive results of the problem solution by a reaction–diffusion medium, it is necessary to convert the distribution of the concentrations of medium components, corresponding to this solution, into an appropriate distribution of some macroscopic physical quantities ("micro–macro" interface). The most convenient way to do it is to record the spatial distribution of the optical characteristics of the medium, such as coloring of the environment, spectral absorption, etc.

In the course of the reaction, when the medium goes from one state to another, the catalyst of the reaction changes its electronic state. As a consequence, the reagents change their color (from red to blue and vice versa). Thus, it is easy to visualize the process and monitor the spatiotemporal evolution of the system. In this case, the available optical tools such as a video camera facilitate entering the registered distribution in computer memory for further processing.

5.1.4 Control and Energy Supply of the Processor

The same physical stimuli that are used to input information can also be control actions. Thus control can be understood as the change of the medium's state (relative concentrations of its components) by a given stimulus that leads to a

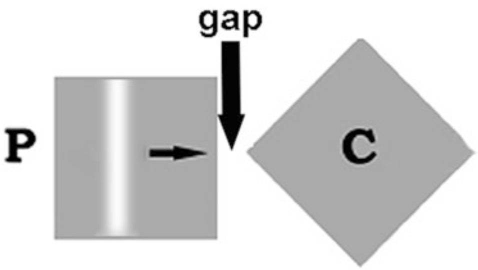


Fig. 5.4 Simulation of a chemical diode

local change in the dynamic mode in the selected areas of the medium. As an example let us consider a concentration pulse bending around a hindrance and passing through a hole (see Fig. 4.9 in the previous chapter). In this case a homogeneous Belousov–Zhabotinsky medium was used in which two divergent linear impulses were initially excited. After that a hindrance was projected onto the surface of the medium by intense radiation that was supposed to be bent around by the left pulse and a band with two holes through which the right impulse was supposed to pass. Intense radiation, projected over the entire duration of the impulses, transferred the medium in these areas into an inactive state (see zero isoclines in Fig. 5.2), which was equivalent to a solid impermeable hindrance introduced into the medium.

A second example is the creation of the "chemical diode" configuration in a reaction–diffusion medium. In 1996 the Japanese researcher Prof. Yoshikawa with colleagues proposed the idea of a device in which unidirectional passage of concentration waves generated in the Belousov–Zhabotinsky reaction takes place. The device, called chemical diode, is a thin gap, inactive for the Belousov–Zhabotinsky reaction, between linear (P-side) and converging to the gap (C-side) boundaries (Fig. 5.4). Experimentally, this system was implemented as a set of two square plates made of microporous glass, with the apex of one of them directed toward the middle of the side of the other plate. The plates were impregnated with a solution of the reaction catalyst—ferroin—and placed into the solution of the remaining components of the reaction. Since ferroin was contained only in the porous glass, the reaction could only take place on the surface of the plates. It was shown that the passage of concentration waves in the direction P→C and C→P is not adequate. Given a gap of a certain width, the wave moves in the direction P→C and does not pass in the opposite direction C→P. Yoshikava and colleagues also showed that a variety of logic circuits can be constructed based on elementary chemical diodes.

A major shortcoming of the experimental device used in this work was high labor intensity of its manufacturing and rather low reliability. However, if one controls the modes of the Belousov–Zhabotinsky reaction by light radiation, a simple experimental scheme can be proposed that facilitates forming both a geometrically arbitrary structure of the diode and a variety of devices based on it. In a homogeneous photosensitive medium, a concentration wave can be excited, and during the entire time of its passage through the device, the structure of the diode

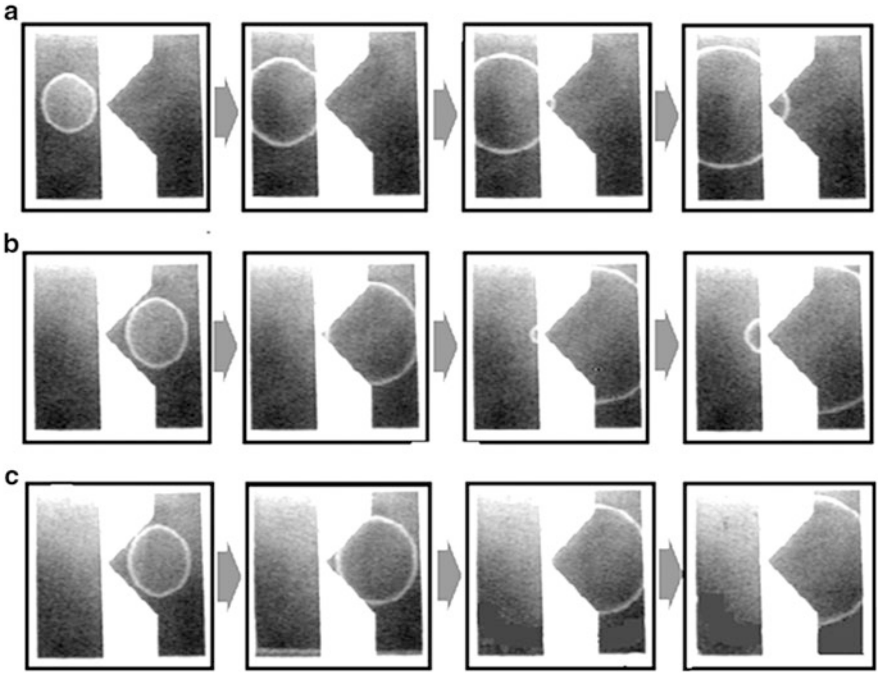


Fig. 5.5 Scheme of a chemical diode based on the Belousov–Zhabotinsky medium: (**a**) and (**b**) the pulse passes through the hole, (**c**) the pulse does not pass through the hole

can be formed by intense light radiation projected onto the surface of the medium. The main options for the passage of a wave through such a system are shown in Fig. 5.5. It should be noted that in spite of its extreme simplicity, the developed technology allows for quickly and accurately changing the geometrical parameters of the diode and organizing their various combinations.

Control actions can also be complex. If, for example, a film consisting of the protein bacteriorhodopsin is introduced into the reaction–diffusion medium, local illumination of the film leads to an increase in hydrogen ion concentration near the illuminated regions of the film. In turn, local change in the acidity of the medium can lead to a local change in its condition.

Energy supply of a biomolecular processor must, of course, be chemical. This means that the environment, working for a long time, should be a continuous flow chemical reactor. Substrates of the reaction must be continuously fed into it, and reaction products must be removed.

5.1.4.1 Some Details: Technical Characteristics of the Reaction–Diffusion Processor

The block diagram of the device and its basic modes of operation are shown in Fig. 5.6. For inputting the image into the medium, the SANYO PLC-510M video

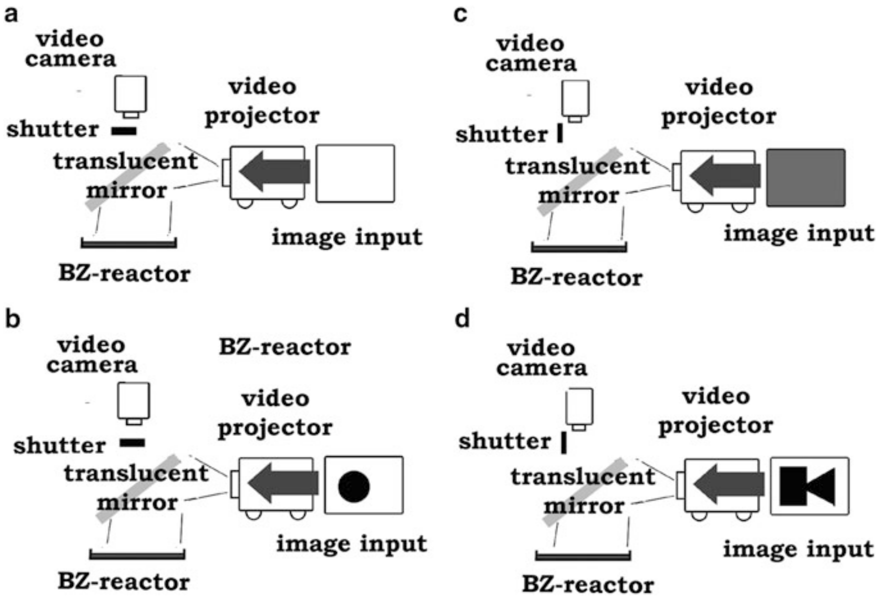


Fig. 5.6 Modes of a reaction–diffusion processor

projector (VGA compatible, 279 ANSI lumens) controlled by a personal computer is used. High homogeneity of the light emission background of this projector essentially reduced the experimental error. At the same time, the computer-controlled projector allowed much greater control over the input information (change the brightness and contrast, add or delete some of the details of the information entered, control the state of the medium during the evolution of the image). Registration of the evolution of the image in the medium was made by a black-and-white video camera Mintron OS-045D (0.02 lux sensitivity, resolution 600 TV lines) in combination with blue filters that increased image contrast. Digitized images were recorded in the memory of a personal computer (VidCap software package) using the mode of randomly selected individual slides or in movie mode.

A closed static reactor based on a thermostatted petri dish 80–120 mm in diameter was employed. Three variants of the media were used:

- A layer of liquid reagent 0.5–1.5 mm thick, comprising the catalyst of the reaction, was used for preliminary study of the image processing operations. In this experiment it was found that the mode corresponding to the original composition of the reagent is retained in the reactor for 15–20 min.
- A medium in which the catalyst of the reaction was immobilized in a thin (~0.2 mm) layer of solid silica gel was applied to aluminum foil (standard plates for liquid chromatography were immersed in a 0.0001 M solution of the catalyst

for 40–50 min), the remaining components of the reaction remained in solution, and the reaction occurred in the boundary layer over the surface of the silica gel.

- A medium in which the catalyst was immobilized in a layer of hydrogel silica gel, 1.0–1.5 mm thick. The catalyst was injected into a layer of silica gel during its formation, other components of the reaction diffused into the layer from solution over silica gel, and the reaction took place in the bulk layer.

Reactors with an immobilized catalyst eliminate pattern distortion caused by external physical influences (vibration, random shocks, etc.). They can accommodate a much larger (compared to the liquid-phase reactor) amount of the initial reagent. Therefore, the running time of the closed reactor without changing the nature of the regime through depletion of reaction components was 1–2 h.

The utilized light-sensitive media (catalyst $Ru(bpy)_3Cl_2$) functioned in excitable and oscillatory modes.

The initial composition of the excitable medium is as follows: H_2SO_4—0.3 M, $KBrO_3$—0.3 M, malonic acid—0.2 M, KBr—0.05 M.

The initial composition of the vibration environment is as follows: H_2SO_4—0.6 to 0.8 M, $KBrO_3$—0.4 M, malonic acid—0.2 M, KBr—0.075 M.

The device for studying image processing operations by Belousov–Zhabotinsky media provides four basic modes of operation.

When reusing a reaction–diffusion medium it is necessary to remove the traces of previous experiments before the start of each variant. To this end the medium was illuminated by intense (white) light for 1.0–1.5 min (Fig. 5.6a).

To enter the initial information the light image was projected onto the surface of the medium. Utilization of a half-transmitting mirror in the optical circuit allowed for observing and controlling the information input process by a video camera (Fig. 5.6b).

The evolution of the entered image was recorded by a video camera. Moreover, since external light sources were excluded, an image projector was used for uniform illumination. Spectral composition and intensity of illumination were set by the Photoshop 7.0 software (Fig. 5.6c).

The optical scheme utilized allows to efficiently control the process of evolution of the entered image. When, during the evolution of the original image, another image defining the configuration of the medium is projected on the surface of the medium, individual sections of the medium can be easily excluded from the process of evolution or the process can be slowed down (Fig. 5.6d). Figure 4.9 demonstrates a concentration pulse bending around an obstacle, annihilation of pulses, and the passing of a pulse through holes (the transformation of a linear wave into a circular one).

Along with the experimental study of image processing operations, software capable of performing numerical simulation was developed and utilized. It was designed as two independent blocks. One of them is intended to enter the initial data and visualize the results of the calculations, the second one to actually perform calculations. Such an arrangement simplifies the addition of data-processing options by changing the configuration files without recompilation of the visualization module.

Calculation modules do not use the graphical interface and do not require user intervention during calculations. This allows to run them on remote servers, and if necessary to recompile them for other platforms. Modules are written in C++ and compiled for PC by the Microsoft Visual C++ 6.0 compiler. The total volume of the executable files is about 1 Mb.

5.2 Image Processing by Belousov–Zhabotinsky Media

5.2.1 What Is Image Processing

In modern research and industrial practice one is often confronted with the fact that crucial information upon which the solution of a given task is based represents an image. These are two-dimensional (or sometimes three-dimensional) distributions of optical characteristics—color, brightness, etc., describing the physical situation characteristic for the problem under consideration. They can represent slices of tissue in medicine, cross-sectional views in materials science or geology, the environment in which an autonomous robot operates, and consequently observed parts on the assembly line during automated quality control of manufactured articles. These are only random examples taken from the vast range of situations characteristic of the modern human activity. For this reason, image processing and recognition has become today an independent practically important area.

Universal proliferation of digital computers in virtually all areas of modern activity has naturally resulted in utilization of this technology for image processing. Nevertheless, high computational complexity of the problems associated with images and their intrinsic association with human vision motivated the development of approaches and tools based on biological principles of information processing.

Transmission, storage, and processing of information by biological systems are fundamentally different from the same operations carried out by modern digital computers. In this case quite complex fragments, and not simple symbols transmitted in a bit-wise fashion, serve as elementary units of information. Those may be the phonemes in processing of verbal information, images with which vision operates, etc. Apparently the first attempts to use this fundamental feature of biological systems were made in the 1950s of the last century. One of the principal efforts among them was the creation of “cellular logic”—the field of computational geometry, which is essentially the algebra of binary images defined on the matrices of binary numbers. Later, in the 1960s Blum proposed the ideology of “wildfire,” based on the wave principles of simultaneous processing of images as a whole.

For example, by exciting the wave at all points of the contour of an arbitrary closed polygon and measuring the time dependence of the area emerging as the contour moves, one can determine the number of its corners. Later it was shown that the Blum algorithm is sufficiently effective to perform a number of image processing operations.

In recent years, mathematical foundations and techniques of image processing have been developed, called "mathematical morphology." This technique, suitable for modern digital computers, is being successfully employed for a multitude of tasks.

Binary mathematical morphology operates with complex two-dimensional (in principle, with multidimensional) objects defined in a discrete space with discrete coordinates (with points being pixels). An object "A" can be viewed as a set of pixels "a" satisfying the condition:

$$\mathbf{A} = \{\mathbf{a} \mid \mathbf{property}\ (\mathbf{a}) = \mathbf{TRUE}\}.$$

Despite the digital representation of the original data mathematical morphology operates with images as a whole. Elementary operations in binary mathematical morphology are dilation and erosion. Herewith the notions of image (object A) and structural element (object B) are introduced, determining the nature of changes in the shape of the object and at its borders (Fig. 5.7). In general, the operation of dilation

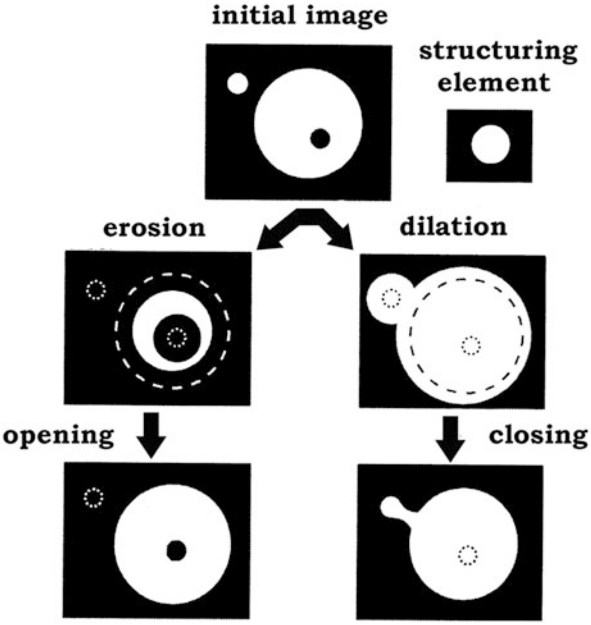


Fig. 5.7 Main operations of the technique of mathematical morphology

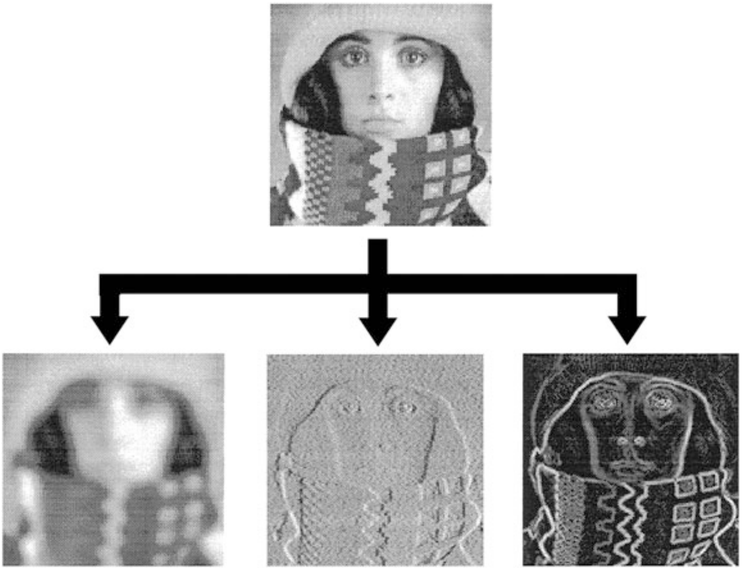


Fig. 5.8 Image processing using the technique of mathematical morphology

$$\textbf{Dilation}\, A \bigoplus B \;=\; \left\{x :\; \left(\hat{B}\right)_x \cap A \neq 0\right\}$$

magnifies the image, while the operation of erosion

$$\textbf{Erosion}\, A \ominus B = \left\{x : (B)_x \subseteq A\right\}$$

diminishes it.

Two fundamental operations of mathematical morphology are built on these elementary operations: opening

$$\textbf{Opening}\, A \circ B = (A \ominus B) \bigoplus B$$

and closing

$$\textbf{Closing}\, A \bullet B = \left(A \bigoplus B\right) \ominus B.$$

A detailed examination shows that the totality of operations “open” and “close” allows to perform all basic operations of image processing—selection of image contour, its skeleton, dividing the image into its simplest parts, etc. As an example, shows some of the operations performed by the technique of mathematical morphology: image smoothing (Fig. 5.8) and the selection of its individual regions and the boundaries between them (Fig. 5.9).

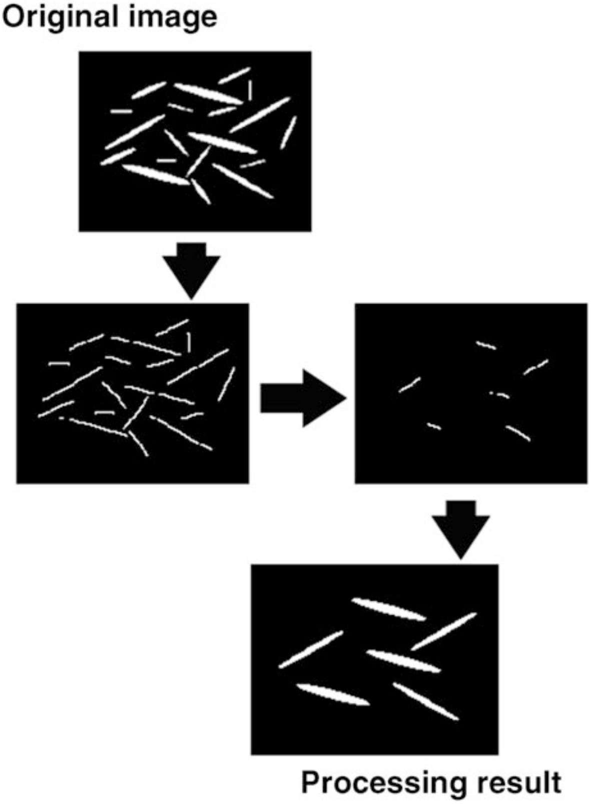


Fig. 5.9 Selection of the individual regions of images and the boundaries between them using the technique of mathematical morphology

A common feature of all these implicit or explicit attempts to simulate the biological characteristics of information processing was that they were being developed as numerical computational methods for information processing. Attempts were also undertaken to build digital computing systems maximally adapted to these computational approaches, in particular to perform cellular logical operations. Nevertheless, they did not lead to significant results, apparently because of the fundamental differences between the discrete digital methodology and biological principles.

5.2.2 Image Processing: Reaction–Diffusion Media and Mathematical Morphology

A reaction–diffusion medium is a physical information processing unit fundamentally different from digital devices. Information is entered into the environment in the form of natural fragments—images—and is processed by the medium without sampling and digitizing them. In essence, this is the analog form of information processing in which the capability of information processing by the human cerebral cortex is simulated.

It should be noted that the main methodological difficulty in using reaction–diffusion devices lies in the fact that the understanding of the mechanisms used by the brain for solving specific problems leaves a great deal from being desired. If they were known, modeling these mechanisms by the media with an architecture similar to the Grossberg neural networks would allow to create effective devices functionally similar to the cerebral cortex. Our ignorance of these mechanisms, combined with the von Neumann principles of information processing entrenched in the minds, often leads to inefficient algorithms. Therefore, devices based on them do not exploit the basic fundamental advantages of reaction–diffusion media. At the same time, a correctly guessed algorithm for information processing by a reaction–diffusion medium leads, on the one hand, to an effective device and, at the same time, hopefully sheds light on the mechanisms of functioning of the cerebral cortex.

In 1986 the German researcher Lothar Kuhnert discovered a remarkable property of light-sensitive Belousov–Zhabotinsky media. For a sufficiently long time (minutes) they retained the image that appears in the medium when this image is projected onto its surface. It was also found that during its existence in the medium, the picture entered undergoes evolution, in the course of which its contour is periodically detected and transitions between the negative and the positive forms of the image take place. This effect was described in an article published by L. Kuhnert with the Pushchino Research Center researchers K. I. Agladze and V. I. Krinsky, which the authors called “Image processing using light-sensitive chemical waves.” Later, physical and informational aspects of this problem were studied in detail by the International Institute of Control Sciences and the physical faculty of the Moscow State University.

These studies have shown that, indeed, in many cases, the results of the image evolution in the medium are equivalent to one of the image processing operations used in practice. Nevertheless, several issues remained open which play an important role in the practical application of chemical reaction–diffusion media for image processing:

- What is the basis of the similarity between the results of image evolution in the medium and the operation of its processing?
- Does this evolution always select only the graphic elements of the image previously entered into the medium, or can features not present in the original image emerge in the course of its evolution?

Answering these questions was made possible through a detailed comparative analysis of image processing by chemical reaction–diffusion media and the method of mathematical morphology.

Let us consider the results of this analysis, starting with the processing of black-and-white images by the Belousov–Zhabotinsky medium, whose composition corresponds to its functioning in an excitable regime.

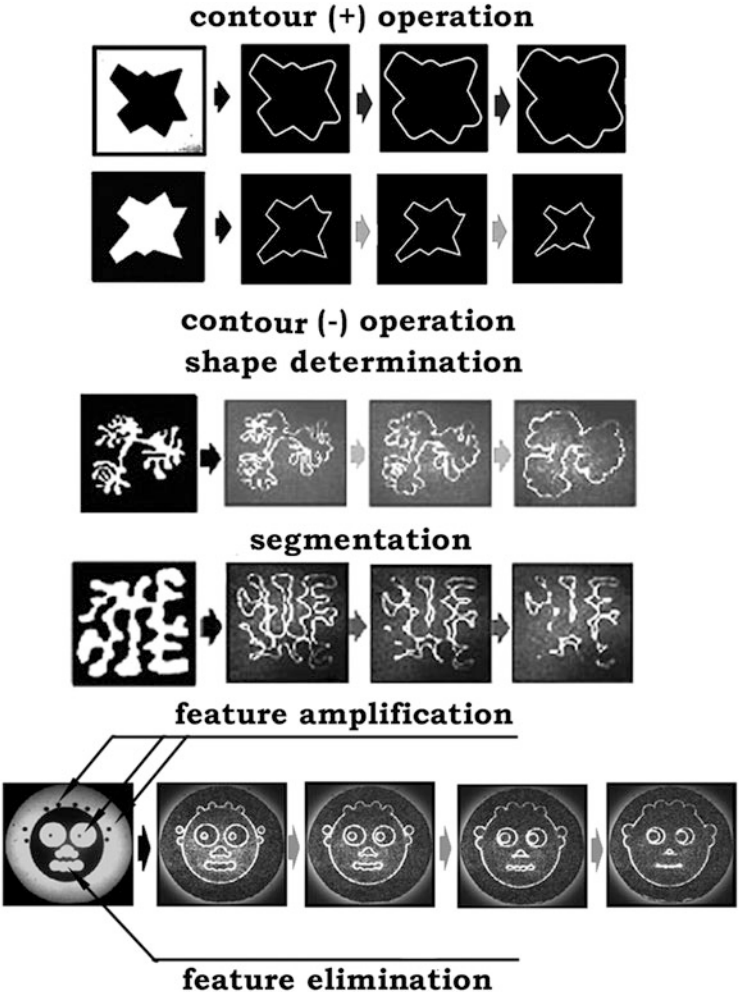


Fig. 5.10 Simulation of operations of the Belousov–Zhabotinsky medium

5.2.2.1 Excitable Regime

The most basic operations in image processing by a reaction–diffusion medium functioning in an excitable regime are the two operations of contour enhancement in the input image that can be defined as “contour(+)” and “contour(−)” (Fig. 5.10). It is easy to see that the contour enhanced at the image border is always distributed in the area illuminated by light radiation. Therefore, in the case of contour(+) the image contour diverges in the course of image evolution in the medium from its center, while in the latter case the contour converges to the center of the image. The choice of an operation is determined by entering positive or negative forms of the

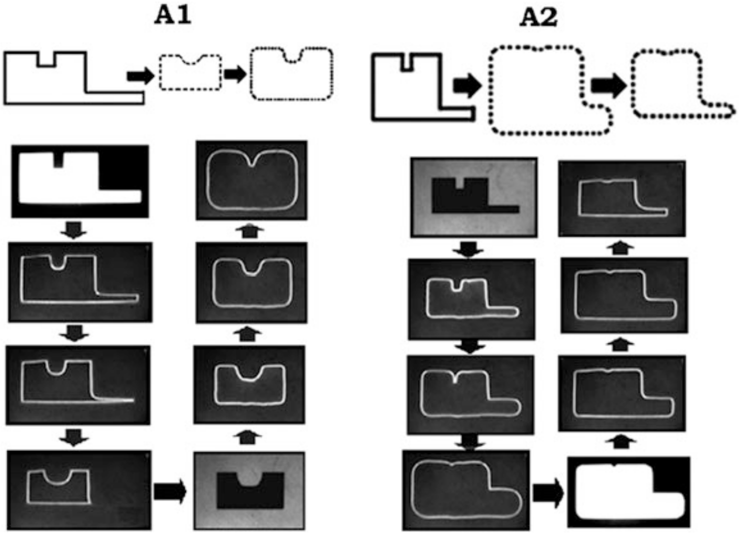


Fig. 5.11 Initial basic operations of mathematical morphology by the Belousov–Zhabotinsky medium

image being processed into the medium. The results of the pilot to perform these operations are in good agreement with numerical simulations of these operations.

It is easy to see that these operations are equivalent to the operations "dilation" and "erosion" in mathematical morphology. Naturally, they do not coincide completely. The evolution of an image in a reaction–diffusion medium corresponds in some sense to a particular case in which the transforming structural element of mathematical morphology has the form of a circle. In this approximation the reaction–diffusion medium easily reproduces the major operations of mathematical morphology "open" and "close." Figure 5.11 shows the execution of these operations on simple images using mathematical morphology and the evolution of these images in a reaction–diffusion medium. In both cases the results of image processing are virtually identical, and the reaction–diffusion medium behaves as though it were simulating the process of performing these operations.

In view of the similarity between the initial operations performed by the method of mathematical morphology and by the reaction–diffusion media, these media provide an opportunity to execute all operations practically used for image processing. Figure 5.10 shows an operation of determining the shape of a complex figure when features of the image small enough compared with the size and shape of the figure are excluded and the basic characteristics of its shape become apparent. The thinning of the image elements is shown in the same figure. Different directions of motion of contours of the individual image fragments lead to the medium eliminating or conversely magnifying specific details of the image.

Implementation by the reaction–diffusion medium of the operations "open" and "close" allows for more complex operations over images. Figure 5.12 shows the

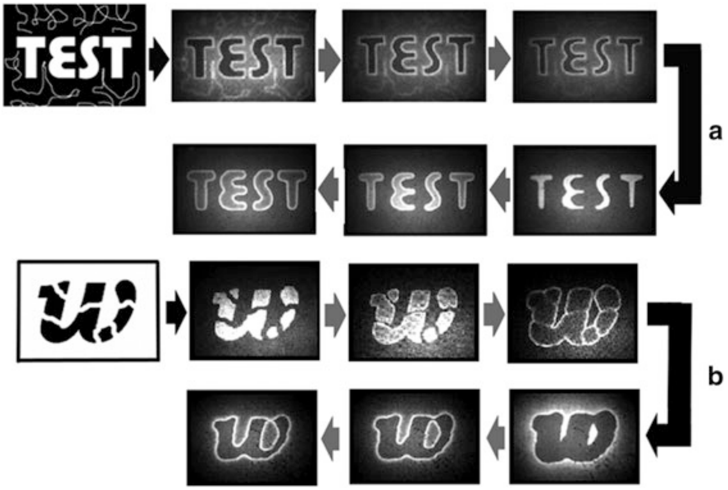


Fig. 5.12 Recovery of images and removal of non-Gaussian noise by the Belousov–Zhabotinsky medium functioning in the excitable mode: (**a**) removal of non-Gaussian noise, (**b**) recovery of the image

recovery of an image in which random defects have been formed, and the removal of non-Gaussian noise clogging the image.

Note that in the excitable mode of the functioning of the medium, determining the skeleton of the image was possible for only a limited class of extended images formed by sufficiently thin fragments. In general, this operation is executed by a medium that operates in the trigger mode. Numerical simulation of this operation is shown in Fig. 5.6. Unfortunately, a medium of the Belousov–Zhabotinsky type in the trigger mode is not sensitive to light radiation, which does not allow to determine experimentally the skeleton of a figure of an arbitrary shape.

A more complete comparison of the results of image processing by the method of mathematical morphology and the evolution of the images in a reaction–diffusion medium of Belousov–Zhabotinsky are given in Table 5.1. The primary conclusion to be drawn from this comparison is that the results of the image evolution in the medium actually coincide with the operations of image processing. The basis of this is apparently constituted by nonlinear mechanisms inherent to both approaches.

Nevertheless, differences between these approaches should be emphasized, although they do not affect the main conclusions.

The method of mathematical morphology allows working with a variety of structural elements of different shape. This makes it possible to fix the more subtle features of the image being processed, which is its distinct advantage.

At the same time, the complexity of the behavior in the Belousov–Zhabotinsky media is very high. On the one hand, as will be shown below, this may result in image processing artifacts that do not reflect the actual structural characteristics of the image. To avoid this, the mode of image processing by the Belousov–

Table 5.1 Comparison of image processing by the method of mathematical morphology and by Belousov–Zhabotinsky reaction–diffusion media

Mathematical morphology	Chemical reaction–diffusion medium
Numerical method of image processing based on nonlinear transformations of their shape	Image processing of the chemical reaction–diffusion medium based on nonlinear dynamic mechanisms
Object processing—an image represented by a set of pixels	Object processing—an image entered as a single entity in the medium
Sequential pixel by pixel image processing by a digital computer	Parallel image processing simultaneously at all its points by a distributed chemical environment
Different modes of image processing based on the structural elements of arbitrary shape	A single circular structural element used almost exclusively
"Dilation" and "erosion"—the basic operations of image processing The two basic operations—"open" and "close"—involve joint application of elementary operations	"Contour (+)" and "contour (−)"—the basic operations of image processing Operations "open" and "close" can be reduced to the joint application of these elementary operations
Virtually all processing operations of black-and-white images can be performed	Virtually all processing operations of black-and-white images can be performed
A large number of gray-scale image processing operations can be performed	A large number of gray-scale image processing operations equivalent to the operations of mathematical morphology can be performed

Zhabotinsky medium should be carefully chosen. On the other hand, as a result of the highly complex behavior of the medium, new opportunities to discriminate information about the structure of the image emerge that are too difficult if not impossible to obtain by the method of mathematical morphology.

In general, there is no doubt that reaction–diffusion media (and in particular the Belousov–Zhabotinsky media) allow for obtaining information about the structure of the image safely and fairly easy.

5.2.2.2 Vibrational Mode

Evolution of images in a medium of the Belousov–Zhabotinsky type functioning in the vibrational mode turns out to be much more complex compared to the excitable regime. Great opportunities for a detailed analysis arise in the case of halftone image processing. In this case, a positive halftone image is first converted into a negative black-and-white image. After that, in a general case, contour enhancement of individual fragments takes place, and then the image is converted to the original gray-scale image. The duration of each phase of this evolution and the information obtained about the structural features of the image may vary depending on the contrast of the original image (Fig. 5.13). When a negative form of the image appears, its gray-scale fragments, the blackening of which varies from point to

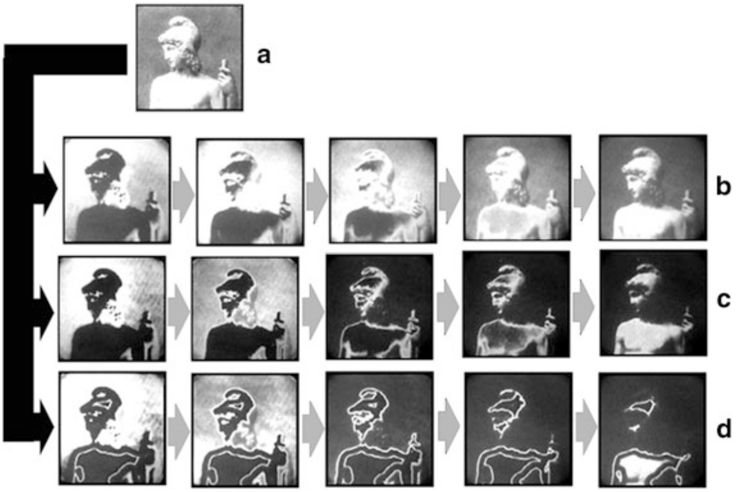


Fig. 5.13 Halftone image processing by the Belousov–Zhabotinsky medium functioning in the vibrational mode: (**a**) original image, (**b**)–(**d**) evolution as a function of the image exposure

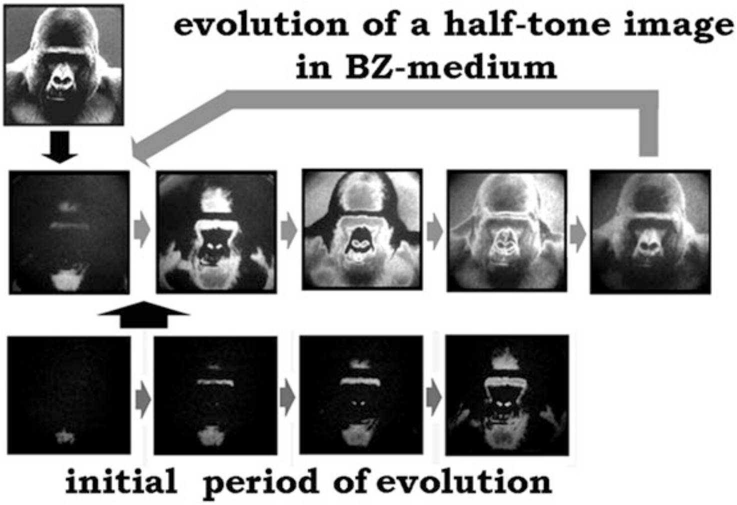

evolution of a half-tone image in BZ-medium
initial period of evolution

Fig. 5.14 Details of image processing by the Belousov–Zhabotinsky medium functioning in the vibrational mode

point, appear as a continuous sequence of fragment details with different blackening, beginning with the darkest parts. Detailed registration of this process (Fig. 5.14) creates an opportunity to build a histogram of the distribution of the blackening in each of the image fragments.

Belousov–Zhabotinsky media operating in the oscillatory mode are characterized by high complexity of behavior. First of all, this is manifested during the

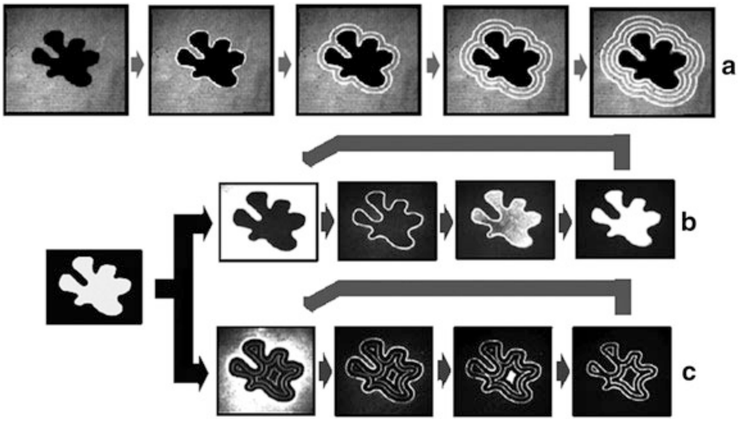


Fig. 5.15 Vibrational mode of the Belousov–Zhabotinsky medium under conditions of high illumination

optical input of the image. When the exposure of the image exceeds a certain value that depends on the composition of the medium and the intensity of radiation, the medium begins to operate in a mode similar to that of the guiding center. When a black-and-white image is entered, contours appear on the borders of its fragments that propagate along the illuminated areas of the image (Fig. 5.15). The contours appear sequentially at intervals of 20–40 s and cannot be suppressed by light radiation. A remarkable feature of this process is that if the flashing exposure (image input) is arrested at some, the medium immediately begins to operate in a vibrational mode (Fig. 5.16). In this case all the appearing contours are involved in the oscillations. When the illumination is turned on again the medium begins to function in the guiding center mode.

Qualitatively, the process of optical information input can be described as moving the point of intersection of zero isoclines (Fig. 5.2) as the isocline inhibitor shifts under the influence of light radiation. At a certain magnitude of illumination this point undergoes transition from the vibrational to the excitable mode and, apparently, becomes a guiding center.

This feature of the environment functioning in the vibrational mode plays an important role in image processing. The number of emerging contours is not connected to the structure of the image and is solely determined by the specific features of the dynamics of the medium—the composition and the exposure of the image. Therefore, when the medium performs image processing operations in the vibrational mode, the maximum exposure value must first be determined while the medium still remains in the vibrational mode.

Let us return to image processing by Belousov–Zhabotinsky media in the oscillatory mode. They can perform virtually any operations of black-and-white image processing. But the most interesting possibilities arise in the case of halftone images. As already evident from the examples shown before, a unique property of the Belousov–Zhabotinsky media is that they convert the spatial distribution of

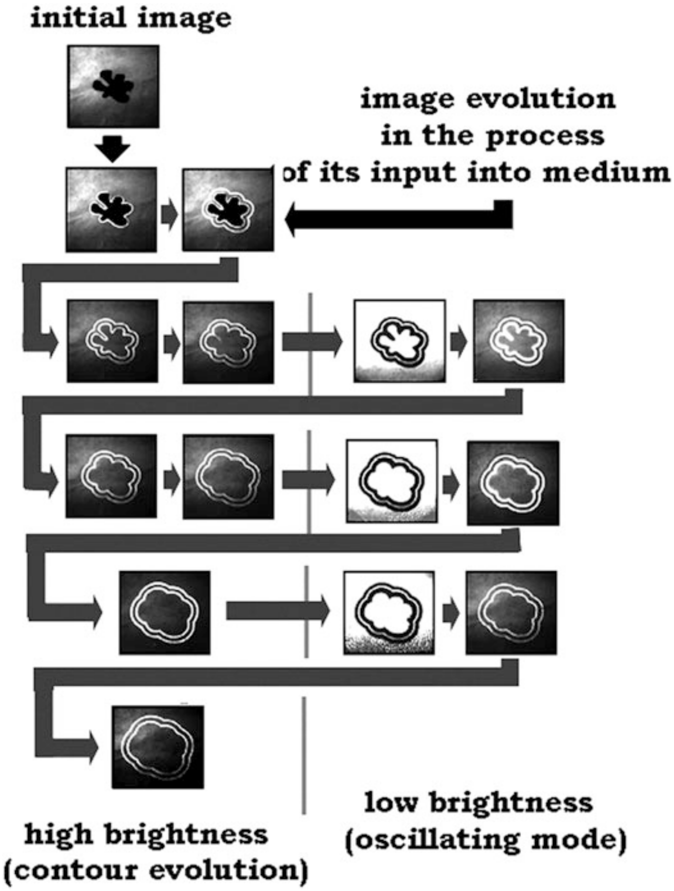


Fig. 5.16 Switching of modes of the Belousov–Zhabotinsky medium by light irradiation

complex information into a time sequence of fragments of this information, such that at each stage certain parts of fragments with the same blackening of the image get detected. In other words, a Belousov–Zhabotinsky medium in an oscillatory mode represents a natural realization of the space–time processor. This allows using them for performing complex tasks of image processing.

Here are some examples.

A “latent image” is a piece of an image with the brightness only slightly different from the brightness of the background picture. An example of such a situation is shown in Fig. 5.17. The difference in brightness between the image of the eagle and the background (Fig. 5.17a) constitutes 10 units of the HSB model in the Photoshop software. A manifold increase in brightness and contrast of the image does not significantly improve the situation (Fig. 5.17b). At the same time the evolution of the image in the Belousov–Zhabotinsky medium allows to detect the image of an

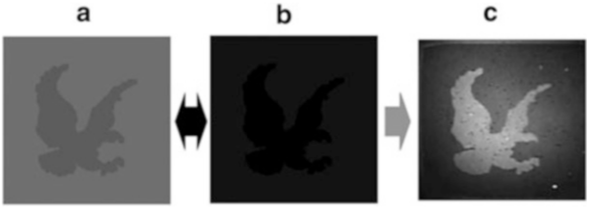


Fig. 5.17 Selection of a “latent image” by the Belousov–Zhabotinsky medium

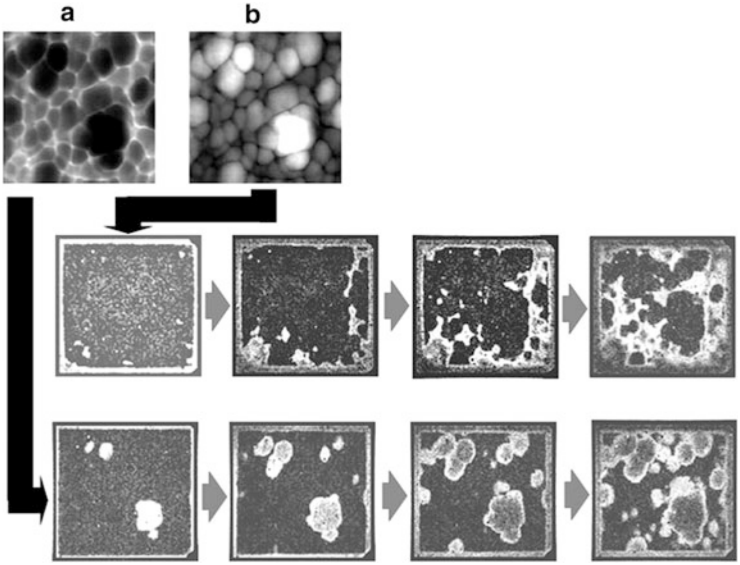


Fig. 5.18 Example of solving the problem of detecting fragments in the images of the same brightness for a positive (**a**) and negative (**b**) form of the image

eagle, despite a very small difference in brightness of the fragment and the background (Fig. 5.17c).

An important role in medicine, materials science, geology, and other areas is played by. Figure 5.18 shows an example of the solution of this problem for a positive (a) and negative (b) form of the image.

One of the problems often encountered in practice for halftone images is “watershed operations” which provide information on the shape of the relief shown in the image. An example of such a problem is shown in Fig. 5.19.

Analysis and decoding of images obtained from satellites is an important problem whose solution finds important applications in various fields of human activity. Evolution of the images of this type (Fig. 5.20) in the Belousov–Zhabotinsky medium allows to split them into separate parts and to simplify their analysis.

A related problem is the detection of the road network on an aerial photograph. Attempts have been made to use the technique of mathematical morphology to

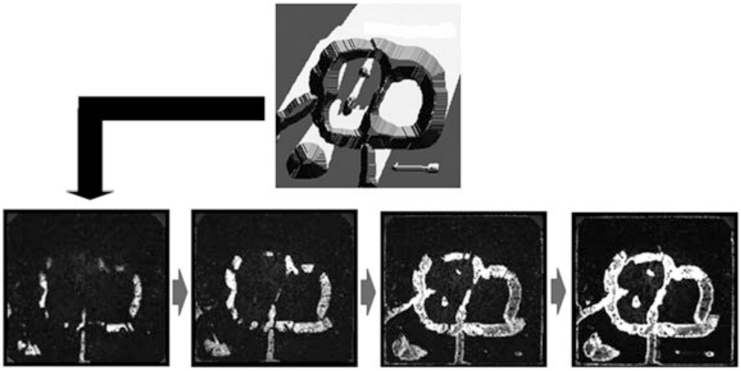


Fig. 5.19 "Watershed" operation performed by the Belousov–Zhabotinsky medium

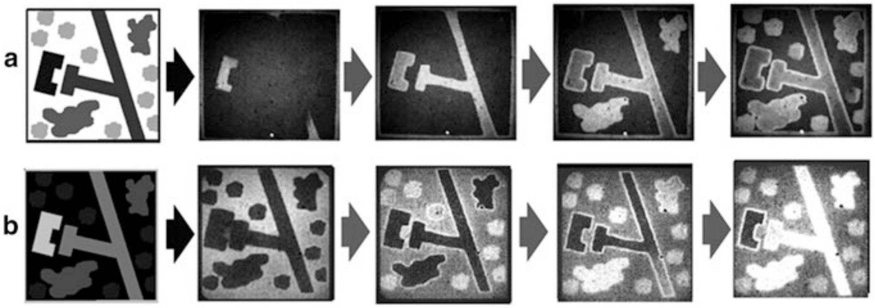


Fig. 5.20 Processing of aerial photographs by the Belousov–Zhabotinsky medium: positive (a) and negative (b) photographs

extract the network of roads in the aerial photographs of cities. The same problem can be solved by the Belousov–Zhabotinsky media. Figure 5.21 shows the original image of the city center (Fig. 5.21a), a preliminary detection of the road network by mathematical morphology (Fig. 5.21b), and an analogous operation performed by a medium of the Belousov–Zhabotinsky type (Fig. 5.21c).

5.2.3 Images, Optical Illusions, and Reaction–Diffusion Media

The vision of mammals and, in particular, human vision are the most complex and sophisticated phenomenon of the activity of the cerebral cortex. The image of the reality surrounding a man is projected onto the eye retina, and after a series of transformations in the optic tract, it enters the visual cortex (Fig. 5.22). This is a

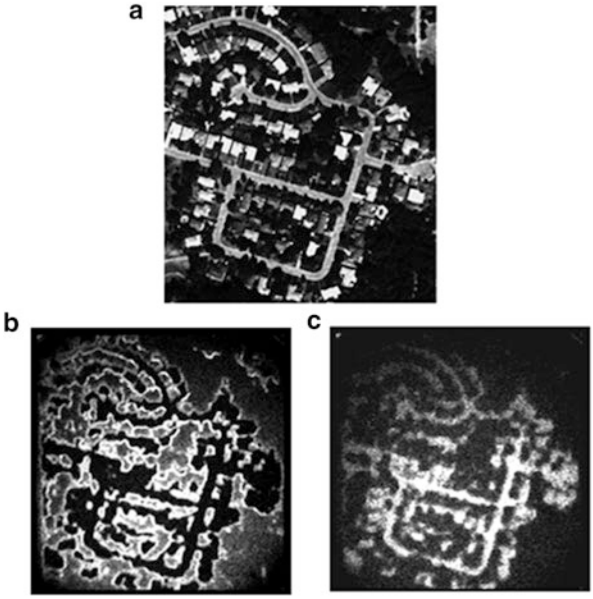


Fig. 5.21 Selection of the road network (**a**) by the technique of mathematical morphology (**b**) and the Belousov–Zhabotinsky medium (**c**)

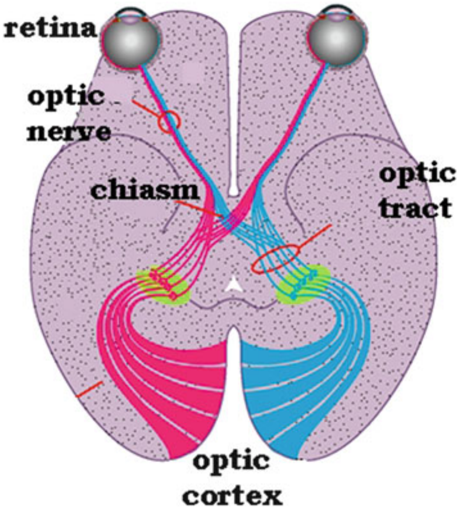


Fig. 5.22 Scheme of image processing by human brain

division of the cerebral cortex that performs primary image processing. It is believed that the visual cortex performs simple operations—enhancement of contours of image fragments, fragment boundaries, lines of one direction, etc.

Above, attention was drawn to the fact that to explain particular functions of the cerebral cortex, especially those associated with vision, Stephen Grossberg proposed a concept of specialized neural networks. According to Grossberg, their

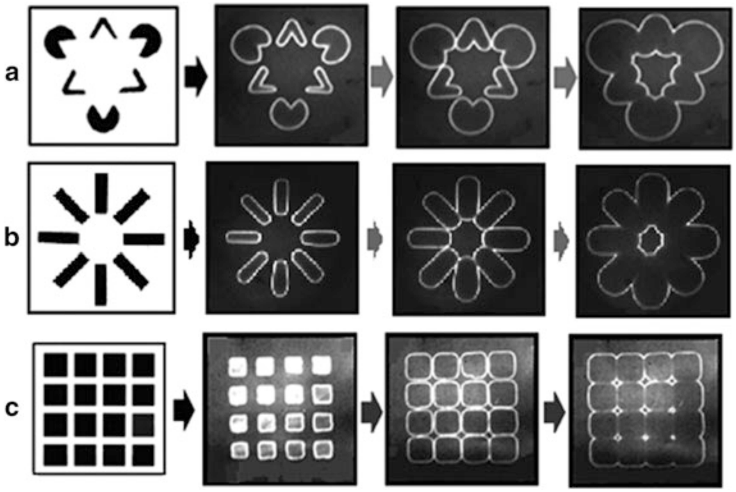


Fig. 5.23 Simulation of the illusions of Canis (**a**, **b**) and phantom points (**c**) by the Belousov–Zhabotinsky medium

functional properties are in many respects equivalent to specific chemical reaction–diffusion media, in particular the Belousov–Zhabotinsky media. For this reason the quest to understand whether it is possible to simulate certain features of human vision, using a chemical reaction–diffusion media, emerges.

One of the features of human interaction with the outside world is the false perception of its individual phenomena—illusions. A large variety of illusions associated with vision and tactile perception are known, but the most interesting among them are optical illusions.

One of the most frequently mentioned ones is the illusion of Canis. On the image depicted in Fig. 5.23, one clearly sees a triangle, which in fact is not present in the picture. In the case of a system of black squares, dark blurry formations—an illusion of phantom points—appear in the spaces between their corners. If these images are entered into a Belousov–Zhabotinsky medium that is functioning in the excitable regime, it turns out that as a result of the evolution of the enhanced image contours either a triangle (Canis illusion) or a system of points in the intervals between the corners of squares (the illusion of phantom points) appears. It therefore appears that the information capabilities of the visual cortex modeled by the Grossberg neural networks, and as a consequence, by the dynamics of reaction–diffusion media, include not only the basic operations of image processing.

The vibrational mode of Belousov–Zhabotinsky chemical environments apparently allows for making assumptions about the mechanisms of other optical illusions. The illusion called the “vase of Rubin” involves the human eye seeing either a vase or two human profiles (Fig. 5.24a). In this case perception is often

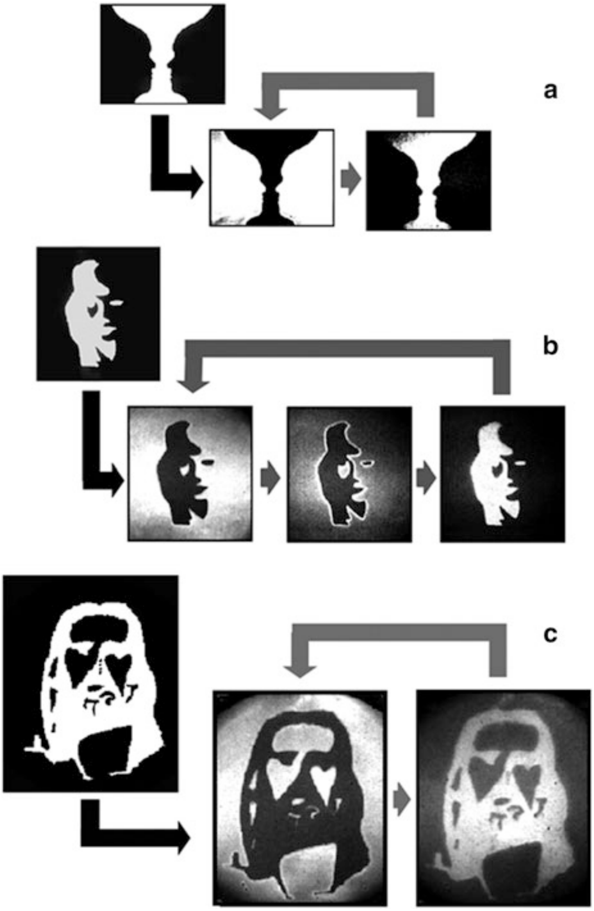


Fig. 5.24 Simulation of the illusions of the "vase of Rubin" (**a**), "saxophone player" (**b**), and "Jesus Christ" (**c**) by the Belousov–Zhabotinsky medium

blurred—the observer has difficulty choosing between these two possibilities (the dark central part of the image appears to be closer in perception to a vase, and the dark periphery to the two profiles). In the process of evolution in a reaction–diffusion medium these two images become equal and constantly replace one another. This can probably explain the uncertainty in the perception of the Rubin's vase.

And finally, the transformation of the image in an oscillatory Belousov–Zhabotinsky medium may explain why, after protracted observation of an image that does not cause any reasonable associations, a sensible figure (illusion of Jesus Christ) can be distinguished if one subsequently looks at a white surface.

5.3 Reaction–Diffusion Processors: Defining the Shortest Path in a Maze

5.3.1 Characteristics of the Problem of Finding the Shortest Path in the Maze

The problem of finding the shortest path in the maze, determined by certain conditions, is one of the best known contemporary problems of high computational complexity. A number of attempts to find efficient algorithms for its solution were undertaken starting from the 1960s of the past century, including the use of nonlinear reaction–diffusion media (mainly the media of the Belousov–Zhabotinsky type). Especially important was the work of American researchers Steinbock, Toth, and Showalter who offered a way to determine the shortest path in complex mazes based on the use of trigger waves propagating in a Belousov–Zhabotinsky medium.

Nevertheless, conclusions about the practical application of reaction–diffusion media were for a long time quite pessimistic, because the propagation velocity of the trigger waves excited in these media is small (~3 mm/min), with the characteristic size of the maze in the order of centimeters. Only relatively recently it has been shown that:

- Known photosensitive reaction–diffusion media of the Belousov–Zhabotinsky type can be effectively used to solve at least not very complex maze problems.
- An effective technique can be developed for finding a path in a maze based on the information about successive stages of wave propagation through it.

5.3.2 Basic Solution Principles of the Problem of Finding the Shortest Path in the Maze

Let us define a maze as an object whose topological properties can be described with a limited directed graph. This means that the object is composed of an arbitrary number of vertices and edges joining them. In accordance with the specifics of the maze let us subdivide the vertices into four types: starting points, which are entry points into the maze (index of such vertices is equal to 1), intermediate points (index of the vertex greater than or equal to 2), dead ends, and destination points (index of such points is equal to 1).

The simplest graph in terms of the structure is a tree (Fig. 5.25). It has one starting point and an arbitrary number of branches and destination points. More complex are multigraphs containing cyclical combinations of edges. In this case at least two routes connecting the chosen starting point and the destination can be determined.

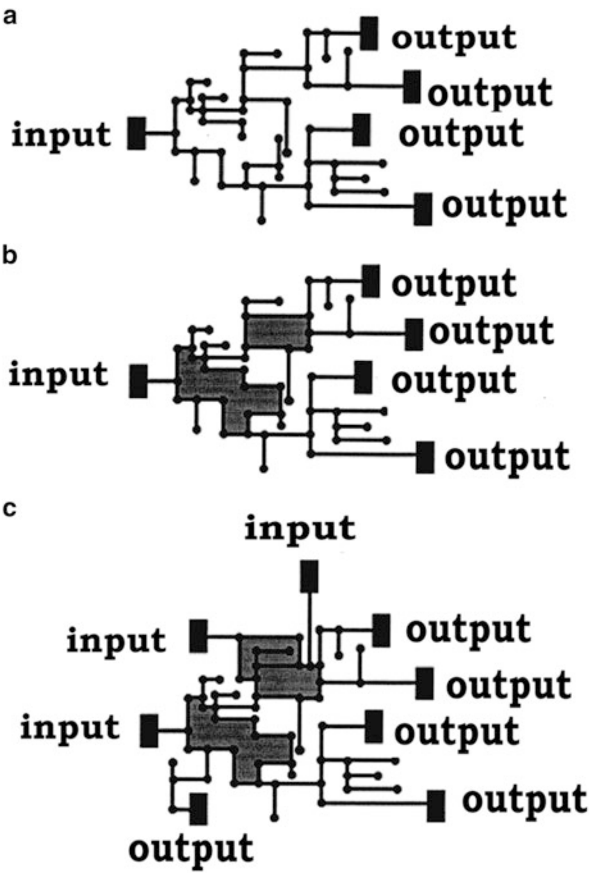


Fig. 5.25 Mazes of different complexities: (**a**) simplest treelike; (**b**) treelike containing cycles; (**c**) complex maze with cycles and the arbitrary number of entrance points and exit points

When developing the technique to determine the shortest path in a maze based on reaction–diffusion media, three basic principles were used at the faculty of physics of the Moscow State University:

1. Information systems that operate on the basis of reaction–diffusion media and capable of solving maze problems must have hybrid architecture, i.e., be a combination of a reaction–diffusion medium and a general-purpose digital computer. In this case the operations of high computational complexity, such as parallel wave propagation in a maze, are performed by the medium, and post-processing of data, which is a task of relatively low computational complexity, is performed by a general-purpose computer.

   Below follow several observations that are important for further understanding of the possibility of creating an efficient computational procedure for finding the shortest path in the maze.

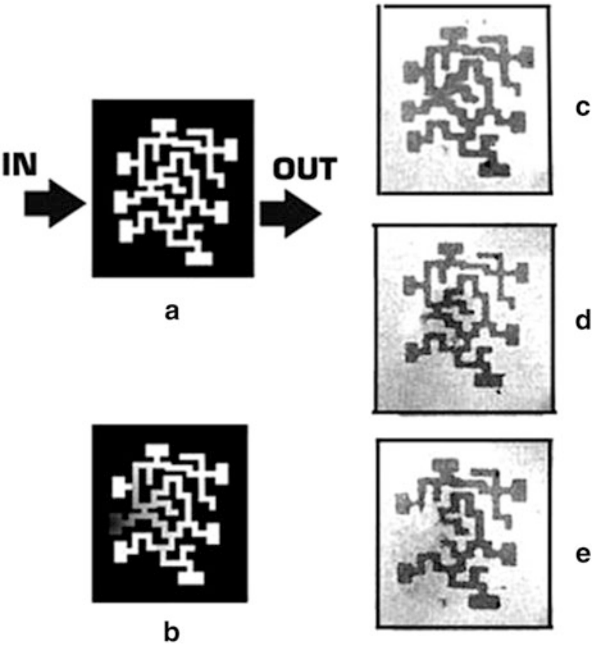


Fig. 5.26 Organization of the passage of the wave in the maze: (**a**) original image of the maze; (**b**) uniform variation of the intensity of the path background in the original image; (**c**)–(**e**) evolution of the image (**b**) in the Belousov–Zhabotinsky medium

The basis of the procedure is the representation of the maze in a reaction–diffusion media and in computer memory in the form of its image (in the simplest case—a black-and-white image, Fig. 5.26). Let us assume that only the entry point into the maze is defined and there exists a technique for recording successive stages of wave propagation, initiated at the entry point, into the computer memory. When the wave propagates along any path in the maze, the black color of the maze will change into the background color (white). After recording the wave movement in the computer memory, a more or less complex calculation method can be applied that allows to trace the movement of the wave front and to determine the point at which the black image of the path in the maze will disappear. But in this case it is impossible to determine the path from the entrance to the labyrinth to its exit, since there is no criterion that would allow to distinguish dead ends from exit points of the maze.

Thus, in order to find a way through the maze, a starting point and a destination must first be specified. More precisely, the image of the labyrinth must have a structure that highlights destination points and helps determine the moment when the wave reaches them.

Let us further assume that the starting and the destination points are given and that the trigger wave propagates through the maze. After some time the wave reaches the first destination point, nearest to the entrance of the maze. After that the wave will successively reach other destination points.

It is easy to see that in this case relative lengths of the paths, and not the paths themselves, can be determined based on the wave transit time. Therefore, to find the shortest path in the maze, an additional algorithm must be specified.

These considerations determine the basic features of the computational procedure for finding a path in the maze based on the wave processes inherent in nonlinear reaction–diffusion media.

Wave propagation through a maze is a parallel operation of high computational complexity. Reaction–diffusion media can effectively carry out such an operation, and its successive stages can be stored in the memory of a digital computer.

It will be shown below that for finding the shortest path in the maze between the specified entry and exit points, a procedure of low computational complexity implemented by a digital computer can be proposed. The basis for this procedure is constituted by processing of images describing wave propagation through the maze, recorded in computer memory.

2. A fundamentally important issue is the formation and storage of the maze image in the medium as the wave passes through it. Attempts are known to solve it by carving the image of a given maze out of an ion-exchange membrane in the catalyst of the Belousov–Zhabotinsky reaction is immobilized. Efforts have also been made to form a maze by printing its image on the surface of a membrane, using a catalyst solution instead of printer ink. Nevertheless, these methods are not effective for creating data-processing devices capable of quickly switching from one type of the maze to another one and transform the maze during the search for the shortest path. Media of the Belousov–Zhabotinsky type excited by light are probably particularly suitable for solving maze problems. Their main property is that they store input information for a sufficiently long time. Therefore, the appearance of the image of the maze in the medium and the process of its evolution during the passage of the wave can be recorded by a camcorder and input into a digital computer. After that, finding the shortest path from the maze entrance to the selected exit can be reduced to standard digital image processing operations.
3. The crucial point in solving maze problems using reaction–diffusion media is sufficiently rapid excitation of the wave process.

   Two types of wave processes propagating in reaction–diffusion media are known:

The first one represents trigger waves arising as a result of the interaction of chemical processes and diffusion of medium components. The speed of propagation of trigger waves is small—0.05 mm/s.

The second process represents phase waves that propagate independently of the diffusion along the phase gradient created in some way. Phase waves are fast, but difficult to control.

To determine the path in the maze phase waves excited by light radiation were used. If projected onto the surface layer of the medium, an image whose intensity varies monotonically along the surface of the medium will undergo evolution which will occur sequentially, with a shift over time, starting with the darkest parts of the image. Therefore, by imposing monotonically changing light background of a given shape on the original black-and-white image of the maze, it is easy to initiate a phase wave at the desired point of the maze.

The procedure for finding the shortest path in the maze was implemented in two main stages.

On the first stage the negative—white on black background—image of the maze was entered into the reaction–diffusion medium (more precisely, the image of the maze plus the background monotonically changing along the maze was projected onto the surface of the medium with a certain selected exposure, Fig. 5.26). A positive image of the entered image appeared in the medium, in which in the course of its further evolution, a wave began to propagate with a certain delay (color change of the path in the maze from black to white). Successive stages of its propagation were recorded by a video camera and then entered into the memory of a PC.

Images recorded in the computer memory were subsequently used for finding the shortest path, which was implemented as a sequence of standard digital operations performed by a personal computer.

The developed method turned out to be rapid and efficient in the case of linear mazes, where one entrance is connected to an arbitrary number of exits, and the direction of the path varies by no more than 90°.

5.3.3 The Procedure for Finding the Shortest Path in the Maze

The procedure for finding the shortest path in the maze consists of two main stages.

The first stage is the excitement of a phase wave at the selected point of the maze and the recording of successive steps of wave propagation through the maze into the computer memory.

The second stage is the numerical analysis of these images to determine the shortest path between the start and the end point of the maze.

Let us discuss this procedure, starting with the case of a simple, linear, treelike maze with one entrance and multiple exit points (Fig. 5.27).

Suppose that a monotonically decreasing background is superimposed on the original image at the selected entry point into the maze. After the projection of the combined image onto the plane of the Belousov–Zhabotinsky reagent, a negative image appears in the medium. At first the image of the original maze appears (black image on white background). Then a propagating phase wave arises which successively changes the black color of the maze into the background color.

The time of the wave propagation through the maze depends on the gradient of intensity of the imposed background. By varying the gradient it is easy to make this time sufficiently small (about 3–5 s), i.e., smaller than the lifetime of the negative phase of the image in the process of its evolution in the medium.

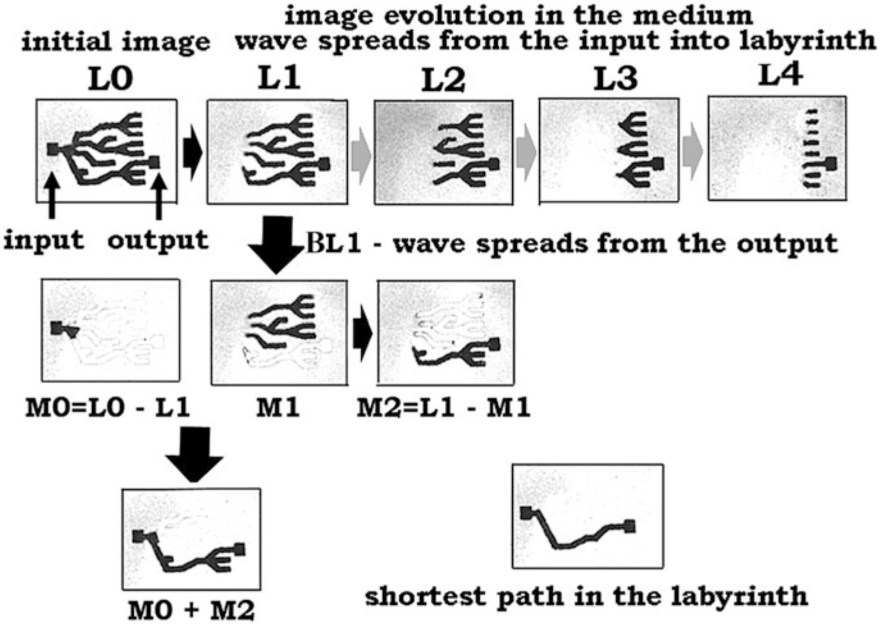


image evolution in the medium
initial image
wave spreads from the input into labyrinth
L0
L1
L2
L3
L4
input
output
BL1 - wave spreads from the output
M0=L0 - L1
M1
M2=L1 - M1
M0 + M2
shortest path in the labyrinth

Fig. 5.27 Determination of the shortest path in a simple treelike maze

The propagation of waves through the maze from the starting point of the maze to its end point is shown in Fig. 5.27. Since this process takes about 3–5 s, it is easy to record successive stages of the propagation of the wave by a video camera and to store them in computer memory. Some of these consecutive images are shown in Fig. 5.27.

Digital processing of the images recorded in the memory was used for finding the shortest path from the entry point into the maze to the destination point.

Various algorithms are known for determining the shortest path from the entrance to the labyrinth to its exit. In this study a different procedure was used which is apparently more effective for the technique based on wave propagation in a maze.

In the process of wave propagation through branch points the maze is split into two (or more) fragments (Fig. 5.28). One of them is connected with an exit from the maze while the other one is not. It is easy to identify a fragment that is connected with the end point of the maze by initiating a backward wave from the exit of the maze. As a result fragments connected with the exit change their color (from black to the background color), while the color of the fragments not connected with the exit remains unchanged.

If the maze is not very complicated, it is possible to replace this auxiliary reaction–diffusion process by a standard image processing procedure, i.e., to use the "fill" of black fragments by the background color (Paint Bucket operation of the Photoshop software package), initiated at the maze exit point. Subtraction of the

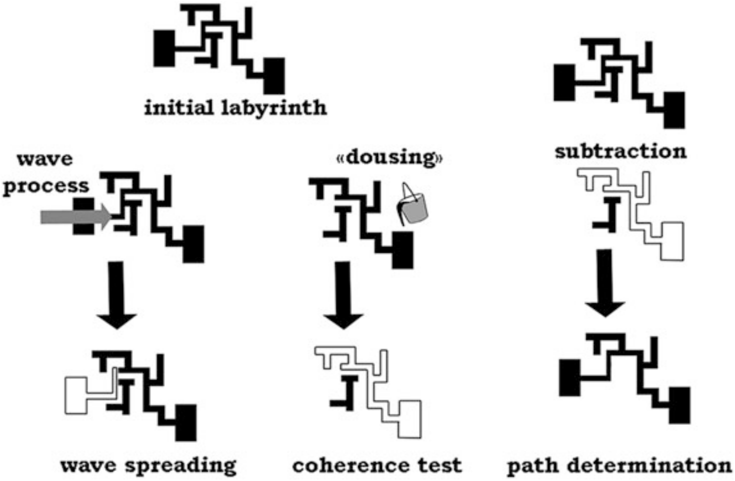


Fig. 5.28 Main stages of the algorithm of the determination of the shortest path in the maze

image obtained after the Paint Bucket operation from the initial image of the maze allows for removing the fragment not associated with the exit.

All image processing operations used to implement the discussed procedures can be performed using the software package Adobe Photoshop 5.0.

Successive repetition of this procedure for each passage of the branch point allows to eliminate all dead-end branches (and the paths to other possible end points) and to determine the path from the maze entrance to the selected end point.

We make a remark about the applicability of the procedure for finding the shortest paths in linear treelike mazes.

The advantage of this procedure is that there is no need to determine the location of the branch point by a human operator. The shortest path can be found as a result of processing of individual images, sequentially recorded during wave propagation.

Moreover, for each image the following sequence of operations is performed (assuming that the recording process begins with the first of them, L1; see Fig. 5.27):

- Subtraction of the L1 image from the initial image of the labyrinth L0, in order to determine the path that the wave passes from the beginning up to the point of wave propagation being considered (L0–1).
- Filling in the fragments L1 connected with the end point by the background color (L1–1).
- Subtraction of L1–1 from L1, i.e., clipping the paths not connected with the end point (L1–2). The subtraction of any fragment of the background intensity must correspond to the zero level of intensity.
- Addition of L0–1 and L1–2 to determine the current image of the maze (L01), which is used in the next step instead of L0.

- Changing L0 into L01.

The result of the operation corresponding to each step, i.e., each current image of the labyrinth, will be:

- The same as the current image in the previous step, if the wave does not pass through the branch point
- Modified current image, with some parts of the maze discarded, if the wave passes through branch point

Generally speaking, the initiation of a backward wave with the help of the Paint Bucket operation to determine the portion of the maze not connected with the end point is not quite correct. However it appears possible to use this operation if the maze is not very complicated and the time of passing the maze is not very long.

The above procedure is simple and effective in the case of linear treelike mazes where all possible paths from the entry point to the destination have roughly the same direction, coinciding with the direction of the wave propagation.

In general, this procedure should be changed because light-initiated phase waves propagate along a gradient of background intensity, and not along the path in the maze, which can change its direction.

Changing the proposed procedure enables the use of its basic principles and to develop a technique suitable even for very complex mazes.

Such a step-by-step technique is based on dividing the maze into linear treelike fragments and sequentially processing each of them. An elementary step of this method involves the following operations:

- Determining the direction of the path in the maze at the chosen starting point
- Excitation of the phase wave at this initial point, which propagates along the gradient of the background (in this case, the gradient coincides with the direction of the path in the selected fragment of the maze), and recording sequential steps of wave propagation
- Pruning of possible dead-end branches
- Determination of the turning points of the path

The elementary step ends at the turning point of the path in the maze, which is further used as the next starting point.

5.3.4 Effectiveness of the Method

There are quite stringent conditions of practical realization of the developed technology to find the shortest path in the maze.

The key distinguishing feature of the Belousov–Zhabotinsky system, based on the Ru-catalyst, is its high sensitivity to small changes in experimental conditions. Therefore, the quality of the optical system used to input the initial data should be high. Moreover, the reaction cell must be carefully protected from ambient light radiation.

Moreover, the complexity of the problem necessitated the development of a reaction–diffusion medium with a structure optimal for this problem.

A two-level reaction–diffusion system was formed. The catalyst in this system is immobilized on the surface of a solid substrate (a thin layer of silica gel), and all other reaction components are in a liquid phase and the reaction takes place at the phase boundary. In this case, the reaction of the Belousov–Zhabotinsky type occurs at the boundary between the liquid phase and silica gel, thus significantly improving the quality of the image of the maze (sharpness, resolution, etc.). Immobilization of the catalyst protected the image from distortions that could be caused by random external influences (shock, vibration, etc.).

Let us make some remarks concerning the effectiveness of the proposed method.

When the wave passes through a branch point, some part of the maze not connected to its exit is discarded and does not participate in the further steps of the procedure. The more complex the maze, the larger parts of it are cut off in the process of finding the shortest path. Therefore, a remarkable property of the proposed method is the increase of its efficiency as the maze becomes more complex.

The step-by-step procedure greatly increases the time required for finding the shortest path in the maze. The effectiveness of the procedure depends on the operating speed of the reaction–diffusion medium and the number of turning points in the maze. Nevertheless, it turns out to be higher than the efficiency of the procedure that uses trigger waves. The time required for processing a maze of an average complexity is approximately 5 min, with the duration of one cycle of the medium about 40 s. This time is an order of magnitude smaller than that of the procedure using trigger waves.

An important feature of the proposed method is also the fact that the time required to determine the shortest path depends linearly on the number of branch points of the maze.

5.4 A System of Interconnected Reaction–Diffusion Reactors: Pattern Recognition Devices

In recent decades most of the experimental and theoretical studies of reaction–diffusion systems have been devoted to the complex processes occurring in continuous homogeneous media. Considerably more complex dynamics corresponds to a set of interrelated reaction–diffusion subsystems operating in a general case, in different dynamic regimes.

Let us assume that the system is built from simple chemical fragments (subsystems) with full mixing, each of which described by the one-dimensional kinetic equation:

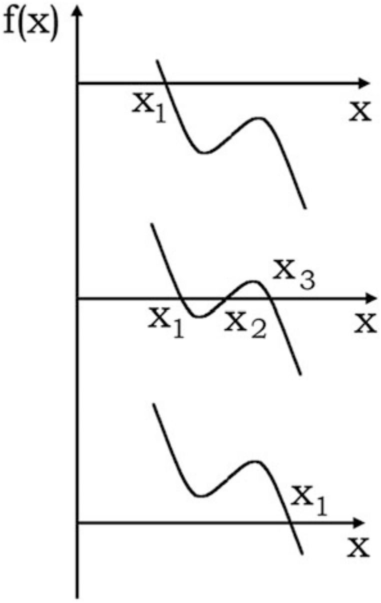


Fig. 5.29 Zero isoclines of the one-dimensional reaction–diffusion equation

$$dx_i/dt = f(x_i).$$

Attractors of this system are stationary states that correspond to the intersections of the zero isocline $f(x_i) = 0$ with the abscissa (Fig. 5.29). These states can be stable or unstable. Three variants of the dynamics are possible for a system described by the considered one-dimensional equation. Two of them apply to the case when the system has one steady state, and the third one to a system with three states. A simple analysis of the stability of the system shows that if $f(x)$ takes only one zero value, the steady state is stable. If three zero points correspond to $f(x)$, then the two steady states ($df(x)/dx < 0$) are stable, while the third one ($df(x)/dx > 0$) is unstable. In this case the system is bi-stable, i.e., it can be in two different states.

Consider a network consisting of subsystems of this type, which are connected by diffusion, mass transfer, or some other mechanism. Then, if all subsystems are mono-stable, i.e., have one stable state, the network also has only one stable state, which is homogeneous. If the subsystems have three states, the network can have up to 2^N stable states. Each subsystem can in principle have two states, and the network structure is given by different distributions of states of all subsystems.

Stationary structures in related reaction–diffusion systems have been studied experimentally in recent years. In particular a system has been studied built on the basis of 16 linearly related reactors with complete mixing. The dynamics of the medium was determined by the well-known chlorite–iodate reaction:

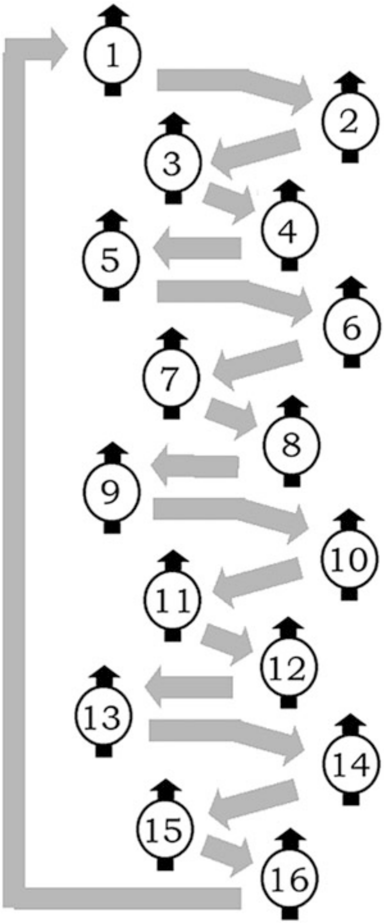


Fig. 5.30 Scheme of a flow system of reactors for the interacting reaction–diffusion systems

$$5ClO_2^- + 2I_2 + 2H_2O \rightarrow 5Cl^- + 4IO_3^- + 4H^+.$$

This is a bi-stable system, whose states depend on the rate of flow of the reagent in a flow reactor. High and low concentrations of iodine correspond to two states of the system. The state of each reactor was determined by the color of the reagent. The state was visualized by introducing into the flow inside the reactor starch, which has a deep blue color when in complex with iodine. Two 16-channel peristaltic pumps (for chlorine and iodate mixtures separately) were used to fill all reactors with an identical mixture of components.

Two more 16-channel peristaltic pumps performed mass transfer. The reactors were divided into two groups, odd and even, to overcome experimental noise. A schematic diagram of the interconnection of reactors is shown in Fig. 5.30.

Two different versions of the dynamics of the simulated neural network were investigated.

The first one involved filling the first reactor only with a solution of sodium chlorite, while all others remained in the states with high concentration of iodine and were colored blue (asymmetric boundary conditions). As a result of the subsequent evolution of the system, after the links between the reactors were activated, one observed the propagation of the wave switching from a state with a high content of iodine into the state with its low content (i.e., color changes from blue to colorless).

In the second version the boundary conditions were symmetrical (i.e., the first and the last reactor were initially filled with a solution of sodium chlorite). In this case one could observe sequential appearance of various stationary structures, depending on the flow speed in the system of reactors (Fig. 5.31).

In the early 1990s of the past century the group of the well-known physical chemist John Ross at the Stanford University (USA) made an important step in understanding the importance of related reaction–diffusion systems for information processing. They used the network architecture of these systems in order to consider theoretically the possibility of creating on their basis of information-logical devices and, in particular, the implementation of Turing machines and neural networks of the Hopfield type.

Later it was shown experimentally that neural networks based on chemical systems have the ability to recognize simple images (see details in reference 5). They used a bi-stable environment based on the reaction:

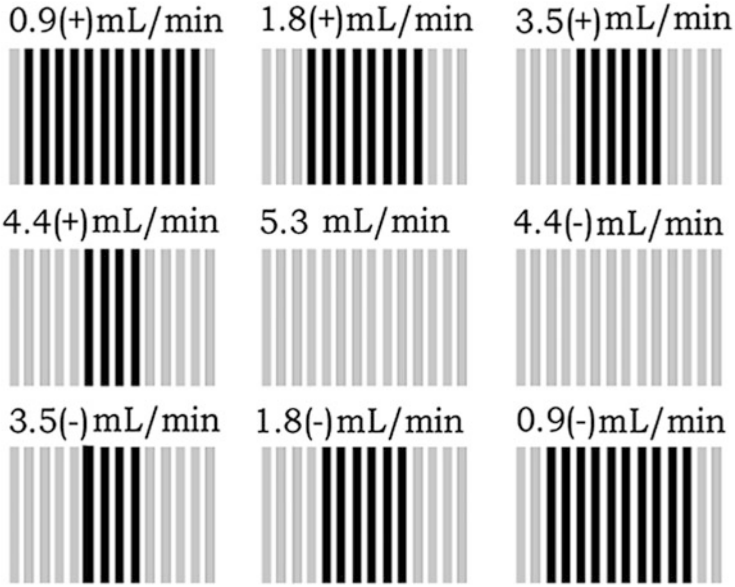


Fig. 5.31 Stationary structures appearing in a system of connected reaction–diffusion reactors

$$2IO_3^- + 5H_3AsO_3 + 2H^+ \rightarrow I_2 + 5H_3AsO_4 + H_2O.$$

Let us call the initial distribution of the states of neurons in the network an initial image. It is transformed by the network during its evolution in accordance with the rules defined by the matrix of neural connections.

A neural network can memorize a certain number of images, whose number is determined by the network structure. One of the most famous memory circuits is the so-called Hebb's rule.

Assume that the following images are stored:

$$V^\mu = \left\{V_1^\mu, \ldots, V_N^\mu\right\} \quad \mu = 1, 2, \ldots, M.$$

Then, for binary vectors, the Hebb's rule is defined as

$$T_{ij} = \sum_\mu (2V_i - 1)\left(2V_j - 1\right).$$

It can be shown that this means that:

- Two elements of the connection matrix of the neurons T_{ij} that are in the same state are connected in most of the memorized patterns, and the power of the connection depends on the number of structures in which these elements are in the same condition.
- Elements that are in different states in the majority of memorized patterns are not connected.

In the case of chemical neural networks, the Ross group used a modified Hebb's rule. In this case, the strength of connection between the elements i and j is defined as

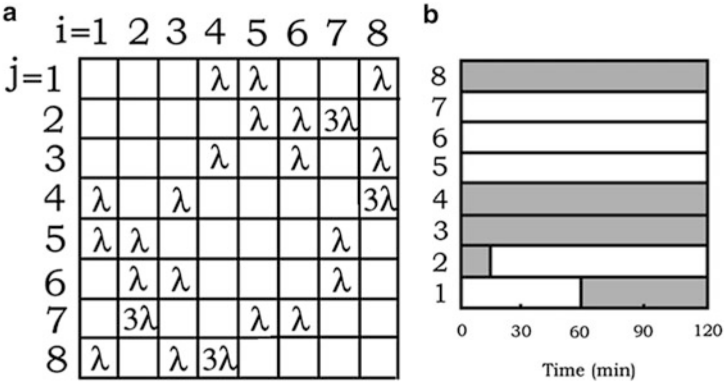
$$T_{ij} = \lambda\vartheta\left\{\sum_\mu \left(2R_i^\mu - 1\right)\left(2R_j^\mu - 1\right)\right\},$$

where λ is the coupling constant, which is determined in the case of chemical networks as a flow rate of reagents through the reactor:

$$\vartheta\{x\} = x, \quad \text{if } x \geq 0,$$
$$\vartheta\{x\} = 0, \quad \text{if } x < 0.$$

It was shown that when using this rule, steady states are generated corresponding to the stored patterns.

These theoretical considerations were the basis for an experimental realization of chemical neural networks. A set of eight reactors of complete mixing with a continuous flow of reagents was used for storing the three images (they were some



	i=1	2	3	4	5	6	7	8
j=1				λ	λ			λ
2					λ	λ	3λ	
3				λ		λ		λ
4	λ		λ					3λ
5	λ	λ					λ	
6		λ	λ				λ	
7		3λ			λ	λ		
8	λ		λ	3λ				

Fig. 5.32 Connection matrix for the chemical neural network (**a**) and recognition of the numerical sequence (**b**)

random sets of numbers). These reactors were connected by mass transfer. The connection matrix for this system is shown in Fig. 5.32.

Assume that some initial distribution of concentrations is introduced into the system that is different from the state of the reactors being stored when their connections are disabled. Then, after the connections are activated, there are two possibilities:

- The network of reactors will have a uniform (homogeneous) distribution of states, if the input image is significantly different from the one being stored.
- One of the stored images appears in the network if the input image is close to one of the images being stored.

The efficiency of such a chemical network was tested with a set of three different images slightly different from the ones stored in the network structures. It was thus shown that the efficiency of recognition of the structures is quite high (see Fig. 5.32).

The investigation of the possibilities of pattern recognition by chemical networks, made by the authors [139], was the first experimental study in this area. It should nevertheless be noted that, despite the undoubted importance of this work, its technical solution was too complex. The equipment used for pattern recognition was cumbersome and inconvenient to operate effectively.

Other solutions for creating chemical multilevel systems for solving pattern recognition problems were proposed. A promising example is the idea of electrochemical connection between subsystems. In general, the possibilities of this approach are far from exhausted, and we can expect in the near future the emergence of new variants of chemical recognition systems.

Chapter 6
Self-Organization: A Common Principle of Information Processing by Distributed Dynamic Systems

> *A new science was born in the last several decades - physics of nonequilibrium processes, the development of which has led to the origin of such new notions as self-organization and dissipative structures used now everywhere in a wide range of disciplines from cosmology, chemistry and biology to ecology and social sciences.*
>
> I. Prigogine "*The End of Certainty*"

6.1 Self-Organization and Its Features in Various Dynamic Systems

The concept or rather the term "self-organization" has been widely used in recent years to describe and explain similar phenomena in physical, chemical, biological, and even economic and sociological systems. All these phenomena have in common that, seemingly contrary to conventional thermodynamic laws, complex ordered structures emerge in a distributed dynamic system consisting of simple parts. The properties of the resulting structures are fundamentally different from the properties of the individual elements of the system. Most surprisingly, self-organization in the system emerges spontaneously from a homogeneous state.

Convincing and consistent examples of self-organization were found in physical systems. The concept of self-organization was further extended to chemical phenomena, in which the term "self-assembly" was also widely used. In biology, self-organization became a central concept over the past half century in describing the dynamics of biological systems, from intracellular processes to the evolution of ecosystems. Today, examples of self-organization can be found in sociology, economics, and even among purely mathematical objects.

At the same time, as the idea of self-organization, as it is understood by the physicists, was being adopted in other natural social sciences, it was becoming increasingly ill defined. Even when self-organization processes really occur in the system, attempts to explain them in physical and statistical terms are not always

© Springer-Verlag Wien 2014

N.G. Rambidi, *Molecular Computing*, DOI 10.1007/978-3-211-99699-7_6

satisfactory. Often, they are rather primitive and cannot explain the whole diversity of processes in complex dynamic systems. As a result, a number of new terms were introduced in the literature with the intention to better define or supplement particular aspects of this complex phenomenon. Besides the already mentioned term “self-assembly,” further frequently used terms include “stigmergy,” “conservative self-organization,” “dissipative self-organization,” etc.

For all these reasons, before considering in detail the essence of self-organization and the mechanisms of the processes responsible for introducing order into the system, let us turn to some concrete examples of scientific and social problems.

6.2 Self-Organization Processes and Their Properties

1. Let us pour in a vessel two solutions containing silver nitrate and sodium chloride. The Ag+ ion is known to be an indicator of the content of halogen ions. Therefore, after mixing these two solutions a thick white silver chloride precipitate drops out.

2. Suppose that a polymer is being synthesized based on two different blocks, oligomers A and B, such that these oligomers consistently alternate in the polymer chain. This process is called block copolymerization. As a result, microphase separation structures arise in the synthesized polymer. These structures vary (Fig. 6.1) depending both on the relative content of oligomers A and B in the synthesized polymer and on the Flory–Huggins value (χ). It is determined by the nature of the interaction between the oligomers and by the temperature. Raising the temperature leads to a decrease in the parameter χ and to disappearance of microphase separation.

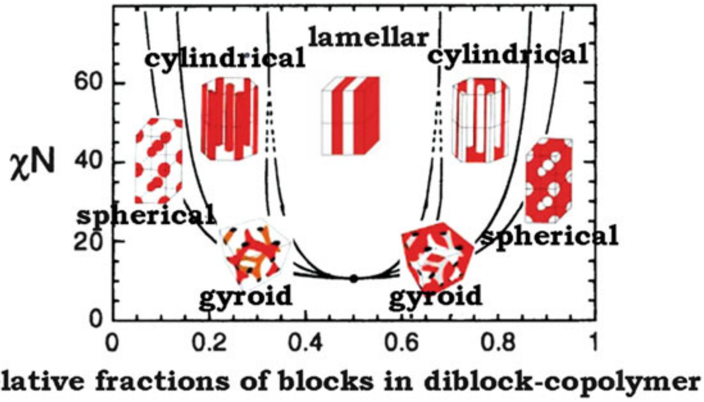


Fig. 6.1 Microphase separation structures in diblock copolymers as a function of the relative content of blocks in the polymer

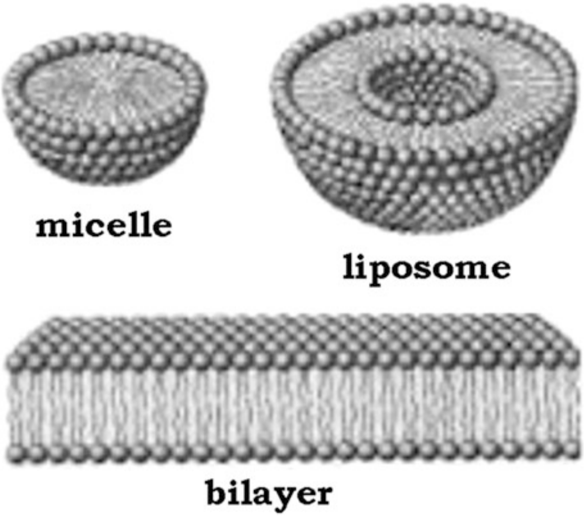


Fig. 6.2 Supermolecular structures of amphiphilic molecules

3. Among the vast number of organic molecules the so-called amphiphilic compounds are particularly important. These molecules represent a combination of fragments with opposite properties—a hydrophilic group (e.g., an acid residue) and a hydrophobic fragment, such as a long hydrocarbon chain. Due to their structure amphiphilic molecules in a polar solvent (water) tend to be on the liquid–gas boundary such that the hydrophilic "heads" are in the water, and the hydrophobic "tails" are pushed into the gas phase. A certain number of amphiphilic molecules, which may vary depending on the structure of the molecule, are found in the solution. In this case, there is a tendency to decrease the energy of interaction of the molecules with the solvent due to spontaneous formation of structural aggregates (Fig. 6.2).

In spherical micelles hydrophilic groups are on the surface of the sphere, and the hydrophobic ones are inside of the sphere. As a result, the repulsive interaction of hydrophobic tails with the solvent is reduced. In a nonpolar solvent (oil) reverse micelles appear in which the hydrophilic heads reside in the center of a spherical formation. A large number of different supramolecular structures—bilayers, vesicles, etc.—are known. Fundamentally important biological amphiphilic molecules are lipids that form the basis of cell membranes.

4. Let us pour into a pan a thin layer of a viscous fluid (e.g., oil) and heat the pan on fire, keeping the surface temperature constant. If the heating is weak, the fluid remains stationary. If, however, the fire is made stronger, increasing the heat flow, then suddenly and spontaneously the entire surface of the oil is broken into regular hexagonal or cylindrical cells (Fig. 6.3a). The structure on the pan resembles a honeycomb.

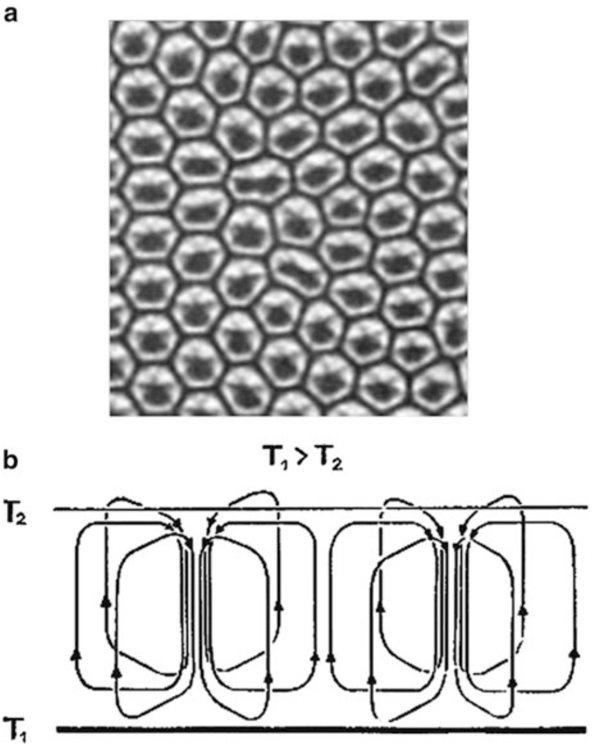


Fig. 6.3 Benard cells (**a**) and scheme of the liquid motion in them (**b**)

In 1900 an article of the French scientist Benard was published. He showed that when a layer of mercury poured into a flat wide vessel was heated from the bottom, the entire layer spontaneously broke up into identical vertical hexagonal prisms. In the central part of each cell the liquid rises and near vertical faces it sinks (Fig. 6.3b). That is, convective flows appear in the vessel, which raise the heated liquid up and lower the cold liquid down. Subsequently, these cells were called Benard cells.

5. Interesting phenomena of spontaneous development of the structure are observed in bacterial environments. One example is the study of the evolution of the colonies of the bacterium *Paenibacillus alvei*. Colonies were grown in petri dishes containing agar with some peptone added. Peptone is a product obtained by partial protein hydrolysis, which is used as a nutrient in bacterial media. During evolution bacterial media spontaneously formed complex spatial distributions of bacteria (Fig. 6.4). In this case, the distribution varied depending on the concentrations of agar and peptone.

6. Suppose that in the closed habitat (e.g., on the island which has no connection to the mainland), a population of creatures leaves. The evolution of its population over time will be determined primarily by the reproducibility parameter α, i.e., the average number of the offsprings produced by one individual. We will use a

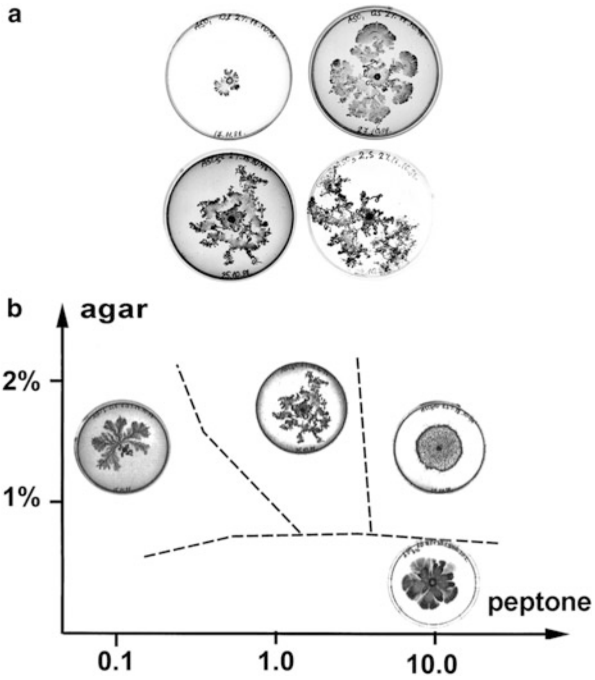


Fig. 6.4 Fractal structures arising during the evolution of the colonies of the bacterium *Paenibacillus alvei* (**a**) and distribution of bacteria in them depending on the concentrations of agar and peptone (**b**)

discrete time scale. Then the number of individuals X_{n+1} at the moment t_{n+1} should be proportional to the number of individuals X_n at the previous moment t_n:

$$X_{n+1} \Rightarrow \alpha X_n.$$

However, since the habitat is limited, overpopulation arising due to uncontrolled fertility should lead to a decrease in population, for example, due to lack of food. Therefore, introducing a restriction on the maximum population size, its evolution can be described by the so-called logistic equation:

$$X_{n+1} = \alpha X_n (N - X_n).$$

Using relative values $x_n = X_n/N$ we obtain a more convenient expression:

$$x_n = \alpha x_n (1 - x_n).$$

Based on the structure of the problem, the first thing that comes into mind is that for large n the solution of this equation tends to some limit. But the logistic equation is nonlinear with respect to x_n and its solution exhibits typical nonlinear features—bifurcations. This term refers to splitting the dependence of the solution from a certain parameter into two (or more) branches. The behavior of solutions of the logistic equation is determined by the parameter α. It is easy to see that with $\alpha < 1$

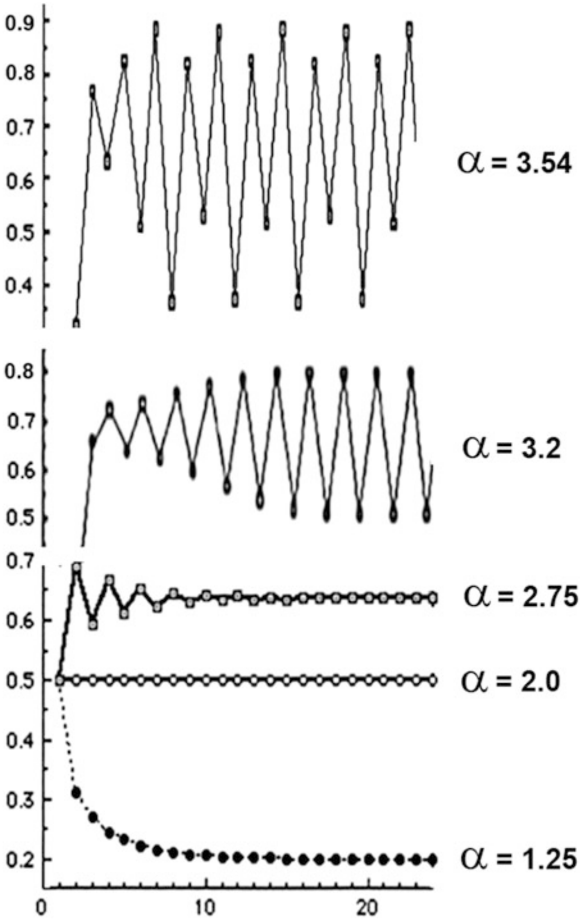


Fig. 6.5 Solutions of the logistic equation depending on the population reproducibility

the population disappears. By direct calculations it is easy to show that if the value is between 1 and 3, the solution actually tends to some asymptotic value (Fig. 6.5). But in the interval $3 < \alpha < 3.45$ the solution oscillates between two values, in the interval $3.45 < \alpha < 3.54$ between four values, and then falls into the region of chaos, when predicting the value of the solution is not possible.

Physically, the behavior of the solution can be qualitatively explained by taking into account that the processes of reproduction of population and its degradation must occur with some time lag.

Yet we should note that the logistic equation used for solving various problems is an ideal model that does not always correspond to actual processes of evolution of a complex system.

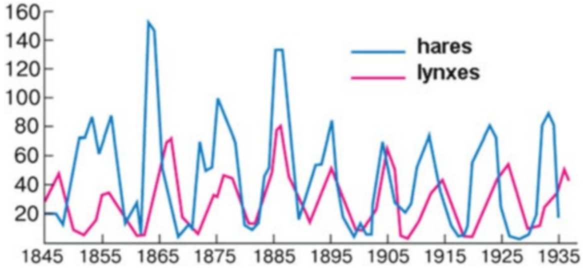


Fig. 6.6 Dynamics of the change in the number of lynx and hares in the limited areal

Similar to the logistics problem but more complex is the "predator–prey" problem. It describes the change in the level of two populations when, in addition to the conditions of the logistic problem, one of the populations (the prey) is being consumed by the other one (predators). This problem is also described by periodic solutions. As a practical example of the evolution of such a system, Fig. 6.6 shows the change in the number of lynx and hares determined based on the pelt-trading records of the Hudson's Bay Company over 90 years.

7. In social sciences, self-organization as the phenomenon that defines the processes taking place in society attracted attention more than 150 years ago. Already the founder of the classical political economy, Adam Smith in his *An Inquiry into the Nature and Causes of the Wealth of Nations* came to the conclusion that the spontaneous order on the market is the result of the interaction of different, often contradictory aspirations, goals, and interests of its participants. Such interaction leads to the establishment of the unplanned order, which is expressed in the balance of supply and demand.

By contrast, high-quality products that are in high demand and are produced in large quantities (positive feedback) increase the order, i.e., reduce the entropy, because the processes of production and exchange are accelerated. These, in turn, increase employment, better meet the needs of society, and lead to a higher living standard for people. After some time, as more goods are produced, the market gets saturated, leading to an equilibrium between supply and demand, but by that time, competing companies have already developed new products of even higher quality. Commodity–money relations become active again. And when the number of producers is sufficiently high, new offerings appear continuously. In this way the nonequilibrium of the market and the effective functioning of the economic system are maintained.

Similar ideas were expressed at that time regarding self-organization of the norms of morality in the society. In this case, the ideas of self-organization of social systems were associated with evolutionary processes. And in the late twentieth century the answer to many questions came from the natural sciences, when a striking similarity of self-organization at various structural levels of matter was discovered.

6.3 Synergetic Principles of Self-Organization

Self-organization is the phenomenon of spontaneous formation of structure in systems of different physical nature. Under the spontaneous emergence of structure we will mean the emergence of an ordered state in an initially random distribution of system components without any apparent external influence. Ordered states in the general case can be spatially uneven distribution of the material components of the system persistent in time, undamped oscillations of concentrations of components of the system when they oscillate between two or more values, more complex forms of ordered collective behavior of components. Structure formation is equally inherent in both the physical devices such as lasers and chemical reaction media, as well as in biological tissues, communities of living organisms, geological and meteorological processes, and social phenomena of human society. Self-organization mechanisms are different in systems of different nature, but nevertheless, they all share some common structural and dynamic characteristics.

Different, and often sharply different from each other, levels of complexity of self-organization can correspond to systems of different nature. This complexity is determined by the nature of the self-organizing system—the complexity of its structure and behavior and the dynamic mechanisms of interaction between the components. Thus, more complex collective behavior of insects (bees, termites, ants), as compared to bacteria and viruses, underlies much more complex processes of self-organizing behavior in the community of collective insects.

Particular manifestations of self-organization at relatively simple levels of its complexity can serve as an integral part of the phenomena at a more sophisticated level.

In a fairly complex in terms of its structure community of ants, several aspects of self-organization at different levels can be distinguished. First of all, this applies to the construction of the dwelling—ant heap. It is known that ant individuals release in the process of their life strongly smelling substances—pheromones. They attract other ants and thus serve, in particular, as a means for controlling the construction process. Placing the initial construction material more or less randomly, the first individuals leave on them traces of pheromones. They serve as a guide for subsequent individuals who also emit pheromones. As a result a complicated structure arises (Fig. 6.7).

Individuals who depart from the ant heap in search of food initially move randomly, emitting faint traces of pheromones. But when the individual finds food and carries a part of it to the anthill, pheromone release increases sharply. The smell attracts other individuals nearby who take part in the delivery of food and further reinforce the smell of the paved trail. Thus, the process of food delivery self-organizes, leading to targeted behavior in the community (Fig. 6.8).

Finally, another result of self-organization with more complex mechanisms can be identified. In the process of evolution of ants random mutation of individuals occurred which, in essence, is analogous to bifurcations in the evolution of physical systems. In the course of the natural selection, the mutated individuals were either

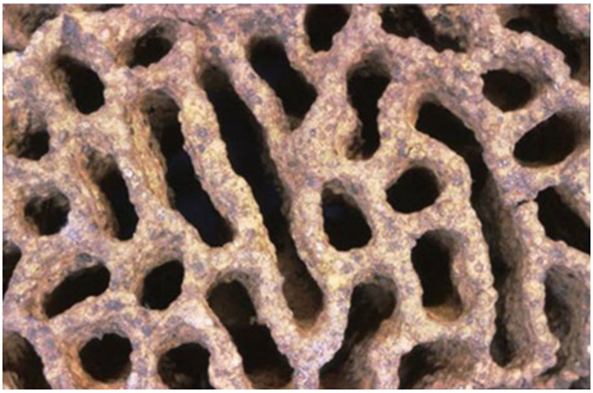


Fig. 6.7 Construction on a community of ants

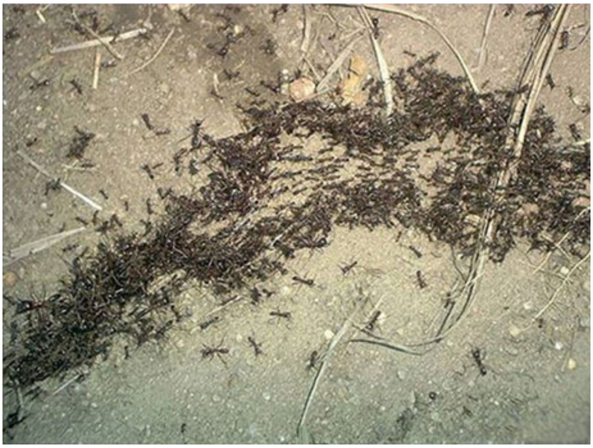


Fig. 6.8 Self-organization of the process of food delivery in the anthill

dying out or were becoming the primary source of new lines of development, fostering procreation. Apparently this division of labor in the ant community can be considered as the emergence of structure in the process of self-organization of the system. It should be noted that the characteristics of self-organization in the ant community fully apply to other collective insects such as bees, termites, etc.

Thus, self-organization is a phenomenon of an interdisciplinary nature belonging to the field of knowledge commonly known as cybernetics, or more narrowly synergetics. Therefore, any particular self-organization process is based on certain dualism. On the one hand, self-organization of the system is realized by specific physical, chemical, or some other mechanisms. On the other hand, to ensure that the system is self-organizing, it is necessary to fulfill cybernetic conditions common to all self-organizing systems—the general principles of self-organization. Let us consider these principles in more detail.

1. Self-organization processes occur in distributed dynamic systems. A distributed system should be a collection of a large number of individual components,

elements that make up the system. These may include individual molecules in a Belousov–Zhabotinsky chemical reaction–diffusion system, individuals in a stock of fish, and individual people in a crowd gathered on the square. These components must interact with each other, i.e., the system must be dynamic, functioning on the basis of the dynamic mechanisms.

2. An important feature of self-organizing processes is that they are carried out in open systems.

In a thermodynamically closed system, the evolution in time leads to a state of equilibrium to which the maximum entropy value of the system corresponds. And, according to Boltzmann, this state is characterized by the maximum degree of randomness.

In open systems, two variants of evolutionary processes are possible:

- Time evolution toward the equilibrium state (in general, this can also be the evolution toward a nonequilibrium but steady state)
- Evolution through a sequence of stationary states, with a change of stationary states due to the slow change in the so-called control parameters (e.g., ambient temperature during the formation of Benard cells)

The well-known Russian physicist Yu. L. Klimontovich cited as a good example the evolution theory of Charles Darwin. It is based on the principle of natural selection. Evolution can lead either to degradation or to be a process of self-organization, in which more complex and more sophisticated structures emerge. Self-organization is therefore not the only result of evolution. "Inner striving" for self-organization is inherent to neither physical nor even biological systems. An alternative way may be degradation, a physical example of which is the temporal evolution of a closed system toward equilibrium. Thus, self-organization is only one of the possible paths of evolution. In order to understand which path will be adopted by a developing system, a criterion for self-organization is required.

A number of systems are known for which such criterion is apparent. Thus, in the case of Belousov–Zhabotinsky chemical reaction–diffusion systems, the initial state corresponds to the uniform, i.e., chaotic, distribution of molecular components of the medium. The order, i.e., self-organization, corresponds to the formation of dissipative structures. It would seem that self-organization must correspond to the maximum degree of order.

But in a general case, the situation becomes much more complicated. Yu. L. Klimontovich refers to the human body as an example. Its stationary state corresponds to a certain degree of randomness, because the equilibrium state (complete randomness) differs in principle from the state of life. Namely, it must be regarded as an ordered state. Thus, a certain standard of chaos must correspond to the order, and deviations from it disrupt vital functions, i.e., the degree of order in the system.

The ground for understanding the spontaneous emergence of order was laid by the great mathematician of the last century, Alan Turing, in his paper "The chemical basis of morphogenesis." He showed that nonlinear dynamic mechanisms in an initially homogeneous medium give rise to ordered structure. Somewhat later.

the founder of the nonequilibrium thermodynamics, Prigogine, examined in detail the formation of ordered structures in the Belousov–Zhabotinsky media. Because these processes require energy inflow or outflow of entropy (its dissipation), Prigogine called such systems and the structures emerging in them dissipative. These processes are also called nonequilibrium phase transitions.

Certain conditions need to be satisfied for the occurrence of nonequilibrium phase transitions that are manifested by the formation of new dissipative structures:

(A) Dissipative structures can be formed only in open systems because only in such systems energy inflow is possible compensating for losses due to dissipation and ensuring the existence of more ordered states.

(B) Dissipative structures arise in macroscopic systems, i.e., systems consisting of a large number of elements (atoms, molecules, macromolecules, cells, etc.). This makes collective interactions possible.

(C) Dissipative structures arise only in systems described by nonlinear equations for macroscopic functions. Examples are kinetic equations, such as the Boltzmann equation, and equations of gas dynamics and hydrodynamics.

(D) For the occurrence of dissipative structures, nonlinear equations must allow for change of symmetry of the solution under certain values of control parameters. This change is manifested, for example, in the transition from the molecular to the convective heat transfer in Benard cells.

3. The system should show both positive and negative feedback. Processes in a dynamic system tend to change the basic relations between the components of the system involved in these processes. Such changes can be called changes at the output of the system. At the same time, these components are required for starting up the processes taking place in the system; they are thus the parameters at the system's input. If changes at the output of the system affect the input parameters such that changes at the output are amplified, we are dealing with positive feedback or autocatalytic growth. Negative feedback is a situation where the dynamic processes in the system maintain a constant state at the output. In a general case dynamic systems with positive and negative feedback are modeled by nonlinear differential equations. This reflects the nonlinear nature of systems capable of self-organization—apparently the fundamental property of the system, which determines its ability to self-organize.

6.3.1 *Some Details: The Nonlinearity of the Surrounding World*

The world around us is complex. This complexity is manifested in scientific research, technology, and everyday life. At the same time, human consciousness has always been characterized by the desire to identify in this complex world a simple component reflecting its essential nature.

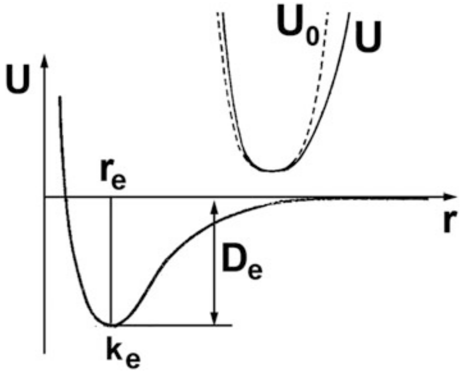


Fig. 6.9 Dependence of the potential energy of the nuclei of a diatomic molecule as a function of the distance between the nuclei. Differences between the harmonic (U_0) and anharmonic (U) functions (*top*)

Let us turn to the simplest physical example—the description of the dynamics of an arbitrary diatomic molecule. Potential energy of the interaction between the nuclei of the molecule is a function with a rather complicated form, which increases sharply at short distances between the nuclei, passes through a minimum at the equilibrium point of the internuclear distance, and asymptotically approaches zero with increasing internuclear distance (Fig. 6.9). It is easy to see that near the minimum, this function can be rather precisely approximated by a parabola, which corresponds to the so-called harmonic approximation of the potential energy of the nuclei of the molecule

$$U = \frac{k_e(r - r_e)^2}{2}.$$

This, in turn, implies that the force acting between the nuclei of a molecule depends linearly on the change of the distance between them

$$F = -k_e(r - r_e).$$

This approximation is important, since the value k_e determines the oscillation frequency of the nuclei

$$\nu = 2\pi\sqrt{\frac{2\mu}{k_e}},$$

where μ is the reduced mass of the nuclei of the molecule.

Thus, three quantities: the equilibrium distance between the nuclei r_e, the depth of the minimum—the dissociation energy—D_e, and the second derivative of the energy of the nuclei of the molecule at the equilibrium point k_e allow for describing the basic properties of a diatomic molecule.

The minimum of the potential energy of the nuclei corresponds to a bound state of the molecule. Therefore, the solution of the Schrödinger equation shows that the nuclei of the molecule are mainly in the vicinity of r_e, and the simple model describes well the dynamics of the nuclei in the molecule. The effect of the asymmetry of the potential energy function is relatively small. It can be taken into account if the expression for the force acting between the nuclei is consecutively supplemented by quadratic, cubic, etc. terms, i.e., moving from a linear dependence of force on the displacement of nuclei to the nonlinear one.

In this simple example, the main features of the physics of the last century become apparent:

- Understanding of high complexity of physical phenomena
- Desire to use a linear model of the phenomenon, if possible
- The belief that taking into account nonlinearity only makes the conclusions of the linear model more precise

Nevertheless, the development of the physical understanding of natural phenomena resulted in the second half of the last century in a gradual understanding of the much more complex role and the possibilities of manifestation of nonlinear processes. Let us begin treating the features of this situation with a few simple and obvious examples encountered in everyday life.

When buying apples on the market, nobody thinks about the linearity of the process, when upon having decided to buy 4 kg instead of 2, the buyer assumes that there will be twice as many of apples. At the same time, few take into account that increasing the speed of the car by a factor of 2, for example, from 50 to 100 km/h, the driver dramatically exacerbates the consequences of a possible accident, because the kinetic energy of the car is proportional to the square of its velocity. These examples show that everyday life is characterized by both linear and nonlinear phenomena, the essence of which one generally ignores.

Examples of nonlinear phenomena are easy to find among biological objects. Leafed trees are characterized by a large number of branched branches, which sharply increases the amount of foliage. This is called fractal structure. A fractal is defined as a mathematical object, in which upon a given transformation, the number of its characteristic details increases nonlinearly. As one possible example consider a geometric structure, the initial state of which includes three segments (see Fig. 6.10). We define as the transformation of the object the drawing of a perpendicular line through the middle of each segment. As a result of each successive transformation, a segment is turned into an x-shaped structure with equal half lines. It is easy to see that as a result of each conversion both the number of crossings and the number of segments increase nonlinearly. A large number of fractal structures are known that differ both in their initial state and in the rules of its transformation. Trees, including leafed trees, can also be considered fractal structures. For them, the fractal structure was, apparently, the decisive factor for survival during the evolution of plants.

The first plants appeared about 500 million years ago in the Paleozoic era. It all started with Rhyniophyta that originated from green algae and were the first to populate the land. During the carboniferous period giant lycopsids, calamites, and horsetails grew on earth with a small degree of fractality of trunks, stems, and leaves assimilating solar energy. In the course of evolution, plants have been

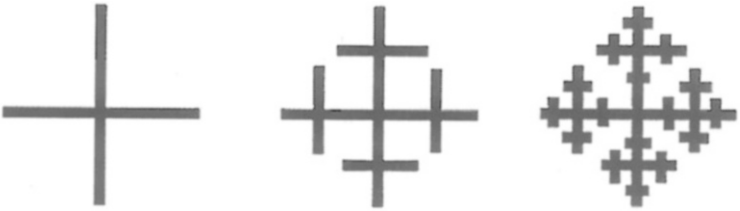


Fig. 6.10 Scheme of constructing the simplest fractal

progressing with a high degree of fractality, which dramatically increases the number of leaves and, consequently, the amount of absorbed solar energy.

An example of a fractal structure is the human blood circulatory system where oxygenated blood passes through a sequence of increasingly branching vessels. This makes it possible to nurture all body tissues.

Apparently, the first nonlinear phenomena have attracted the attention of specialists more than 150 years ago. John Scott Russell, a marine engineer and a lecturer at the Edinburgh University, observed the movement of horse-drawn barges on the canal. In his report, Scott Russell wrote that he discovered that upon a sudden halt of the barge, towed by a pair of horses, a part of the water separated from it: "A mass of water rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed." Scott Russell followed it on horseback for a few miles until he lost the sight of it. Only 50 years later, D. J. Korteweg and G. de Vries derived nonlinear equations, which had a solution in the form of a solitary bell-shaped wave moving at a constant speed along the surface of water in a shallow channel of rectangular cross section.

Interest in nonlinear phenomena resumed in the 50 years in connection with research on plasma physics. This was greatly facilitated by the emergence of opportunities to solve nonlinear equations with the help of powerful computers.

The study of nonlinear phenomena during the last decades of the past century has led to fundamental results. The understanding was achieved of the fact that they not only have high complexity but, in general, cannot be regarded as a refinement of linear models.

4. An important consequence of the nonlinearity of dynamic mechanisms in distributed systems is the manifestation of the so-called emerging properties and "emerging" mechanisms. They were discussed in detail in Chap. 4 using as an example chemical reaction–diffusion media. As in these media, in an arbitrary distributed dynamic system, three levels of dynamics can be distinguished:

- The level of interaction between the elements of the medium, i.e., the nature of their interconnections (microlevel)
- The level at which the system can be in some kind of a stationary state (mesolevel)

- The level of interaction between the states of the system and its environment (macro-level)

In this case, it is the dynamics of the macro-level that is responsible for the apparent properties of the system.

6.3.2 Some Details: Self-Organization, Dissipative and Conservative, Self-Assembly, Stigmergy, etc.

Let us discuss some concepts that are used along with the notion of “self-organization” and, in some cases, against it. The most commonly used of them is the notion of “self-assembly.”

The notion of “self-assembly” has chemical origins. It was introduced in 1987 by the famous French chemist, Nobel laureate Jean-Marie Lehn, to distinguish, among the many phenomena of spontaneous self-organization, the processes of spontaneous structure formation in systems that are in the state of thermodynamic equilibrium. Indeed, we know a large number of such processes of structure formation under the equilibrium or, more precisely, under close to equilibrium conditions. Among them, for example, are helix–coil transitions in polymer molecules, the formation of supramolecular structures of amphiphilic molecules (micelles, liposomes, bilayers), and so on up to crystallization processes. The term “self-assembly” is mainly used in relation to molecular systems. Nevertheless, self-assembly processes were found in other micrometer structures. Thus, electrochemical processes at the border between heptane and a water solution of copper sulfate led to formation of transparent copper films about 1 μm thick on the heptane side and dendrite-like copper formations on the copper sulfate side.

Nevertheless, while being seemingly reasonable, contrasting the equilibrium and nonequilibrium processes of spontaneous pattern formation does not appear justified. First of all, strictly equilibrium processes are rare in practice. It is well known how technically difficult it is to grow large monocrystals under isothermal conditions, maintaining small growth rates. Chemical processes are usually carried out under conditions close to equilibrium. A criterion of this proximity is the reversibility of the process—a necessary condition for equilibrium. Understanding this helps to avoid the uncertainty that arises in some cases, when physical and chemical mechanisms are identical in terms of their physicochemical mechanisms, but differing in their complexity. Thus, the formation of supramolecular assemblies of amphiphilic molecules—for example, micelle formation in solution—undoubtedly belongs to self-assembly according to the definition of this term. But at the same time, lipid bilayers—the basis of cell membranes—can hardly be attributed to this category. Recently, in order to avoid unnecessary confusion, the phenomenon of spontaneous structure formation in equilibrium systems is increasingly frequently called conservative self-organization.

In view of all this opposition between the concept of "self-organization" and a number of concepts such as self-assembly, dissipation, and conservative self-organization, this hardly makes sense. These concepts, of course, can be used as limited terms highlighting some partial aspect of the phenomenon. In general, all of them rather correspond to separate levels of the general concept of "self-organization," which differ in complexity (structural and behavioral), corresponding to the considered level of the process and its dynamic mechanism.

6.4 Reaction–Diffusion Processor: A Self-Organizing Dynamic System

Self-organization plays a decisive role in information processing by biological objects at different levels of their complexity. Therefore, devices that implement biological principles of information processing fundamentally differ in terms of their physical nature from von Neumann devices commonly used today.

Consider as an example a possible variant of the processor based on the Belousov–Zhabotinsky chemical reaction–diffusion medium. Without discussing specific technical solutions, we will pay attention to the basic organizational principles and mechanisms of functioning of the processor. We will use the specific experience gained in the pilot study of the information capabilities of the prototypes of such devices described in Chap. 5.

Let us define the primary characteristics of the processor, which determine its information capabilities. The processor must be an extended reaction–diffusion medium, in the simplest case quasi two-dimensional, i.e., a layer whose thickness is medium, small compared with its length and width. Dynamic regimes of the reaction–diffusion medium are determined by its state—the chemical composition and temperature. Therefore, the environment must be thermostated. The preservation of its initial composition in each regime must also be ensured, because the components of the medium in a closed volume are depleted during the reaction. One of the easiest ways to meet this condition is the use of a flow reactor, in which the composition of the medium is continuously kept constant. Let us also assume that the Belousov–Zhabotinsky medium is light sensitive. Therefore, the illumination of the environment should be strictly regulated. Assume that the initial state of the processor is given by a certain choice of concentrations of the medium components. It will be shown below that it is most convenient to choose this state such that the concentration of medium components would correspond to the vibrational mode at zero illumination of the medium (Fig. 6.11). We will assume that the catalyst of the reaction is situated in the reactor and is immobilized on a suitable carrier. Under these conditions, the medium, whose components are evenly distributed in the system feeding the reactor, upon entering the reactor spontaneously goes over into a dynamic state of concentration fluctuations. Thus, the initial working

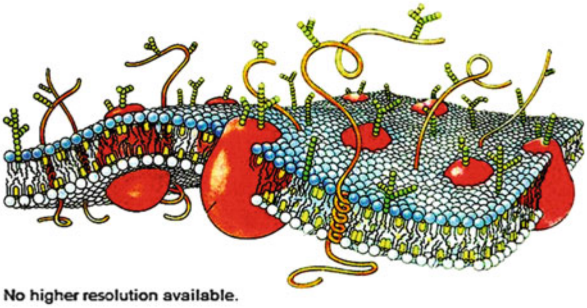


Fig. 6.11 Change in the stationary state of the photosensitive Belousov–Zhabotinsky medium as a function of its illumination

condition of the reactor is created through a process of self-organization in the Belousov–Zhabotinsky medium.

Of course this initial state of the reactor can be used for information processing, i.e., for carrying out operations initiated by the vibrational mode of the medium. They were discussed in detail in Chap. 5. At the same time, the capabilities of the processor can be significantly extended by switching the medium to the other dynamic regimes that allow for performing information processing operations different from those carried out in the vibrational mode.

It would seem that the most natural way to switch the regime is to change the composition of the medium, which is determined by specifying the parameters of the system for filling the reactor. However, this option has a significant drawback. In this case, the switching time of the dynamic regimes turns out to be too large—tens of seconds. The composition of the mixture is determined by the efficiency of the pumps supplying individual components into the medium, and therefore the system for filling the reactor has substantial inertia. A faster and more convenient option to switch modes can be implemented by changing the illumination of the light-sensitive Belousov–Zhabotinsky medium.

Let us return to the zero isoclines of this system (Fig. 6.11). The initial state of the environment in the reactor in this figure corresponds to complete absence of illumination. This is one of the stationary states corresponding to the vibrational mode of the medium. Let us illuminate the medium with uniform light radiation in the visible spectrum. Then a consistent increase in the intensity of the radiation will cause the medium to pass through a sequence of stationary states. In this case, the switching time is in the range of tenths of a second. After switching, the medium can remain in each of the stationary states for any time as long as the chemical composition, temperature, and intensity of the illuminating light are kept constant. Various operations are performed by the medium in its various states, which greatly increases the information capabilities of the reaction–diffusion processor (Fig. 6.12).

Along with the light radiation, controlled switching of the regimes of the medium can be performed by local or global electric field. In this case, various possibilities for exciting the dynamic regimes of the environment and for

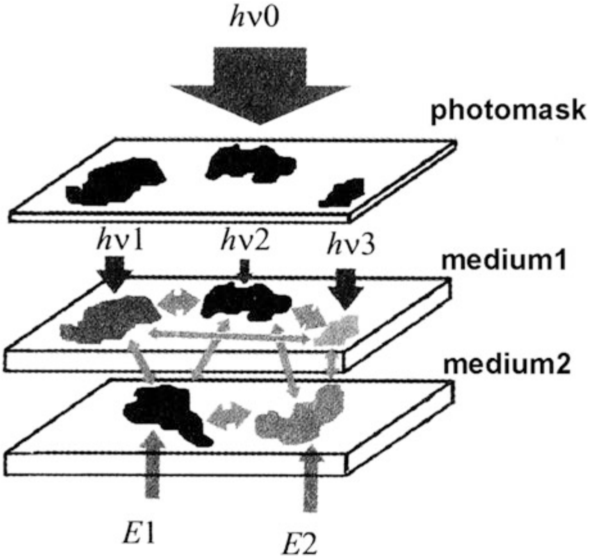


Fig. 6.12 Scheme of the dynamics of the reaction–diffusion processor

influencing them locally arise—the splitting of the waves, their annihilation, and formation of complex autowave structures.

As noted above, in Chap. 5, intense light radiation transfers the medium into an inert state. This creates a number of additional options for constructing information processing systems. By constantly projecting on the medium a given distribution of the intensity of the radiation of high intensity, one can obtain in the medium areas of specified shape and size. The evolution of the medium will occur only in those areas. In this case the required dynamic regime can be selected individually in each area by defining the degree of its illumination (Fig. 6.11). This technique—the selection of the working area of the medium by intense light radiation—was used, for example, to find the shortest path in the maze and for the formation of a "chemical diode," covered in detail in Chap. 5.

Let two spatial distributions of the reaction–diffusion medium, operating in different dynamic regimes, be separated by a thin partition permeable to some components of the medium. In this case a spontaneous process of interaction between media arises, which leads to a change in their modes of operation. This effect can be used to create multilevel devices capable of performing complex logic functions.

Thus, the reaction–diffusion processor is a complex dynamic system, in which from the chemical medium of variable and even of the same composition subsystems can be formed that perform operations of different nature. Their formation is due to the processes of self-organization of the medium, initiated by control stimuli. Diffusion interactions can combine separate subsystems, linking them into a single information-logical device.

Chapter 7
New Ideas. . .New Opportunities?

There is a time for everything, and a season for every activity under heaven . . .

A time to cast away stones, and a time to gather stones together . . .
a time to embrace, and a time to refrain from embracing.
Ecclesiastes

The previous chapter was devoted to the investigations of recent years in information technology, based on two approaches generally accepted today—the von Neumann paradigm and the biological principles of information processing. However, at the same time attempts were made to find new ways of information processing different from the traditional ones and having certain advantages over them. These new ideas hardly have any fundamental novelty. Rather, they involved a combination of digital von Neumann principles and biological principles.

7.1 Development of the Biological Principles of Information Processing: Amorphous Computing

At the turn of the century, in the end of the last decade of the last century, a group of MIT researchers headed by the physicist Harold Abelson proposed the concept of "amorphous computing." This concept was supported by DARPA (Defense Advanced Research Projects Agency) and attracted to its development a significant number of scientists—physicists, chemists, and biologists. The authors of the concept see it as a further, more in-depth elaboration of the biological principles of information processing in biological systems, built from a giant number of locally interacting elements. The main goal of amorphous computing is to develop such elements and methods to control them that could ensure a specified collective interaction. It is assumed that these objects possess some typical properties that are manifested in practice—the spread of characteristics, the nature of the interaction

© Springer-Verlag Wien 2014

N.G. Rambidi, *Molecular Computing*, DOI 10.1007/978-3-211-99699-7_7

which is not always exactly known, and a time-varying pattern of interconnections. The interaction of the elements is considered to be local. In fact, amorphous computing, in terms of its initial assumptions, is the opposite of the idea of using molecular structures as basic elements. Molecular elements are in principle identical, and none of them is "defective" due to imperfections in the process.

The authors of the concept emphasize that as a result of improvement of various technological processes in recent years a real basis for the development of amorphous computing has been created.

Modern technologies of mechano-electronic microsystems allow formation of chips that simultaneously accommodate logic circuits, sensors, actuators, and devices responsible for communication between the elements. The authors of the concept note that today there exist real opportunities to mix such chips with building materials or paints. Therefore a situation is conceivable where the paint covering the wall is sensitive to vibration and prevents the invasion of robbers or reduces external noise.

Even more significant opportunities open up today due to substantially increased understanding of biochemical mechanisms of the processes in living cells. This can be used for creating, based on the methods of cell engineering, sensor cells, actuator cells, programmed cells, cells delivering drugs to specified tissues or organs of the human body at the specified time, etc.

The basic concept of amorphous computing is an element of a system, a particle with a certain set of characteristics (properties). Initially it is assumed that the amorphous system is characterized by a scatter of properties within a certain range. It is also assumed that the particles are distributed randomly on the surface or in the volume. The particles interact with each other on the basis of local mechanisms, which generally must be nonlinear. As a result of this interaction, the internal state of a particle can be changed, as it is done, for example, by the methods of cell engineering.

It is easy to see that the main characteristics of an amorphous system follow the characteristics of distributed reaction–diffusion media. Therefore, naturally, in the amorphous system propagating waves switching the properties of the particles may arise, and the formation of complex spatiotemporal structures may occur.

In order to thoroughly study the properties and dynamics of amorphous systems, the authors of the concept chose an unconventional, original approach. Attempts were made to develop an abstract programming language of the evolution of amorphous systems. In this case, the development of the language actually involved choosing the characteristics of the particles and the operations on them that must be specified so that the system displayed the behavior provided for by the concept. The entire set of the characteristics found was supposed to be incorporated at a later stage into practically created particles. We illustrate this approach on an example that is often mentioned in the literature describing the fundamentals of amorphous computing.

One of the main devices of the planar semiconductor technology is the inverter. A typical version of such a circuit, manufactured using the CMOS technology, is

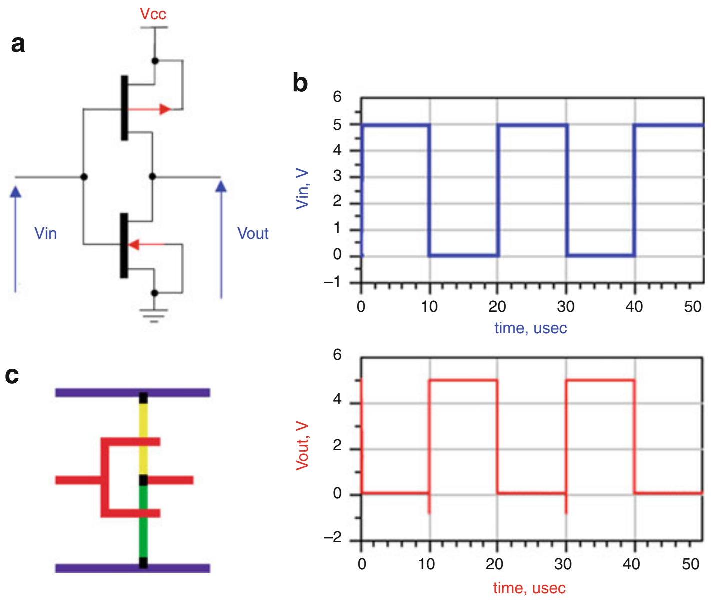


Fig. 7.1 Scheme of a semiconductor signal inverter (**a**), dependence of the voltage at its output on the voltage at its input (**b**), and scheme of the formation of an inverter using amorphous computing methods (**c**)

shown in Fig. 7.1. The signal at the input of such a device removes the signal at its output, and conversely the absence of the input signal causes a signal at the output to appear (Fig. 7.1). Consider the formation of such a device in an amorphous system.

The programming language of amorphous systems was developed by Daniel Coore and was named GPL—Growing Point Language. This language is applicable to particles to which certain properties are attributed. It is assumed that each particle possesses relatively limited logical capabilities and a small memory. Particles operate asynchronously. But at the same time their information capacity is approximately the same as they are created using the same technology. All particles are programmed the same way. But each of them remembers its state and can generate a sequence of random numbers. In general, particles do not memorize their location and orientation. Each particle can pass information to a certain number of its neighbors; this can be accomplished by radio signals or chemical factors. It is assumed that there is some communication radius r, which is considerably larger than the particle size and, at the same time, much smaller than the spatial dimensions of the amorphous medium.

The main concept of GPL is growing points. Suppose there are a large number of particles with identical properties. Let us introduce into the population of particles, at random or according to some rule, a particle with properties different from the environment. This particle—a growing point—initiates a wave of switching states

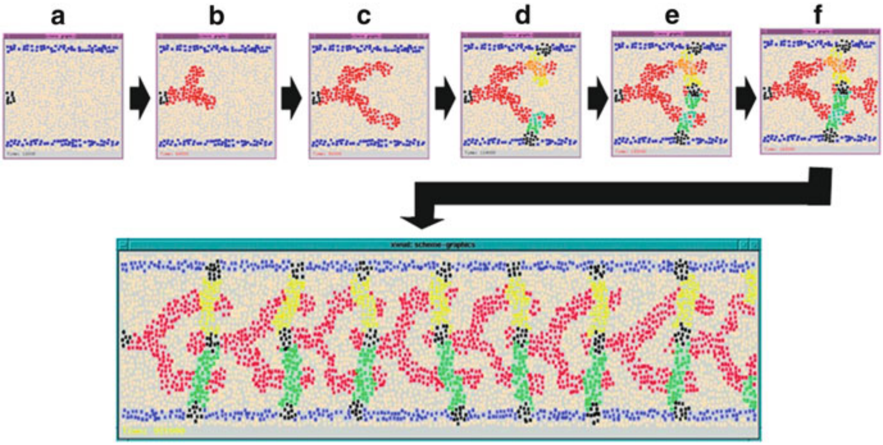


Fig. 7.2 Scheme of the formation of the sequence of the signal inverter using amorphous computing methods, (**a**)–(**f**) sequences of stages

of the neighboring particles within the radius of communication. Switching is initiated by signals (radio, chemical), which are sent by the growing point. This process is expected to be stepwise, with the switching wave propagating across the field of particles, if no limits are imposed on its propagation.

The biological notion of tropism is introduced as such a restriction extending the capabilities of GPL. It involves directed interaction of the waves initiated by growing points. Let us consider a couple of examples. Suppose that there are two growing points at some distance from each other and initiating switching waves. We will assume that the wave emanating from point *B* can switch only those particles not affected by the wave *A*. Then the propagation of the wave *A* inhibits the propagation of the wave *B*. As a second example, consider the situation when the wave *B* can only be switched by the particles located near *A*. In this case, the wave *B* behaves as if it were attracted to the point *A*. In a general case, by analogy with biological phenomena, one can introduce the notion of pheromones that are emitted by each growing point, with the concentration monotonically decreasing with distance from the growing point. Pheromone concentration determines the degree of tropism, i.e., the repulsive force from this growing point or attraction to it. In the GPL language a growing point in its active state lays material and releases pheromones. In fact, both of these processes are implemented by switching the state of the neighboring point, i.e., by specifying its new parameters.

Let us return to our example—the creation of the signal inverter with the circuit shown in Fig. 7.1. Its schematic diagram based on the elements that are formed in an amorphous medium is shown in the same figure. Formation of the inverter is carried out over several stages, as the main elements of the device are created by the particles of the same type. These stages are shown in Fig. 7.2. In order to illustrate the main features of GPL, consider a program of the first stages of the formation of the inverter (Fig. 7.2a–c):

```
(Define-growing-point (make-red-branch length)
   (Material red-stuff)
   (Sixe 5)
   (Tropism (and (away-from red-pheromone)
              (And (keep-constant pheromone-1)
                    (Keep-constant pheromone-2))))
   (Avoids green-pheromone)
   (Actions
     (Secrete 2 red-pheromone)
     (When ((<length 1)
               (Terminate))
             (Default)
                (Propagate (- length 1) )))))
```

In this program the starting growing point is referred to as "red." Throughout the process determined by the program, a new branch of red particles (material red stuff) must be grown. The direction along which the branch grows is determined by tropism—the repulsion from the red starting point, the repulsion from the top (pheromon1), and the repulsion from the bottom (pheromon2) backbone, which are assumed to have already been formed.

The same principle governs the formation of the remaining elements of the inverter, as shown in Fig. 7.2. On the basis of inverter circuits various logical devices can be created. Therefore, the authors of the concept believe that the technology of amorphous computing can be effectively used for the development of information processing devices based on industrially produced particles with the desired properties.

At the same time, the authors of the concept lay the greatest hopes on the symbiosis between the ideas of amorphous computing and cell engineering.

7.1.1 Some Details

Genetic information that determines the structure, function, and evolutionary pathways of a living organism is encoded in the molecules of the deoxyribonucleic acid, DNA, which are located in the nuclei of cells. The basis of the DNA molecule is a polymer chain with alternating groups of phosphoric acid and a sugar—deoxyribose (Fig. 7.3). In view of the structural features of these molecular groups (i.e., the values of angles between the bonds of the atoms within them), these polymer chains form helical structures. Attached to each ribose group is one of the four nitrogenous bases: adenine (A), thymine (T), guanine (G), or cytosine (C). In general, the group consisting of the phosphoric acid residue, deoxyribose, and the base is called nucleotide. A remarkable property of these bases is that by means of hydrogen bonds, the pairs A-T and G-C can form stable groups whose sizes along one of the axes are surprisingly similar (Fig. 7.4). Therefore, these pairs form bridges that connect the strands of DNA into a stable double helix. The A-T and G-C pairs and the structures based upon them are called complementary. During reproduction (replication) of DNA molecules the double helix unfolds and the molecule of the enzyme DNA polymerase attaches to each strand. This enzyme determines the type

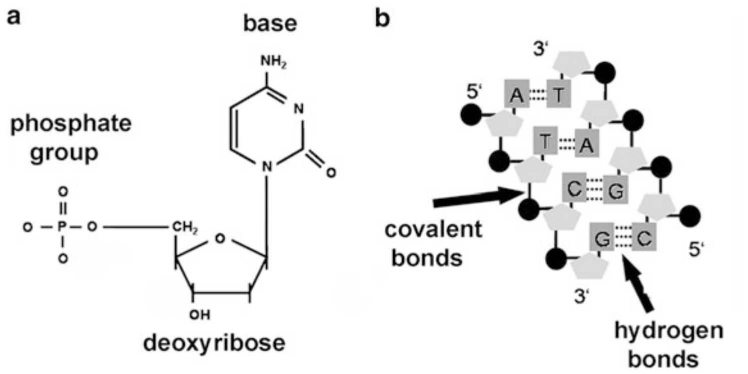
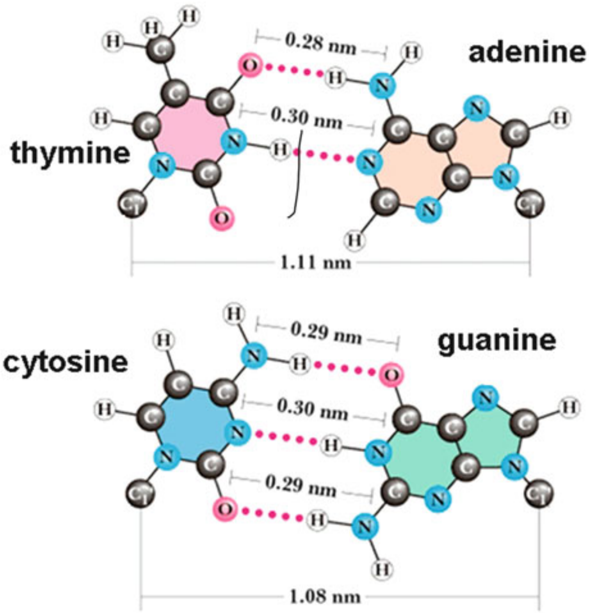


Fig. 7.3 Structure of the DNA molecule, structure of nucleotide (**a**) and connection of nucleotides in a double helix (**b**)

Fig. 7.4 Complementary bases of the DNA molecule



of the base, captures from the environment a complementary base, attaches it to the strand being formed by the enzyme, and moves to the next base of the DNA being replicated. In this fashion another pair of strands complementary to the initial DNA strands is created which can form another double helix of DNA completely identical to the original one.

The processes of protein synthesis, constantly occurring in the cell and maintaining its life activities, are determined by the information recorded by triplets of bases (codons) in the DNA chain. The structure of the synthesized protein is determined by a part of the genome, the stretch of DNA sequence corresponding to this protein, located in the cell nucleus. Protein synthesis is spatially separated in the cell and occurs in other cell formations—ribosomes. The information about the structure to be synthesized is transferred to the ribosome by another molecule—RNA—whose structure exactly matches to the structure of the corresponding DNA region. The synthesis of the RNA molecule, i.e., the readout (transcription) of information, is performed by the enzyme RNA polymerase. It splits the double-stranded DNA molecule and uses it to synthesize a copy of the section of DNA—an RNA molecule corresponding to the synthesized protein. This region is recognized by the enzyme RNA polymerase using a specific label in the DNA chain—a molecular group called the promoter. The enzyme attaches to the promoter and starts transcription at this point of the chain. At the same time, intracellular processes in the self-organizing system of the cell are complex and multifunctional. In particular, the start of transcription depends on the presence or absence of another protein, called operator. It can block the promoter, which prevents the attachment of RNA polymerase to the DNA strand and blocks RNA transcription.

The cells of living beings are complex biological devices whose functioning is based on a large number of biochemical reactions occurring in them. Cells spontaneously divide, reproducing themselves in a set of identical copies. At the same time, the methods of cell engineering techniques developed in recent decades allow to change the program of functioning of the cell, affecting its genetic apparatus. This, in principle, creates the possibility to use the cells as particles of amorphous systems that can perform complex functions such as sensor cells, actuator cells, etc.

Of substantial interest for amorphous computing is the use of cellular material for the formation of information processing devices. One of the possibilities being actively discussed today is the creation of cells serving as signal inverters.

It is known that cell life is ensured by constant synthesis of proteins involved in its metabolism. The synthesis occurs in specialized cell structures—ribosomes. Based on these mechanisms of protein synthesis in the cell, the participants in the project "amorphous computing" G. Sussman and T. Knight proposed a biochemical approach to the creation of digital information processing devices. It is based on the idea of a biochemical signal inverter (Fig. 7.5). Suppose that the protein Z is being synthesized, with the RNA polymerase reading its structure from the corresponding section of the DNA chain. At the same time, if there exists a protein A, which is the operator of the process, the synthesis of protein Z may be terminated. Thus, the system of transcription is an inverter controlled by the protein A. The authors of this idea believe that designing interrelated chains of protein synthesis in the cell, one can build digital logical devices of sufficiently high complexity.

The discussed possibilities of creating digital devices are the main prospective practical applications of the principles and technology of amorphous computing discussed in the literature today. It must be noted that in general they make an ambiguous impression. The starting point of the concept is a distributed dynamic system of particles. Such systems operate with a high degree of parallelism of operations. Nevertheless, based on amorphous systems it is proposed to create digital devices with sequential execution of operations.

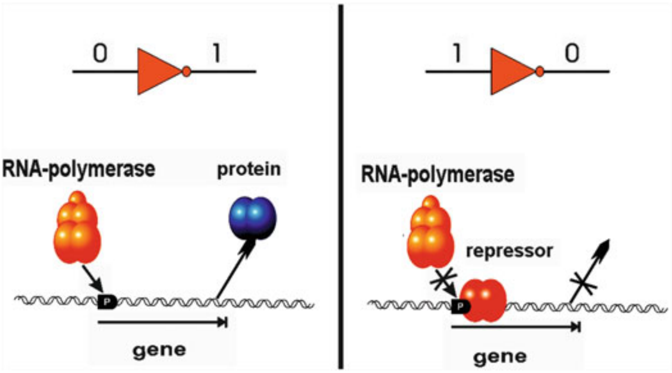


Fig. 7.5 Scheme of a biomolecular signal inverter

The authors of the concept themselves point out that the rate of functioning of the biochemical digital devices should be small. And so it is unlikely they can be used to solve the problems of high computational complexity. Nevertheless, they believe that the principal possibility to create large supramolecular systems with a given structure opens up new avenues in molecular engineering. Thus, the main publication describing the principles and design techniques of amorphous computing ends with the words:

"On the whole, we are in the primitive stage of the development of cell and amorphous computing analogous to the early stages of electronics in the beginning of the 20th century. The further development will open new boundaries of engineering, which will dominate in information technologies of the next century..."

7.2 Semiconductor Reaction–Diffusion Devices: The First Attempts

Autowave processes in chemical reaction–diffusion systems (not to mention the biological ones) are much slower than the same processes in solid-state semiconductor media. And, despite the fact that the speed of the chemical media is sufficient for solving many practical problems, in recent years attempts were made to create solid-state autowave systems. One of the first experimental works was carried out by the Soviet physicists L. L. Golikov, V. N. Nemenuschi, M. I. Elinson, and Y. Balkarei back in 1981.

They used the plates of the ferroelectric crystal (triglycine sulfate) at a temperature slightly below the Curie temperature of the ferroelectric. The voltage applied to the electrodes on the plates of the ferroelectric with a frequency of 10^3–10^5 Hz causes the medium to heat. Theoretical estimates show that in this system, stationary states at three different temperatures $T_1 < T_2 < T_3$ may arise. Two of them (T_1 and T_3) are stable, and one (T_2) is unstable. Initially the medium is in one of the

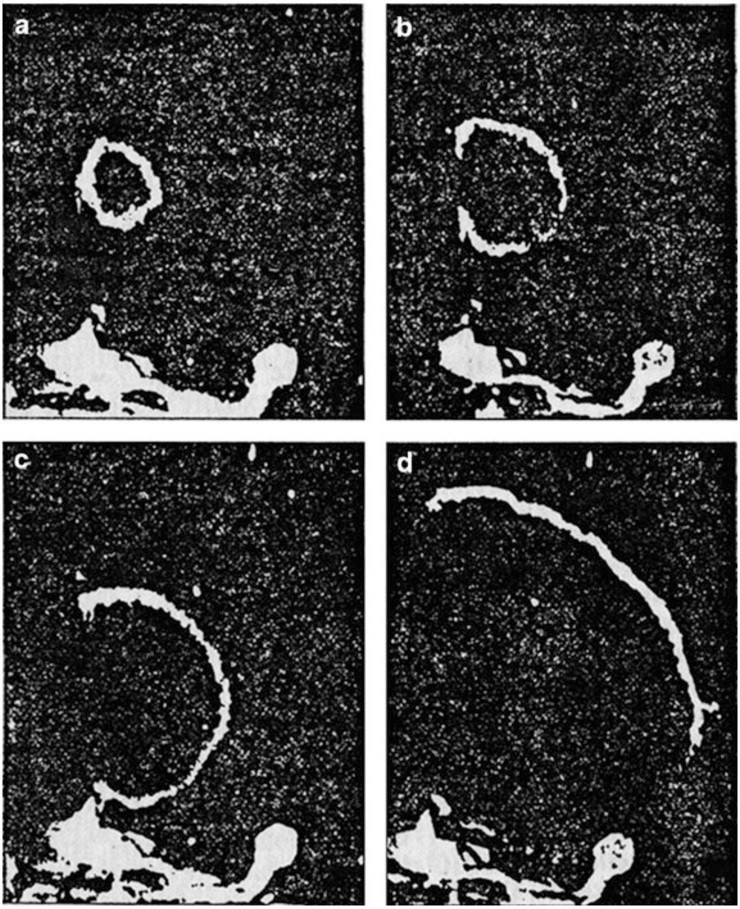


Fig. 7.6 Autowave processes in a ferroelectric: propagation of a switching heat wave

stable states. By local heating or cooling it can be brought to another state, which propagates in the medium as a switching heat wave (Fig. 7.6). To visualize this process, one of the planes of the plate was covered with a layer of cholesteric liquid crystal that changed its color with temperature changes. The speed of wave propagation in such a system was in the range 0–1 cm/s. Subsequently, the study of self-wave processes in semiconductors was carried out mostly by theoretical physicists. For example, in 1990 the Soviet physicists B. S. Kerner and V. V. Osipov published in *Advances in Physical Sciences* an exhaustive description of the spontaneous formation and evolution of dissipative structures. Unfortunately, until now there have been no attempts to use the autowave properties of semiconductors to create information processing devices.

A different approach to creating information-logical semiconductor devices was proposed in the beginning of our century by the Japanese researchers Tetsuya Asai, Yoshihito Amemiya, and their colleagues at the University of Hokkaido. They started to develop semiconductor planar circuits (chips), simulating Belousov–Zhabotinsky chemical reaction–diffusion media. Several variants of circuit design to implement these models were proposed. Apparently the most interesting one is the use of single-electron oscillators.

Earlier, in Chap. 4, it was pointed out that qualitatively a Belousov–Zhabotinsky medium can be represented as a set of unit cells with dimensions smaller than the diffusion length, interacting with each other through the diffusion of medium components. The main regimes of the unit cells are concentration fluctuations and the excitable regime. For semiconductor modeling of Belousov–Zhabotinsky media, a one-electron oscillator constructed on the basis of the tunnel transition and possessing analogous properties was selected as an elementary cell.

Let there be a tunnel junction—a thin layer of nonconductive material between two electrodes. Let us apply a small direct current voltage to the electrodes. Since the transition, as the simplest capacitor, has a capacitance, it starts to charge, and at a certain voltage on the electrodes, electron tunneling takes place. This process is repeated over and over again (Fig. 7.7). This process of reducing the resistance of the device at low bias voltages is called Coulomb blockade. Often such a process, occurring in the tunnel junction, is compared with the sequential formation of droplets in a loosely closed water tap. Figure 7.8 shows a schematic diagram of such an oscillator and its modes, which are characterized by nanosecond times. To create a model of the reaction–diffusion medium, these oscillators were combined by capacitive coupling (Fig. 7.8), which plays the role of diffusion. Real chips implementing this scheme were not produced. Their functional features were determined by computer modeling. It turned out that in a device containing 100 × 100 oscillators, the quality of the recorded wave process leaves much to be desired (Fig. 7.9). Results improved significantly when the scheme was modified. It used oscillators built on several series-connected tunnel junctions.

The developed reaction–diffusion chip is a quasi-flat structure—a layer containing oscillators and their connection elements. The authors of the device then made the next step. They tried to create a multilayer device in which quasi-planar structures were merged into a spatial circuit. Computer modeling of the functionality of this device led to interesting results. The quality of the Voronoi diagram, initiated and designed at the first level of the device, dramatically improved on the third layer of the device, connected with the first one (Fig. 7.10). This is consistent with the views on the importance of multilevel information processing by reaction–diffusion systems.

Today it is difficult to determine the practical significance of the work done by the Japanese researchers. More needs to be done to understand the real capabilities of the developed semiconductor devices and technological features of their manufacturing.

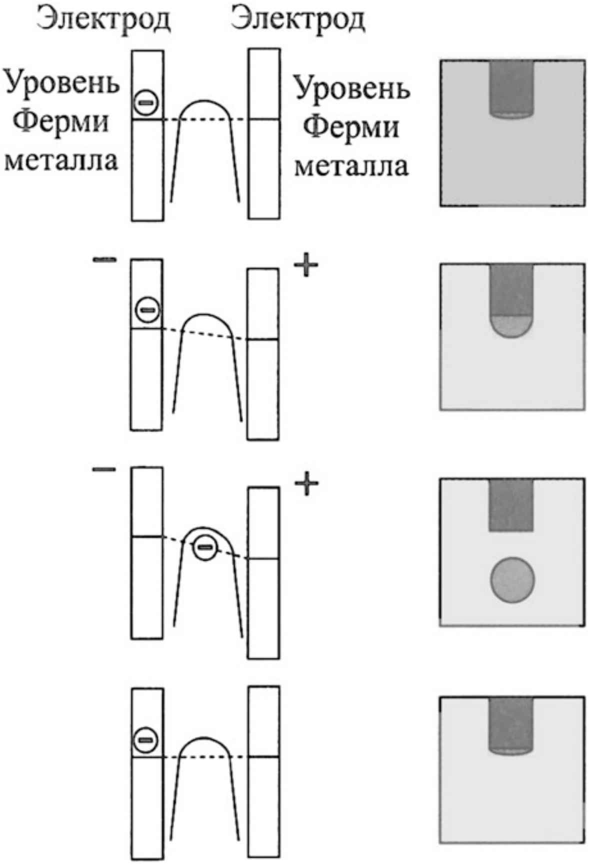


Fig. 7.7 Scheme of the "Coulomb blockade" effect

7.3 DNA Computing: A Sophisticated Combination of Principles and "Tools" of Information Processing

The 1990s of the last century were marked by the unexpected appearance of a new information approach, which did not have any predecessors. In 1994, Leonard Adleman, professor of computer science and molecular biology at the University of Southern California in Los Angeles, published in the journal *Science* an article entitled "Molecular computation of solutions to combinatorial problem." This article immediately gained many supporters and followers and became the starting point of a large number of follow-up works.

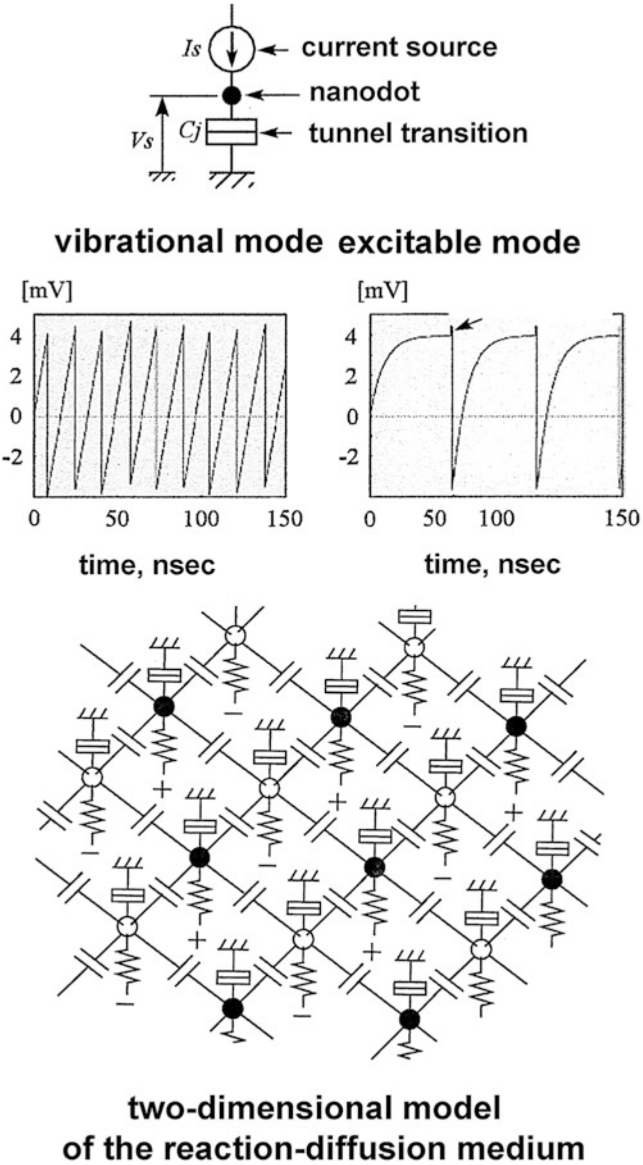


Fig. 7.8 Semiconductor reaction–diffusion scheme on the basis of single-electron oscillators

7.3.1 Some Details

As a teenager, Len Adleman could not decide what occupation to choose. After enrolling in the University of California at Berkeley, he first chose chemistry, then medicine, and then, eventually, mathematics. “Я прошел через множество вещей

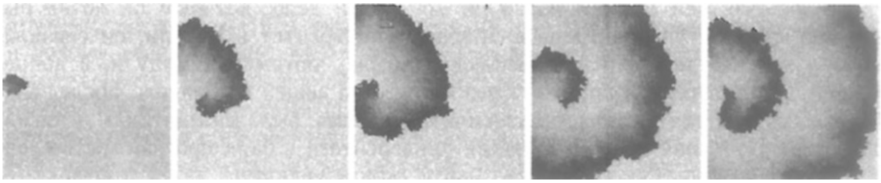


Fig. 7.9 Evolution of the autowave helical structure in the semiconductor reaction–diffusion scheme

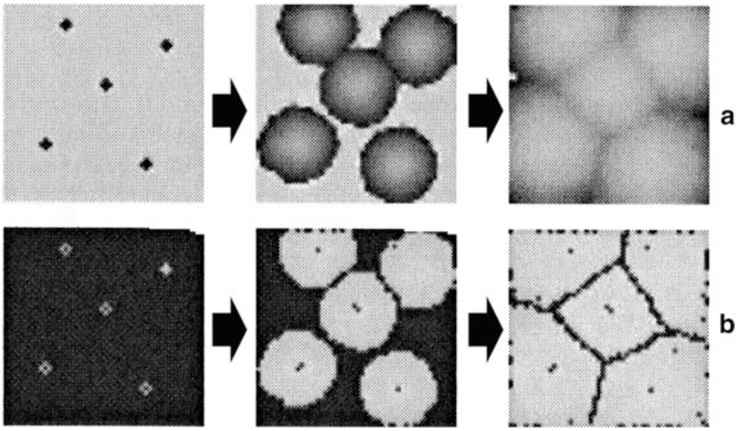
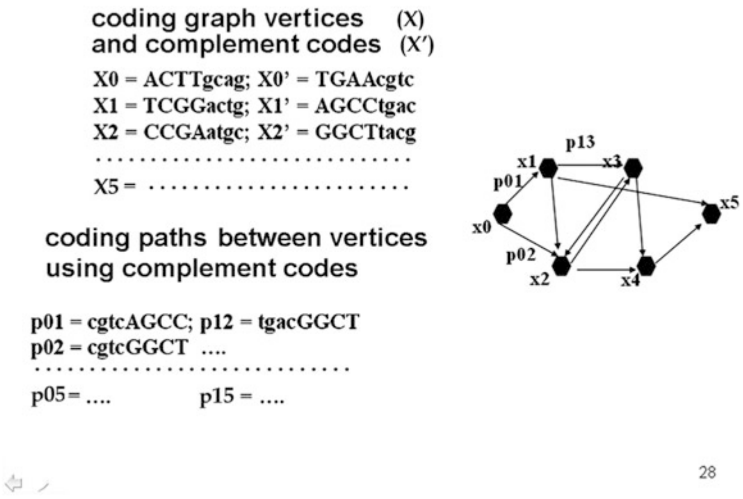


Fig. 7.10 Determination of the Voronoi diagram by the multilayer semiconductor reaction–diffusion device (**a**, *upper level*; **b**, *lower level*)

и, наконец, последнее, что оставалось, для того, чтобы закончить во время, была математика"—he wrote later. After graduating he started to work as a programmer in a bank and continued his search. He attempted again to find himself in medicine and then in physics, while still working in the bank. Afterwards, he returned to Berkeley, where he prepared and defended a thesis on the theoretical aspects of computational complexity. Immediately thereafter he was appointed assistant professor of mathematics at MIT.

At this time, his colleagues at MIT—Ronald Rivest and Adi Shamir—worked on the problem of information security. They devised a coding system, the so-called key, which is a mathematical formula to encrypt and decrypt data. They engaged Adleman in this project. As a result, they created an encryption system called RAS system, according to the first letters of their names. It brought them fame and, consequently, money.

Despite the excellent conditions of work at MIT, in 1980 Adleman returned to the University of Southern California in Los Angeles. It was there that he felt the inclination at first to immunology and then to molecular biology, whose unsolved problems, in his view, "отличались такой же красотой, как и математические."



coding graph vertices (X)
and complement codes (X′)

X0 = ACTTgcag; X0' = TGAAcgtc
X1 = TCGGactg; X1' = AGCCtgac
X2 = CCGAatgc; X2' = GGCTtacg
....
X5 = ...

coding paths between vertices
using complement codes

p01 = cgtcAGCC; p12 = tgacGGCT
p02 = cgtcGGCT ....
....
p05= .... p15 = ...

28

Fig. 7.11 Encoding of graph elements in the problem of determining Hamiltonian paths

In particular, he was attracted to the problems of storing genetic information by DNA molecules, transcription of this information, and its transmission.

A unique feature of Adleman was that his interest in information issues of molecular biology was combined with a deep understanding of the problems of high computational complexity. While developing an encryption system, he perfectly realized that traditional mathematical methods were insufficient. Taken together all this led him to develop an approach called DNA computing.

As a concrete problem of high computational complexity, which could serve for working out the basic principles of the approach, Adleman chose the problem of finding Hamiltonian paths—an integral part of the traveling salesman problem. The problem of finding Hamiltonian paths is to determine all possible paths that pass through each point (the city that the traveling salesman needs to visit) of a set containing *N* points. In this set the start and the end points are defined and thus each point can be visited only once. Adleman chose as an object a system of seven points (Fig. 7.11). His approach was to solve the problem in two stages. The first stage involved simultaneous identification of all the possible paths for this set of points, both passing through all the points and not passing through some points several times and only once, etc. The computation time of this stage is exponential in the conventional numerical techniques for solving the problem.

The second stage is to analyze the obtained mixture of all possible paths. Adleman showed that this process can be performed by methods of modern genetic engineering in polynomial rather than exponential time.

The technical side of the approach was that each point of the system was "modeled" by a segment of single-stranded DNA molecule (Fig. 7.11). In Adleman's methodology a segment containing 20 different nucleotides corresponds to each point (city). Since each point of the path is joined with two others, the DNA

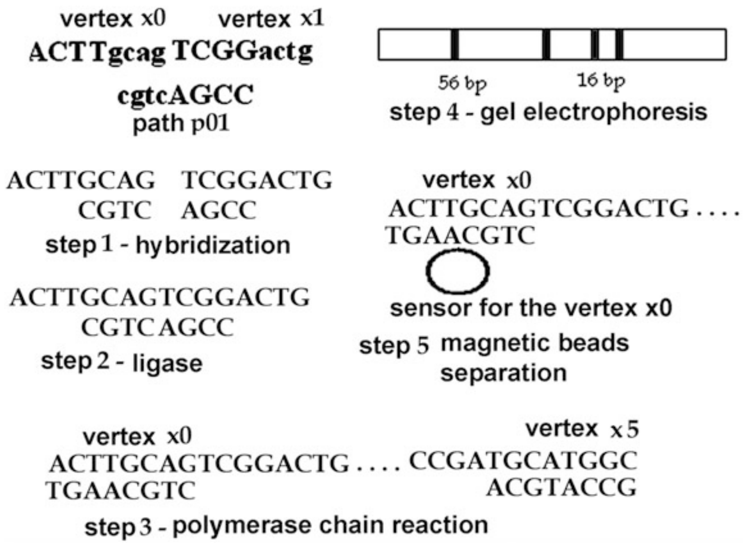


Fig. 7.12 Main stages of solving the problem of determining Hamiltonian paths by the technique of DNA computing

segment is divided into two parts, corresponding to the right and the left neighbor of the point. In addition, 20-nucleotide sequences of the single-stranded DNA are synthesized corresponding to all possible distances between points (1–2, 1–3 ... 1–*N*, ... 2–1, 2–3, 2–4, etc.). Each segment corresponding to the distance between the *i*th and the *j*th points is a sequence of 10 nucleotides complementary to the right part of the code of the point *i* and 10 nucleotides complementary to the left of the code of the point *j*. All segments of nucleotides corresponding both to the points and to the possible distances between them are merged into a reactor. As a result hydrogen bonds are formed between the complementary DNA codes corresponding to each other. In other words, the process called hybridization takes place. After this, a special enzyme (ligase) cross-links the segments connected by hydrogen bonds into single-stranded DNA fragments. Thus, synthesis method implies that a huge number of DNA molecules corresponding to all possible paths between points are produced. In this case, all of them are synthesized simultaneously, i.e., with a huge degree of parallelism.

The second stage of the solution to the problem of Hamiltonian paths developed by Adleman was using genetic engineering techniques in order to determine whether, among the components of the mixture of single-stranded DNA, there are copies corresponding to these paths. Adleman applied an analysis technique based on the principle of successive removal of anything not related to the solution of the problem.

To accomplish this, Adleman proposed the following analysis scheme (Fig. 7.12):

(Step 1)—identify the chains in which the initial nucleotides correspond to the starting point of all routes and the end nucleotides correspond to the end point.
(Step 2)—identify the chains with the length corresponding to a specified number of paths.
(Step 3)—determine whether all points of the paths are included in the chains of the nucleotides of the remaining DNA molecules.

If as a result of these steps DNA molecules still remain that have not been deleted, we can assume that:

- A solution to the problem of Hamiltonian paths exists.
- All solutions are contained in the remaining DNA molecules.

It should immediately be noted that the desired sequences of nucleotides comprise but a small portion of the material obtained at the first stage. Therefore, at the first analysis step Adleman used a technique that allowed for a manifold increase in the number of DNA molecules that remain for the next steps of analysis—PCR, i.e., polymerase chain reaction. This reaction starts with primers attaching to DNA molecules—20-nucleotide sequences corresponding to the start and end points of Hamiltonian paths. Subsequently the enzyme DNA polymerase reproduces a sequence of nucleotides located between the primers. This process is repeated multiple times with the original and the newly derived molecules. As a result, the number of these molecules after 30–40 cycles may be increased by 10^{27} times. This process also has high sensitivity and can detect 10–100 copies of DNA in the mixture.

A second step of the analysis relies on liquid gel electrophoresis, which allows for separating the mixture obtained by PCR into separate components with different molecular weights.

In the third step of the analysis affine separation on magnetic beads was used. Microparticles of a paramagnetic material chemically bonded with the chains of nucleotides corresponding to individual points of the Hamiltonian path were injected into the mixture analyzed. Upon attachment to the corresponding DNA molecules, the obtained complexes were extracted by magnetic field.

Adleman himself notes that according to his estimates:

- The performance of the proposed devices is 10^{14} operations per second.
- The energy efficiency is 2×10^{19} operations/joule.
- The approach can be used to solve combinatorial problems of moderate complexity.
- At the same time, the domain of problems where the approach can be used is substantially limited.
- The results of the calculations depend strongly on the exact fulfillment of the conditions of the proposed approach.

Shortly after the appearance of Adleman’s work, the American mathematician Richard Lipton from Princeton demonstrated how DNA can be used to encode binary numbers and to solve the problem of satisfying a logical expression. Its essence is that, given some logical expression involving n logical variables, one needs to find all combinations of variable values that make the expression true. Traditional techniques reduce this problem to the sorting of $2n$ combinations. Lipton showed that all these combinations can be easily encoded by DNA, and then the Adleman method can be applied. Lipton also proposed a way of breaking DES (the data encryption standard), treated as a kind of a logical expression.

In addition several other applications of DNA computing were proposed. At the University of Wisconsin, the problem of delivering four types of pizza to four destinations, which implied 16 different answers, was solved with the help of DNA. Scientists from the Princeton University solved the combinatorial chess problem: with the help of RNA they found the correct move of a chess knight on the board of nine squares (a total of 512 variants).

In general, the future of DNA computing remains uncertain. The fact that for more than 10 years after the appearance of Adleman’s work no workable computing device has been created based on his suggestions apparently indicates serious difficulties of practical implementation of this approach. At the same time the originality of this approach is attracting new researchers and ensures its support by DARPA (Defense Advanced Research Projects Agency).

Chapter 8
What's Next?

You will remember the beginning of humankind. Our first parents were quick to get themselves into trouble. They were expelled from the garden of Eden. I understand That Adam took Eve's hand, and said! "My dear, we are living in a time of transition

Stafford Beer *"World in Torment"*

Predicting the future is a dubious and ungrateful undertaking. Too many factors that are not apparent today may suddenly change the well-known and seemingly obvious trends.

And yet I would like to try to look slightly ahead, going somewhat beyond the issues discussed in this book, and imagine the future of molecular devices for information processing. Let's start with digital computing devices.

8.1 Do the Developers of Digital Computers with von Neumann Architecture Need Molecular Components?

Maturation and development of the postindustrial society is accompanied by the emergence of serious problems that can only be solved by complex and quite laborious calculations. The computational power of modern commercial computers is a far cry from what is needed in such cases. Therefore, unique computing systems—supercomputers—are being created in the industrialized countries.

One of the best-known and important issues is preservation and development of nuclear deterrence weapons in the context of the international ban on nuclear testing. In the end of the last decade of the past century, the Military Applications Division (DAM) of the French Atomic and Alternative Energies Commission (CEA) began a program of numerical simulation of nuclear processes required for decision-making about the storage time and further development of nuclear

© Springer-Verlag Wien 2014

N.G. Rambidi, *Molecular Computing*, DOI 10.1007/978-3-211-99699-7_8

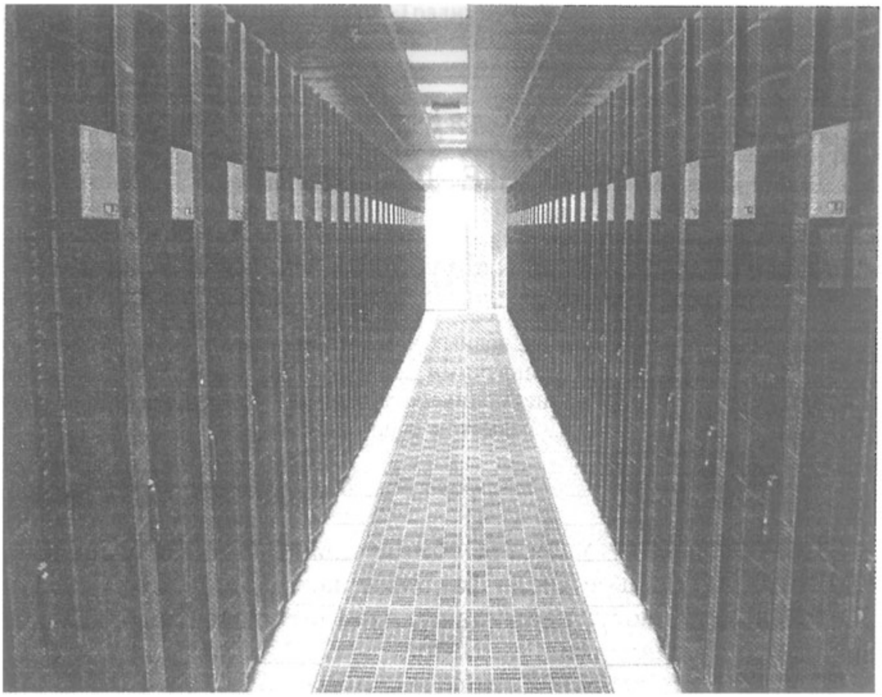


Fig. 8.1 Small part of the Tera-10 supercomputer

weapons. The program provided for creation of a supercomputer “Tera” with consistently expanded computing power. According to the program it was supposed to reach 1, 10, and 100 teraflops (one teraflop equals 10^9 floating-point operations per second) in 2001, 2005, and 2009, respectively. In the current version Tera-10 is the most powerful European and fifth most powerful supercomputer in the world and represents a computer system built around 4352 Dual-Core Intel Itanium 2 processors. The random access memory of Tera-10 is 27 TB, and its disk storage capacity is 1 PB (1 petabyte = 1,000 terabytes = 1,000,000 GB = 10^9 MB = 10^{12} bytes). Figure 8.1 gives some impression of a small part of the supercomputer. Unfortunately, the financial details of the project were not disclosed by Bull, the company that develops the supercomputer.

But one should not forget that in addition to the price of the computer itself, presumably tens of millions of dollars, the total cost of having these unique capabilities is extremely high. To accommodate Tera-10, about 2,000 m^2 of specialized space is needed, not counting the space for auxiliary equipment. And these facilities should be significant, since only for cooling the computer about 3 MW of power is required (an increase up to 5 MW is already envisaged). Tera-10 consumes 1.8 MW of electric power.

Heat dissipation is a serious problem in operating supercomputers. Impressive examples of its consequences were given by the creators of a less powerful supercomputer system—the employees of the Virginia Polytechnic Institute and State University (Blacksburg, Virginia, USA), who created a supercomputer for "merely" $5 million.

The system was assembled from Apple 1100 clusters (the 2 GHz G5 system with 4 GB of memory). The performance of the supercomputer was 17.6 teraflops and the random access memory 176 TB. But despite the less impressive performance, the developers wrote that its power consumption would be sufficient for 3,000 homes of average size.

It is these issues—power and heat dissipation—that dramatically increase the interest in molecular elements that consume much less energy and dissipate much less heat.

Today semiconductor electronics is deeply enrooted in industrial infrastructure. Moreover, it optimally meets the demands of the modern society. However, industries such as supercomputer manufacturing would greatly benefit from the practical use of molecular components. However, it should be noted that the molecular components have semiconductor rivals in the same size range. Above all, these are devices based on quantum wells.

8.1.1 Some Details: Quantum Dots and Cellular Automata

In the early 1970s of the last century a new direction emerged in semiconductor physics—the study of heterostructures formed by semiconductors of varying composition and properties. Especially interesting were the heterostructures with the spatial dimensionality different from three dimensions, i.e., from the usual solid. Those may be thin nanometer films or filaments as well as nanometer ensembles of atoms. Since quantum effects manifest themselves at nanometer scale, these systems were called quantum wells, quantum wires, and quantum dots. Their remarkable property is that the nano-sizes of semiconductor structures restrict the movement of electrons, and therefore the density of electronic states in them is fundamentally different from the macroscopic body. It is easy to see (Fig. 8.2) that quantum effects appear starting from quantum wells, i.e., when the motion of electrons is limited to nanometer sizes, at least in one dimension.

Without going into detail of extremely interesting properties and practical applications of quantum wells and quantum wires, let us discuss quantum dots, which are considered today as one of the possible alternatives to the molecular elements of computing devices.

Quantum dots are sometimes (particularly in the popular press) called artificial atoms. In fact, a quantum dot is a collection of atoms with nano-sizes in all three spatial dimensions. The motion of electrons in such a system can be approximated by a simple quantum mechanical model known as "particle in a rectangular potential field." This model is described by the Schrödinger equation:

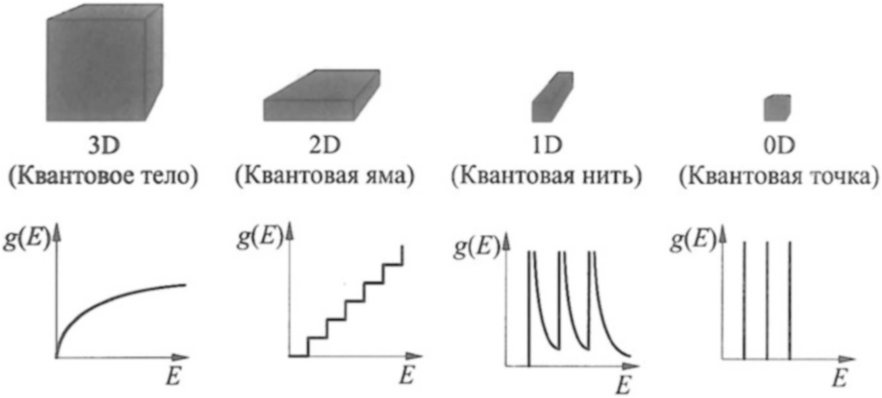


Fig. 8.2 Density of the energy states for a three-dimensional solid, quantum well, quantum wire, and quantum dot

$$\left\{-\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right]-U(x,y,z)\right\}\Psi=E\Psi,$$

where the potential energy in a box with sides a, b, c

$$U(x,y,z)=U(x)+U(y)+U(z).$$

Here
$U(x)=U(y)=U(z)=0$, if $-\frac{a}{2}\le x\le\frac{a}{2}$, $-\frac{b}{2}\le y\le\frac{b}{2}$, $-\frac{c}{2}\le z\le\frac{c}{2}$
or
$U(x)=U(y)=U(z)=U_0$ at all other x, y, and z values.
The solution of this equation is

$$E_{n_1,n_2,n_3}=\frac{\pi^2\hbar^2}{2m}\left(\frac{n_1^2}{a^2}+\frac{n_2^2}{b^2}+\frac{n_3^2}{c^2}\right),\quad \text{where } n_1,n_2,n_3=1,2,3\ldots$$

Thus, a discrete spectrum, generally similar to the spectrum of the atomic system, corresponds to the quantum point.

In a quantum dot there can exist from a single to a large number of electrons, whose distribution is determined by the Pauli principle.

Quantum dots can be created by the method of molecular beam epitaxy. Another substance with a structure similar to that of the substrate is sprayed on a well-prepared surface. Everything should happen in a high vacuum to avoid inclusion of impurities in the object to be formed. The deposition rate must be carefully controlled in order to avoid the formation of structural defects. Spontaneous growth of quantum dots in a so-called Stranski–Krastanov mode has been well studied on the example of InAs/GaAs. During the growth of the first monomolecular layer of InAs on the GaAs surface, elastic stresses arise due to differences of permanent

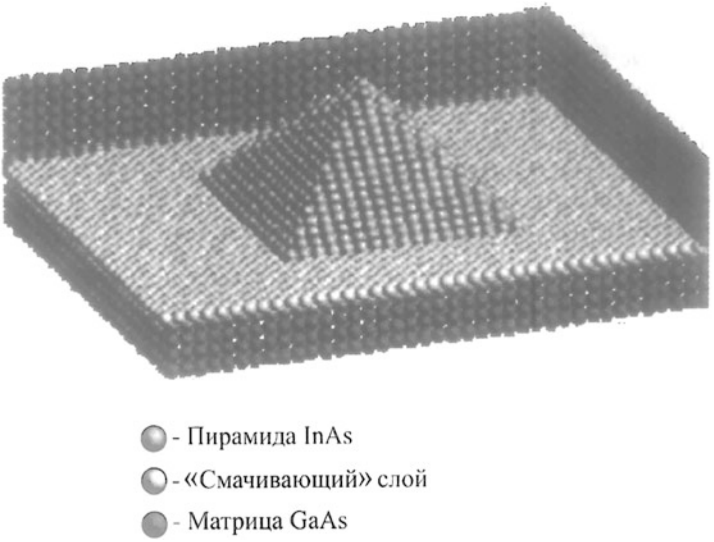
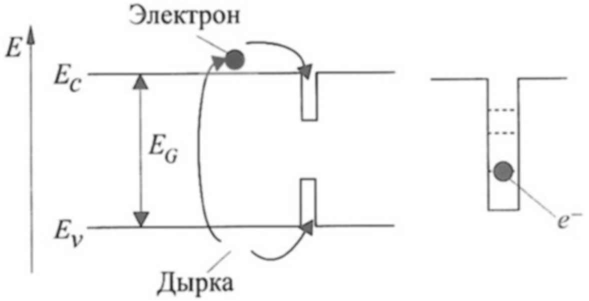


Fig. 8.3 Experimental semiconductor implementation of a quantum dot

Fig. 8.4 Band structure of a semiconductor containing a quantum dot



crystal lattices. If the deposition continues, they grow, making formation of individual “drops” on the surface of the first layer (called the “wetting” layer) more advantageous than the uniform distribution of matter on the surface of the first layer. Thus little “pyramids” with the properties of quantum dots arise (Fig. 8.3).

Qualitatively, these pyramids can be regarded as defects on the surface of the semiconductor core. In this case impurity levels appear in the band structure (Fig. 8.4)—slightly above the valence band (holes) and slightly below the conduction band (electrons). Levels corresponding to the impurity level of the conduction band are characterized by a discrete spectrum, i.e., they correspond to the quantum dot.

The unique properties of quantum dots laid the foundation of the concept of building semiconductor information processing tools—cellular automata—developed by a team of physicists at the University of Notre Dame (France).

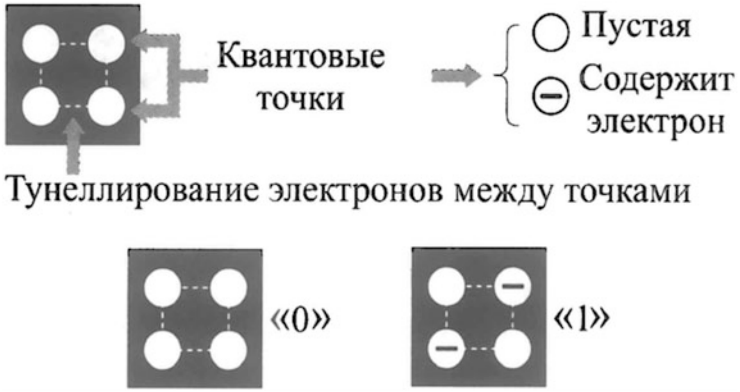


Fig. 8.5 Main elements of a cellular automaton based on quantum dots

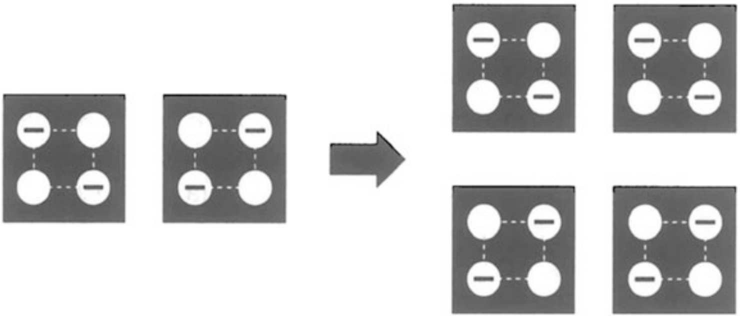
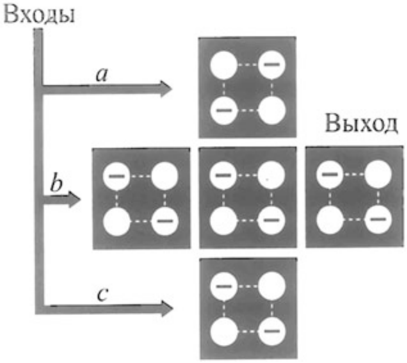


Fig. 8.6 Reconstruction of linear fragments of a cellular automaton based on quantum dots

The unit cell of a cellular automaton is a system of four quantum dots (Fig. 8.5). Two of them contain each an electron that is able to move from one point to another. The distance between the quantum dots should be small enough for the electrons to interact with each other. Due to their mutual repulsion the electrons will be located at the most distant points. Let us assume that one of the configurations encountered in this case can be attributed to the state "0" and the other one to the state "1" (Fig. 8.5). Thus, we can construct a cellular automaton with elements that can exist in two states. Because of the weak interaction between the elements, the interaction radius of 1 corresponds to the automaton. It is easy to see that if a certain distribution of electrons in the initial element (the left element in Fig. 8.6) is defined, then due to the electrostatic interaction of electrons in the neighboring cells, restructuring of states will occur in these elements. On the basis of this effect complex logic circuits can be constructed, one of which is exemplified by Fig. 8.7. Thus, cellular automata based on quantum dots have the potential to become progenitors of information processing devices with a high degree of integration of elements, if a reliable technology to manufacture them industrially is developed.



a	b	c	
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1

Fig. 8.7 Logic element on the basis of quantum dots

8.2 Biological Principles of Information Processing and Their Role in the Development of Information Technology

Let us assume that one of the major problems hindering further progress of information technologies today is the need to create intelligent systems for collecting and processing information as well as control systems that could be mass-produced and would be able to effectively solve the problems of artificial intelligence. Let us consider in more detail the main factors that determine the possibilities of their development on the basis of chemical reaction–diffusion media.

8.2.1 Some Operational Requirements for a Reaction–Diffusion Processor

Because of the nature of the problems considered, information processing devices should be, above all, more mobile, with a high degree of reliability. Since they should be mass-produced, their cost should not be high. They must solve problems in real time. In a large number of applications, this time scale should be determined by the reaction time of man, i.e., 0.1–1.0 s. In a significant number of developments (e.g., in the autonomous guidance systems), a long operating life of the device is not required.

Chemical reaction–diffusion media largely satisfy these requirements. At the same time they allow to put into practice a fundamentally new, unconventional approach to creation of information-logical devices, facilitating their practical implementation. The high performance of the device is achieved not by maximal miniaturization and the increase in processing speed limit, but rather due to the highest possible logical complexity of the elementary operations that optimally meet the challenges of intelligent decisions.

The basis of high efficiency of the reaction–diffusion devices is constituted by their inherent characteristics of information processing.

8.2.1.1 High Natural Parallelism

A distributed reaction–diffusion system is a continuous biochemical, chemical, or physical medium where in each microvolume information processing by the same algorithm takes place in parallel (simultaneously). The size of a typical microvolume is in the order of the diffusion length, i.e., 0.01–0.1 mm. Therefore, the most primitive device—a flat layer of the medium (with the linear dimensions of 100 by 100 mm and the thickness of 1 mm)—corresponds to the degree of parallelism 10^8–10^6. The degree of parallelism is dramatically higher in three-dimensional devices.

8.2.1.2 Nonlinear Mechanisms of Information Processing

Nonlinear mechanisms of the dynamics of the distributed reaction–diffusion systems enable them to perform complex logical operations as elementary operations, and not simple binary operations, as in the modern digital computers.

Suppose, for example, that our task is to detect the contour of a sufficiently simple geometric shape on a grid of $10^3 \times 10^3$ elements. The numerical technique of contour detection implemented on a modern personal computer implies that in a general case, it is necessary to perform 3–5 floating-point operations in each point of the grid. Together, this makes 3×10^6–5×10^6 elementary operations that are equivalent to one elementary operation of a reaction–diffusion medium.

The number of operations of q digital computer increases sharply with the complication of the contour shape, which requires a transition to a more narrow-meshed grid $10^4 \times 10^4$, $10^5 \times 10^5$, etc.

The fundamental advantage of the reaction–diffusion medium in this case is that increased complexity of the contour may necessitate an increase of the linear dimensions of the medium (the resolution of the medium does not change), but the time of contour detection remains the same.

8.2.1.3 Speed of Reaction–Diffusion Information Processing Devices

Execution time of elementary operations currently used by biochemical and chemical media is rather large, as well as that of biological organisms which operate on similar information principles. Nevertheless, the high logical complexity of the elementary operations executed by media dramatically increases computing capabilities.

Consider again one of the simplest operations of high computational complexity—contour detection of an arbitrary shape, which is executed by the medium in 1–5 s. Compare this time with the time required for the same operation by a personal digital computer.

Suppose that the required error of contour detection of the shape is provided by a grid with $10^3 \times 10^3$ points. We assume that, in general, the number of operations required is ~5×10^6. The time of floating-point addition by a personal computer based on the 600 MHz Pentium III processor is ~$(2.5–3) \times 10^{-9}$ s. Hence, the contour detection time is ~10^{-2} s. That is, on a $10^3 \times 10^3$ grid, the contour detection time by a PC is two orders of magnitude smaller than in the case of a reaction–diffusion medium. If the contour of a complex shape is detected, which requires improved accuracy, the situation changes radically. Execution time of this operation by the medium remains the same, i.e., 1–5 s. In the case of a personal computer, it increases to 1 s (for a $10^4 \times 10^4$ grid) and to 100 s (for a $10^5 \times 10^5$ grid).

This example illustrates a fundamental feature of reaction–diffusion media. Their advantages become the more tangible with more complex tasks.

Note that the detection of the contour of a shape is one of the simplest operations of high computational complexity. In the case of other elementary operations performed by media, their advantages become more pronounced.

The speed of reaction–diffusion media is determined by specific chemical reactions occurring in the medium. A preliminary analysis shows that, apparently, media can be developed with execution time of elementary operations 10^{-1} to 1 s, which will significantly improve their performance.

I would also like to emphasize that the most promising areas of possible application of the media considered may be technical devices for which the real time scale is relatively large. Image analysis in medicine and materials science does not require high execution speed of operations. An autonomous robot moving on rough terrain can be controlled with reaction time of the control system ~1 s.

8.2.1.4 Multilayered Architecture

The multilayered architecture has not been used so far in the development of the models of reaction–diffusion information processing devices. Nevertheless, from general considerations it follows that it offers significant potential for increasing the performance of these devices.

Multilayered architecture will enable more efficient implementation of biologically motivated principles of information processing. In a general case multilayered devices should be characterized by:

- Processing and compression of information at each level of processing
- Transmitting attractors of the previous level, i.e., the results of data compression, to the next level

This can increase manifold information capabilities of reaction–diffusion devices.

8.2.2 Technological Characteristics of Manufacturing Reaction–Diffusion Devices

Geometric dimensions of a reaction–diffusion device should be determined by the size of the active elements of the medium (10^{-1} to 10^{-2} mm). Therefore, a device containing 10^6 acting elements may represent the simplest case of a quasi-flat layer of the reagent with dimensions of 100 × 100 mm (10 × 10) mm. The micrometer dimensions of acting elements and relatively low sensitivity of the medium to foreign matter allows for:

- Dramatically reducing the cost of raw materials for manufacturing the devices compared with semiconductors devices, because they do not require ultrahigh purification from impurities
- Dramatically simplifying and reducing the cost of the industrial production technology of the devices, since it does not require an extremely high degree of purification of both gas and liquid media from dust and micro-inclusions.

Operations performed by such a device are characterized by the specific chemical reactions occurring in the active medium, the spatial structure of the device, and the control stimuli. Based on past experience, it may be suggested that instead of a complex miniature system of transistors and interconnections on a chip, a multilayer reaction–diffusion device will consist of a system of active layers with linear dimensions on the order of tens of millimeters on a polymer base and with the size of structural features in a layer of 0.1 mm, separated by a semipermeable membranes.

Thus, the manufacturing complexity of reaction–diffusion devices should be significantly lower and the technological equipment much simpler than those required for manufacturing of advanced semiconductor integrated circuits.

8.2.3 Closer to Nature: An Offensive of Polymeric Materials

The volume and reliability of data that can be obtained from experimental studies of processes in reaction–diffusion media have multiplied over the past decade, with a crucial role played by two main directions of development of experimental techniques. One of them—utilization of light-sensitive media—was discussed in detail above. Let us consider the second, no less important direction—the use of polymeric materials for the formation of reaction–diffusion media.

There are various options for the use of polymeric materials for studying the processes occurring in reaction–diffusion media. Naturally, their use is dictated by the problem to be solved, by the characteristics of the data input and output, and by the methods used to control the medium.

Today optical methods are the main method of inputting and outputting information when dealing with reaction–diffusion media.

In this case, the projected light image is converted into a distribution of chemical components of the medium. Two main factors determine the adequacy of the distribution of concentrations of medium components with respect to the original light pattern.

First of all, it is the minimum value and, most importantly, the uniformity of the light background of the optical device used to input the original image. Any heterogeneity may cause a dynamic process in the medium not associated with the input data.

Secondly, it is the constancy of the thickness of the medium, which may be disrupted due to a manufacturing error of the reactor, in which the process takes place, its inaccurate leveling, etc. This results in additional gradients of concentration of the chemical components of the medium, which also leads to random dynamic processes interfering with the investigated dynamics.

A planar polymer layer of a given thickness that does not interact chemically with the components of the medium is a spatial matrix of the polymer substance containing 80–90 % water. If the water is replaced by the reagent of the reaction–diffusion system, the thickness of the reagent layer will be maintained by the polymer matrix. As a result, the reagent layer will be affected neither by the errors of the reactor and of leveling nor by random mechanical effects (shock, vibration, etc.).

Polymeric materials allow for creating spatially inhomogeneous media with a given structure.

Let us consider just two examples of the formation of such systems.

The problem of finding the shortest path in a maze was considered above. The characteristic property of the formation of the medium used for this purpose was immobilization of the catalyst of the chemical reaction in a thin silica gel layer. This allowed to fix the distribution of the reaction components, corresponding to the labyrinth, and to organize the wave process propagating along it.

A somewhat more complex design of the medium was used by Steinbock and colleagues who modeled logic devices, switchable by a wave process in a reaction–diffusion medium. The picture of the device was applied by a printer to the surface of a thin ion exchange membrane catalyst using the solution of the reaction's catalyst instead of the printer ink. The membrane itself was placed on a layer of agar gel, which contained the remaining components of the chemical reagent.

Polymer matrices can control the processes in the environment.

N. Kazanskaya with coworkers developed two systems in which the polymer matrix plays an active role.

The first of these is a combination of two polymeric membranes. One of them contained a photosensitive component—spiropyran. Additionally, the ionophore nonactin was injected into the membrane to couple the membrane photoresponse with urea hydrolysis occurring on another membrane and catalyzed by urease.

The second system was a two-level spatially combined one. It consisted of a polymer matrix, saturated with urease, which hydrolyzed urea. Enzyme activity was a function of temperature near the point of reversible collapse of the gel of the matrix.

There exists the possibility of designing a polymer matrix in which polymer fragments could serve as one of the components of the chemical reaction occurring in a reaction–diffusion system.

In this case, systems with long-term memory can be developed. For example, when inhibiting a photosensitive reaction by radiation in some fragments of the medium and carrying out the reaction in the remaining fragments of the polymer matrix, the component associated with the matrix will be depleted only in dark areas. After repeated use of the system, these areas will turn out to be not active, i.e., they will preserve the memory of the previous process.

I would like to emphasize that these examples are only a part of the opportunities afforded by the formation of reaction–diffusion systems based on polymer matrices. The rapid development of this technique will undoubtedly expand the experimental opportunities, particularly in the design of advanced multilayer systems.

8.2.4 The Brain and the Reaction–Diffusion Computer

As mentioned above, one of the main objectives of this book is to tell about what really is today's incarnation of the biological principles of information processing.

Naturally, the most weighty arguments in their favor are the experimental works performed in recent years, in fact over the last decade. It becomes clear, even on the basis of these first timid attempts that biological principles are not a system of some scholastic reasoning, but rather a practically oriented concept that is in the process of development and the capabilities of which are far from exhausted.

Therefore, an important question arises—what are the limits to information processing capabilities of devices built on biological principles? What is their possible role in the prospective future system of computing and information-logical tools?

In order to try to find the answer to this question, let us return to the 1990s of the last century.

In the late 1980s one of the leading theoreticians in the field of informatics, Michael Arbib, made a proposal to expand the concept of computation, so that it organically included the "style" of information processing used by the human brain. In his article "The Brain as a metaphor of sixth generation computing," he wrote: "Этот стиль основан на постоянном совместном взаимодействии систем, активность которых выражается как взаимодействие пространственн-о-временных образов в многоуровневой системе нейронов."

The remarkable features of the Arbib's approach were that:

- The brain is a computer oriented toward action. This is reflected by the fact that the system constituted by a human, an animal, or a robot must be able to correlate the internal actions with the interactions with the environment so as to build an "internal model" of the world.

- The brain has multilevel hierarchical organization. The most important fact is that no single-layer model is capable of reproducing the functions of the brain.
- The brain is not a system with consecutive information processing.

In essence, Arbib's ideas underlie the interest in neural networks that emerged in that time. Moreover, this was the interest in complex variants of semiconductor computer architecture that would allow for a manifold increase in the parallelism of computation.

In those same years, Michael Conrad, the leading US expert in the field of molecular electronics, published a detailed comparative analysis of information characteristics of the brain and the von Neumann computer. A remarkable feature of Conrad's approach was that it was based on general information concepts, without tying them to any physical implementation of the devices. Conrad compared the fundamental differences in information processing by the brain and by computer, based on the exclusion principle he had previously introduced:

"A system cannot at the same time be effectively programmable at the level of structure, amendable to evolution by variation and selection, and computationally efficient."

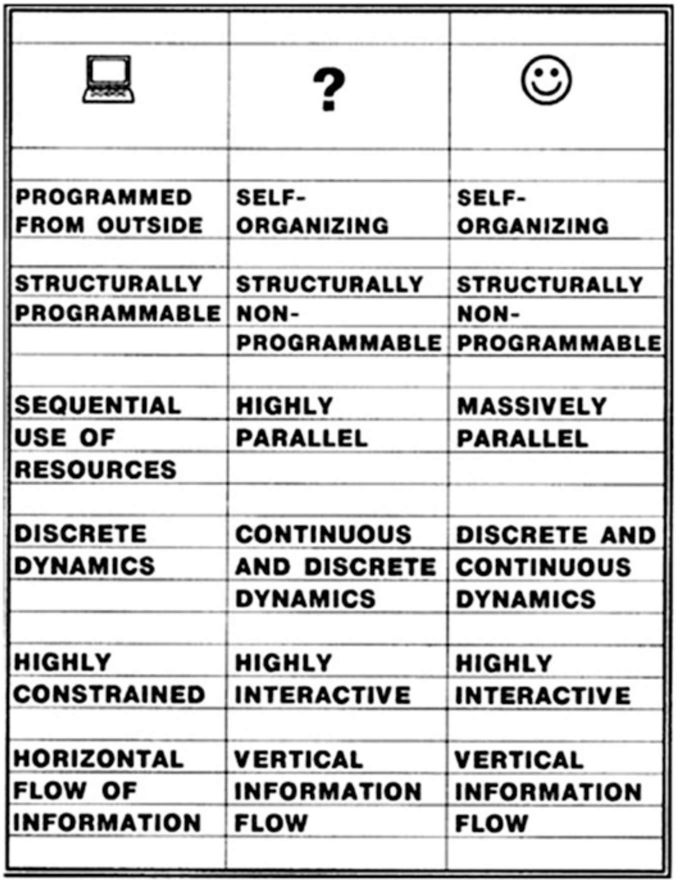
These differences are summarized in Fig. 8.8. They include the possibility of structured programming of the systems, the parallel or serial nature of information processing, the vertical or horizontal flows of information, etc. In essence, in this analysis Conrad demonstrated that the brain and the von Neumann computer are two extreme alternatives for information processing systems.

Since Conrad's approach is quite general, not tied to any physical implementation, it appeared attractive to use it to assess the place of distributed reaction–diffusion devices in the general system of information processing devices. Consider the basic information characteristics of reaction–diffusion devices. In contrast to the von Neumann computer, reaction–diffusion devices are not externally programmable. Their dynamics is determined by the state (and structure) of the medium as well as by control stimuli.

Even the simplest devices exhibit a very high degree of parallelism, mixed continuous–discrete dynamics, and vertical fluxes of information transmission and processing. Even in a simple system one can identify:

- The level of macro–micro data transformation, i.e., the level of data input
- Dynamics at the molecular (micro) level which implements the method of information processing
- The level of micro–macro transformation of information, i.e., physicochemical readout of the solution to the problem

The degree of self-organization of chemical reaction–diffusion devices is high. Moreover, they manifest gradualism, i.e., small changes in the state of the medium (the concentrations of its components and temperature) lead, in a certain area of the states, only to a relatively small quantitative rather than sharp qualitative change of dynamic regimes. This feature, in essence, is the basis for constructing systems with training.



💻	?	☺
PROGRAMMED FROM OUTSIDE	SELF-ORGANIZING	SELF-ORGANIZING
STRUCTURALLY PROGRAMMABLE	STRUCTURALLY NON-PROGRAMMABLE	STRUCTURALLY NON-PROGRAMMABLE
SEQUENTIAL USE OF RESOURCES	HIGHLY PARALLEL	MASSIVELY PARALLEL
DISCRETE DYNAMICS	CONTINUOUS AND DISCRETE DYNAMICS	DISCRETE AND CONTINUOUS DYNAMICS
HIGHLY CONSTRAINED	HIGHLY INTERACTIVE	HIGHLY INTERACTIVE
HORIZONTAL FLOW OF INFORMATION	VERTICAL INFORMATION FLOW	VERTICAL INFORMATION FLOW

Fig. 8.8 Information characteristics of the information processing devices based on various principles

And finally, reaction–diffusion systems possess other properties that determine the system's ability of adaptive behavior. Among them is the nature of the interaction with the environment, deep negative feedback, etc.

Thus, reaction–diffusion media have almost all the characteristics necessary to build on their basis devices with high behavioral complexity, ability to learn and to solve problems of high computational complexity.

In general, the comparison of the information characteristics of von Neumann computer, the brain, and the reaction–diffusion device leads to the conclusion that in terms of information characteristics reaction–diffusion devices are substantially closer to the brain than to the digital device (even if the digital device is implemented as a multiprocessor parallel system).

In terms of its informational characteristics the brain is immeasurably richer than any man-made device. The completeness of the solutions of intellectual problems

(problems of high computational complexity) achieved by the brain is striking. Nevertheless, the surprising similarity of the information characteristics of distributed systems, operating on the basis of nonlinear dynamic mechanisms, suggests that it may be possible to create devices that mimic the functions of the brain, at least in some limited areas of the intellectual activity of the brain.