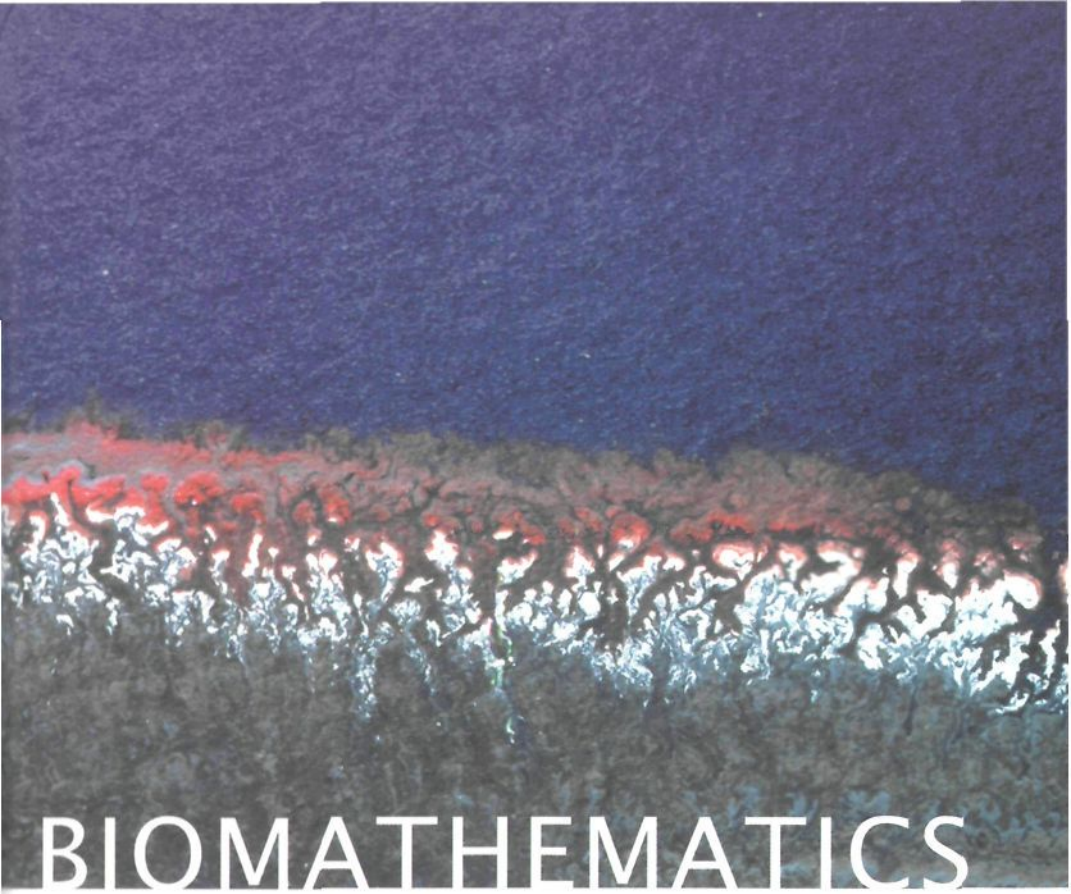


BIOMATHEMATICS

modelling and simulation

editor J C Misra

BIOMATHEMATICS

modelling and simulation

This page is intentionally left blank

BIOMATHEMATICS

modelling and simulation

Editor

J C Misra

Indian Institute of Technology, Kharagpur

World Scientific

NEW JERSEY · LONDON · SINGAPORE · BEIJING · SHANGHAI · HONG KONG · TAIPEI · CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

BIOMATHEMATICS
Modelling and Simulation

ISBN 981-238-110-4

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed in Singapore by Mainland Press

Dedicated to my parents

Late Ganes Chandra Misra and
Mrs. Pramila Misra
who served throughout their lives for the
cause of education with a missionary zeal

This page is intentionally left blank

THE EDITOR

Dr. Jagadis Chandra Misra, Senior Professor at the Department of Mathematics, Indian Institute of Technology Kharagpur, received his Ph.D. degree in Applied Mathematics from Jadavpur University in 1971 and the highly coveted D.Sc. degree in Applied Mathematics from the University of Calcutta in 1984. For over 35 years he has been engaged in teaching and research at several distinguished institutions in India, Germany and North America. He has been the Chairman of the Department of Mathematics, IIT Kharagpur during the period 1998–2001. As a recipient of the prestigious *Humboldt Fellowship*, he was at the University of Hannover during the period 1975–1977 and also in 1982, where he carried out his research in the field of Biomechanics in collaboration with Professor Oskar Mahrenholtz at the Institute of Mechanics, University of Hannover, in collaboration with Professor Christoph Hartung from the Biomedical Engineering Division of the Medical University at Hannover. He has held the position of Visiting Professor at the University of Calgary in Canada, at the Indian Institute of Technology, Kanpur in India and at the University of Bremen in Germany. In 1984 he received the prestigious *Silver Jubilee Research Award* from IIT Kharagpur, his research publications having been adjudged to be *outstanding*. In 2004 Prof. Misra bagged the highly prestigious *Raja Ram Mohan Roy Prize* for outstanding contributions in the field of Mathematics and the *Rashtriya Gaurav Award* (National Glory) for outstanding academic contributions. He has been the recipient of the coveted *Indian Science Congress Platinum Jubilee Lecture Award* in Mathematical Sciences in 2005.

Professor Misra was elected a *Fellow* of the National Academy of Sciences in 1987, the Institute of Mathematics and its Applications (UK) in 1988 and the Institute of Theoretical Physics in 1988. He was also elected a Fellow of the Royal Society of Medicine, London in 1989 in recognition of the significant impact of his research work in the field of Biomedical Engineering and a Fellow of the Indian National Academy of Engineering, New Delhi in appreciation of the impact of his researches on Engineering and Technology in 1999. He was elected a Member of the International Society of Biorheology New York, GAMM (Germany), Biomechanical Engineering Society (USA) and an Active Member of the New York Academy of Sciences.

Professor Misra published 12 advanced level Books, a monograph and over 150 research papers in international journals in the areas of Biomathematics, Biomechanics, Mathematical Modelling and Theoretical Solid and Fluid Mechanics. His research results have appeared in highly prestigious journals like *Journal of Mathematical Biology* (USA), *Bulletin of Mathematical Biology* (UK), *Journal of Biomechanics* (UK), *Journal of Biomedical Engineering* (UK), *Blood Vessels* (Switzerland), *Rheologica Acta* (Germany), *Biorheology* (UK), *International Journal of Solids and Structures* (UK), *International Journal of Nonlinear Mechanics* (UK), *Fluid Dynamics Research, ZAMP* (Switzerland), *Mathematical and Computer Modelling* (USA), *Journal of Mathematical Analysis and Applications* (USA), *Computers and Mathematics with Applications* (USA), etc. His publications have been well cited in scientific literatures and referred to in several text books. He has made pioneering research on mathematical modelling in each of the areas of Cardiovascular Mechanics, Mechanics of Brain Injury, Mechanics of Fracture and Remodelling of Bones and Peristaltic Flows in Physiological Systems. His theoretical findings on the compressibility of vascular tissues is a major breakthrough in the study of arterial biomechanics and were subsequently verified experimentally by Prof. Y. C. Fung at the Bioengineering Laboratory of the University of California, San Diego. The model developed by him for the study of arterial stenosis bears the potential to provide an estimate of the variation of blood viscosity as the height of the stenosis increases. The observations of the study were used by later investigators in the quantification of Doppler colour flow images from a stenosed carotid artery. Misra's theoretical study on the mechanics of cerebral concussion caused due to rotational acceleration of the head has opened up new vistas in neurological research and neurosurgery. On the basis of the study he could make some valuable predictions regarding the threshold of cerebral concussion for humans, in

terms of angular acceleration. He was the first to account for the effect of material damping of osseous tissues on bone remodelling induced due to the surgical procedure of intra-medullary nailing. Misra's study on the effect of a magnetic field on the flow of a second-grade electrically conducting fluid serves as a very important step towards the perception of MHD flow of blood in atherosclerotic vessels. It throws sufficient light on the quantitative features of blood flow in constricted arteries.

Professor Misra has been a Member of Expert Committees of the National Science Foundation of USA, Imperial Press of UK and World scientific of Singapore. He has also been a Member of the expert committees of the Council of Scientific and Industrial Research, New Delhi, of the Indira Gandhi National Open University as well as a member of the Technical Advisory Committee of the Indian Statistical Institute, Calcutta.

Professor Misra is an Associate Editor of the *International Journal of Innovative Computing, Information and Control* (Japan) and a Member of the Editorial Board of the *International Journal of Scientific Computing.*

On various occasions he delivered invited lectures at the Courant Institute of Mathematical Sciences, New York and at the Cornell Medical Center, New York, Universities of California at Los Angeles and San Diego in USA, Imperial College London, Universities of Cambridge, Oxford, Manchester and Glasgow in UK, University of Hamburg-Harburg and University of Kassel in Germany, University of Paris and Ecole Polytechnique in France, Graz University in Austria, Delft University in the Netherlands, Universities of Tokyo, Osaka and Kobe in Japan, National University of Singapore and Hong Kong University of Science and Technology. He delivered invited plenary lectures and chaired plenary sessions at the Fourth International Congress of Biorheology held in Tokyo and International Conferences on Computational and Applied Mathematics held in Belgium in 2002, Applied Mathematics and Mathematical Physics held at Sylhet in 2003, and Information Technology held in Kathmandu in 2003. Professor Misra delivered the prestigious *Bhatnagar Memorial Lecture* in 2001 and the *Indian Science Congress Platinum Jubilee Lecture* in Mathematical Sciences in 2005. In 2006, Prof. Misra was elected the President of Mathematical Sciences Section (including Statistics) of the Indian Science Congress.

Professor Misra built up an active school of research at IIT Kharagpur and guided 25 research scholars towards their Ph.D. degrees.

This page is intentionally left blank

PREFACE

This really is the golden age of Mathematics. It has been said that half the Mathematics ever created has been in the last 100 years and that half the mathematicians who have ever lived are alive today. We have seen such achievements as the resolution of the four-colour problem and Fermat's last theorem, with the latter being a special manifestation of a much more general result!

This book consists of chapters that deal with important topics in Biomathematics. A glance through any modern textbook or journal in the fields of ecology, genetics, physiology or biochemistry reveals that there has been an increasing use of mathematics, which ranges from the solution of complicated differential equation in population studies to the use of transfer functions in the analysis of eye-tracking mechanisms. This volume deals with Applied Mathematics in Biology and Medicine and is concerned with applied mathematical models and computer simulation in the areas of Molecular and Cellular Biology, Biological Soft Tissues and Structures as well as Bioengineering.

In this volume an attempt has been made to cover biological background and mathematical techniques whenever required. The aim has been to formulate various mathematical models on a fairly general platform, making the biological assumptions quite explicit and to perform the analysis in relatively rigorous terms. I hope, the choice and treatment of the problems will enable the readers to understand and evaluate detailed analyses of specific models and applications in the literature.

The purpose of bringing out this volume on Biomathematics dealing with interdisciplinary topics has been twofold. The objectives are to promote research in applied mathematical problems of the life sciences and to enhance cooperation and exchanges between mathematical scientists, biologists and medical researchers. This volume has both a synthetic and

analytic effect. The different chapters of the volume have been mostly concerned with model building and verification in different areas of biology and the medical sciences.

I believe people in the entire spectrum of those with interest in ecology, from field biologists seeking a conceptual framework for their observations to mathematicians seeking fruitful areas of application, will find stimulation here. It may so happen that some readers may find some parts of this volume trivial and some of the parts incomprehensible. Keeping this in view the extensive bibliographies given at the end of each chapter do attempt to provide an entry to the corresponding areas of study.

For over 35 years I have been engaged in teaching and research at several well-known institutions of India, Germany and North America. Publication of the series of books has been the fruit of a long period of collaboration together with relentless perseverance. My labour will be deemed amply rewarded if at least some of those for whom the book is meant derive benefit from it.

I feel highly indebted to the contributors of this volume who have so kindly accepted my invitation to contribute chapters. The enormous pleasure and enthusiasm with which they have accepted my invitation have touched me deeply, boosting my interest in the publication of the book.

I constantly remember the extent of care my parents have taken to impart proper education to me. I am highly indebted to Srimat Swami Shankaranandaji Maharaj, seventh President of the Ramakrishna Math and the Ramakrishna Mission, Belur Math, Swami Tejasanandaji and Swami Gokulanandaji, the then Principal and Vice-Principal of the Ramakrishna Mission Vidyamandira, Belur Math and to the monastic members of the Ramakrishna Mission Calcutta Students' Home, Belgharia for their kind guidance and suggestions and for instilling in me, while I was still a college and university student, a deep sense of total involvement in pursuing academic goals and a strong commitment to human values.

It is a pleasure to acknowledge the moral support, help and encouragement that I have been receiving constantly in all my academic activities from my wife Shorasi and my children Subhas, Sumita and Sudip.

I.I.T. Kharagpur J. C. Misra

January, 2005

CONTENTS

CHAPTER 1

DETECTING MOSAIC STRUCTURES IN DNA SEQUENCE ALIGNMENTS

DIRK HUSMEIER

This article first provides a concise introduction to the statistical approach to phylogenetics. It then describes a new method for detecting mosaic structures in DNA sequence alignments, which is based on combining two probabilistic graphical models: (1) a taxon graph (phylogenetic tree) representing the relationships among the taxa, and (2) a site graph (hidden Markov model) representing spatial correlations between nucleotides.

1. Introduction

The recent advent of multiple-resistant pathogens has led to an increased interest in interspecies recombination as an important, and previously underestimated, source of genetic diversification in bacteria and viruses. The discovery of a surprisingly high frequency of mosaic RNA sequences in HIV-1 suggests that a substantial proportion of AIDS patients have been coinfected with HIV-1 strains belonging to different subtypes, and that recombination between these genomes can occur *in vivo* to generate new biologically active viruses [25]. A phylogenetic analysis of the bacterial genera *Neisseria* and *Streptococcus* has revealed that the introduction of blocks of DNA from penicillin-resistant non-pathogenic strains into sensitive pathogenic strains has led to new strains that are both pathogenic *and* resistant [16]. Thus interspecies recombination, illustrated in Figs. 8 and 9, raises the possibility that bacteria and viruses can acquire biologically important traits through the exchange and transfer of genetic material.

In the last few years, a plethora of methods for detecting interspecies recombination have been developed — following up on the seminal paper by John Maynard Smith [16] — and it is beyond the scope of this article to provide a comprehensive overview. Instead, the focus will be on a novel approach, in which two probabilistic graphical models are combined: (1) a taxon graph (phylogenetic tree) representing the relationships among the species or strains, and (2) a site graph (hidden Markov model) representing

interactions between different sites in the DNA sequence alignments. While at present this approach is still limited to deal with only small numbers of species or strains simultaneously, it has two advantages over existing (mostly heuristic) methods: first, it can predict the locations and breakpoints of recombinant regions more accurately than what can be achieved with most existing techniques. Second, it provides a proper probabilistic generative model. This implies that well-known methods from statistics, like maximum likelihood, can be applied to estimate the parameters. It also renders the model amenable to established statistical methods of hypothesis testing and model selection.

The article is organized as follows. Section 2 provides a brief introduction to the statistical approach to phylogenetics. Section 3 explains the biological process of interspecific recombination. Section 4 provides a short recapitulation of hidden Markov models. Section 5 discusses a hybrid model — combining phylogenetic trees with hidden Markov models — for detecting recombination in DNA sequence alignments. Also, different ways of estimating the model parameters are discussed. Section 6 describes several DNA sequence alignments, on which the proposed model and training algorithms are tested. The results of these tests are discussed in Sec. 7. Finally, Sec. 8 contains a short summary and an outlook on future work.

2. A Brief Introduction to Phylogenetics

2.1. *Topology and parameters of a phylogenetic tree*

The objective of phylogenetics is to infer the evolutionary relationships among different species or strains and to display them in a tree-structured graphical model called a *phylogenetic tree*. An example is given in Fig. 1. The leaves of the (unrooted) phylogenetic tree represent contemporary species, like chicken, frog, mouse, etc. The inner or hidden nodes represent hypothetical ancestors, where a splitting of lineages occurs. These so-called speciation events lead to a diversification in the course of evolution, separating, for example, warm-blooded from cold-blooded animals, birds from mammals, primates from rodents, and so on. A phylogenetic tree conveys two types of information. The *topology* defines the branching order of the tree and the way the contemporary species are distributed among the leaves. For example, from Fig. 1 we learn that the mammals — human, chicken, mouse, and opossum — are grouped together, and are separated from the group of animals that lay eggs — chicken and frog. Within the

Frog	G C T T G A C T T C T G A G G T T
Chicken	G C G T A A C T T C A C A T G A T
Human	G C G T C A C T T G A G A C G C T
Rabbit	G C G T C A C T T G A G A C G C T
Mouse	G C G T C A C T T G A C A G G C T
Opossum	G C G T C A C T T G A G A C G C T

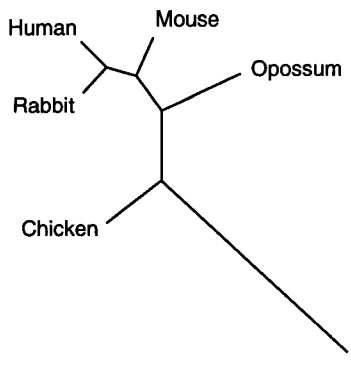


Fig. 1. **Phylogeny and DNA sequence alignment.** The figure shows a phylogenetic tree for six species and a subregion of the DNA sequence alignment from which it is inferred. The *topology* of the tree is the branching order, that is, the way the species are distributed across the leaf nodes. The *parameters* of the tree are the branch lengths, which represent phylogenetic time.

former group, opossum is grouped out, since it is a marsupial and therefore less closely related to the other "proper" mammals. Exchanging, for instance, the leaf positions of opossum and rabbit changes the branching order and thus leads to a different tree topology. For n species there are, in total, $(2n-5)!!$ different (unrooted) tree topologies, as can easily be proved by induction (see, for instance [4, Chap. 7]). In what follows, we will use the integer variable $S \in \{1, 2, \ldots, (2n-5)!!\}$ to label the different tree topologies.

The second type of information we obtain from a phylogenetic tree are the *branch lengths*, which represent phylogenetic time, measured by the average amount of mutational change. For example, Fig. 1 shows a comparatively long branch leading to the leaf with *frog*. This suggests that the splitting of the lineages separating *frog* from the other animals happened comparatively long ago, that is, earlier than the other speciation events. This is a reasonable conjecture as *frog* is the only cold-blooded animal, whereas all the other animals are warm-blooded. A (unrooted) tree

for n species has $n-2$ inner nodes, and thus $m = n + (n-2) - 1 = 2n-3$ branches. In what follows, individual branch lengths will be denoted by w_i, and the total vector of branch lengths will be denoted by $\mathbf{w} = (w_1, \ldots, w_{2n-3})$.

2.2. *DNA sequences and sequence alignments*

We now need a method to infer the correct topology of a tree and its branch lengths for a given set of species. As the driving force for evolution are mutations, that is, errors in the replication of DNA, it is reasonable to base our inference process on this information. This approach has recently become viable by major breakthroughs in DNA sequencing techniques. In July 1995, the entire 1.8 million base pairs of the genome of *Haemophilus influenzae*, a small Gram-negative bacterium, was published. Since then, the amount of DNA sequence data in publicly accessible data bases has been growing exponentially and is now about to claim its biggest triumph: the complete 3.3 billion base-pair DNA sequence of the entire human genome (for which a first draft was already released in June 2000).

DNA is composed of an alphabet of four nucleotides, which come in two families: the purines *adenine* (A) and *guanine* (G), and the pyrimidines *cytosine* (C) and *thymine* (T). DNA sequencing is the process of determining the order of these nucleotides. After obtaining the DNA sequences of the taxa of interest, we want to compare homologous subsequences, that is, regions of the genome that have been acquired from the same common ancestor. Also, one has to allow for nucleotide insertions and deletions. For example, a direct comparison of the sequences

```
A C G T T A T A
A G T C A T A
```

gives the erroneously small count of only a single site with identical nucleotides. This is due to the insertion of a *C* in the second position of the first strand, or, equivalently, the deletion of a nucleotide at the second position of the second strand (the insertion of a so-called *gap*). A correct comparison leads to

```
A C G T T A T A
A - G T C A T A
```

which suggests that the sequences differ in only two positions. The process of (1) finding homologous DNA subsequences and (2) correcting for

insertions and deletions is called DNA sequence alignment. A standard algorithm is Clustal-W, discussed in [28]. The details are beyond the scope of this article.

Figure 1, top, shows a small section of the DNA sequence alignment used for inferring the tree at the bottom of Fig. 1. Rows represent different species or strains (generic name: taxa), columns represent different sites or positions on the DNA. At the majority of sites, all nucleotides are identical, which reflects the fact that the compared sequences are homologous. At certain positions, however, differences occur, resulting from mutational changes during evolution. In the fifth column, for instance, human, rabbit, mouse, and opossum have a C, chicken has an A, and frog has a G. This reflects the fact that the first four species are mammals and therefore more closely related to each other than to the two remaining species. Note, however, that the process of nucleotide substitution is intrinsically stochastic. We will therefore discuss, in the following two subsections, a mathematical model for statistical phylogenetic inference.

2.3. *A mathematical model of nucleotide substitution*

The driving force for evolution are nucleotide substitutions, which can be modelled as transitions in a 4-state state space, shown in Fig. 2. $P(Y|X,w)$, where $X,Y \in \{A,C,G,T\}$, denotes the probability of a transition from nucleotide X into nucleotide Y, conditional on the elapsed phylogenetic time w. The latter is given by the product of an unknown mutation rate λ with physical time t: $w = \lambda t$. To rephrase this: $P(Y|X,w)$ is the probability that nucleotide Y is found at a given site in the DNA sequence given that w phylogenetic time units before, the same site was occupied by nucleotide X.

An intuitively plausible functional form for these probabilities is shown on the right of Fig. 2. For $w = 0$, there is no time for nucleotide substitutions to occur. Consequently, $P(A|A,w=0) = 1$, and $P(C|A,w = 0) = P(G|A,w=0) = P(T|A,w=0) = 0$. As w increases, nucleotide substitutions from A into the other states lead to an exponential decay of $P(A|A,w)$, and, concurrently, an increase of $P(C|A,w)$, $P(G|A,w)$, and $P(T|A,w)$. This increase is faster for a mutation within a nucleotide class (purine $\rightarrow$ purine, pyrimidine $\rightarrow$ pyrimidine), than between nucleotide classes (purine $\leftrightarrow$ pyrimidine). For $w \rightarrow \infty$, the system "forgets" its initial configuration as the result of the mixing caused by an increasing number of nucleotide substitutions. Consequently, $P(Y|X,w) \rightarrow \Pi(Y)$,

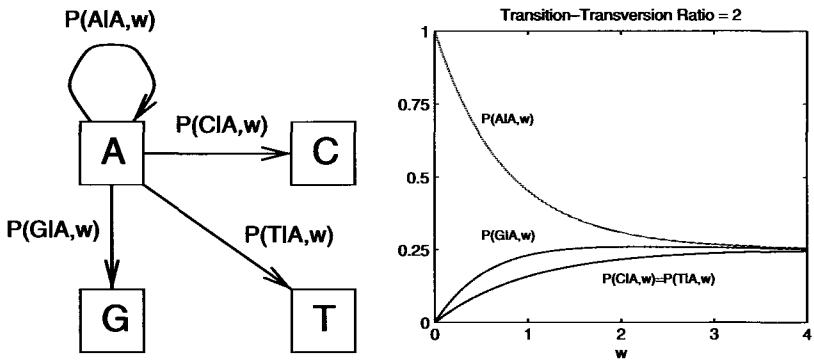


Fig. 2. **Mathematical model of nucleotide substitutions.** *Left:* Nucleotide substitutions are modelled as transitions in a 4-state state space. The transition probabilities depend on the phylogenetic time $w = \alpha t$, where t is physical time and α is a mutation rate. *Right:* Dependence of the transition probabilities (vertical axis) on w (horizontal axis). The graphs were obtained from the Kimura model with a transition-transversion ratio of 2.

where $X, Y \in \{A, C, G, T\}$, and $\Pi(Y)$ is the equilibrium distribution (here $\Pi(Y) = 1/4 \,\forall\, Y$).

Let $y_i(t) \in \{A, C, G, T\}$ denote the nucleotide at site i and at physical time t. This notation will be used throughout this chapter: the subscript refers to the position in the alignment, while the expression in brackets denotes physical or (later) phylogenetic time. The total length of the alignment is N, that is, $i \in \{1, \ldots, N\}$. The derivation of the aforementioned results is based on the theory of homogeneous Markov chains and the following assumptions:

- The process is Markov:

$$P\big(y_i(t+\Delta t)|y_i(t), y_i(t-\Delta t), \ldots\big) = P\big(y_i(t+\Delta t)|y_i(t)\big).$$

- The Markov process is homogeneous:

$$P\big(y_i(s+t)|y_i(s)\big) = P\big(y_i(t)|y_i(0)\big).$$

- The Markov process is the same for all positions:

$$P\big(y_i(t)|y_i(0)\big) = P\big(y_k(t)|y_k(0)\big) \quad \forall\, i, k \in \{1, \ldots, N\}.$$

- Substitutions at different positions are independent of each other:

$$P\big(y_1(t), \ldots, y_N(t)|y_1(0), \ldots, y_N(0)\big) = \prod_{i=1}^{N} P\big(y_i(t)|y_i(0)\big).$$

This implies that the nucleotide substitution process at a given site is completely specified by the following 4-by-4 transition matrix:

$$\mathbf{P}(t) = \begin{bmatrix} P(y(t) = A|y(0) = A) & \cdots & P(y(t) = A|y(0) = T) \\ P(y(t) = G|y(0) = A) & \cdots & P(y(t) = G|y(0) = T) \\ P(y(t) = C|y(0) = A) & \cdots & P(y(t) = C|y(0) = T) \\ P(y(t) = T|y(0) = A) & \cdots & P(y(t) = T|y(0) = T) \end{bmatrix}. \tag{1}$$

Because of the site independence, the site label (that is, the subscript) has been dropped to simplify the notation. Equation (1) obviously implies that

$$\mathbf{P}(0) = \mathbf{I}, \tag{2}$$

where $\mathbf{I}$ is the unit matrix. We now make the ansatz

$$\mathbf{P}(dt) = \mathbf{P}(0) + \mathbf{R}dt, \tag{3}$$

where $\mathbf{R}$ is the so-called rate matrix. From the theory of homogeneous Markov chains it is known that

$$\mathbf{P}(t + dt) = \mathbf{P}(dt)\mathbf{P}(t), \tag{4}$$

which follows from the Chapman–Kolmogorov equation; see [10] or [22]. Inserting Eqs. (2) and (3) into (4) gives:

$$\mathbf{P}(t + dt) = (\mathbf{I} + \mathbf{R}dt)\mathbf{P}(t) \tag{5}$$

and

$$\frac{d\mathbf{P}}{dt} = \mathbf{RP}. \tag{6}$$

This is a system of linear differential equations with the solution

$$\mathbf{P}(t) = e^{\mathbf{R}t}. \tag{7}$$

To make sure that $\mathbf{P}(t)$ is a proper transition matrix, that is, has columns that sum to 1, the columns of the rate matrix $\mathbf{R}$ have to sum to 0. A possible design for $\mathbf{R}$, the so-called Kimura model [15], is of the form

$$\mathbf{R} = \begin{bmatrix} -2\beta - \alpha & \beta & \alpha & \beta \\ \beta & -2\beta - \alpha & \beta & \alpha \\ \alpha & \beta & -2\beta - \alpha & \beta \\ \beta & \alpha & \beta & -2\beta - \alpha \end{bmatrix}. \tag{8}$$

Here, the rows (from top to bottom) and columns (from left to right) correspond to the nucleotides A, C, G, T (in the indicated order). The positive parameters α and β denote the rates of transitions (mutations within a

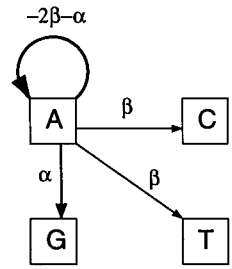


Fig. 3. **Kimura model of nucleotide substitutions.** The figure presents a partial graphical display of the rate matrix of Eq. (8), showing mutations out of nucleotide A. The positive parameter α denotes the rate of a transition (purine $\rightarrow$ purine, pyrimidine $\rightarrow$ pyrimidine), while β denotes the rate of a transversion (purine $\leftrightarrow$ pyrimidine).

nucleotide class: purine $\rightarrow$ purine, pyrimidine $\rightarrow$ pyrimidine) and transversions (mutations between nucleotide classes: purine $\leftrightarrow$ pyrimidine), respectively.[a] An illustration is given in Fig. 3.

It can now be shown [15] that inserting (8) into (7) leads to

$$\mathbf{P}(t) = e^{\mathbf{R}t} = \begin{bmatrix} d(t) & f(t) & g(t) & f(t) \\ f(t) & d(t) & f(t) & g(t) \\ g(t) & f(t) & d(t) & f(t) \\ f(t) & g(t) & f(t) & d(t) \end{bmatrix}, \tag{9}$$

where

$$\begin{aligned} f(t) &= \frac{1}{4}(1 - e^{-4\beta t}), \\ g(t) &= \frac{1}{4}(1 + e^{-4\beta t} - 2e^{-2(\alpha+\beta)t}), \\ d(t) &= 1 - 2f(t) - g(t). \end{aligned}$$

Defining $\lambda = 4\beta$, which implies that the phylogenetic time is given by

$$w = 4\beta t, \tag{10}$$

this results in

$$f(w) = \frac{1}{4}(1 - e^{-w}) \tag{11}$$

[a]Unfortunately this terminology, which is used in molecular biology, leads to a certain ambiguity in the meaning of the word *transition*. When we talk about transitions between *states*, a transition can be any nucleotide substitution event. When we talk about *transitions* as opposed to *transversions*, a transition refers to a certain type of nucleotide substitution.

$$g(w) = \frac{1}{4}(1 + e^{-w} - 2e^{-\frac{\tau+1}{2}w}) \tag{12}$$

$$d(w) = 1 - 2f(w) - g(w) \tag{13}$$

in which τ denotes the transition-transversion ratio:

$$\tau = \frac{\alpha}{\beta}. \tag{14}$$

Denoting by $P(Y|X,w)$ the probability that at a given site in the alignment nucleotide Y is observed given that nucleotide X was at this site w phylogenetic time units before, we can re-write $\mathbf{P}$, the transition matrix of (1), as follows:

$$\begin{aligned}\mathbf{P}(w) &= \begin{bmatrix} P(A|A,w) & P(A|C,w) & P(A|G,w) & P(A|T,w) \\ P(G|A,w) & P(G|C,w) & P(G|G,w) & P(G|T,w) \\ P(C|A,w) & P(C|C,w) & P(C|G,w) & P(C|T,w) \\ P(T|A,w) & P(T|C,w) & P(T|G,w) & P(T|T,w) \end{bmatrix} \\ &= \begin{bmatrix} d(w) & f(w) & g(w) & f(w) \\ f(w) & d(w) & f(w) & g(w) \\ g(w) & f(w) & d(w) & f(w) \\ f(w) & g(w) & f(w) & d(w) \end{bmatrix},\end{aligned} \tag{15}$$

where $d(w)$, $f(w)$, and $g(w)$ are given by (11)–(13).

Setting $\tau = 2$ leads to the graphs on the right of Fig. 2 and the results discussed at the beginning of this section.

2.4. *Likelihood of a phylogenetic tree*

A phylogenetic tree is a directed acyclic graph (DAG), which allows the expansion of the joint probability of the nodes in terms of the transition probabilities of (15). This expansion is based on the factorization rule for directed graphical models (see, for instance [12]), according to which the joint probability of a set of random variables $x_1, \ldots, x_n$ can be factorized as

$$P(x_1, \ldots, x_N) = \prod_{i=1}^{N} P(x_i|\text{parents}[x_i]), \tag{16}$$

where parents$[x_i]$ is the set of random variables corresponding to the subset of nodes with an arrow that feeds into x_i.

Consider Fig. 4, left. The black nodes, labelled by y_1, y_2, y_3, and y_4, represent contemporary species. The white nodes, labelled by z_1 and z_2, represent hypothetical ancestors. We are interested in the probability $P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S)$, where y_1, y_2, etc. represent nucleotides at the

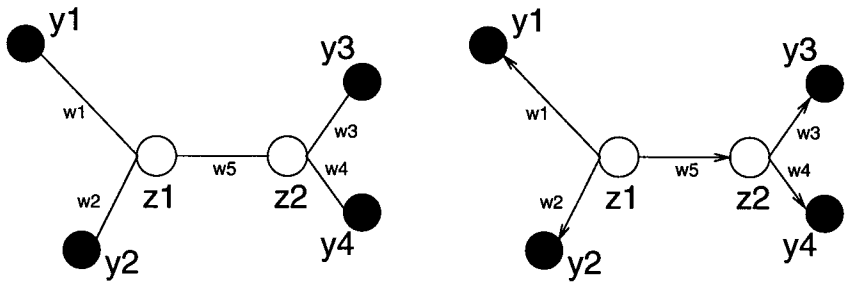


Fig. 4. **Phylogenetic trees.** Black nodes represent contemporary or extant species. White nodes represent hypothetical ancestors, where lineages bifurcate (speciation). *Left:* Undirected graph. *Right:* Directed graph. Node z_1 is the root of the tree, and arrows are directed.

respective nodes, $\mathbf{w}$ is the vector of all branch lengths, and S is a label defining the tree topology. Choosing, arbitrarily, z_1 to be the root of the tree, see Fig. 4, right, the application of (16) gives:

$$P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S) = P(y_1|z_1, w_1)P(y_2|z_1, w_2)P(z_2|z_1, w_5)P(y_3|z_2, w_3)P(y_4|z_2, w_4)\Pi(z_1). \tag{17}$$

The equilibrium distribution over the four nucleotides, $\Pi(z_1)$, is a parameter vector of the model. For example, in the Kimura model we have $\Pi(z_1 = A) = \Pi(z_1 = C) = \Pi(z_1 = G) = \Pi(z_1 = T) = 0.25$. The other factors represent transition probabilities, which are defined in (15).

Now, we assume that the transition matrix (15) is reversible:

$$P(Y|X, w)\Pi(X) = P(X|Y, w)\Pi(Y), \tag{18}$$

where $X, Y \in \{A, C, G, T\}$. Obviously, this holds true for the Kimura model discussed above. It can then be shown that the expansion of the joint probability $P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S)$ is independent of the root position.

Compare, for instance, the three directed graphs in Fig. 5. We have just derived the expansion for the tree on the left; see (17). Applying the expansion rule (16) to the tree in the middle, we obtain:

$$P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S) = P(y_1|z_1, w_1)P(y_2|z_1, w_2)P(z_1|z_2, w_5)P(y_3|z_2, w_3)P(y_4|z_2, w_4)\Pi(z_2). \tag{19}$$

Now, reversibility implies that $P(z_1|z_2, w_5)\Pi(z_2) = P(z_2|z_1, w_5)\Pi(z_1)$, hence the expansions in (17) and (19) are identical. By the same token,

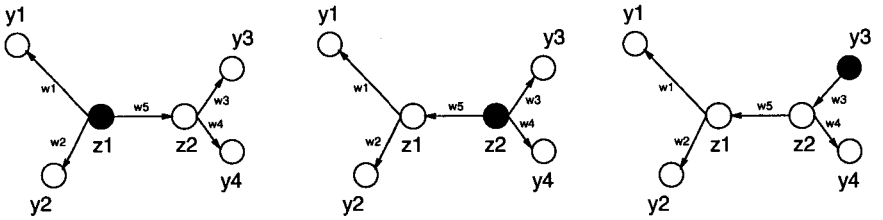


Fig. 5. **Different root positions.** The figure shows three directed graphs with different root positions (shown in black).

expanding the joint probability $P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S)$ according to the tree on the right of Fig. 5 gives

$$P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S) = P(y_1|z_1, w_1)P(y_2|z_1, w_2)P(z_1|z_2, w_5)P(y_4|z_2, w_4)P(z_2|y_3, w_3)\Pi(y_3). \tag{20}$$

Applying reversibility, $P(z_2|y_3, w_3)\Pi(y_3) = P(y_3|z_2, w_3)\Pi(z_2)$, this expansion is seen to be identical to (19) and hence (17). In the terminology of graphical models, the three directed graphs in Fig. 5 are *distribution equivalent* [12], that is, they represent the same joint probability distribution. In fact, a more rigorous proof [6] generalizes this finding to any phylogenetic tree: if the transition matrix is reversible, trees that only differ with respect to the position of the root and the directions of the arcs are equivalent. Consequently, we can choose the position of the root arbitrarily.[b]

The factorization (17) allows us to compute the probability of a complete configuration of nucleotides. However, while we obtain the nucleotides of the extant species, y_i, from the DNA sequence alignment, the nucleotides at the inner nodes, z_i, are never observed. This requires us to marginalize over them, as illustrated in Fig. 6:

$$P(y_1, y_2, y_3, y_4|\mathbf{w}, S) = \sum_{z_1}\sum_{z_2} P(y_1, y_2, y_3, y_4, z_1, z_2|\mathbf{w}, S). \tag{21}$$

There are efficient message-passing algorithms to carry out this marginalization and decrease the computational complexity of the summation; see [6].

The upshot of this procedure is that for a given column $\mathbf{y}_t$ in the alignment, a probability $P(\mathbf{y}_t|\mathbf{w}, S)$ can be computed, which depends on the

[b]In more recent phylogenetic models, this reversibility constraint has been relaxed. See, for instance [8].

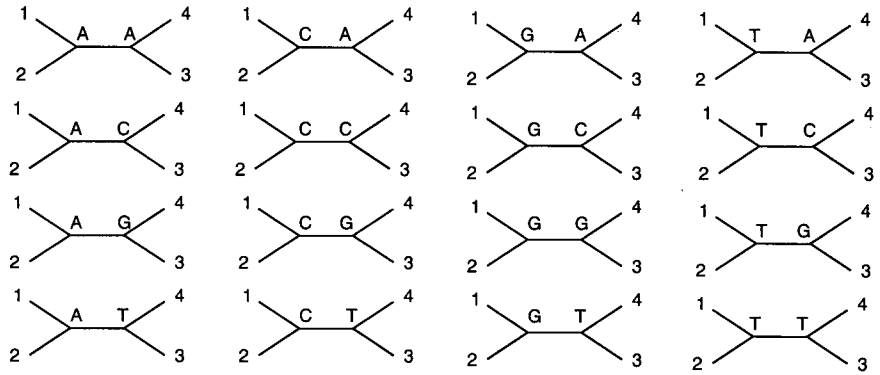


Fig. 6. **Marginalization over hidden nodes.** Leaf nodes represent extant taxa, which are observed (nucleotides in the DNA sequence alignment). Hidden nodes represent hypothetical ancestors, which are *not* observed (nuisance parameters). To obtain the probability of an observation, that is, the probability of observing a given column of nucleotides at a certain position in the DNA sequence alignment, one has to sum over all possible configurations of hidden nodes.

tree topology, S, and the vector of branch lengths, $\mathbf{w}$. This can be done for every site, $1 \leq t \leq N$, which allows, under the assumption that mutation events at different sites are independent of each other, the computation of the likelihood $P(\mathcal{D}|\mathbf{w}, S)$ of the whole DNA sequence alignment $\mathcal{D}$:

$$P(\mathcal{D}|\mathbf{w}, S) = \prod_{t=1}^{N} P(\mathbf{y}_t|\mathbf{w}, S). \tag{22}$$

This, in principle, opens the way to a maximum likelihood optimization of the tree: given a DNA sequence alignment $\mathcal{D}$, the tree $(\hat{S}, \hat{\mathbf{w}})$ most supported by the data is the one that maximizes the likelihood:

$$(\hat{S}, \hat{\mathbf{w}}) = \underset{S,\mathbf{w}}{\operatorname{argmax}}\{P(\mathcal{D}|\mathbf{w}, S)\}. \tag{23}$$

More precisely, one should also state the dependence of the likelihood on the nucleotide substitution model and its parameters, which also need to be optimized so as to maximize the likelihood. For the Kimura model, discussed above, we have one parameter: the transition-transversion ratio τ. Two more complex model, the HKY85 model [11] and the Felsenstein 84 model [5], have three further parameters: the equilibrium probabilities for the nucleotides, $\Pi(A), \Pi(C), \Pi(G), \Pi(T)$ (due to the constraint $\Pi(A) + \Pi(C) + \Pi(G) + \Pi(T) = 1$, there are three rather than four free parameters). Recently, more complex nucleotide substitution models have been

developed, as reviewed in [23]. These details are beyond the scope of this article. To keep the notation simple, the dependence of the likelihood on the nucleotide substitution model will not be stated explicitly.

A principled difficulty in applying the maximum likelihood method outline here is that the optimization problem is NP hard. As mentioned in Sec. 2.1, n taxa give rise to $(2n-5)!!$ different (unrooted) tree topologies, that is, the number of different tree topologies increases super-exponentially with the number of taxa. In practice this means that for large numbers of taxa one has to resort to iterative, greedy search algorithms, which usually find only a local rather than the global maximum of the likelihood. Effective algorithms have been proposed in [6] and [5], and are implemented in the program DNAML of the PHYLIP software package [7]. For an introductory text, see also [4]. The details of these optimization algorithms will not be summarized here. Instead, this article will focus on a fundamental problem inherent in the phylogenetic analysis of certain bacteria and viruses.

3. Recombination

Conventional phylogenetic analysis, as described in the previous section, assumes that all sites in a DNA multiple alignment have the same evolutionary history. This is a reasonable approach when applied to DNA sequences obtained from most species. However, this assumption is violated in certain bacteria and viruses due to interspecific *recombination*, which is a process that leads to the transfer or exchange of DNA subsequences between different strains. The resulting mixing of the genetic material by the formation of so-called *mosaic* sequences is likely to be an important source of genetic variation and is a process through which, for example, disease-causing bacteria may acquire resistance to antibiotics. Figure 8 shows an example in which the incorporation of the genetic material from another strain leads to a change of the branching order (topology) in the affected region, which results in conflicting phylogenetic information from different regions of the alignment. If undetected, the presence of mosaic sequences can lead to errors in phylogenetic tree estimation. Their detection, therefore, is a crucial prerequisite for inferring the evolutionary history of a set of DNA sequences.

Figure 9 shows an example of recombination in HIV-1 [25]. The left subfigure shows a phylogenetic tree for eight established strains of HIV-1. The subfigure on the right shows a so-called circulating recombinant strain, denoted by ZR-VI 191. If the phylogenetic analysis is done on the basis of the *env* gene, this strain is found to be most closely related to the A strain.

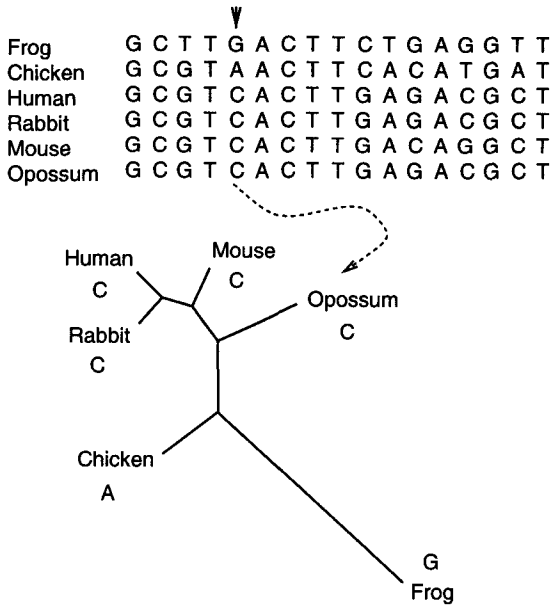


Fig. 7. **Statistical approach to phylogenetics.** For a given column $\mathbf{y}_t$ in the alignment, a probability $P(\mathbf{y}_t|\mathbf{w}, S)$ can be computed, which depends on the tree topology, S, and the vector of branch lengths, $\mathbf{w}$. This can be done for every site, $1 \leq t \leq N$, which allows the computation of the likelihood $P(\mathcal{D}|\mathbf{w}, S)$ of the whole DNA sequence alignment $\mathcal{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$.

For a phylogenetic analysis based on the *gag* gene, ZR-VI 191 is most closely related to the G strain. Ignoring recombination and treating the sequence of ZR-VI 191 as a monolithic entity will adversely affect the estimation of the branch lengths in the phylogenetic tree. For medical applications, determining a strain as a mosaic sequence of well-established strains can be important for vaccine development [25].

In the last few years, a plethora of methods for detecting interspecies recombination have been developed — following up on the seminal paper by John Maynard Smith [16] — and it is beyond the scope of this article to present a comprehensive overview. Many detection methods for identifying the nature and the breakpoints of the resulting mosaic structure are based on moving a window along the alignment and computing a phylogenetic divergence score for each window position. Examples are the bootstrap support for the locally optimal topology [26], the likelihood ratio between the locally and globally optimal trees [9], and the difference in the fitting

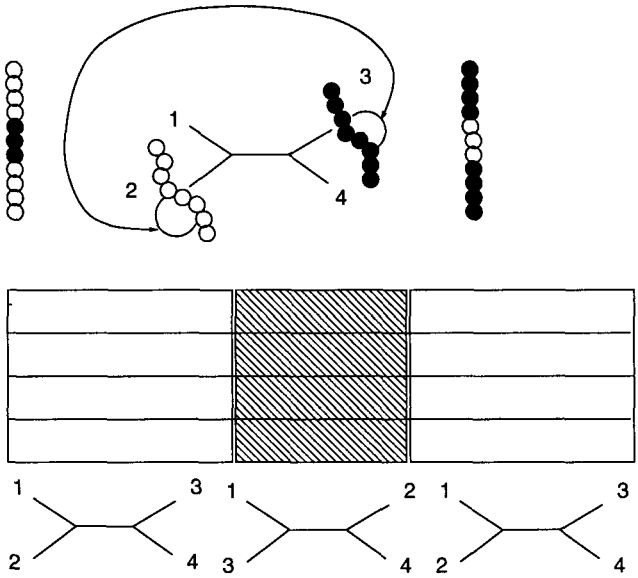


Fig. 8. **Influence of recombination on phylogenetic inference.** The figure shows a hypothetical phylogenetic tree of four strains. Recombination is the exchange of DNA subsequences between different strains (top diagram, middle), which results in two so-called mosaic sequences (top diagram, margins). The affected region in the multiple DNA sequence alignment (shown by the shaded area in the middle diagram) seems to originate from a different phylogenetic topology, in which two branches of the phylogenetic tree have been exchanged (bottom diagram, where the numbers at the leaves represent the four strains). Reprinted from [14], with permission from Mary Ann Liebert.

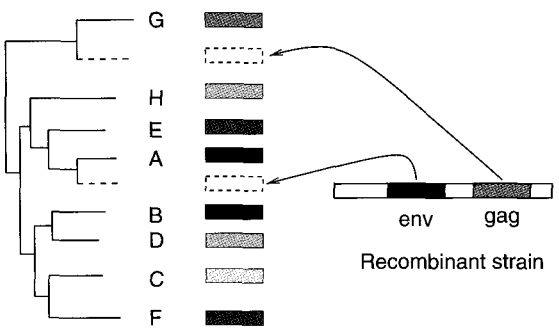


Fig. 9. **Recombination in HIV-1.** The left subfigure shows a phylogenetic tree for eight established strains of HIV-1. The subfigure on the right shows a so-called circulating recombinant strain, denoted by ZR-VI 191. If the phylogenetic analysis is done on the basis of the *env* gene, this strain is found to be most closely related to the A strain. For a phylogenetic analysis based on the *gag* gene, ZR-VI 191 is most closely related to the G strain.

scores between two adjacent locally optimized trees [17]. The determination of the breakpoints of the mosaic structure is then based on an analysis of the signals thus obtained, using bootstrapping to estimate their significance. While these methods are useful for a preliminary scan of a DNA sequence alignment, the spatial resolution for the identification of the breakpoints is typically of the order of the window size and, consequently, rather poor.

This chapter discusses a different approach, which was first suggested in [13]. The idea is to introduce a hidden state, which represents the tree topology at a given site. A state transition from one topology into another corresponds to a recombination event. To introduce correlations between adjacent sites, a site graph is introduced, representing which nucleotides interact in determining the tree topology. To keep the mathematical model tractable and the computational costs limited, interactions are reduced to nearest-neighbour interactions. The natural framework for modelling such a system is a hidden Markov model, whose application to the detection of recombination was first suggested in [18]. The next section provides a brief introduction to hidden Markov models.

4. A One-Minute Introduction to Hidden Markov Models

Assume you are in a casino and take part in some (hopefully legal) gambling game involving a die. You are playing against two players: a fair player, who uses a fair die, and a corrupt player, who uses a loaded die. The situation is illustrated in Fig. 10. Unfortunately, the other players are hidden behind a brick wall, and all you observe is a sequence of die faces; see Fig. 11. The

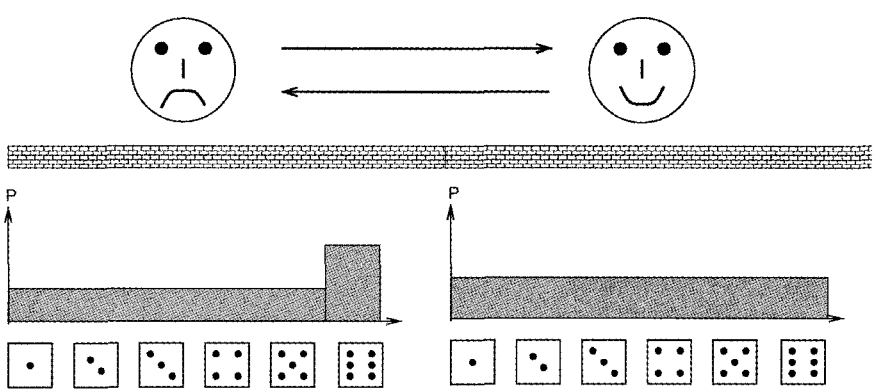


Fig. 10. **Corrupt casino 1.** Two players are in a casino: a fair player (right) using a fair die, and a corrupt player (left) using a loaded die.

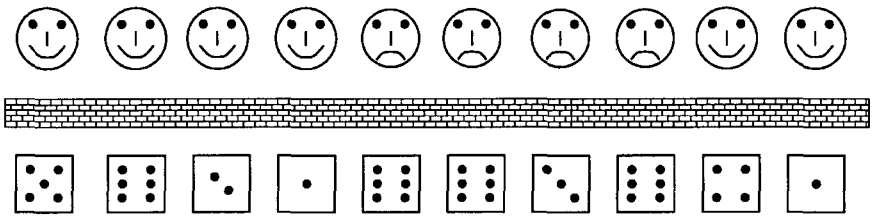


Fig. 11. **Corrupt casino 2.** The player is hidden behind a brick wall, and only the die faces are observed. The problem is to predict which player is rolling the die at a given time t.

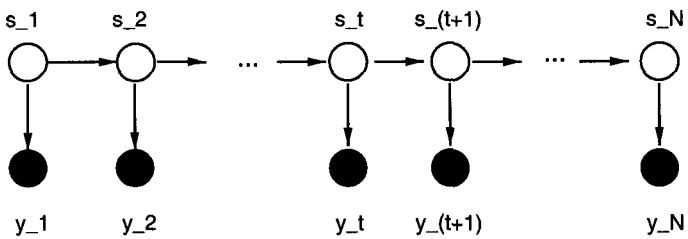


Fig. 12. **Hidden Markov model.** Black nodes represent observed random variables (the die faces), white nodes represent hidden states (the players), and arcs represent conditional dependencies. The joint probability factorizes into a product of emission probabilities (vertical arrows) and transition probabilities (horizontal arrows). The prediction task is to find the most likely sequence of hidden states given the observations.

task is to predict which player is rolling the die at a given time, and to predict the breakpoint where the corrupt player is taking over (in order to nab him).

If the decision of a player to pass the die on to the other player is made instantaneously on the basis of the current situation without considering the earlier past, the process corresponds to a hidden Markov model (HMM), shown in Fig. 12. Here, black nodes represent observed random variables y_t (the die faces) at different moments in time t, white nodes represent hidden states S_t (the players) at different times, and arcs represent conditional dependencies. The task is to find the most likely sequence of hidden states given the observations, that is, the mode of

$$P(\mathbf{S}|\mathbf{y}) = P(S_1, \ldots, S_N|y_1, \ldots, y_N). \tag{24}$$

At first, this task seems to be intractable: for K different states (here: $K = 2$ for "fair" and "corrupt") and a sequence of length N, there are

K^N different state sequences. Hence, an exhaustive search seems to be impossible for all but very short sequence lengths N. Fortunately, there is a dynamic programming method, the so-called *Viterbi algorithm*, which reduces the computational complexity to $\mathcal{O}(N)$ (that is, linear in N) by exploiting the sparseness of the connectivity of the graph in Fig. 12.

Recall that in a directed graphical model the joint probability of the random variables $x_1, \ldots, x_N$ can be factorized according to (16). The application of this formula to the graph in Fig. 12 gives:

$$P(y_1, \ldots, y_N, S_1, \ldots, S_N) = \prod_{t=1}^{N} P(y_t|S_t) \prod_{t=2}^{N} P(S_t|S_{t-1}) P(S_1). \tag{25}$$

We refer to $P(y_t|S_t)$ as the *emission probabilities* (corresponding to the vertical edges), $P(S_t|S_{t-1})$ as the *transition probabilities* (which correspond to the horizontal edges), and $P(S_1)$ as the *initial probability*. From (25) we obtain the recursion:

$$\begin{aligned}
\gamma_n(S_n) &= \max_{S_1,\ldots,S_{n-1}} \ln P(y_1, \ldots, y_n, S_1, \ldots, S_n) \\
&= \max_{S_1,\ldots,S_{n-1}} \left[\sum_{t=1}^{n} \ln P(y_t|S_t) + \sum_{t=2}^{n} \ln P(S_t|S_{t-1}) + \ln P(S_1) \right] \\
&= \ln P(y_n|S_n) + \max_{S_{n-1}} \left[\ln P(S_n|S_{n-1}) + \max_{S_1,\ldots,S_{n-2}} \left[\sum_{t=1}^{n-1} \ln P(y_t|S_t) \right.\right. \\
&\quad \left.\left. + \sum_{t=2}^{n-1} \ln P(S_t|S_{t-1}) + \ln P(S_1) \right] \right] \\
&= \ln P(y_n|S_n) + \max_{S_{n-1}} [\ln P(S_n|S_{n-1}) + \gamma_{n-1}(S_{n-1})].
\end{aligned} \tag{26}$$

Obviously:

$$\begin{aligned}
\max_{S_1,\ldots,S_N} P(S_1, \ldots, S_N|y_1, \ldots, y_N) &= \max_{S_1,\ldots,S_N} \ln P(y_1, \ldots, y_N, S_1, \ldots, S_N) \\
&= \max_{S_N} \gamma_N(S_N)
\end{aligned} \tag{27}$$

and the mode, $P(\hat{S}_1, \ldots, \hat{S}_N|y_1, \ldots, y_N)$, is obtained by recursive backtracking:

Initialization:

$$\hat{S}_N = \operatorname*{argmax}_{S_N} \gamma_N(S_N). \tag{28}$$

Recursion:

$$\hat{S}_{n-1} = \underset{S_{n-1}}{\operatorname{argmax}} \left[\ln P(\hat{S}_n | S_{n-1}) + \gamma_{n-1}(S_{n-1}) \right]. \tag{29}$$

The computational complexity of a single step of the recursions (26) and (29) is $\mathcal{O}(K^2)$, that is, it only depends on the number of different states K but is independent of the sequence length N. The total computational complexity of the algorithm is thus linear in N, which is a considerable improvement over the naive method, which was K^N. For a more detailed exposition of this topic, see [24].

5. Detecting Recombination with Hidden Markov Models

5.1. *The model*

Let us now study how HMMs can be applied to model mosaic structures in DNA sequence alignments. Here, the hidden state represents the phylogenetic tree topology at a given site. For four taxa, for instance, there are three possible tree topologies, shown in Fig. 13. The subscript t now represents sites in the DNA sequence alignment rather than time, hence S_t is the hidden state corresponding to the tth site in the alignment. The observations $\mathbf{y}_t$ are the columns of the DNA sequence alignment, that is, $\mathbf{y}_t$ is the vector with the nucleotides of all the taxa at the tth site in the alignment. For a given tree, we can compute the probability of $\mathbf{y}_t$, as discussed in Sec. 2.4 and illustrated in Fig. 7. Hence for a given DNA sequence alignment $\mathcal{D} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, we can apply the Viterbi algorithm to find the most likely sequence of hidden states, $S_1, \ldots, S_N$, that is, the mode of $P(S_1, \ldots, S_N | \mathbf{y}_1, \ldots, \mathbf{y}_N)$. Recombination events then correspond to state transitions in the Viterbi path.

Recall that in an HMM, the joint probability factorizes into the product of the emission probabilities, $P(\mathbf{y}_t | S_t)$, and the transition probabilities, $P(S_t | S_{t-1})$, where the latter correspond to recombination events. With K

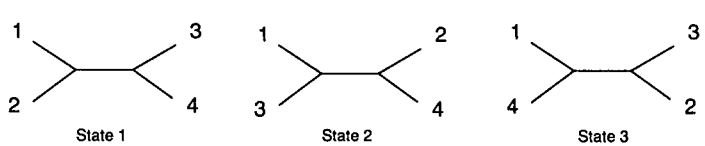


Fig. 13. **Different tree topologies for four taxa.** Shown are the three possible phylogenetic tree topologies for four taxa. Species 1 can be clustered with species 2, 3, or 4. Reprinted from [14], with permission from Mary Ann Liebert.

different tree topologies, there are, in principle, $K(K-1)$ transition probabilities to be specified. However, given that recombination is likely to be a rare event, it would hardly be possible to reasonably infer these parameters from the DNA sequence alignment (over-fitting), nor is it likely that detailed prior knowledge is available to decide on these parameters in advance. For this reason, only *one* free parameter was used in [18]: the overall probability that no recombination occurs. This is similar to an approach taken in [5] for modelling rate variation among sites. Let ν be the probability that the tree topology remains unchanged as we move from a given site in the alignment, t, to an adjacent site, $t+1$ or $t-1$. We then obtain for the state transition probabilities:

$$\begin{aligned} P(S_t|S_{t-1}) &= \nu\delta(S_t, S_{t-1}) + \frac{1-\nu}{K-1}[1-\delta(S_t, S_{t-1})] \\ &= \nu^{\delta(S_t,S_{t-1})}\left(\frac{1-\nu}{K-1}\right)^{1-\delta(S_t,S_{t-1})}, \end{aligned} \tag{30}$$

where $\delta(S_t, S_{t-1})$ denotes the Kronecker delta function, which is 1 when $S_t = S_{t-1}$, and 0 otherwise. It is easily checked that this satisfies the normalization constraint $\sum_{S_t} P(S_t|S_{t-1}) = 1$. For the emission probabilities, recall from Sec. 2.4 and Fig. 7 that for a given nucleotide substitution model, the probability of a column vector $\mathbf{y}_t$ depends both on the tree topology, S_t, and the vector of branch lengths corresponding to this topology, $\mathbf{w}_{S_t}$. To simplify the notation, let us introduce the accumulated vector of all branch lengths in all possible topologies, $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_K)$, and define: $P(\mathbf{y}_t|S_t, \mathbf{w}_{S_t}) = P(\mathbf{y}_t|S_t, \mathbf{w})$. This means that S_t indicates which subvector of $\mathbf{w}$ applies. We can depict the dependence of the probability distribution on the parameters $\mathbf{w}$ and ν in an extended graphical model, shown in Fig. 14. Applying the Viterbi algorithm gives us the most likely hidden state sequence conditional on the observations (that is, the DNA sequence alignment) and the parameters $\mathbf{w}$ and ν:

$$\underset{S_1,\ldots,S_N}{\operatorname{argmax}} P(S_1, \ldots, S_N|\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{w}, \nu). \tag{31}$$

We thus need a way to estimate these parameters.

5.2. *Naive parameter estimation*

A straightforward way to estimate the branch lengths $\mathbf{w}$ seems to be a separate maximum likelihood optimization for each possible tree topology. This can be accomplished with the methods described at the end of Sec. 2.4, and was applied in [18]. However, Fig. 8 points to a serious shortcoming

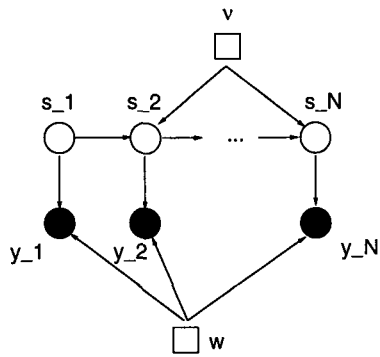


Fig. 14. **Modelling recombination with a hidden Markov model.** Positions in the model, labelled by the subscript t, correspond to positions in the DNA sequence alignment. Black nodes represent observed random variables; these are the columns in the DNA sequence alignment. White nodes represent hidden states; these are the different tree topologies, as shown in Fig. 13. Squares represent parameters of the model: the vector of branch lengths $\mathbf{w}$, and the recombination parameter ν. Arcs represent conditional dependencies. The probability of observing a column vector $\mathbf{y}_t$ at position t in the DNA sequence alignment depends on the tree topology S_t and the vector of branch lengths $\mathbf{w}$. The tree topology at position t depends on the topologies at the adjacent sites, S_{t-1} and S_{t+1}, and the recombination parameter ν.

of this approach. For a proper estimation of the branch lengths of the recombinant tree, that is, the tree that corresponds to the shaded centre region of the alignment, one would have to base the parameter estimation on this very region of the alignment. Unfortunately, its location is not known in advance. Estimating the branch lengths from the whole DNA sequence alignment leads to seriously distorted values — see Fig. 15 — since the estimation includes data for which the tree topology is incorrect. A heuristic way to address this problem, suggested in [18], is to estimate the branch lengths from a subregion of the alignment. The length of this region should be matched to the length of the recombinant region, which, however, is not known in advance. Also, this approach does not offer a way to estimate the recombination parameter ν.

5.3. *Maximum likelihood*

A solution to this problem, proposed in [14], is a proper maximum likelihood estimation of the parameters so as to maximize

$$L(\mathbf{w}, \nu) = \ln P(\mathcal{D}|\mathbf{w}, \nu) = \ln \sum_{\mathbf{S}} P(\mathcal{D}, \mathbf{S}|\mathbf{w}, \nu) \tag{32}$$

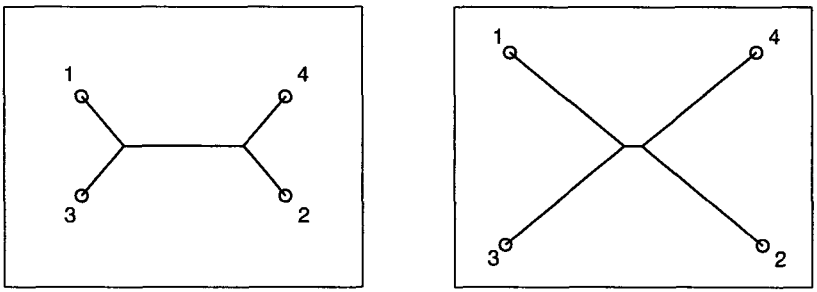


Fig. 15. **Effect of naive parameter estimation.** The left figure shows the correct recombinant tree, corresponding to the recombinant region in the alignment of Fig. 8. The right figure shows the tree that results from a maximum likelihood estimation of the branch lengths from the whole DNA sequence alignment. This includes the flanking regions — shown in white in Fig. 8 — where the recombinant tree topology is incorrect. Obviously, the branch lengths have been significantly distorted, with a contraction of the internal branch and an extension of the external branches. Reprinted from [14], with permission from Mary Ann Liebert.

with respect to the vector of branch lengths $\mathbf{w}$ and the recombination parameter ν. This requires a summation over all state sequences $\mathbf{S} = (S_1, \ldots, S_N)$, that is, over K^N terms. For all but very short sequence lengths N this is intractable. A viable alternative, however, is the expectation maximization (EM) algorithm [3]. Let $Q(\mathbf{S})$ denote an arbitrary probability distribution over the hidden state sequences, and define

$$U(\mathbf{w}, \nu) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln P(\mathcal{D}, \mathbf{S}|\mathbf{w}, \nu) - \sum_{\mathbf{S}} Q(\mathbf{S}) \ln Q(\mathbf{S}). \tag{33}$$

We are interested in the posterior distribution of the hidden state sequences, $P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \nu)$, given the DNA sequence alignment, $\mathcal{D}$, and the parameters, $\mathbf{w}$ and ν. The difference between $Q(\mathbf{S})$ and $P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \nu)$ is measured by the Kullback–Leibler divergence

$$KL(Q, P) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \left(\frac{Q(\mathbf{S})}{P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \nu)} \right), \tag{34}$$

which is always non-negative, and zero if and only if $Q = P$. The proof, which is based on the concavity of the logarithm, is straightforward. Now, combining (33) and (34), we can rewrite the likelihood of (32) as

$$L(\mathbf{w}, \nu) = U(\mathbf{w}, \nu) + KL(Q, P). \tag{35}$$

This decomposition was first suggested in [20], and can easily be proved by recalling that $P(\mathcal{D}, \mathbf{S}|\mathbf{w}, \nu) = P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \nu) P(\mathcal{D}|\mathbf{w}, \nu)$ and $\sum_{\mathbf{S}} Q(\mathbf{S}) = 1$.

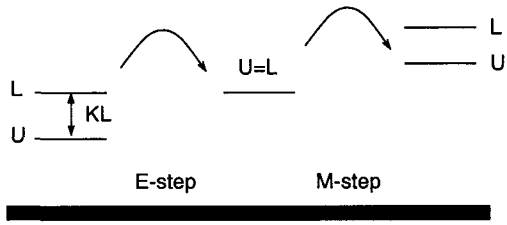


Fig. 16. **Illustration of the EM algorithm.** U is a lower bound on the log likelihood L, with a difference given by the Kullback–Leibler divergence KL. The E-step sets KL to zero. Since the model parameters are kept constant, the log likelihood L is not changed. The M-step adapts the model parameters so as to maximize U. Since U is a lower bound on L, this also increases L.

Since $KL(Q,P)$ is non-negative, U is a lower bound on L: $U(\mathbf{w},\nu) \leq L(\mathbf{w},\nu)$. The EM algorithm alternates between optimizing the distribution over the hidden states $Q(\mathbf{S})$ (the E-step) and optimizing the parameters given $Q(\mathbf{S})$ (the M-step). The E-step holds the parameters fixed and sets Q to the posterior distribution over the hidden states given the parameters, $Q(\mathbf{S}) = P(\mathbf{S}|\mathcal{D},\mathbf{w},\nu)$. This sets $KL(Q,P) = 0$ and, consequently, $L(\mathbf{w},\nu) = U(\mathbf{w},\nu)$. The M-step holds the distribution $Q(\mathbf{S})$ fixed and computes the parameters $\mathbf{w},\nu$ that maximize U. Since $L(\mathbf{w},\nu) = U(\mathbf{w},\nu)$ at the beginning of the M-step, and since the E-step does not affect the model parameters, each EM cycle is guaranteed to increase the likelihood unless the system has already converged to a (local) maximum (or, less likely, a saddle point). An illustration of the algorithm is given in Fig. 16.

Now, similar to the discussion in Sec. 4, we can exploit the sparseness of the connectivity of the underlying graphical model and simplify the maximization of U considerably. From the factorization (25) we have:

$$\begin{aligned} P(\mathcal{D},\mathbf{S}|\mathbf{w},\nu) &= P(\mathbf{y}_1,\ldots,\mathbf{y}_N,S_1,\ldots,S_N|\mathbf{w},\nu) \\ &= \prod_{t=1}^{N} P(\mathbf{y}_t|S_t,\mathbf{w}) \prod_{t=2}^{N} P(S_t|S_{t-1},\nu) P(S_1). \end{aligned} \tag{36}$$

Inserting (36) into (33) gives

$$\begin{aligned} U(\mathbf{w},\nu) = &\sum_{\mathbf{S}} Q(\mathbf{S}) \sum_{t=1}^{N} \ln P(\mathbf{y}_t|S_t,\mathbf{w}) \\ &+ \sum_{\mathbf{S}} Q(\mathbf{S}) \sum_{t=2}^{N} \ln P(S_t|S_{t-1},\nu) + C, \end{aligned} \tag{37}$$

where C is independent of the parameters $\mathbf{w}$ and ν. Equation (37) simplifies considerably. The first term allows immediate marginalization over all but one state S_t in the state sequence $\mathbf{S}$:

$$\sum_{S} Q(\mathbf{S}) \sum_{t=1}^{N} \ln P(\mathbf{y}_t|S_t, \mathbf{w}) = \sum_{t=1}^{N} \sum_{S_t=1}^{K} Q(S_t) \ln P(\mathbf{y}_t|S_t, \mathbf{w}). \tag{38}$$

For the second term, recall the definition of the transition probabilities $P(S_t|S_{t-1}, \nu)$ in (30), define

$$\Psi = \sum_{\mathbf{S}} \sum_{t=2}^{N} Q(\mathbf{S}) \delta(S_t, S_{t-1}) = \sum_{t=2}^{N} \sum_{S_t=1}^{K} Q(S_t, S_{t-1} = S_t) \tag{39}$$

and note that

$$\sum_{\mathbf{S}} \sum_{t=2}^{N} Q(\mathbf{S})[1 - \delta(S_t, S_{t-1})] = N - 1 - \Psi. \tag{40}$$

This gives

$$\sum_{\mathbf{S}} Q(\mathbf{S}) \sum_{t=2}^{N} \ln P(S_t|S_{t-1}, \nu) = \Psi \ln \nu + (N - 1 - \Psi) \ln \left(\frac{1 - \nu}{K - 1} \right). \tag{41}$$

Inserting (38) and (41) into (37), we obtain:

$$\begin{aligned} U = & \sum_{t=1}^{N} \sum_{S_t=1}^{K} Q(S_t) \ln P(\mathbf{y}_t|S_t, \mathbf{w}) + \Psi \ln \nu \\ & + (N - 1 - \Psi) \ln \left(\frac{1 - \nu}{K - 1} \right) + C. \end{aligned} \tag{42}$$

Note that U only depends on the marginal univariate probability $Q(S_t)$, and the marginal two-variate probability $Q(S_t, S_{t-1})$ (via (39)), but no longer on the multivariate joint probability $Q(\mathbf{S})$.

5.3.1. *E-step*

The probabilities $Q(S_t)$ and $Q(S_t, S_{t+1})$ are updated in the E-step, where we set:

$$Q(S_t) \rightarrow P(S_t|\mathcal{D}, \mathbf{w}, \nu) \tag{43}$$

$$Q(S_{t-1}, S_t) \rightarrow P(S_{t-1}, S_t|\mathcal{D}, \mathbf{w}, \nu). \tag{44}$$

These computations are carried out with the *forward-backward* algorithm for HMMs [24], which is a dynamic programming method that reduces the

computational complexity from $O(K^N)$ to $O(N)$. The underlying principle is similar to that of the Viterbi algorithm, discussed in Sec. 4, and is based on the sparseness of the connectivity in the HMM structure. Details are beyond the scope of this chapter, and the interested reader is referred to the tutorial [24], or textbooks like [1] and [4], which also discuss implementation issues.

Now, all that remains to be done is to derive update equations for the parameters $\mathbf{w}$ and ν so as to maximize the function U (M-step).

5.3.2. *M-step: Optimization of the recombination parameter*

Setting the derivative of U with respect to ν to zero, $\frac{\partial U}{\partial \nu} = 0$, we obtain

$$\nu = \frac{\Psi}{N-1}. \tag{45}$$

This optimization is straightforward since, as seen from (39), Ψ only depends on $Q(S_{t-1}, S_t)$, which is obtained by application of the forward-backward algorithm (see above).

5.3.3. *M-step: Optimization of the branch lengths*

Only the first term on the left-hand side of (42) depends on the branch lengths $\mathbf{w}$. This requires a maximization of

$$\sum_{t=1}^{N} \sum_{S_t=1}^{K} Q(S_t) \ln P(\mathbf{y}_t | S_t, \mathbf{w}), \tag{46}$$

which can be achieved with standard phylogenetic programs, like PHYLIP (mentioned in Sec. 2.4). The only modification required is the introduction of a weighting factor $Q(S_t)$ for each site, as illustrated in Fig. 17.

5.3.4. *Reason for not optimizing the prior probabilities*

In principle, U has a further set of parameters that need to be optimized: the $K-1$ prior probabilities $P(S_1)$ (see (36)). Due to the rarity of recombination events, however, a maximum likelihood approach would most probably lead to over-fitting. Also, since DNA sequence alignments are usually sufficiently long, $N \gg K$, the influence of $P(S_1)$ on the mode of $P(S_1, \ldots, S_N | \mathcal{D})$ is negligible. It therefore seems to be reasonable to keep the prior probabilities constant: $P(S_1) = \frac{1}{K} \forall S_1 \in \{1, \ldots, K\}$.

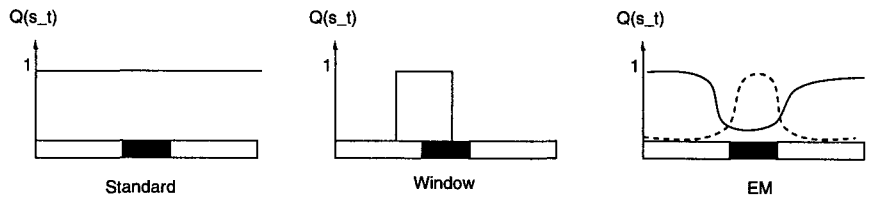


Fig. 17. **Nucleotide weighting schemes.** The figure shows three nucleotide weighting schemes for estimating the branch lengths of the phylogenetic trees. The bottom of each figure represents a multiple DNA sequence alignment with a recombinant zone, printed in grey, in the middle. *Left:* Naive approach, suggested in [18], where the tree parameters are estimated from the whole alignment. This corresponds to constant weights, $Q(S_t) = 1 \; \forall t$. *Middle:* Heuristic window method, also suggested in [18], where the tree parameters are estimated from a subregion of the alignment. The length of this region should be matched to the length of the recombinant region, which, however, is not known in advance. *Right:* Maximum likelihood with the EM algorithm. The dashed line shows the site-dependent weights $Q(S_t = T_R)$ for the recombinant topology T_R, the solid line represents the weights for the non-recombinant topology T_0 : $Q(S_t = T_0)$. Note that in this scheme the weights $Q(S_t)$ are updated automatically in every iteration of the algorithm as a natural consequence of the optimization procedure (E-step). Reprinted from [14], with permission from Mary Ann Liebert.

5.3.5. *Algorithm*

The implementation of the parameter update scheme is straightforward and can be accomplished with the following algorithm:

(1) Initialize the parameters $\mathbf{w}$ and ν. This can be done as in [18], that is, by choosing a plausible recombination rate and by estimating $\mathbf{w}$, for each of the topologies, with a phylogenetic program like DNAML from the whole alignment.
(2) Compute $Q(S_t)$ and $Q(S_{t-1}, S_t)$ with the forward-backward algorithm for HMMs.
(3) Compute Ψ from (39) and adapt ν according to (45).
(4) For $t = 1$ to N: weight the tth column in the multiple sequence alignment, $\mathbf{y}_t$, by $Q(S_t)$, and optimize the branch lengths $\mathbf{w}$ so as to maximize $U(\mathbf{w})$ in (46). This can, in principle, be achieved with a standard phylogeny program, like DNAML of the PHYLIP package [7]. The only change required is the introduction of a weighting scheme for the sites in the alignment.
(5) Test for convergence. If the algorithm has not yet converged, go back to step 2.

Note that this algorithm can be interpreted as a modified version of the Baum–Welch algorithm; see [24].

6. Test Data

The viability of the proposed HMM scheme was tested on the following three DNA sequence alignments.

6.1. *Synthetic data*

DNA sequences, 1000 nucleotides long, were evolved along a 4-species tree, using the Kimura model of nucleotide substitution, which was described in Sec. 2.3. The transition-transversion ratio was set to $\tau = 2$. Two recombination events were simulated by exchanging the indicated lineages, as shown in Fig. 18.

6.2. *Gene conversion in maize*

When looking at the distribution of genes within genomes, one finds that many genes, rather than existing as individual copies, are part of a larger family of related genes called a *multigene family.* A special form of recombination, which takes place in multigene families and contributes greatly to their evolution, is *gene conversion.* This process occurs when the DNA sequence of one gene is replaced (or "converted") by the DNA sequence from another; for further details, see, for instance [21], Chapter 3. Evidence for gene conversion between a pair of maize actin genes (involving Maz56 and Maz63; see below) has been reported in [19]. In the present study, the following four maize sequences were analyzed: Maz56 (GenBank/EMBL accession number U60514), Maz63 (U60513), Maz89 (U60508), and Maz95 (U60507). As discussed in Sec. 2.2, prior to any phylogenetic analysis the DNA sequences need to be aligned. This was done with the program Clustal-W [28], using the default parameter settings and discarding columns with gaps. The three hidden states of the HMM are defined as follows. State 1: ((Maz56,Maz63),(Maz89,Maz95));

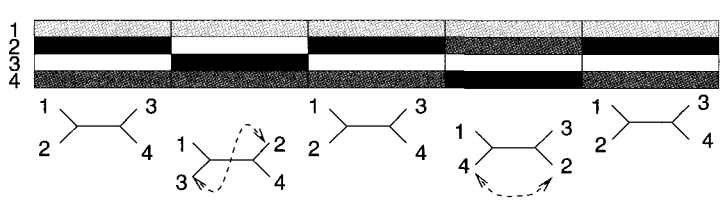


Fig. 18. **Synthetic DNA sequence alignment.** Two recombination events are simulated by swapping the indicated lineages. Defining the predominant tree topology as state 1, the first recombination event corresponds to a transition into state 2, while the second event corresponds to a transition into state 3.

state 2: ((Maz56,Maz89),(Maz63,Maz95)); state 3: ((Maz56,Maz95), (Maz63,Maz89)).

6.3. *Recombination in Neisseria*

One of the first indications for interspecific recombination was found in the bacterial genus *Neisseria* [16]. The analysis in this study was done on a subset of the 787 nucleotide *Neisseria argF* DNA multiple alignment studied in [29], selecting the following four strains: (1) *N. gonorrhoeae* (X64860), (2) *N. meningitidis* (X64866), (3) *N. cinera* (X64869), and (4) *N. mucosa* (X64873) (GenBank/EMBL accession numbers are in brackets). Zhou and Spratt [29] found two anomalous, or more diverged regions in the DNA alignment, which occur at positions $t = 1 - 202$ and $t = 507 - 538$.[c] In the rest of the alignment, *N. meningitidis* clusters with *N. gonorrhoeae* (defined as state 1 in our HMM), while between $t = 1$ and $t = 202$, they found that it is grouped with *N. cinera* (defined as state 3 in our HMM). Zhou and Spratt [29] suggested that the region $t = 507 - 538$ was more diverged as a result of rate variation. An illustration is given in Fig. 19.

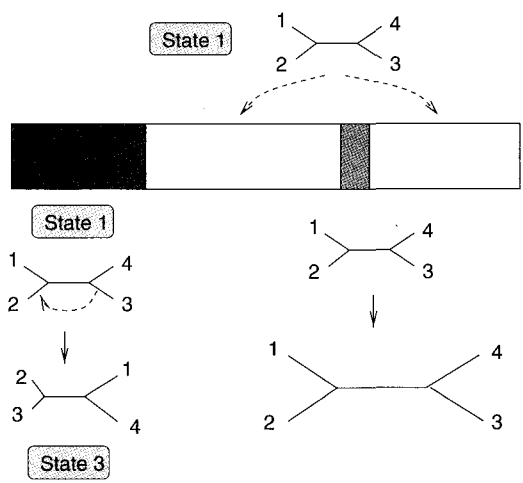


Fig. 19. **Recombination in Neisseria.** According to [29], a recombination event corresponding to a transition from state 1 into state 3 has affected the first 202 nucleotides of the DNA sequence alignment. A second more diverged region seems to be the result of rate variation.

[c]Note that Zhou and Spratt [29] used a different labeling scheme, with the first nucleotide at $t = 296$, and the last one at $t = 1082$.

7. Simulation

Both training schemes, the heuristic method described in Sec. 5.2, and the maximum likelihood approach described in Sec. 5.3, were tested on the three DNA sequence alignments. The application of the heuristic method was similar to [18]. For each of the three possible tree topologies, the branch lengths were estimated separately with maximum likelihood on the whole alignment, using the Kimura model of nucleotide substitution, which was described in Sec. 2.3. The practical computation was carried out with the program DNAML of the PHYLIP package [7]. The transition-transversion ratio τ was optimized with maximum likelihood, using the program package PUZZLE [27]. The recombination parameter was set to $\nu = 0.8$. As opposed to [18], the optimization was not restricted to subsets of the alignments, since the subset size is a parameter that cannot be properly optimized.

The maximum likelihood approach followed the procedure described in Sec. 5.3, optimizing all the parameters simultaneously with the EM algorithm. The initial recombination parameter was set to $\nu = 0.8$, as for the heuristic approach, and the initial probabilities for the three tree topologies were set to equal values: $P(S_1 = 1) = P(S_1 = 2) = P(S_1 = 3) = 1/3$. The EM algorithm typically took about 10–30 EM steps to converge, depending on the data set. Further details can be found in [14].

After parameter estimation, the classification of a site can be based on the mode of the posterior probability $P(S_t|\mathcal{D})$, that is, set $S_t = k$ if $P(S_t = k|\mathcal{D}) \geq P(S_t = i|\mathcal{D}) \forall i \neq k$. A problem of this approach is that even if $S_t = k_t$ maximizes $P(S_t|\mathcal{D})$ for all $t \in \{1, \ldots, N\}$, it is not guaranteed that $(k_1, k_2, \ldots, k_N)$ maximizes $P(S_1, S_2, \ldots, S_N|\mathcal{D})$ [24].[d] Therefore, a better approach is to base the classification of the sites S_t directly on the mode of the joint posterior probability $P(S_1, S_2, \ldots, S_N|\mathcal{D})$, which can be computed with the Viterbi algorithm, described in Sec. 4. However, the deviation between the predictions based on the mode of the marginal posterior probabilities $P(S_t|\mathcal{D})$ and the joint posterior probability $P(S_1, S_2, \ldots, S_N|\mathcal{D})$ was found to be negligible in the simulation studies described here, and the marginal posterior probability $P(S_t|\mathcal{D})$ has the advantage that it can be graphically displayed.

[d] Assume, for instance, that $S_t = k_t$ maximizes $P(S_t|\mathcal{D})$ and $S_{t+1} = k_{t+1}$ maximizes $P(S_{t+1}|\mathcal{D})$, but that $P(S_{t+1} = k_{t+1}|S_t = k_t) = 0$. Then $P(S_t = k_t, S_{t+1} = k_{t+1}|\mathcal{D}) = 0$, so (k_t, k_{t+1}) is not the mode of $P(S_t S_{t+1}|\mathcal{D})$.

This visualization has been done in Figs. 20–22, which show the results obtained with the two training methods on the three DNA sequence alignments. Each figure contains two subfigures: the left subfigure shows the results obtained with the heuristic training scheme, and the right subfigure shows the results obtained with the maximum likelihood scheme. Each subfigure is composed of three graphs. These graphs show the posterior probabilities for the three topologies, $P(S_t = 1|\mathcal{D})$ (top), $P(S_t = 2|\mathcal{D})$ (middle),

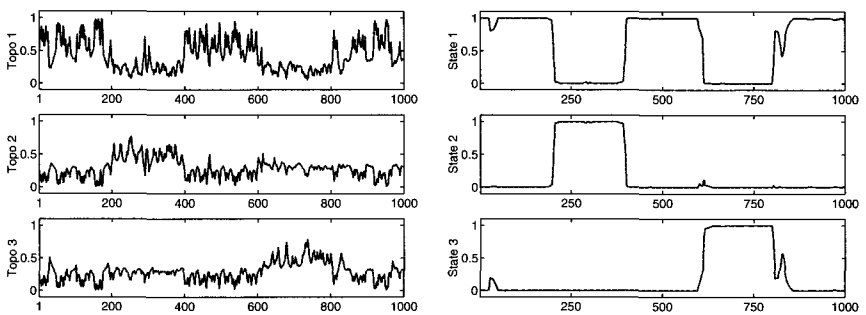


Fig. 20. **Detection of recombination in the synthetic DNA sequence alignment.** The figure contains two subfigures, where each subfigure is composed of three graphs. These graphs show the posterior probabilities for the three topologies, $P(S_t = 1|\mathcal{D})$ (top), $P(S_t = 2|\mathcal{D})$ (middle), $P(S_t = 3|\mathcal{D})$ (bottom), plotted along the DNA sequence alignment (the subscript t denotes the position in the alignment). *Left:* Heuristic training scheme. *Right:* Parameter estimation with maximum likelihood.

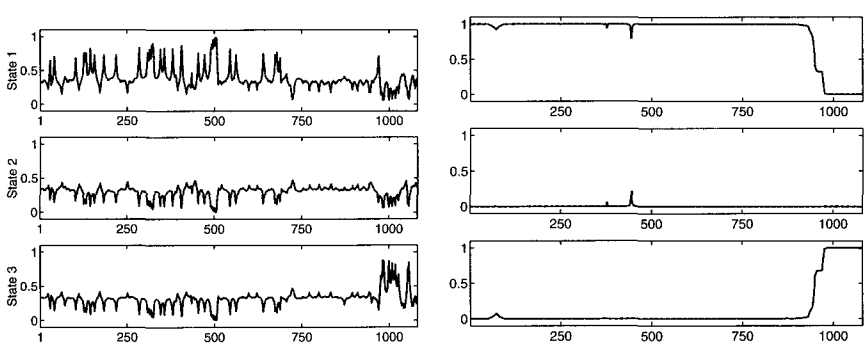


Fig. 21. **Detection of gene conversion between two maize actin genes.** The figure contains two subfigures, where each subfigure is composed of three graphs, as explained in the caption of Figure 20. *Left:* Heuristic training scheme. *Right:* Parameter estimation with maximum likelihood.

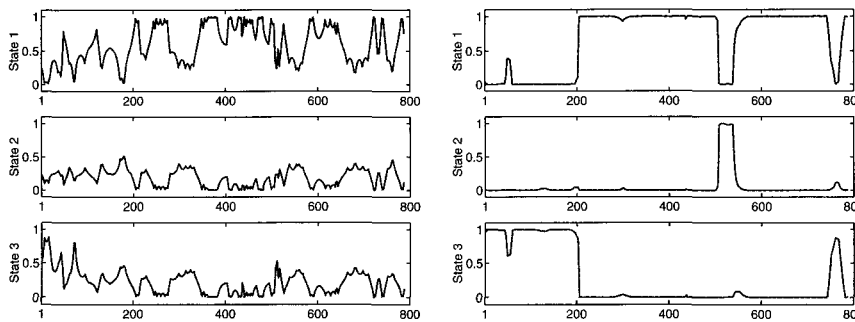


Fig. 22. **Detection of recombination in the Neisseria DNA sequence alignment.** The figure contains two subfigures, where each subfigure is composed of three graphs, as explained in the caption of Fig. 20. *Left:* Heuristic training scheme. *Right:* Parameter estimation with maximum likelihood.

and $P(S_t = 3|\mathcal{D})$ (bottom), plotted along the DNA sequence alignment (recall that the subscript t denotes the position in the alignment). The probabilities are computed with the forward-backward algorithm, which was mentioned in Sec. 5.3, and is discussed at length in [24].

7.1. *Synthetic DNA sequence alignment*

Figure 20 shows the results obtained on the synthetic DNA sequence alignment. For the heuristic training scheme (left subfigure) the overall pattern of the posterior probabilities is correct, showing an increase for state $S_t = 2$ in the region $200 < t < 400$, and an increase for state $S_t = 3$ in the region $600 < t < 800$. However, the signals are very noisy, and an automatic classification based on the mode of the posterior probability would incur a high proportion of erroneously predicted topology changes. This shortcoming is significantly improved as a result of using the maximum likelihood scheme. The predicted state transitions coincide with the true breakpoints, and the tree topologies are predicted correctly. The posterior probabilities for the states, $P(S_t|\mathcal{D})$, are mostly close to zero or one. This indicates a high confidence in the prediction, which is reasonable: since the DNA sequence alignment results from the *simulation* of a recombination process, the transitions between topologies are, in fact, well defined. The estimated recombination parameter is $\nu = 0.992$. With four breakpoints in an alignment of length 1000 nucleotides, the correct value for the recombination parameter is $\nu = 0.996$, which deviates from the prediction by only 0.4%.

7.2. *Gene conversion in maize*

The prediction on the maize DNA sequence alignment is shown in Fig. 21. When using the heuristic parameter estimation method (left), the overall pattern of the graphs $P(S_t|\mathcal{D})$ captures the gene conversion event in that the final section shows a clear increase of the posterior probability for state $S_t = 3$. However, the signals are very noisy and unsuitable for an automatic detection of gene conversion without human intervention. The application of the maximum likelihood scheme leads to a clear improvement: a sharp transition from state $S_t = 1$ to state $S_t = 3$ is predicted in accordance with the gene conversion event found in [19].

7.3. *Recombination in Neisseria*

Figure 22 shows the prediction obtained on the *Neisseria* DNA sequence alignment. The heuristic training method (left) leads to a signal that is very noisy and only gives a vague indication of a topology change at the beginning of the alignment. Estimating the parameters with maximum likelihood leads to a considerable reduction in the noise. A topology change from state $S_t = 3$ to $S_t = 1$ with a breakpoint at site $t = 202$ is predicted, which is in accordance with the findings in [29]. Also, the second anomalous region between sites $t = 507$ and $t = 538$ is clearly detected in that the posterior probability for state 1, $P(S_t = 1|\mathcal{D})$, is significantly decreased, with sharp transitions at the sites predicted in [29]. However, while the HMM predicts a recombination event corresponding to a transition from state 1 into state 2, the findings in [29] suggest that this mosaic segment is more likely the result of rate variation than recombination. This will be discussed in more detail below.

8. Discussion

We have combined two probabilistic models for detecting interspecific recombination in DNA sequence alignments: (1) a taxon graph (phylogenetic tree) representing the relationships among the taxa, and (2) a site graph (HMM) representing which nucleotides interact in determining the tree topology. The parameters of the combined model can be estimated in a maximum likelihood sense with the EM algorithm, and this leads to a significant improvement on an older heuristic parameter estimation scheme. In fact, the simulation study carried out here suggests that recombinant regions can be accurately located, in agreement with the true location

(simulation study) or the location predicted in previous, independent work (maize actin genes, *Neisseria*).

Two limitations of the approach presented here, however, have to be discussed.

Each possible topology constitutes a separate hidden state of the HMM. Now recall, from Sec. 2.1, that for n taxa there are $(2n-5)!!$ different unrooted tree topologies. This implies that the number of states K increases super-exponentially with the number of taxa, which limits our algorithm to alignments of small numbers of taxa. In practical applications, the HMM method is therefore at best combined with a fast low-resolution preprocessing step that can analyze more taxa simultaneously. A useful approach is to conduct the initial search for recombination with split decomposition [2], a method that represents evolutionary relationships among sequences by a network if there are conflicting phylogenetic signals in the data. Split decomposition itself does not allow individual recombination events to be identified nor the statistical support for them to be assessed. It is, however, a useful preprocessing step in that a network that strongly deviates from a bifurcating tree is suggestive of recombination and gives hinds as to which sequences might belong to candidate recombinant strains. This can then be further investigated with the high-resolution method discussed in the present paper.

The second limitation is that the hidden states represent different tree topologies, but do not allow for different rates of evolution. However, if a region has evolved at a drastically different rate, employing a new state for modelling this region might increase the likelihood even though the new state itself — representing a different (wrong) topology — is ill-matched to

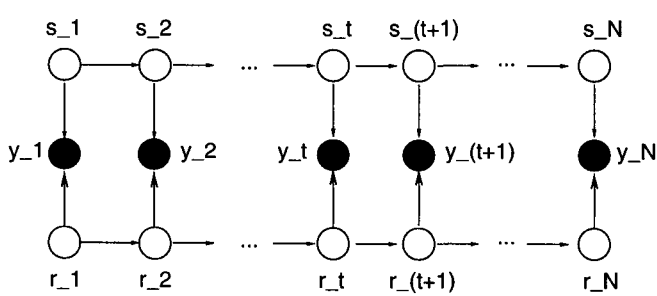


Fig. 23. **Factorial Hidden Markov Model.** In generalization of the standard HMM of Fig. 12, a factorial HMM has two separate families of hidden states: one represents different topologies (S_t), the other represents different evolutionary rates (r_t).

the data. Consequently, a *differently diverged* region might be erroneously classified as *recombinant*, which seems to have happened on the *Neisseria* sequence alignment, as discussed in the previous section. A way to redeem this deficiency is to employ a *factorial hidden Markov model*, shown in Fig. 23, and to introduce two separate hidden states: one representing different topologies, the other representing different evolutionary rates. This effectively combines the method of the present paper with the approach in [5]. A detailed investigation of this idea is the subject of future research.

References

[1] P. Baldi and P. Brunak, *Bioinformatics — The Machine Learning Approach* (MIT Press, Cambridge, MA, 1998).

[2] H. Bandelt and A. W. M. Dress, Split decomposition: A new and useful approach to phylogenetic analysis of distance data, *Molecular Phylogenetics Evolution* **1** (1992) 242–252.

[3] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Stat. Soc.* **B39**(1) (1977) 1–38.

[4] R. Durbin, S. R. Eddy, A. Krogh and G. Mitchison, *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge, UK, 1998).

[5] J. Felsenstein and G. A. Churchill, A hidden Markov model approach to variation among sites in rate of evolution, *Molecular Biology Evolution* **13**(1) (1996) 93–104.

[6] J. Felsenstein, Evolution trees from DNA sequences: A maximum likelihood approach, *J. Molecular Evolution* **17** (1981) 368–376.

[7] J. Felsenstein, Phylip. Free package of programs for inferring phylogenies, available from http://evolution.genetics.washington.edu/phylip.html, 1996.

[8] N. Galtier and M. Gouy, Inferring patterns and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis, *Molecular Biology Evolution* **15**(7) (1998) 871–879.

[9] N. C. Grassly and E. C. Holmes, A likelihood method for the detection of selection and recombination using nucleotide sequences, *Molecular Biology Evolution* **14**(3) (1997) 239–247.

[10] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, 3rd edn. (Oxford University Press, New York, 1985).

[11] M. Hasegawa, H. Kishino and T. Yano, Dating the human-ape splitting by a molecular clock of mitochondrial DNA, *J. Molecular Evolution* **22** (1985) 160–174.

[12] D. Heckerman, A tutorial on learning with Bayesian networks, in *Learning in Graphical Models*, ed. M. I. Jordan, Adaptive Computation and Machine Learning (MIT Press, Cambridge, Massachusetts, 1999), pp. 301–354.

[13] J. Hein, A heuristic method to reconstruct the history of sequences subject to recombination, *J. Molecular Evolution* **36** (1993) 396–405.

[14] D. Husmeier and F. Wright, Detection of recombination in DNA multiple alignments with hidden Markov models, *J. Computational Biology* **8**(4) (2001) 401–427.

[15] M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, *J. Molecular Evolution* **16** (1980) 111–120.

[16] J. Maynard Smith, Analyzing the mosaic structure of genes, *J. Molecular Evolution* **34** (1992) 126–129.

[17] G. McGuire, F. Wright and M. J. Prentice, A graphical method for detecting recombination in phylogenetic data sets, *Molecular Biology Evolution* **14**(11) (1997) 1125–1131.

[18] G. McGuire, F. Wright and M. J. Prentice, A Bayesian method for detecting recombination in DNA multiple alignments, *J. Computational Biology* **7**(1/2) (2000) 159–170.

[19] M. Moniz de Sa and G. Drouin, Phylogeny and substitution rates of angiosperm actin genes, *Molecular Biology Evolution* **13** (1996) 1198–1212.

[20] R. M. Neal and G. E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in *Learning in Graphical Models*, ed. M. I. Jordan (MIT Press, Cambridge, MA, 1999), pp. 355–368.

[21] R. D. M. Page and E. C. Holmes, *Molecular Evolution — A Phylogenetic Approach* (Blackwell Science, Cambridge, UK, 1998).

[22] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd edn. (McGraw-Hill, Singapore, 1991).

[23] D. Posada and K. A. Crandall, Selecting the best-fit model of nucleotide substitution, *Syst. Biology* **50**(4) (2001) 580–601.

[24] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE* **77**(2) (1989) 257–286.

[25] D. L. Robertson, P. M. Sharp, F. E. McCutchan and B. H. Hahn, Recombination in HIV-1, *Nature* **374** (1995) 124–126.

[26] M. O. Salminen, J. K. Carr, D. S. Burke and F. E. McCutchan, Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning, *Aids Res. Human Retroviruses* **11**(11) (1995) 1423–1425.

[27] K. Strimmer and A. von Haeseler, Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies, *Molecular Biology Evolution* **13** (1996) 964–969.

[28] J. D. Thompson, D. G. Higgins and T. J. Gibson, CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice, *Nucleic Acids Res.* **22** (1994) 4673–4680.

[29] J. Zhou and B. G. Spratt, Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: Interspecies recombination within the *argF* gene, *Molecular Microbiology* **6** (1992) 2135–2146.

This page is intentionally left blank

CHAPTER 2

APPLICATION OF STATISTICAL METHODOLOGY AND MODEL DESIGN TO SOCIO-BEHAVIOUR OF HIV TRANSMISSION

JACOB OLUWOYE

The common scientific approaches to the reasoning of problems are mathematical reasoning or statistical reasoning. Mathematical or formal reasoning is mostly deductive, in that, one reasons from general assumptions to specifics using mathematical logic and axioms for multi-criteria decision-making. The purpose of this chapter is to relate statistical methodology and model design to planning and policy making for a viable solution to the ravaging Human Immune Deficiency Virus (HIV)/Acquired Immune Deficiency Syndrome (AIDS). Demonstrating the benefits that can be derived from adapting the concept and model building approach in planning and decision-making of public health and urban development. Discussion in this chapter is presented the following sequence: (a) introduction, (b) deductive/inductive approach, (c) statistical methodology and model design, (d) adaptation of "Seldom Do" models to human behaviour, (e) the discrete choice modelling and its application to the socio-behaviour of HIV transmission. The chapter concludes that the "Seldom Do" model approach offers potential for addressing the development of planning and multi-criteria decision processes associated with health and urban development problems in our society.

Keywords: Deductive/inductive approach; HIV/AIDS behaviour; modelling; sociomedicine; urban development policy.

1. Introduction

The common scientific approaches to the reasoning of problems are mathematical reasoning or statistical reasoning. Mathematical or formal reasoning is mostly deductive, in that one reasons from general assumptions to specifics using mathematical logic and axioms for multi criteria decision-making [1]. Mathematical probability, which is the basis of all statistical theory, had its beginning in ancient times. Certain mathematical patterns were developed as pastimes by the Greeks, and others were first found to coincide with chance happenings, such as occur in card games and later found to coincide with actual happenings. According to [2] and in quote

"mathematical methods are gaining wide acceptance in the study of infectious diseases and putting this powerful tool in the hands of public health community is an extremely important development". It was not until the Seventeenth Century that one of the first practical uses was made of probability when life expectancy tables were published for use in computing life insurance premiums and benefits.

Thus, this chapter will be based on the principles of applied research which attempt to use existing knowledge as an aid to the solution of a given problem or set of problems. When considering the problem of predicting the rate of HIV infection, it is important to factor in geographic dimensions that have been totally ignored due to ignorance and an undue concern for confidentiality.

[3] reported that at a global level, there are considerable differences between regions, states, localities, cities, towns and villages in the levels of prevalence, and rate of transmission, of the Human Immune Deficiency Virus (HIV)/Acquired Immune Deficiency Syndrome (AIDS). Furthermore, there are differences between Regions, States and Localities in the social and demographic characteristics of HIV carriers/AIDS suffers (e.g., the relative proportions of heterosexuals/homosexuals, injecting drugs users, male/female infant HIV carriers) [3]. However, measures to improve health and quality of life in developing countries now need greater attention, together with the need to protect and improve the environment.

It should be noted that diverse and complex environmental health problems cross national boundaries and often need to be dealt with internationally. It is therefore not surprising that large organizations try to pool their efforts in the context of environmental and health policies and research [4].

As AIDS is projected to remain of critical importance in this century, attempts to forecast and predict its developments are urgently needed. A vast amount of literature describing many different aspects of the disease has already been investigated. But as [5] points out: "Rarely can one find an attempt to model the spread of AIDS incorporating the basic spatial dimensions of human existence. Most modelling seems to be focused completely within the temporal domain". One of Kabel's main lines of argument is that modelling the geographical distribution of AIDS can contribute to both educational intervention and the planning of heath care delivery systems.

Medical cartography can play an important role in both areas, as it is an excellent means of communication. In order to be useful to resource planners, predictions of AIDS should include a spatial component.

2. Deductive and Inductive Approach

The detailed methodology, which a researcher should adopt, is a function of the problem, which she/he has set for themselves; different intellectual problems have to be tackled in different ways [6]. Traditionally, however, approaches to problem solving are classified into one of two categories, deductive or inductive. The qualities of each general methodology interact with the intellectual problem, so that sometimes the inductive approach and other times the deductive approach is preferable.

The deductive approach to investigation implies the deduction of a series of events or states from a set of pre-established axioms, and often a comparison of observed phenomena with the deduced events or states (Fig. 1(a)). The inductive approach, in contrast, starts with the observation of a set of phenomena and concludes with attempts to recognise patterns and logical structures in these phenomena, often with some suggestions or conclusions as to their cause (Fig. 1(b)). The former thus starts with a hypothetical cause and then attempts to identify an effect, while the latter observes an effect and then searches for a cause.

If the deductive approach is to be implemented, then a set of axioms must be created or must already be in existence. The deductive approach thus implies some prior knowledge of the problem or of the reasons for the causes in the cause-effect equation. The inductive approach, on the other hand, suggests that investigation proceeds from a state of ignorance. If

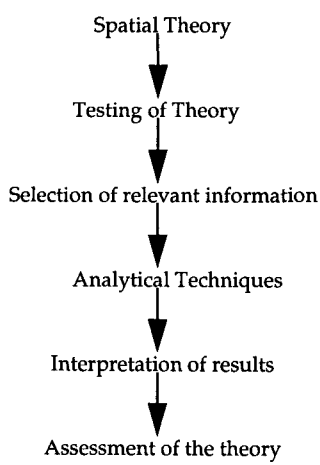


Fig. 1(a). Deductive research methods.

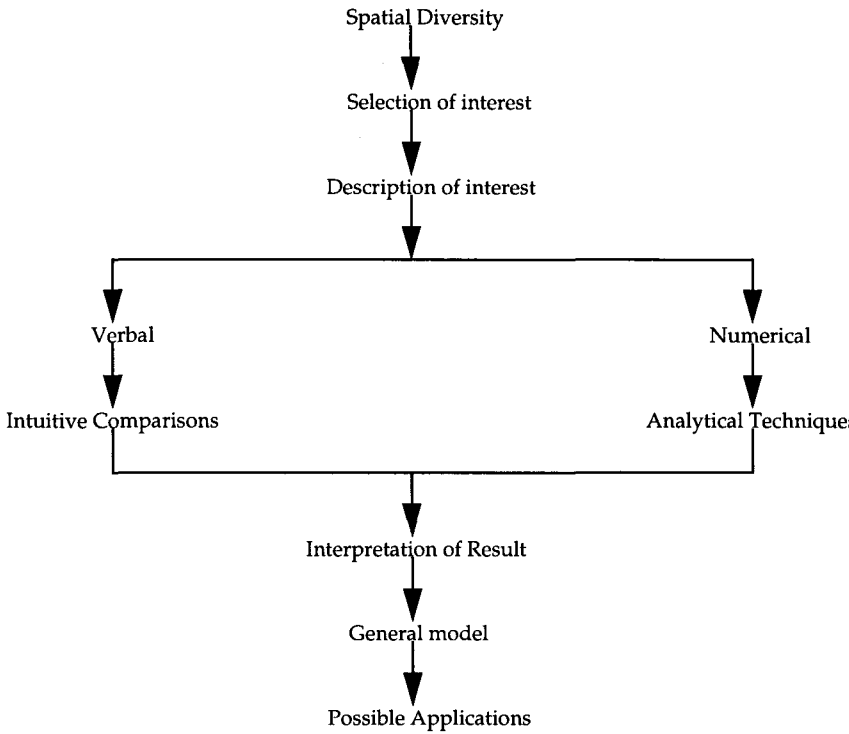


Fig. 1(b). Inductive research methods.

a body of knowledge about a problem already exists then the deductive approach is often adopted. When there is no body of knowledge, which can be built upon or criticised, as the starting point for an investigation, then the inductive approach is to be preferred [5].

3. Statistical Methodology and Model Design

Some scholars claim that the purpose of science is prediction. This is the practical person's viewpoint even when it is endorsed by such scholars as Knight [7]. Neo-Machians (after Ernst Mach) go even further. Just as Mach [8] focused attention on economy of thought without regard for the special role of logical order, they claim that practical success is all that counts; understanding is irrelevant. No doubt if science had no utility for the practical person, who acts on the basis of predictions, scientists (both physical and social) would now be playing their little game only in private clubs. However, even though prediction is the touchstone of scientific

knowledge, "in practice man proves the truth", [9, p. 76] the purpose of science in general is not prediction, but to gather knowledge which can be used as a means of enhancing prediction.

There are two distinct levels of individual analysis: "descriptive" analysis and "predictive" analysis. Both are behavioural in the sense of involving individuals at the aggregate level, but are distinctly different in objective.

The primary distinction between description and prediction can be illustrated diagrammatically by the [10] Schema of Scientific Explanation (see Fig. 2).

The difference between the two is of a pragmatic character. If E has been observed, i.e., an individual behaviour, and a suitable set of statements C1, C2, ..., Ck, L1, L2, ..., Lr is provided afterwards then we have "explanation".[a] If the later statements are given and E is derived from the C's and L's before E is observed, then we have "prediction". For a model to have explanatory power it must, together with other requirements, contain empirical propositions in the explanatory variables which must be confirmed by all available relevant evidence. Such a model would reflect the understanding theoretical constructs of the habit and decision periods.

Mathematical or formal reasoning is mostly deductive in that one reasons from general assumption to specifics using mathematics precision. Models built with this approach are usually larger and more complex than

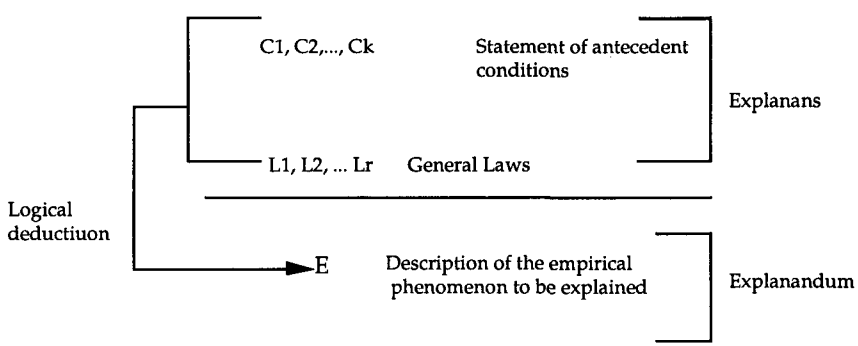


Fig. 2. Schema of scientific explanation.

[a]Given the confirmation of E, certain philosophers give a certain confirmation to statements of antecedent conditions plus general laws. This does not deductively follow from E, but the existence of E gives inductive support to a statement that the initial conditions plus general laws do, in fact, hold.

those developed using statistical reasoning. A mathematical model is developed on the basis of axiomatic assumptions, which are useful, precise, reasonable and concerned with simplicity. Based on these general assumptions, the mathematician reasons a precise mathematical structure, thus establishing a system of relationships referred to as a theory or a model. The properties of the model should be examined and the theory should also evaluate and tested in terms of what it tells the researcher in the light of the consequences of the axiomatic assumptions.

The statistical or factual approach deals directly with empirically derived factual data using inductive reasoning. Data should be collected and attempts should be made to find whether patterns of regularities exist within the data. Besides usually simplifying the data, the three common uses of statistical reasoning are description, induction and hypothesis testing [11].

Description could involve finding and explaining the distribution of some phenomena. Induction aims to establish empirically an association between variables such as a simple correlation, linear or non-linear relationship, and so on. Hypothesis testing involves making a decision to reject or accept a given generalisation within a probabilistic framework. In simple terms, the common uses of such reasoning are either to describe or infer something from an assumption [12].

3.1. *Five steps in model building*

Whichever of the above approaches or combination of approaches is used, model building is likely to involve five steps (Fig. 3).

Step 1. Step 1 is to make assumptions regarding the data. Such assumptions will help reduce or remove possible uses of the shotgun method of problem solving.
Step 2. The second step is to reason based on the prior assumptions. At this step the initial properties and results of the model are examined. Up to this point the model is usually non-operational.
Step 3. Step 3 is making the model operational. This involves assuming functional forms for the relationships. Are they linear, quadratic? Theoretical variables must be defined and a means established to estimate model parameters (i.e., fitting the model).
Step 4. In Step 4 the procedure involves estimating and testing the model by actually plugging data into it. Predictive accuracy is reviewed

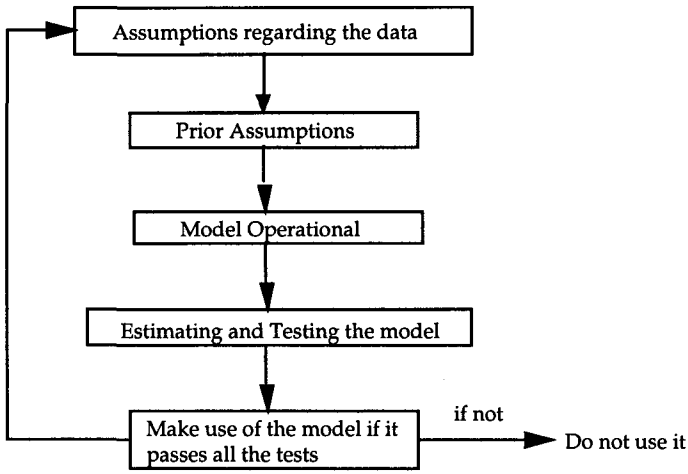


Fig. 3. Five steps in model building.

by analysing trial prediction accuracy. Goodness-of-fit of the whole model is examined and residuals are analysed.

Step 5. Step 5 is the use of the model if it passes all the tests.

3.2. *Model building approach*

What are "Seldom do" models? "Seldom do" models involve all five steps discussed above, in that seldom are they deduced axiomatically and made operational.

However, mathematical models, which are derived from sound reasoning, can be widely used for such purposes (for example the study of polypartnerism) as in Fig. 4. Furthermore, the six dimensions of mathematical models are discussed below.

Understanding or explaining the often complex relationships that exist in the real world are fundamental objectives of mathematical models. Complexities usually arise from the fact that many variables act to produce some reaction (i.e., no single cause) and that there are interdependencies among variables. Variables in mathematical models are usually designated as either endogenous or exogenous variables. In a given problem, exogenous variables are the predetermined variable or independent variables while endogenous variables are determined by exogenous and other endogenous variables.

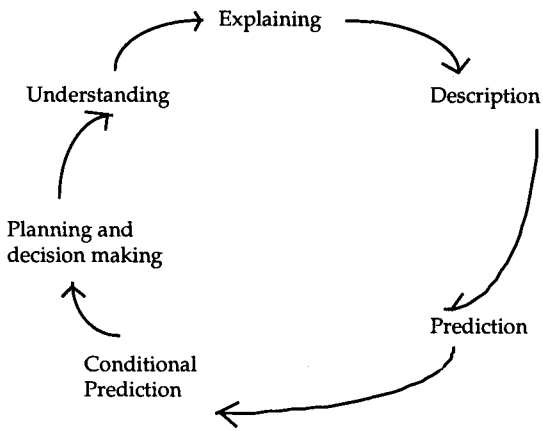


Fig. 4. Six dimensions of mathematical models.

4. Adaptation of "Seldom Do" Models to Human Behaviour

As a simple example, for a particular point in time land-use will affect socio-human behaviour and both in turn will influence trip generation rates.

However, in a temporal sense, trip generation rates will influence land-use which will affect occupation.

It is evident that in developing mathematical models an extremely important assumption is made when determining which variables are endogenous and which are exogenous. This, of course, implies causal relationships and if the assumptions regarding the variables are incorrect, the model will not work, or, at the very least, will be misleading.

In addition to understanding or explaining, models of this type have the additional property of description and can be used to generate data for which the researcher has no measurements. Simply stated:

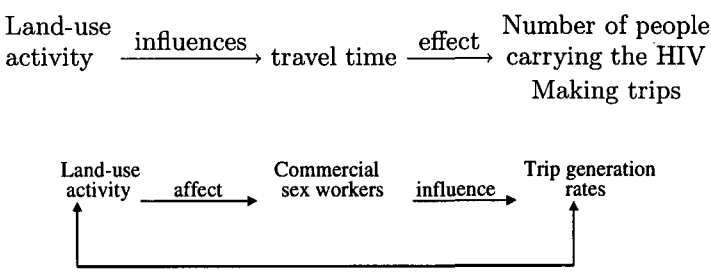


Fig. 5. Land use/commercial sex workers/trip generation interactions.

For example, models employing land use variables (as easily measured exogenous variables) could inferentially describe travel time. This, in turn, could be used to imply something about the relative levels of congestion or conflict existing in a particular area.

The third property is that of prediction. The predictive power of a model builds upon description in that it can predict values of variables for which there are not yet measurements. These predictions can then be used for some future point in time assuming a single future form.

Graphically a predictive model might look like:

Fig. 6. Predictive models.

(Within the model [box] endogenous variables are linked and influence other endogenous variables.)
Set in the simplest terms, for example:

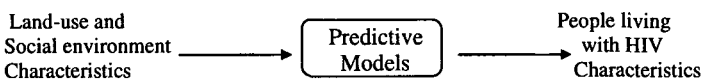


Fig. 7. Predictive models.

Variables must be confirmed by all available relevant evidence. Such a model would reflect the underlying theoretical constructs for the habit and decision periods.

Descriptive analysis looks at "individual drug users", investigating current behaviours in terms of various factors influencing individual drug users' behaviour.

When prediction is considered, some adjustments are required. A changing situation (decision period analysis) can be evaluated in a descriptive framework, just like evaluation of a situation in the behaviour period. Also, prediction can occur under conditions where change is occurring, or where behaviour period conditions exist.

Condition prediction, the fourth property or type of model, deals with alternative future predictions based on conditional assumptions.

The usefulness in planning and public policy-making are quite clear, however the properties of the model become increasingly more complex

in modelling alternative forms of the future. Because of this, some input variables have to be made control or policy/actions variables. This can be illustrated diagrammatically:

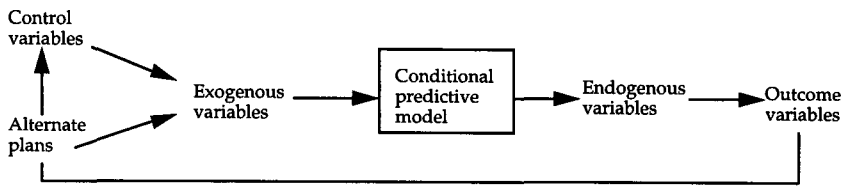


Fig. 8. Policy/Actions model.

The characteristics of the variables are important to the decision-making process. Outcome variables can be any type of variable, control, non-control, exogenous or endogenous. Also alternative plans specify the control and exogenous variables. As an example, set within the context of the epidemiology research, alternative plans might aim to reduce vulnerability of road users (Driver on long distance route) along commercial roads. Control variables might include reducing the number of stops, including educational intervention and policy on long distance road stops. Exogenous variables might include social activities at that location, land use characteristics, and traffic generation. The endogenous variable predicted by the model might be potential risk, in terms of, say, travel time, or it might be the number, and purpose of stops.

If there is only one outcome when testing alternative plans the theory is simple. However, when you have several outcome variables the theory becomes complex, as does decision-making. With more than one alternative and outcome, a criterion or pay-off function probably is necessary. What is needed is a method to evaluate trade-offs, for example, in spread of HIV problem, trade-offs such as number of infected people versus changes in time or other cost/benefit type ratios. Such pay off functions should usually be dictated by the immediate situation as viewed by the local governing body in terms of their variables of interest.

Another form of trade-off is predictive trade-off.

Observations are expected to cluster around the point of indifference. There is a cluster of observations of non-definite behaviour generated. Discriminant analysis can handle this relationship because it is specifically designed to either minimise misclassification with respect to some presumed threshold, or to obtain the greatest separation of the two populations relative to the within-population variance.

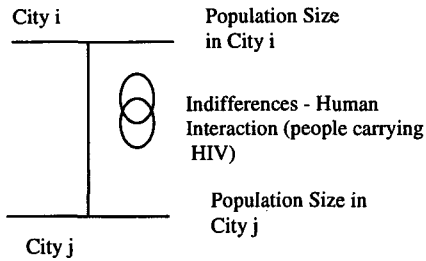


Fig. 9. Predictive trade-off.

The fifth type of model is the decision-making, planning model. This is often referred to as the maximising, minimising or optimising model. It is used to generate alternative plans and optimise the input variables. Linear or non-linear (dynamic) programming methods are usually employed. Graphically, this model might be portrayed:

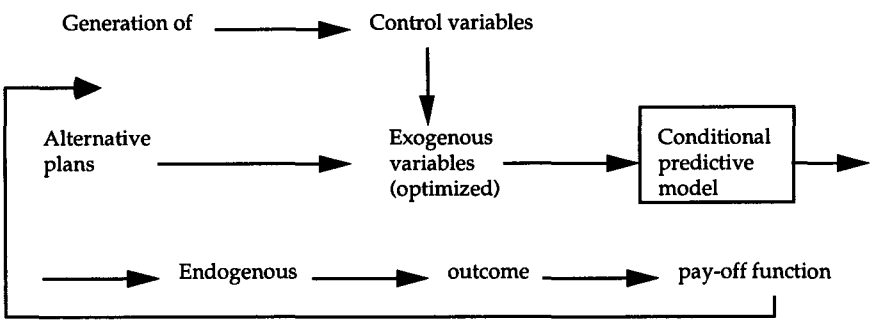


Fig. 10. Decision making and planning model.

This type of model generates alternative plans and optimises the exogenous variables using linear or non-linear programming techniques. Since such techniques are very time-consuming the development of this type of model is usually quite costly.

5. The Discrete Choice Modelling

Discrete choice is a type of regression technique that uses a choice set of mutually exclusive and collectively exhaustive alternatives to describe an outcome.

Looking at different theories of people's behaviour, there seems to be recognition of a conceptual framework with a number of characteristics. In recent years, the interaction-oriented approach has been articulated in the form of systems framework. In its application to people living with HIV, one should point to the definition of a system and the notion of interaction as being most useful in providing a relevant framework. It should be noted here that a system is defined as a set of **elements** having definite attributes, together with the **relationships** between the elements and between their attributes. Since the systems usually exist in some kind of **environment**, one can define the environment as a set of those elements, which do not belong to the system and whose attributes influence the system, or are influenced by it. Finally, each system has a specific **function**, which imposes a defined standard of performance.

Discrete choice modelling is a well-established regression technique that has been used extensively in several disciplines related to psychology, economics, mathematics, and transportation engineering. Several books and papers have been written on this subject [13–18]. However due to reasons outlined below, the author has decided to use the Multinomial Logit (MNL).

The Multinomial Logit (MNL) model can be used to calculate the probabilities of choosing different alternatives in sampling people living with HIV. In the MNL model, individuals are assumed to choose the alternative that yields the highest utility. Some authors [19–21] have emphasised that data derived from binomial counts should be analysed to take into account the binomial denominator, so that the proportion (percentage) of the population already infected can be analysed in order to accommodate the variance while at the same time retaining the binomial probability distribution inherent in the data.

In order to understand the **logit** approach as a representation of an alternative behavioural hypothesis, the author considers the case of a number of alternative outcomes. It is likely that sex trade individuals act to maximise utility (V), and that they constantly evaluate alternative ways of achieving outcomes (s) consistent with this behavioural postulate. An alternative outcome is closer if and only if it provides the highest (indirect) utility.

The discrete/continuous model, the utility of the ith alternative for the qth individual, U_{iq}, should be calculated on the basis of variables affecting the choice of the decision makers (e.g., homosexuals)

$$U_{iq} = \sum_{k=1}^{k} ß_{ik} X_{ikq}. \tag{1}$$

The characteristics are likely to be socioeconomic variables (attributes of the decision-maker). In addition, alternative specific dummy variables can also be used as part of the decision process.

Thus the author decided to use population with only two choice alternatives: one infected population and the other population at risk (e.g., Homosexuals, Bisexuals, IV Drug users, Heterosexuals, Hemophiliacs, Blood transfuses). It should be noted here that the value of an explanatory variable could not be the same for all alternatives. Because of this, one has to set each variable to zero for one alternative. The choice of which variable should be set to zero for each alternative has no effect on the final results of the model, but it naturally alters the form of the utility functions.

Let ρ be the probability that HIV infected population will continue to grow or spread and hence $(1-\rho)$ the probability that population at risk will react to educational intervention; then one may want to apply a linear specification of the form:

$$\frac{\rho}{1-\rho} = a + b\frac{T_1}{T_2} + c\frac{C_1}{C_2} + d_1Q_1 + d_2Q_2 + d_3Q_3. \tag{2}$$

Where T_1 and T_2 are the times (minutes per day for social activities) needed for the two populations, C_1 and C_2 their costs of living (dollars per day), and Q_1, Q_2, and Q_3, population characteristics which are considered to be relevant to the choice (income, family size, age, etc.).

A difficulty with respect to Eq. (2) is that the left-hand side (the probability ρ is constrained to the interval from zero to one, whereas the right-hand side can in principle take arbitrary real values. This defect can be remedied by replacing the left-hand ρ by a more suitable variable, e.g.,

$$\frac{\rho}{1-\rho} = e^{\alpha}\left(\frac{T_1}{T_2}\right)^{\beta}\left(\frac{C_1}{C_2}\right)^{\gamma}\prod_{i=1}^{3}Q_i^{\delta_i} \tag{3}$$

or in logarithmic form:

$$\log\frac{\rho}{1-\rho} = \alpha + \beta\log\left(\frac{T_1}{T_2}\right) + \gamma\log\left(\frac{C_1}{C_2}\right) + \sum_{i=1}^{3}\delta_i\log Q_i. \tag{4}$$

The left-hand variable in Eq. (4) is known as the *logit* corresponding to the probability that people living with HIV/AIDS are dying. The *logit* is monotonically increasing function of the probability ρ varying between $-\infty$ and ∞. Note that it is numerically equal to the logit of the complementary event but of opposite sign:

$$\log\frac{\rho}{1-\rho} = -\log\frac{\rho}{1-\rho}.$$

This implies that the linear logit specification has the convenient property that it is perfectly symmetric in the two alternatives (population infected versus population at risk). If one interchanges the roles of the two alternatives, each term in the equation remains as it is except that it takes the opposite sign. Suppose, however, that we do not have two alternatives but three or more. Then Eq. (4) is not sufficient.

Formalising the problem of choice, one expects the utility provided by the population (infected and at risk) to be a function of individual social-behaviour characteristics, socio-economic characteristics, attraction (land-use) and a disturbance term (i.e., on-street prostitutes, drugs). If one uses a reduced form structure, the utility will also be dependent on the continuous variable, utilisation. Therefore, the unobservable could be characteristics population (infected and at risk) and/or attributes of social activities. This concept, therefore, combines two ideas — the idea of a variation in taste among individuals in a population and the idea of unobserved variables in land-use/urban social behaviour models. These components of the utility function will be denoted by the M-dimensional vector ε, and the utility function will be written $U(x, b, z, s, \varepsilon)$. For the individual infected and at risk, ε is a set of fixed constants (or functions) but for the investigator ε is a random variable with some joint density function, denoted

$$f\varepsilon(\varepsilon_1, \ldots, \varepsilon_m), \text{ which includes a density on } U.$$

Assuming, that the individual infected and at risk has decided to travel for social activity j. Conditional on this decision, his/her utility as a function of x_j and z, the remaining choice variables, is

$$U_j = U(0, \ldots, 0, x_j, 0, \ldots, 0, b_1, \ldots, b_N, z, s, \varepsilon).$$

By virtue of assumption, this conditional direct utility can be written as

$$U_j = U_j(x_j, b_j, z, s, \varepsilon).$$

The population infected and at risk maximises U_j subject to the conditional time constraints.

$$P_j x_j + z = y, \text{ and the non-negatively conditions } x_j \geq 0, z \geq 0.$$

For the purpose of explaining the logit method and its relationships to the choice of HIV infections, one needs to note that the objective is to construct a model to find the probability, ρ, which one can calculate in preference to another aspect of unprotected sex. This probability of choice can be explained in terms of combination of explanatory variables.

It should be noted that the resultant probability function, almost identical to the cumulative normal curve, is a symmetrical sigmoid curve diverging from the normal curve at the extremes only. In developing the model when the regress and (or dependent) variables are dichotomous, and they are on the values 1 (for infected population) and 0 (for non-infected population).

A qualitative dependent variable, such as the binary choice of social behaviour, imposes an automatic restriction on the range of variation of its conditional distribution, constraining the probability of choice to take values between 1 and 0. Discrete choice models are similar in appearance to least square linear regression models (refer to Eq. (5)), but are different in that the value of the discriminant function (P^*) is substitute in Eq. (6). Equation (6) is called the logit function and takes on the values between zero and one. Other functions with similar properties can also be used such as the probit, urban and Geompertz.

$$P^* = a_0 + a_1X_1 + a_2X_2 + \cdots + a_mX_m \tag{5}$$

where, $X_1, X_2, \ldots, X_m$ = independent variables.

$$PPI = e^{P^*}/(1 + e^{P^*}) \tag{6}$$

where, PPI = Probability of proportion of the population already infected
e = exponential function.

Figure 11 shows a graphical representation of the logistic function. Notice that it has as "S" shaped appearance and asymptotically approaches zero and one as the value of P^* approaches negative infinity and positive infinity, respectively.

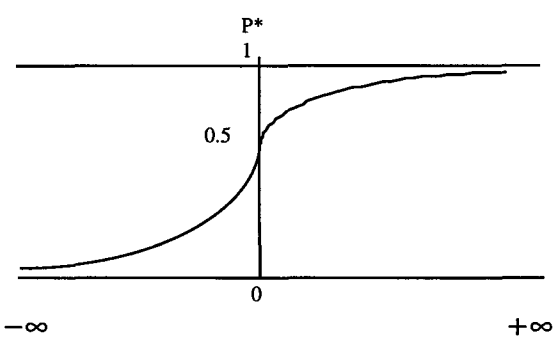


Fig. 11. The logit as a function of the probability.

6. Application of the Use of Logit Specification to Socio-Behaviour of HIV Transmission

To illustrate the use of logit specification, consider the two major factors of transmission of HIV that reported by [3] and [22]. The two major factors are:

(1) The Pool — The Pool comprises the total availability, during a certain period of a transmissible and infectious agent within a particular population — in short the proportion of the population already infected; and
(2) Polypartnerism — Polypartnerism is the number of different contacts with whom an individual engages in unprotected coitus (in the case of HIV, needle/syringe sharing as well as sexual contacts).

Examples of Secs. 6.1–6.3 below were modified based upon G.V. Crockett, *Introduction to Statistical Technique in the Social Sciences*, pp. 124–129.

6.1. *Binary choice models*

As discussed above one can see that, all models contained a dependent or endogenous variable, which was continuous. However, an increasing area of interest is in models in which the dependent variable can take only a limited range of values. For example, one might want to model "Polypartnerist" behaviour where there are only three unprotected coitus; or one might want to study variables influencing the "Pool" — the proportion of the population already infected. In this section the author will concentrate on models where are only two categories — for example, "Sexual Contacts", or "Needle/Syringe Sharing", and so on.

Binary choice models are couched in probabilistic terms; as an example, given a sample of population with HIV in a particular city or country and data on their attributes (age, sex, income etc.), the choice models to be described can predict the likelihood (or the probability) of an individual engages in unprotected coitus on the basis of their individual incomes. That is, other things equal, low-income individuals with HIV more or less likely to engage in needle/syringe sharing than high-income individuals with HIV?

This section looks only at two of a range of possible models, the first being the linear probability model, and the second, the logistic regression model.

6.2. *The linear probability model*

This model is a simple extension of the linear multiple regression models, and takes the form:

$$Y_i = b_0 + X_i + e_i \tag{8}$$

where $Y_i = 1$ if the choice is made in target group for both sexual partners using condoms and $Y_i = 0$ if the choice is made in target group for only sexual partners using condoms;

X_i = the set of attributes (age, sex, social location, distance to location, etc.) e_i = error term.

When Y_i is a dichotomous variable, the regression can be interpreted as describing the probability that an of HIV risk behaviours among an individual engages unprotected coitus, given information about the individual's age, sex, income, etc.

Since Y_i can only take the value one or zero, it is not difficult to show that the variance of the error term is not constant [23, pp. 226–227], and that observations where the probability of choosing to use condom (for example), are close to zero or close to one, will have relatively low variances, while observations where the probability is close to 0.5 will have high variances. As discussed in Sec. 5 above, this characteristic of the error term not displaying a constant variance (called heteroscedasticity) results in a loss of efficiency, nevertheless, the parameter estimates are still unbiased and consistent.

Consider the following problem:

A sample of hypothetical data on age versus whether an individual engages unprotected coitus HIV risk behaviours. The model then is:

$$Y = b_0 + b_1 \text{ Age} + e \tag{9}$$

where, $Y = \begin{cases} 1 & \text{If both sexual partners used condoms} \\ 0 & \text{If one sexual partner used condoms} \end{cases}$

and AGE = the age in years of HIV risk behaviours among an individual engages unprotected coitus by fitting an OLS regression to the data, resulting in the following fit:

$$\hat{Y} = -0.23 + 0.025 \text{ Age}. \tag{10}$$

Since $\hat{Y}$ is interpreted as the probability of both sexual partners used condoms, Eq. (10) shows that as the age of the individual with HIV increases, the probability of both sexual partners used condoms increases,

while conversely the probability of one sexual partner used condoms decreases.

Hence, if Age = 20,

$$\begin{aligned}\hat{Y} &= \text{ Probability of both sexual partners used condoms} \\ &= -0.23 + (0.025 * 20) \\ &= 0.27\end{aligned}$$

If, however, Age = 40

$$\begin{aligned}\hat{Y} &= \text{ Probability of both sexual partners used condoms} \\ &= -0.23 + (0.025 * 40) \\ &= 0.77\end{aligned}$$

The fitted Eq. (11) is depicted in Fig. 12 below.

6.3. *The logit model*

An obvious problem arises with Eq. (10), Fig. 12 shows that $\hat{Y}$ can exceed 1, or be less than zero, both impossibilities.

One way of avoiding this would be to set all $\hat{Y}$ greater than 1, equal to 1, and all $\hat{Y}$ less than 0, equal to 0 forming the flattened Z shape ABCD. This specification, however, is still not satisfactory in that, being linear; it suggests that equal changes in age result in the same change in probability, regardless of the age. Thus a unit increase in age of usage of condoms is

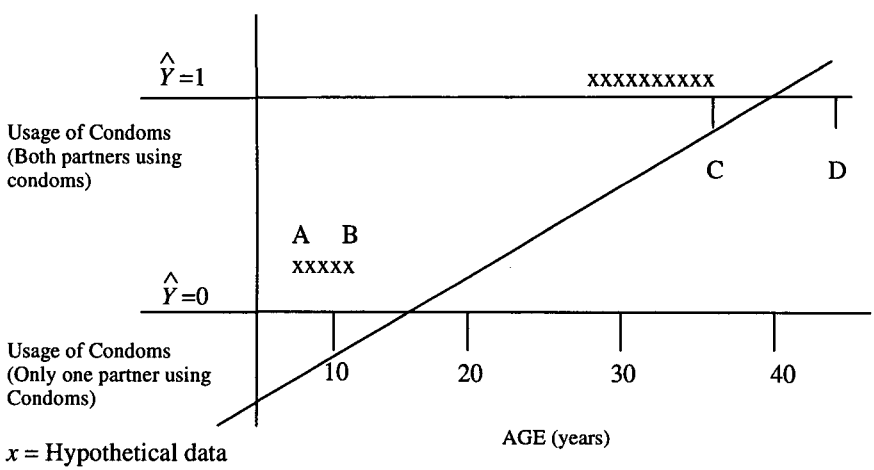


Fig. 12. Usage of condoms versus age.

predicted to cause a 0.025 increase in the probability of both partners using condoms, regardless of whether the sexual behaviours of individual is in his or her early 20's, or in his or her late 50's. It is much more likely that a constant change in age will produce a relative change in $\hat{Y}$; i.e., the upper and lower tails would more likely be much flatter than the linear function. At the lower end of the Y values, this flattening can be obtained by fitting the log of Y versus Age:

$$\log \hat{Y} = \hat{b}_0 + \hat{b}_1 \text{ Age}.$$

Similarly, at the high end of Y values (i.e., as Y approaches 1), the flattening can be achieved by taking the log of $(1 - Y)$. Combining both ends of the scale, results in the model:

$$\log \hat{Y} - \log(1 - \hat{Y}) = \hat{b}_0 + \hat{b}_1 \text{ Age} \tag{11}$$

or,

$$\text{more generally, } \log(P/(1 - \hat{P})) = \hat{b}_0 + \hat{b}_1 X.$$

Solving Eq. (11) for P gives:

$$P = \frac{1}{1 + e^{-(b_0 + b_1 X)}}$$

which is called a logistic curve, as graphed in Fig. 13

Using application of the previous authors it would appear that Eq. (11) could be easily estimated by OLS with a suitable log transformation of the dependent variable. This is not possible, however, because if $P = 1$, the

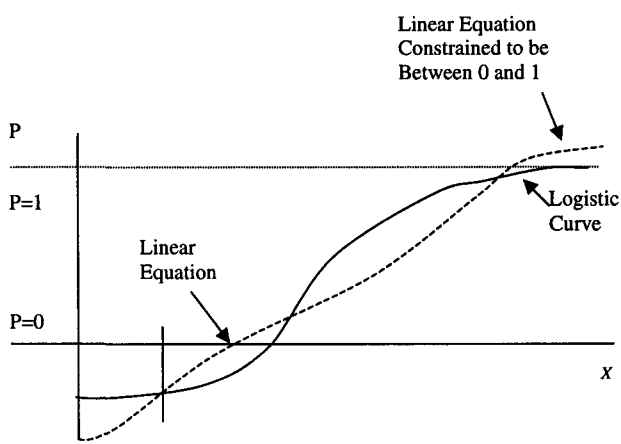


Fig. 13. Graph of the logistic curve.

expression $P/(1-P)$ is infinitely large, while if $P = 0$, the expression equals zero; hence the logarithm in each case would be undefined. Fortunately, a method called maximum likelihood estimation (MLE) is available which can cope with this problem; briefly, in this method, different values of $\hat{b}_0$ and $\hat{b}_1$ are tried until those values of b_0 and b_1 are discovered which maximize the likelihood of their having come from the sample of (X, Y) values given. The parameter values that make this likelihood largest are therefore called the maximum likelihood estimator's $\hat{b}_0$ and $\hat{b}_1$. The computer using an iterative procedure finds them; i.e., the computer keeps trying different values until it converges on the maximum likelihood values. Hence the user should be aware that this procedure is likely to be quite costly in terms of computer time, and accordingly limit the number of independent variables, or use random samples of the original data in order to limit the number of independent variables when the full data set is used.

7. Conclusion

This paper has discussed the major five steps of a "Seldom do" model for the purpose of improving multi-criteria decision making for health and urban problems, in which the role of models and other statistical tools of analysis in the information system are important. In particular, we have noted the relationship between variables which have considerable value to the planner in the understanding and development of planning and decision processes.

The planners must also know how such variables change over time and the way they respond to intervention. Models of health, urban development and other statistical techniques structure these relationships between the major variables; helping planners to analyse urban and environmental health problems. They are therefore very useful tools in understanding the complexity of the health (HIV/AIDS preventive) activities and urban development policies. Data banks and models are therefore very much interconnected within the urban information system.

8. General Comments

The discussion of model types and model building was based in part on courses (i.e., quantitative methods, statistical methods, methods of sociological inquiry, etc.) taken by the author at the University of Wisconsin-Madison, USA; Howard University, Washington, D.C.; and Independent Research (Ph.D.) at the UNSW Kensington, Australia.

Acknowledgments

This chapter by design is for original statistical application and concept. The author hereby acknowledges the in-depth ideas of the previous authors mentioned in the chapter and other researchers in the field of the Built Environment and Public Health. A great deal of debt is owed to David Hensher, Lester Johnson, GV Crockett and Moshe Ben-Akiva.

References

[1] J. O. Oluwoye, 'Seldom-Do' models approach for multiple criteria decision-making in environmental design management, in *13th International Conference on Multiple Criteria Decision Making (MCDM)*, 6–10 Jan (1997), p. 43.

[2] R. M. Anderson, in *NIHM/ASIST "AIDS Strategic Intervention Simulation Tool. HIV/STD Prevention and Translational Research Center for Mental Health Research on AIDS, NIH Verson 0.5B* (2000).

[3] N. Ford and S. Koetsawang, The socio-cultural context of the transmission of HIV in Thailand, *Soc. Sci. Med.* **33**(4) (1991) 405–414.

[4] J. O. Oluwoye, Spatial interaction and social-behaviour of HIV transmission, in *Applicable Mathematics — Its Perspective's and Challenges*, ed. J. C. Misra (Narosa Publishing House New Delhi, India, 2001), pp. 399–406.

[5] R. Kabel, AIDS: Predicting the next map, *Interfaces* **21**(3) (May–June, 1990) 80–92.

[6] J. O. Oluwoye, *Research Methods & Statistical Methodology/Model Design — A "Teach Yourself" Guide* (School of Building Studies, UTS, 1992).

[7] F. Knight, *The Ethics of Competition* (Harper and Brothers, New York, 1935), pp. 109–110.

[8] E. Mach, *Popular Scientific Lectures* (Chicago, 1895), p. 195.

[9] F. Engels, *Quotation of Marx*, Ludwig Feuenbach and the Outcome of Classical Germal Philosophy, London (1947), p. 76.

[10] C. G. Hempel and P. Oppenheim, *Studies in the Logic of Explanation* (Philosophy of Science, 1948), p. 15.

[11] J. O. Oluwoye, Statistical applications in real estate appraisal, *Austral. J. Property Res.* **1** (1991) 72–79.

[12] J. O. Oluwoye, Assessment of pedestrian crossing activity in the determination of reducing conflict between pedestrians and vehicles along a strip of commercial streets in Nigeria, unpublished PhD Thesis (UNSW, 1988).

[13] M. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Applications to Travel Demand* (MIT Press, Cambridge, MA, 1985).

[14] D. A. Henhser and L. W. Johnson, *Applied Discrete Choice Modelling* (Groom Helm Ltd., London, 1981).

[15] K. Train, *Qualitative Choice Analysis: Theory, Econometrics, and An Application to Automobile Demand* (MIT Press, 1986).

[16] H. Rintamaki, Paakanpunkiseudun Liikennetutkimus L1Tu-T6 otaniemi Tekmillinen Korkeakoulu, Liikennetekniika, *Julkaisu* **50** (1980).

[17] P. R. Stopher, E. G. Ohstrom, K. D. Kalten back and D. L. Clause, Logit mode — choice models for nonwork trips, *Transpn. Res. Rec.* **987** (1984) 75–81.
[18] J. H. Aldrich and F. D. Nelson, *Linear Probability, Logit and Probit Models*, Quantitative Applications in the Social Sciences (A Sage University Paper, Beverly Hills, 1984).
[19] D. R. Cox, *The Analysis of Binary Data* (Menthema Co. Ltd., Lougon, 1970), p. 12.
[20] A. J. Dobson, *Introduction to Statistical Modelling* (Chapman and Hall, 1983).
[21] P. McCullagh and J. A. Nelder, *Generalised Linear Models*, Monographs on Statistics and Applied Probability (Chapman and Hall, 1983).
[22] A. J. Nahmias, F. Lee and D. Danielsson, Epidemiological principles for understanding the prevalence of HIV infections and their possible control. In *AIDS in Children, Adolescents and Heterosexual Adults*, eds. R. F. Schinazi and A. J. Namias (Elsevier, New York, 1988), p. 154.
[23] R. Pindyk and D. Rubinfield, *Econometric models and Economic Forecasts*, 2nd edn. (McGraw-Hill, 1981).

Bibliography

[1] D. McFaddenn, Econometric models of probabilistic choice, in eds. C. F. Manski and B. McFadden, *Structural Analysis of Discrete Data with Econometric Applications* (MIT Press, Cambridge, 1981).
[2] W. H. Greene, *Econometric Analysis* (Macmillian Publishing Company, New York, 1990).
[3] J. Berkson, Application of the logistic function of Bio-Assay, *J. Amer. Stat. Assoc.* **39** (1944) 357–365.
[4] An approach to developing transport improvement proposals. Bureau of Transport Economics Occasional Paper 24 (Australian Government Publishing Service Canberra, 1978).
[5] G. V. Crockett, *Introduction to Statistical Techniques in the Social Sciences* (Quoll Enterprises, Western Australia, 1988).

CHAPTER 3

A STOCHASTIC MODEL INCORPORATING HIV TREATMENTS FOR A HETEROSEXUAL POPULATION: IMPACT ON THRESHOLD CONDITIONS

ROBERT J. GALLOP, CHARLES J. MODE and CANDACE K. SLEEMAN

During recent years the medical community has been aggressively searching for a cure of the HIV disease, but so far a cure has not been found. Consequently, the goal of HIV/AIDS treatments has been to impact the health of infected individual and to extend their life-expectancy with the hope that in time the medical community may find a cure. Much success has been reported in improving the life-span of infected individuals. How this increased life-span for infected individuals effects the overall impact of the spread of the disease remains unknown. Thus, while success of HIV treatments, such as the HAART therapy, is beneficial in the infected population, consequences must be considered for the susceptible audience. The investigation described in this article will focus on the heterosexual population; therefore, a thorough investigation must consider the multiple facets present in the heterosexual population. A stochastic model for the heterosexual population with sufficient parameters to model the multiple facets impacting the spread of the disease in this audience is considered. Proximity to threshold conditions specify when the disease will spread if a small number of infected people are introduced into a susceptible population. Determination of distance to the threshold condition for a population participating in HIV treatments and a population without HIV treatments is investigated. Simulations suggest that from the perspective of only successful implementation of HIV treatments to this audience, a more rapid spread of the disease throughout the susceptible sector occurs. The merits of HIV treatments is not in question but, in this era of more advances in HIV treatments, HIV treatments must be coupled with attention to general awareness and further education and prevention efforts, for proper control of the epidemics spread.

1. Introduction

Since public recognition of the acquired immunodeficiency syndrome (AIDS) epidemic during the mid 1980's and the causal agent of AIDS, human immunodeficiency virus (HIV), a great amount of effort has been devoted to constructing and analyzing statistical models of the epidemic. When focus is placed solely on heterosexual transmission, certain facets

of the disease transmission and audience characteristics must be considered. As discussed by Dietz and Hadeler [5], if two susceptible individuals form a marital couple, then they can be considered temporarily immune as long as they do not separate and have no sexual contacts outside the couple. While understanding the idiosyncrasies of the heterosexual community and that there is a need to model the heterosexual spread of HIV/AIDS, questions arise as how to model the epidemic. Recently, Mode and Sleeman [13] formulated a modeling structure which incorporates key facets of the heterosexual population such as couple formation, couple dissolution, selectivity of partners in couple formations, and selectivity of partners for extra-marital contacts. It is formulated in a stochastic framework where semi-Markovian life cycle models for single females, single males, and couples are outlined based on the theory of competing risks, where the disease may progress among stages of severity. Much attention has been given to the effect of HIV/AIDS treatments such as the highly active antiretroviral therapy (HAART) treatment of protease inhibitors [2, 4, 7, 8], which illustrate subjects may move among the stages of the disease with both improvement and deterioration possible. The Mode–Sleeman model accounts for both possible transitions among stages of disease.

By operating on conditional expectations of the present, given the past, deterministic models, expressed as non-linear difference equations, may be embedded in the stochastic process. By letting the time increment approach zero, the embedded non-linear difference equations give rise to a system of differential equations. As will be illustrated in examples, by exploiting the stability properties of this embedded system of differential equations, it is possible to provide insights on threshold conditions as to whether an epidemic spreads in the population according to the stochastic model.

The concept of threshold conditions is one of the most important concepts in mathematical epidemiology [10], and is used to specify conditions in terms of the parameters such that the disease will spread if a small number of infected people are introduced into a large susceptible population. As discussed by Hyman *et al.* [11], an analysis of the stability of the infection-free equilibrium gives rise to an epidemic threshold condition. Briefly, a stability analysis consists of linearizing the embedded differential equations around the infection-free equilibrium and determining when the largest real part of the eigenvalues crosses zero gives rise to a threshold condition. When the value is positive, it indicates the introduction of a few infectives into a susceptible population will result in a spread of the disease throughout the susceptible population with positive probability. When

negative, it indicates the susceptible population is resilient to the introduction of a few infectives; therefore, a minor epidemic may develop but will eventually become extinct with positive probability. When zero, it indicates a threshold condition based on the multi-dimensional parameter space has been met. The magnitude of the largest real part of the eigenvalues indicates the rate of spread, if positive, or the rate of restoration to an infection-free system, if negative. For two systems with the same sign for the largest real part of the eigenvalues, the system with the larger magnitude will have the quicker rate of infection, if positive, or quicker rate of restoration to an infection-free system, if negative.

The main focus of this article will be to illustrate the effect of HIV/AIDS treatments on the spread of the disease in a susceptible population with the introduction of a few infectives. To illustrate this point, threshold investigations for a system with HIV/AIDS treatments and a system without HIV/AIDS treatments will be compared. To enhance the understanding of the epidemic process, fifty Monte Carlo realizations of the stochastic processes will be computed on monthly time intervals of 720 months. Monte Carlo samples will be summarized statistically on a monthly basis, by derivation of the minimum, maximum, 25th quantile, 50th quantile, and 75th quantile. To provide a basis of comparison of the systems, computer generated graphs of the two systems will be simultaneously compared. The coupling of advances in education and prevention efforts and HIV treatments will also be investigated and comparisons to the other systems will be made.

2. Parameters for a Heterosexual Population

The heterosexual population is partitioned into three groups: X, single females, Y, single males, and Z, coupled individuals. Each group is partitioned in time and severity of the disease. Stages of the disease are expressed in terms of $CD4^+$ counts, with stages represented by intervals of $CD4^+$ counts, with higher stages representing more severe immune deficiency. By definition, stage 0 indicates an individual is susceptible. Thus, at time t, $X(t;i)$ represents the number of single females in stage i at time t, $Y(t;j)$ represents the number single males in stage j at time t; and $Z(t;i,j)$ represents the number of couples with a female in stage i and a male in stage j at time t. In order to formulate a model which realistically captures the multiple facets of the heterosexual population, parameters representing the characteristics of the disease infection and progression and population transition must be defined. For disease infection, parameters are needed to describe

extra-marital infection and intra-marital infection. For disease progression, parameters are needed to describe the declination of $CD4^+$ counts and the potential reconstitution of $CD4^+$ counts through effective HIV treatments. For population transitions, parameters are needed to describe recruitment, couple formation, couple dissolution, and deaths.

Focusing first on the infection of a susceptible individual, one recognizes that infection can occur through sexual contact with an infected individual. Potentially unsafe sexual contact may involve all individuals regardless of marital status. We will classify all sexual contacts outside wedlock as extra-marital sexual contacts.

The probability a susceptible is infected within a couple during a given time period, depends on the expected number of marital sexual contacts per unit time, γ_{mc}, and the probability the susceptible is infected per marital contact. Let $q_{fm}(k)$ denote the probability a susceptible female is infected per marital contact when her partner is in stage k of the disease. The parameter $q_{mm}(k)$ is defined similarly for males. As discussed by Hyman *et al.* [11], the probability of infection per sexual contact may be differ across gender and severity of the disease; therefore, as illustrated above, a model must accommodate this characteristic of disease transmission.

The probability a susceptible is infected through extra-marital contacts during a given period with a given partner depends on the expected number of sexual contacts per unit time, η_f for females and η_m for males, and the probability the susceptible is infected per extra-marital contact, $q_{fem}(k)$ for females and $q_{mem}(k)$ for males. Unlike the marital case, where there is only one person with whom a susceptible engages in sexual activity, in the extra-marital case, there may be multiple partners with whom a susceptible can engage in extra-marital sexual contacts. Let the parameters λ_f and λ_m denote the expected number of extra-marital sexual partners per unit time for females and males respectively.

While there are many social-demographic variables which may influence choice of partners for sexual activity, our attention will be on the impact of HIV/AIDS awareness on choice of partner. During the later part of the twentieth century more people became aware of HIV/AIDS as a heterosexually transmitted disease with no discrimination as to age, race, health, or sexual preference [15]. Through increased public awareness and education, it seems reasonable to suppose that individuals may possess more skills in screening of potential partners prior to the initiation of sexual intimacy. Screening skills may consist of individuals asking about potential partner's past sexual history, drug use, and previous HIV test

results. Non-negative parameters quantifying an individual's inclination to accept a potential partner for extra-marital sexual contacts will be denoted by β_{fem} for females and β_{mem} for males. Given that a female is in stage i of disease, the generic form of the acceptance probabilities used in this article are $\alpha(i,j) = \exp(-\beta|i-j|)$ denote the conditional probability that she finds a male in stage j acceptable as a sexual partner. Observe that the larger the value of β_{fem}, the smaller is the probability that a female in stage i of disease will find a male acceptable as a sexual partner. Beta values of zero indicate individuals randomly select partners with no caution or screening for HIV. A more comprehensive discussion of acceptance probabilities may be found in the Mode–Sleeman [14], including a discussion of other functional forms and extensions to higher dimensions. In the next section, formulas showing how these probabilities enter into the formulation will be given. Turning to population transitions, the mortality rate parameters per unit time will be denoted by μ_{f0} for females and μ_{m0} for males. Incremental change in death rates due to stages of disease will be defined as follows: let μ_{fk} denote the incremental change in risk of death for females in stage k of the disease and define the parameter μ_{mk} similarly for males in stage k of disease. Parameters accounting for transition to the next more severe stage of disease and the next less severe stage of the disease must be present in the formulation. When an infective is in stage 1, there is no transition to susceptible and when an infective is in the final stage of the disease, there is no transition to the next more severe stage of the disease. The parameter $\gamma_f(k,k+1)$ is the risk of the transition $k \to k+1$ for an infected female per unit time in stage $k = 1,2,\ldots,n-1$ of disease, and the parameter $\gamma_m(k,k+1)$ has the same interpretation for a male in stage k. The parameter $\gamma_f(k,k-1)$ is the risk per unit time of the transition $k \to k-1$ for an infected female in stage $k = 2,\ldots,n$, and the parameter $\gamma_m(k,k-1)$ has a similar interpretation for males.

Similar to the extra-marital contacts, singles may enter into a marital status, where choice of partner is from a collection of non-married individuals. Acceptance parameters for choice of marital partner are defined as β_{fm} for females and β_{mm} for males. Rate of couple formation is given as ρ. Rate of couple dissolution is given as δ.

Recruits may enter the population over time. Recruits are often thought of as adolescents reaching the age of sexual activity. The parameter μ_f denotes the expected number of single females into the population per unit time and $\rho_f(j)$ is the probability a recruit is of type $j = 0,1,2,\ldots,n$. Similar definitions are made for single males entering the population.

3. Latent Risks for Transitions in Population

Having defined the parameters of the model, the next step is to set down those functions that appear in the stochastic population process. Let $\Theta_f(f) = (\theta_f(t;j,k))$ denote a matrix of latent risks for life cycle model of single females. Later in this section, this matrix, as well as other matrices of latent risks, will be defined explicitly in terms of the parameters of the model. A function of basic importance in studying discrete time approximations to the life cycle models in continuous time is the following conditional probability, which arises in computing Monte Carlo realizations of the process, using chains of multinomial distributions. Given that a female is in stage j at time t, let $\pi_f(t;j,k;h)$ be the conditional probability there is a jump to stage $k \neq j$ during the time interval $(t, t+h]$. Then, it can be shown by using the classical theory of competing risks that

$$\pi_f(t;j,k;h) = (1 - \exp[-\theta_f(t;j)h]) \frac{\theta_f(t;j,k)}{\theta_f(t;j)} \tag{1}$$

where

$$\theta_f(t;j) = \sum_k \theta_f(t;j,k)$$

is the total latent risk for a transition from stage j at time t, and $\theta_f(t;j,k)$ is the latent risk for a transition from j to k at time t. It follows that

$$\pi_f(t;j,j;h) = 1 - \sum_{k \neq j} \pi_f(t;j,k;h) = \exp[-\theta_f(t;j)h] \tag{2}$$

is the conditional probability that there is no transition from stage j during $(t, t+h]$, given the process was in stage j at time t.

Most latent risks are constant over time with the exception of the those for couple formation and extra-marital sexual contacts, which depend on the stage of the population at some time $\gamma_{fm}(t;j,k)$ denote the conditional probability a single female in stage j finds a single male in stage k acceptable for matrimony at time t. Then, in the stochastic component of the model, it can be shown by applying the total law of probability and Bayes' formula that

$$\gamma_{fm}(t;j,k) = \frac{Y(t;k)\alpha_{fm}(j,k)}{\sum_{k=0}^{n} Y(t;k)\alpha_{fm}(j,k)}. \tag{3}$$

A similar formula for the conditional probability $\gamma_{mm}(t;k,j)$ that a single male in stage k finds a single female in stage j acceptable for matrimony can be derived by substituting X's for Y's and α_{mm} for α_{fm}. Mode and

Sleeman [14] may be consulted for further details on the derivation of this formula.

Let $\gamma_{fem}(t;j,k)$ denote the conditional probability a female in stage j finds a male in stage k acceptable for extra-marital sexual contact at time t. Then, by using a similar argument, it can be shown that

$$\gamma_{fem}(t;j,k) = \frac{Y_T(t;k)\alpha_{fem}(j,k)}{\sum_{k=0}^{n} Y_T(t;k)\alpha_{fem}(j,k)} \tag{4}$$

where

$$Y_T(t;k) = Y(t;k) + \sum_{j=0}^{n} Z(t;j,k) \tag{5}$$

is the total number of males in stage k in the population at time t. Similarly, let $\gamma_{mem}(t;j,k)$ denote the conditional probability that male in stage k finds a female in stage j acceptable for extra-marital sexual contact at time t. An analogous procedure was used to derive a formula for this probability.

In the deterministic model embedded in a stochastic process introduced by Mode and Sleeman [13], the random function $N_{CF}(t;j,k)$ represents the potential couples of type (j,k) that may be formed in the time interval $(t,\ t+h]$ and was calculated as:

$$N_{CF}(t;j,k) = \min[X(t;j)\gamma_{fm}(t;j,k), Y(t;k)\gamma_{mm}(t;k,j)]. \tag{6}$$

The random function $N_{EMP}(t;j,k)$ represents the potential number of extra-marital social contacts of type (j,k) occurring during the time interval $(t,t+h]$ was and was estimated in the embedded deterministic model as:

$$N_{EMP}(t;j,k) = \min[X_T(t;j)\gamma_{fem}(t;j,k), Y_T(t;k)\gamma_{mem}(t;k,j)]. \tag{7}$$

Observe that in [6] and [7], the arguments in the min-function are conditional multinomial expectations, given the state of the population at time t.

For a female in stage j at time t, let the random function $\rho_{fem}(t;j,k)$ denote the conditional probability that she has an extra-marital sexual contact with a male in stage k. The random function for males $\rho_{mem}(t;k,j)$ is defined similarly. For females this function has the form

$$\rho_{fem}(t;j,k) = \frac{N_{EMP}(t;j,k)}{\sum_v N_{EMP}(t;j,v)}. \tag{8}$$

A similar formula is used for $\rho_{mem}(t;k,j)$.

Finally, let the random function $Q_{femc}(t)$ represent the conditional probability that a susceptible female at time t becomes infected during $(t,\ t+h]$ through infectious extra-marital sexual contacts. Similarly,

$Q_{memc}(t)$ denotes the conditional probability a susceptible male at time t becomes infected during $(t, t+h]$ through infectious extra-marital sexual contacts. These random functions are calculated by

$$Q_{femc}(t) = \sum_{k=0}^{n} \rho_{fem}(t; 0, k) q_{fem}(k) \tag{9}$$

and

$$Q_{memc}(t) = \sum_{j=0}^{n} \rho_{mem}(t; j, 0) q_{mem}(j). \tag{10}$$

Having defined the latent risk that are elements of the matrix $\Theta_f(f)$ and other matrices of latent risk for the life cycle models of single males and couples, the next step is to set down an explicit form of $\Theta_f(f)$. Let Υ_f denote the state space for the life cycle model for single females. Death will terminate the life cycle for individual; therefore, the state space will consist of a subset, Υ_{f1} of two absorbing states: E_{11} (death from causes other than the disease) and E_{12} (death from causes attributable to the disease). The set of transient states, denoted by Υ_{f2}, will consist of $(n+1)$ states, where E_{20} corresponds to the female is susceptible to the disease and E_{2r} corresponds to the female is in stage r of the disease for $r = 1, 2, \ldots, n$. Thus, the state space for females, $\Upsilon_f = \Upsilon_{f1} \cup \Upsilon_{f2}$, consists of $(n+3)$ states. The state space for males is defined similarly.

The matrix of latent risks for females can be laid out in partitioned form as follows:

$$\Theta_f(t) = \begin{bmatrix} 0_{11} & 0_{12} \\ \Theta_{f,21}(t) & \Theta_{f,22}(t) \end{bmatrix} \tag{11}$$

where 0_{11} and 0_{12} are 2×2 and $2 \times (n+1)$ zero matrices indicating there are no transitions out of absorbing states, $\Theta_{f,21}(t)$ is an $(n+1) \times 2$ matrix governing transitions from transient states to absorbing states, and $\Theta_{f,22}(t)$ is an $(n+1) \times (n+1)$ matrix governing transitions among the transient states. The jth row of $\Theta_{f,21}(t)$, denoted by $\Theta_{f,21}(t)_j$ has the following form:

$$\Theta_{f,21}(t)_j = (\mu_{f0}, \mu_{fj}). \tag{12}$$

$\Theta_{f,21}(t)_j$ corresponds to a single female in stage j governing transitions to one of the two absorbing states: death not due to the disease and death

due to the disease. The jth row of $\Theta_{f,22}(t)$ denoted by $\Theta_{f,22}(t)_j$ has the following form:

$$\begin{aligned}\Theta_{f,22}(t)_j = (&0, \lambda_f Q_{femc}(t)\delta_{j0} + \delta_{j2}\gamma_f(2,1), \ldots, \gamma_f(j, j-1),\\ &0, \gamma_f(j, j+1), 0, \ldots, 0),\end{aligned} \tag{13}$$

where $\gamma_f(j, j-1)$ corresponds to transitions from stage j to $j-1$. The entry of $\gamma_f(j, j-1)$ corresponds to the $j-1$ column of $\Theta_{f,22}(t)_j$. Note, by assumption, there is no transition from a state of infection back to a susceptible state. Similarly, $\gamma_f(j, j+1)$ corresponds to transitions from stage j to $j+1$. The entry of $\gamma_f(j, j+1)$ corresponds to the $j+1$ column of $\Theta_{f,22}(t)_j$. Note there is no transition from stage n to stage $n+1$, because the disease is not assumed to have n stages of disease. In the above equation, Kronecker's delta, δ_{ij}, is utilized to illustrate susceptibles becoming infected via extra-marital contacts and also, the transition from stage 2 to stage 1 of infection. Kronecker's delta have the following properties:

$$\begin{aligned}&\delta_{ij} = 1 \quad \text{if } i = j, \quad \text{and}\\ &\delta_{ij} = 0 \quad \text{if } i \neq j.\end{aligned} \tag{14}$$

For singles males,

$$\Theta_m(t) = \begin{bmatrix} 0_{11} & 0_{12} \\ \Theta_{m,21}(t) & \Theta_{m,22}(t) \end{bmatrix} \tag{15}$$

where 0_{11} and 0_{12} are 2×2 and $2 \times (n+1)$ zero matrices indicating there are no transitions out of absorbing states, $\Theta_{m,21}(t)$ is an $(n+1) \times 2$ matrix governing transitions from transient states to absorbing states, and $\Theta_{m,22}(t)$ is an $(n+1) \times (n+1)$ matrix governing transitions among the transient states. The kth row of $\Theta_{m,21}(t)$, denoted by $\Theta_{m,21}(t)_k$ has the following form:

$$\Theta_{m,21}(t)_k = (\mu_{m0}, \mu_{mk}). \tag{16}$$

$\Theta_{m,21}(t)_k$ corresponds to a single male in stage k governing transitions to one of the two absorbing states: death not due to the disease and death due to the disease. The kth row of $\Theta_{m,22}(t)$ denoted by $\Theta_{m,22}(t)_k$ has the following form:

$$\begin{aligned}\Theta_{m,22}(t)_k = (&0, \lambda_m Q_{memc}(t)\delta_{k0} + \delta_{k2}\gamma_m(2,1), \ldots, \gamma_m(k, k-1),\\ &0, \gamma_m(k, k+1), 0, \ldots, 0)\end{aligned} \tag{17}$$

where $\gamma_m(k, k-1)$ corresponds to transitions from stage k to $k-1$. The entry of $\gamma_m(k, k-1)$ corresponds to the $k-1$ column of $\Theta_{m,22}(t)_k$. Note, there is no transition from a state of infection back to a susceptible state.

Similarly, $\gamma_m(k, k+1)$ corresponds to transitions from stage k to $k+1$. The entry of $\gamma_m(k, k+1)$ corresponds to the $k+1$ column of $\Theta_{m,22}(t)_k$. Note there is no transition from stage n to stage $n+1$, because the disease is not assumed to have n stages of disease. In the above equation, Kronecker's delta, δ_{ij}, is as described in Eq. (14).

Following the formation of a couple, if either a male or female is susceptible, he or she may become infected through sexual contacts with infected persons, or, if any member is infected, he or she may experience a transition with respect to stages of the disease. Furthermore, there may be dissolution of the couple if either the female or male die, or there is a separation. Similar to the description for females and males, state spaces for a semi-Markovian process describing the evolution of life cycles of couples following their formation may be derived as follows.

Denote the set of absorbing states as Υ_{c1}, which has five elements that signal the conclusion of the partnership. E_{sep} denoted the partnership ends in separation or divorce. E_{f11} and E_{f12} denote that the female member of the partnership dies due to causes other than the disease or dies due to causes of the disease, respectively. E_{m11} and E_{m12} are defined similarly for the male member of the partnership. Thus the set of absorbing states of the life cycle model for couples is:

$$\Upsilon_{c1} = \{E_{sep}, E_{f1}, E_{f2}, E_{m1}, E_{m2}\}. \tag{18}$$

A couple of type $\tau = (j, k)$ is defined, such that the female is of type $j \in \Upsilon_{f2}$ and the male is of type $k \in \Upsilon_{m2}$, where Υ_{f2} and Υ_{m2} are the sets defines in the life cycle model for single females and single males. The set of all couple types will constitute the set of transient states of the life cycle model for couples, symbolically represented as:

$$\Upsilon_{c2} = \{\tau = (j, k) | (j, k) \in \Upsilon_{f2} \times \Upsilon_{m2}\} \tag{19}$$

so that the state space for the evolution of partnership is:

$$\Upsilon_c = \Upsilon_{c1} \cup \Upsilon_{c2}. \tag{20}$$

To simplify notation, denote elements of the set Υ_c by the Greek letter τ.

To make a comparable matrix representation of $\Theta_c(t)$, as was performed for $\Theta_f(t)$ and $\Theta_m(t)$, when the disease has n stages, the set Υ_{c2} of transient states contains $(n+1)^2$ elements so that the space Υ_c for the evolution of couples contains $5 + (n+1)^2$ elements. Because no exits from absorbing

states are possible, the $(5+(n+1)^2) \times (5+(n+1)^2)$ matrix $\Theta_c(t)$ of latent risks for evolution of couples may be represented in the partitioned form:

$$\Theta_c(t) = \begin{bmatrix} 0_{11} & 0_{12} \\ \Theta_{c,21}(t) & \Theta_{c,22}(t) \end{bmatrix} \tag{21}$$

where 0_{11} and 0_{12} are 5×5 and $5 \times (n+1)^2$ zero matrices, $\Theta_{c,21}(t)$ is an $(n+1)^2 \times 5$ matrix governing transitions from transient states to absorbing states, and $\Theta_{c,22}(t)$ is an $(n+1)^2 \times (n+1)^2$ matrix of latent risks governing transitions among the transient states. The row corresponding to couple of type (j,k) of $\Theta_{c,21}(t)$, denoted by $\Theta_{c,21}(t)_{(j,k)}$ has the following form:

$$\Theta_{c,21}(t)_{(j,k)} = (\delta, \mu_{f0}, \mu_{fj}, \mu_{m0}, \mu_{mk}). \tag{22}$$

For the representation of the couple combinations, male stages of disease iterate within female stages of the disease; therefore, to understand the structure of $\Theta_{c,22}(t)$ consider the $(n+1) \times (n+1)$ submatrix where the female is of j and the male stage of disease range from 0 to n, denoted by $\Theta_{c,22}(t)_{j\bullet}$:

$$\Theta_{c,22}(t)_{j\bullet} = \text{super-diag}[\lambda_m Q_{memc}(t) + \gamma_{mc} q_m(j), \gamma_m(1,2), \ldots, \gamma_m(n-1,n)] \tag{23}$$

is along the upper quasi-diagonal and

$$\Theta_{c,22}(t)_{j\bullet} = \text{super-diag}\,[0, \gamma_m(2,1), \ldots, \gamma_m(n,n-1)] \tag{24}$$

is along the lower quasi-diagonal.

The above considers the latent risk for male member of a couple experiencing a transition. To consider the latent risk for a female member of the couple experiencing a transition consider the following. Let $\theta_c(t;0,k)$ represent the latent risk for the female member of a couple governing a transition to stage 1 of the disease when the female is a susceptible. Then,

$$\theta_c(t;0,k) = \lambda_f Q_{femc}(t) + \gamma_{mc} q_f(k). \tag{25}$$

Let $\theta_c(t;j,k)_L$ denote the latent risk of the female in the couple governing transitions from stage j to stage $j-1$. Thus,

$$\theta_c(t;j,k)_L = \gamma_f(j,j-1). \tag{26}$$

Let $\theta_c(t;j,k)_U$ denote the latent risk of the female in the couple governing transitions from stage j to stage $j+1$. Thus,

$$\theta_c(t;j,k)_U = \gamma_f(j,j+1). \tag{27}$$

Based on Eqs. (25) through (27), the remaining non-zero entries for $\Theta_{c,22}(t)$ can be defined.

Given the latent risks and all parameters of the model, formulas for the total risk functions may be derived by summation across the rows of the latent risk matrix. Formulas similar to those in [1] and [2] may also be set down for the matrices of latent risks $\Theta_m(t)$ for single males as well as that for couples $\Theta_c(t)$.

4. Stochastic Evolutionary Equations

Now with the latent risks per stage completely defined, a discrete time stochastic process can be defined. At time $t + h$, the number of single females in stage j is a sum of three components: recruitments to stage j, undergo a transition to stage j, and couple dissolution by divorce or separation or death of the male partner when the female is in state j. This is represented by

$$X(t+h;j) = X_R(t+h;j) + \sum_{v \neq j} X_T(t+h,v,j) + X_{DIS}(t+h,j) \qquad (28)$$

for every $j = 0, 1, 2, \ldots, n$, where the subscript R indicates recruitment into the population, T represents transition to stage j, and DIS represents dissolution of a couple formation involving a female partner in stage j. A similar equation can be written for single males.

The random function $Z(t + h; \tau_1; \tau_2)$, denoting the number of couples of type $\tau_2 = (j, k)$ at time $t + h$, is the sum of two components. One component consists of those couples who were of type τ_1 at t and made a transition to type τ_2 during $(t, t + h]$ and the number $Z_{CF}(t + h; \tau_1; \tau_2)$ formed during $(t, t + h]$ from a single female of type j and a single male of type k, respectively, at time t; therefore,

$$Z(t+h,\tau_2) = \sum_{\tau_1 \neq \tau_2} Z_T(t+h;\tau_1;\tau_2) + Z_{CF}(t+h;\tau_2). \qquad (29)$$

When considering the discrete time approximations to processes in continuous time, it is of interest to investigate what happens when the increment becomes small, $h \to 0$. Under this assumption and using the following the equation, $1 - \exp(xh) = -xh + o(h)$, which can be illustrated via Taylor series expansions, Eqs. (1) and (2) can be rewritten as:

$$\pi_f(t;j,k;h) = \theta_f(t;j,k)h + o(h), \qquad (30)$$

$$\pi_f(t;j,j;h) = 1 - \theta_f(t;j)h + o(h). \qquad (31)$$

This result is very useful when embedding a set of differential equations in the stochastic partnership model, where latent risk appear as constant or coefficient functions. Similar equations may be written down for the π-probabilities for single males and couples. In the stochastic model these conditional probabilities are random functions, but in the embedded differential equations these random functions are estimated by a procedure described in Mode and Sleeman [13] and elsewhere. In what follows all estimates of random functions will be denoted by a hat over the functions. For example, an estimate of $X(t;j)$ will be denoted by $\hat{X}(t;j)$.

As $h \to 0$, the embedded non-linear difference equations represented in Eq. (28) may be expressed in a modified form. The difference equations for single females become:

$$\begin{aligned}\hat{X}(t+h;j) = \mu_f \rho_f(j)h &+ \sum_{v \in \Upsilon_{f2}} \hat{X}(t;v)\hat{\pi}_f(t;v,j;h) \\ &- \sum_{k \in \Upsilon_{m2}} \hat{N}_{CF}(t;j,k)q_{CF}(j,k;h) \\ &+ \sum_{k \in \Upsilon_{m2}} \sum_{\tau_2 \in DIS_f} \hat{Z}(t;j,k)\hat{\pi}_c(t;j.k;\tau_2;h) \qquad (32)\end{aligned}$$

for every $j \in \Upsilon_{f2}$. For single males we have:

$$\begin{aligned}\hat{Y}(t+h;k) = \mu_m \rho_m(k)h &+ \sum_{v \in \Upsilon_{m2}} \hat{Y}(t;v)\hat{\pi}_m(t;v,k;h) \\ &- \sum_{v \in \Upsilon_{f2}} \hat{N}_{CF}(t;v,k)q_{CF}(v,k;h) \\ &+ \sum_{j \in \Upsilon_{f2}} \sum_{\tau_2 \in DIS_m} \hat{Z}(t;j,k)\hat{\pi}_c(t;j.k;\tau_2;h) \qquad (33)\end{aligned}$$

for every $k \in \Upsilon_{m2}$. For couples we have:

$$Z(t+h,\tau_2) = \sum_{\tau_1 \neq \tau_2} \hat{Z}(t+h,\tau_1,\tau_2)\hat{\pi}_c(t;\tau_1,\tau_2 h) + \hat{N}_{CF}(t;\tau_2)q_{CF}(\tau_2;h). \qquad (34)$$

Given the above relationships described in Eqs. (30)–(34), for every $j \in \Upsilon_{f2}$ the non-linear difference equations for single females may be

written in the form

$$\hat{X}(t+h;j) = \mu_f \rho_f(j)h + \hat{X}(t;j)(1-\hat{\theta}_f(t;j)h) + \sum_{j\neq v\in\Upsilon_{f2}} \hat{X}(t;v)\hat{\theta}_f(t;v,j)h - \sum_{k\in\Upsilon_{m2}} \hat{N}_{CF}(t;j,k)\rho(j,k)h + \sum_{k\in\Upsilon_{m2}} \sum_{\tau_2\in DIS_f} \hat{Z}(t;j,k)\hat{\theta}_c(t;j,k;\tau_2)h + o(h). \quad (35)$$

By forming the ratio $\frac{\hat{X}(t+h;j)-\hat{X}(t;j)}{h}$ and letting $h \to 0$, it can be seen that for all $j \in \Upsilon_{f2}$ the following system of differential equations for single females arise

$$\frac{d\hat{X}(t;j)}{dt} = \mu_f \rho_f(j) - \hat{X}(t;j)\hat{\theta}_f(t;j) + \sum_{j\neq v\in\Upsilon_{f2}} \hat{X}(t;v)\hat{\theta}_f(t;v,j) - \sum_{k\in\Upsilon_{m2}} \hat{N}_{CF}(t;j,k)\rho(j,k) + \sum_{k\in\Upsilon_{m2}} \sum_{\tau_2\in DIS_f} \hat{Z}(t;j,k)\hat{\theta}_c(t;j,k;\tau_2). \quad (36)$$

Similarly, for every $k \in \Upsilon_{m2}$ the differential equation for single males take on the following form:

$$\frac{d\hat{Y}(t;k)}{dt} = \mu_m \rho_m(k) - \hat{Y}(t;k)\hat{\theta}_m(t;k) + \sum_{j\neq v\in\Upsilon_{m2}} \hat{Y}(t;v)\hat{\theta}_m(t;v,k) - \sum_{j\in\Upsilon_{f2}} \hat{N}_{CF}(t;j,k)\rho(j,k) + \sum_{j\in\Upsilon_{f2}} \sum_{\tau_2\in DISm} \hat{Z}(t;j,k)\hat{\theta}_c(t;j,k;\tau_2). \quad (37)$$

For couples, an analogous system of differential equations may be derived. For every $\tau \in \Upsilon_{c2}$, these equations have the form

$$\frac{d\hat{Z}(t;\tau)}{dt} = -\hat{Z}(t;\tau)\hat{\theta}_c(t;\tau) + \sum_{\tau\neq\tau_1} \hat{Z}(t;\tau_1)\hat{\theta}_c(t;\tau,\tau) + \hat{N}_{CF}(t;\tau)\rho(\tau). \quad (38)$$

5. Form of the Embedded ODE for Given Parameters

Under the assumption of n stages of disease, the embedded differential equation for a single female in stage j based on the parameters of the

model is as follows:

$$\begin{aligned}\frac{d\hat{X}(t;j)}{dt} &= \mu_f\rho_f(j) - (\mu_{f0} + I(j>0)\mu_{fj})\hat{X}(t;j) + \delta\sum_{k=0}^{n}\hat{Z}(t;j,k)\\ &+\left(\sum_{k=1}^{n}\mu_{m0} + I(k>0)\mu_{mk}\right)\hat{Z}(t;j,k) + (\mu_{m0})\hat{Z}(t;j,0)\\ &-I(j=0)\lambda_f Q_{femc}(t)\hat{X}(t;0) + I(j=1)\lambda_f Q_{femc}(t)\hat{X}(t;0)\\ &+I(j>1)\gamma_f(j-1,j)\hat{X}(t;j-1) - I(j>0)\gamma_f(j,j+1)\hat{X}(t;j)\\ &+I(j>0)\gamma_f(j+1,j)\hat{X}(t;j+1) - I(j>1)\gamma_f(j,j-1)\hat{X}(t;j)\\ &-\sum_{k=0}^{n}\hat{N}_{CF}(t;j,k)\rho. \end{aligned} \tag{39}$$

The hat superscript indicates parameters are estimates of the recursive equations of the stochastic formulation. The indicator function I(•) is defined as one when the expression in the parentheses is true and zero otherwise. Recall there is no transition to a susceptible stage from and infective stage and there is no transition beyond stage n. There is no incremental change in mortality rate due to the disease when a person is in the susceptible stage. A similar equation can be derived for a single male in stage k:

$$\begin{aligned}\frac{d\hat{Y}(t;k)}{dt} &= \mu_m\rho_m(k) - (\mu_{m0} + I(k>0)\mu_{mk})\hat{Y}(t;j) + \delta\sum_{j=0}^{n}\hat{Z}(t;j,k)\\ &+\left(\sum_{j=1}^{n}\mu_{f0} + I(j>0)\mu_{fj}\right)\hat{Z}(t;j,k) + (\mu_{f0})\,\hat{Z}(t;0,k)\\ &-I(k=0)\lambda_m Q_{memc}(t)\hat{Y}(t;0) + I(k=1)\lambda_m Q_{memc}(t)\hat{Y}(t;0)\\ &+I(k>1)\gamma_m(k-1,k)\hat{Y}(t;k-1) - I(k>0)\gamma_m(k,k+1)\hat{Y}(t;k)\\ &+I(k>0)\gamma_m(k+1,k)\hat{Y}(t;k+1) - I(k>1)\gamma_m(k,k-1)\hat{Y}(t;j)\\ &-\sum_{j=0}^{n}\hat{N}_{CF}(t;j,k)\rho. \end{aligned} \tag{40}$$

A similar equation can be given for couples:

$$\begin{aligned}\frac{d\hat{Z}(t;i,j)}{dt} &= -\delta\hat{Z}(t;i,j) - (\mu_{f0} + I(i>0)\mu_{fi})\,\hat{Z}(t;i,j)\\ &-(\mu_{m0} + I(j>0)\mu_{mji})\,\hat{Z}(t;i,j)\\ &+I(i>1)\gamma_f(i-1,i)\hat{Z}(t;i-1,j)\end{aligned}$$

$$
\begin{aligned}
&- I(i>0)\gamma_f(i,i+1)\hat{Z}(t;i,j)\\
&+ I(i>1)\gamma_f(i+1,i)\hat{Z}(t;i+1,j)\\
&- I(i>1)\gamma_f(i,i-1)\hat{Z}(t;i,j)\\
&+ I(j>1)\gamma_m(j-1,j)\hat{Z}(t;i,j-1)\\
&- I(j>0)\gamma_m(j,j+1)\hat{Z}(t;i,j)\\
&+ I(j>0)\gamma_m(j+1,j)\hat{Z}(t;i,j+1)\\
&- I(j>1)\gamma_m(j,j-1)\hat{Z}(t;i,j) + \gamma_{mc}q_m(i)I(j=1)\hat{Z}(t;i,j-1)\\
&+ \gamma_{mc}q_m(i)I(i=1)\hat{Z}(t;i-1,j) - \gamma_{mc}q_m(i)I(j=0)\hat{Z}(t;i,j)\\
&+ \gamma_{mc}q_f(j)I(j=0)\hat{Z}(t;i,j) + \hat{Z}(t;i-1,j)I(i=1)\lambda_f\hat{Q}_{femc}(t)\\
&+ \hat{Z}(t;i,j-1)I(j=1)\lambda_m\hat{Q}_{memc}(t)\\
&- \hat{Z}(t;i,j)I(i=0)\lambda_f\hat{Q}_{femc}(t)\\
&+ \hat{Z}(t;i,j)I(j=0)\lambda_m\hat{Q}_{memc}(t) + \hat{N}_{CF}(t;i,j)\rho. \qquad (41)
\end{aligned}
$$

6. Determination of the Spread of the Disease

For n stages of disease, let $V(t)$ be the $(2(n+1)+(n+1)^2)\times 1$ vector with components:

$$
V(t) = \begin{bmatrix} \hat{X}(t;i) \\ \hat{Y}(t;j) \\ \hat{Z}(t;i,j) \end{bmatrix}. \qquad (42)
$$

The vector-matrix form of the embedded differential equation is

$$
\frac{dV(t)}{dt} = R + AV(t) + B(V)V(t) + F(V) \qquad (43)
$$

with R the constant vector of susceptible recruits; A the matrix of constant latent risks; $B(V)$ the matrix of latent risk due to extra-marital contacts; $F(V)$ the vector arising in couple formations. Note that the constant latent risks include death rates, the effect of HIV/AIDS treatment, the progression of the disease, couple dissolution rates, and intra-marital infection.

According to Spiegel [19], solutions of the embedded differential equation in the vicinity of the infection-free equilibrium behave like the solutions of the linearization of the system. Linearization of the system necessitates deriving the Jacobian of the system at the infection-free equilibrium. Stability is determined by the maximum real part of the eigenvalues. The more positive the maximum real part, the faster infection will spread throughout the population. Methods were established to determine the Jacobian evaluated at the infection-free equilibrium for the system in Eq. (43). These

methods, while not achieving a closed form solution based on the parameters of the model, welcome software implementation. Algorithms were written in APL2000 [6] on the IBM-PC platform. Immediate identification of system stability is available by reviewing the list of eigenvalues. Again further details on the derivation of Jacobian matrices can be found in Mode and Sleeman [14].

7. Results of Monte Carlo Simulation Experiments

One of the costs of the quest for realism in formulating the stochastic model of a heterosexual population under consideration is that the number of parameters can be quite large. However, when the formulation is restricted to the case of four stages of disease both the stochastic model and the embedded differential equations become computationally tractable. In this connection, it may be of interest to consult some of the more recent investigations of staged models of infectious diseases with four stages [11, 12, 18]. To answer questions regarding the main focus of this article, to investigate of threshold conditions for a sample incorporating HIV treatments versus a sample not incorporating HIV treatments, the chosen parameter assignments should reflect current behavior for both the population and the disease. Current reported behavior will be based on a review of the recent literature describing relevant characteristics such as mortality, infection rates, sexual contact rates, marriage rates, and divorce rates. Because full conditions for agreement between deterministic and expected stochastic solution are at present unknown, comparison between the deterministic and stochastic facets of the model will be based on statistical summaries of Monte Carlo simulation samples [14]. Fifty Monte Carlo realizations of the stochastic processes will be computed on monthly time intervals of 360 months. Monte Carlo samples will then be summarized statistically on a monthly basis, by derivation of the minimum, maximum, 25th quantile, 50th quantile, and 75th quantile at each month. The deterministic solution, which serves as a measure of central tendency for the process, will be computed on a monthly time scale. Both systems, the deterministic and stochastic, will give us a comprehensive understanding of the impact of the introduction of a few infectives into a susceptible population over time. To provide a basis of comparison of the systems, computer generated graphs of the two systems will be simultaneously compared. Parameter assignments are as indicated in Sleeman and Mode [18]. The parameter assignments are displayed in Tables 1–5.

Table 1. Initial parameter assignments for four-stage model.

Initial numbers, single females per stage	(902, 1, 0, 0, 0)
Initial numbers, single males per stage	(690, 1, 0, 0, 0)
Initial numbers, couples with both susceptible	4828

Table 2. Expectations for sexual contacts.

Expected number of extra-marital male partners per female	$\lambda_f = 0.25$
Expected number of extra-marital female partners per male	$\lambda_m = 0.25$
Expected number of marital sexual contacts	$\gamma_{mc} = 8$

Table 3. Probability of infections per sexual contact.

Prob. infection per extra-marital contact, stage 1	$q_{fem}(1) = q_{mem}(1) = 0.10$
Prob. infection per extra-marital contact, stage 2	$q_{fem}(2) = q_{mem}(2) = 0.05$
Prob. infection per extra-marital contact, stage 3	$q_{fem}(3) = q_{mem}(3) = 0.05$
Prob. infection per extra-marital contact, stage 4	$q_{fem}(4) = q_{mem}(4) = 0.10$
Prob. infection per marital contact, stage 1	$q_{fm}(1) = q_{mm}(1) = 0.10$
Prob. infection per marital contact, stage 2	$q_{fm}(2) = q_{mm}(2) = 0.05$
Prob. infection per marital contact, stage 3	$q_{fm}(3) = q_{mm}(3) = 0.05$
Prob. infection per marital contact, stage 4	$q_{fm}(4) = q_{mm}(4) = 0.10$

Table 4. Mortality rates.

Mortality rates for females, stage 0	$\mu_{f0} = 1/720$
Mortality rates for females, stage 1, stage 2, stage 3	$\mu_{f1} = \mu_{f2} = \mu_{f3} = 1/240$
Mortality rates for females, stage 4	$\mu_{f4} = 1/23.8$
Mortality rates for males, stage 0	$\mu_{m0} = 1/660$
Mortality rates for males, stage 1, stage 2, stage 3	$\mu_{m1} = \mu_{m2} = \mu_{m3} = 1/180$
Mortality rates for males, stage 4	$\mu_{m4} = 1/23.8$

Table 5. Acceptance parameters and coupling parameters.

Couple formation-dissolution	$\rho = 1/12; \delta = 1/120$
Acceptance parameters for extra-marital parameters	$\beta_{fem} = \beta_{mem} = 0$
Acceptance parameters for marital partners	$\beta_{fm} = \beta_{mm} = 0$

As discussed by Sleeman and Mode [18], duration of marital partnerships were long in the sense that the latent expectation of a marital partnership was $1/\delta = 120$ months or 10 years but the expected latent waiting time among marital partner was $1/\rho = 12$ months or one year. Thus, the samples consist of a heterosexual population with long duration of partnerships. Initial conditions, as seen in the first three rows of Table 1,

were determined by Sleeman and Mode [18]. The β-parameters are all set to 0, $\beta_{fem} = \beta_{mem} = \beta_{fm} = \beta_{mm} = 0$, which indicate that marital partners and extra-marital sexual partners are chosen at random, with no regards to medical status or disease severity of the potential partner. Extra-marital sexual contacts occurred at a rate of 3 per year or $\lambda_f = \lambda_m = 0.25$. It was assumed the probability of infection per stage were the same for marital and extra-marital contacts. Not illustrated in Table 1 are the γ-parameters for duration of stay in stages 1, 2, and 3 before transition to the next more severe stage. Parameter assignments were determined by results of Sleeman and Mode [18] and Longini *et al.* [12]. The parameters are as follows:

$$\gamma_f(1,2) = \gamma_m(1,2) = 1/12, \quad \gamma_f(2,3) = \gamma_m(2,3) = 1/52.62, \quad \text{and}$$
$$\gamma_f(3,4) = \gamma_m(3,4) = 1/62.89.$$

Applying the Jacobian methodology, the Jacobian is unstable with maximum real part of the eigenvalues is 0.0313; therefore, the introduction of a few infectives at the parameter settings in Table 1, an epidemic will develop with positive probability. Presented in Fig. 1 are the trajectories based on the summary statistics of the 50 Monte Carlo simulation runs for the cumulative number of infected females in couples along with the trajectory for the embedded deterministic model for projections of 720 months.

One of the most striking features of Fig. 1 is the level of stochasticity exhibited in the projections. For example in every epoch, the minimum

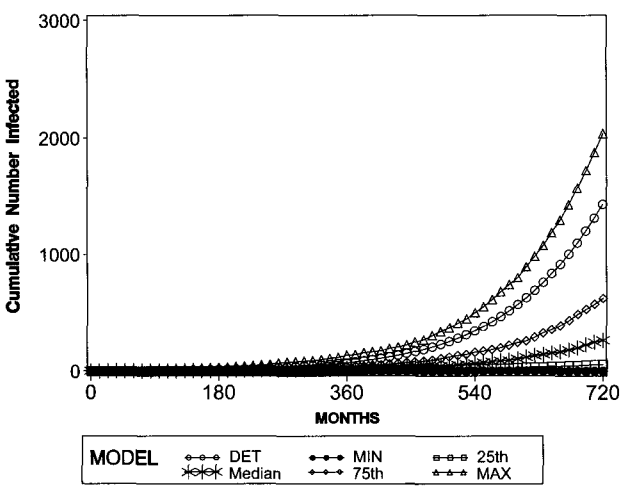


Fig. 1. HIV/AIDS treatments not implemented.

of the 50 realizations is on the horizontal axis, indicating that in some realizations of the process no new infection occurred but according to the maximum of the 50 realizations, nearly 2500 coupled females have been infected by 720 months. Nonetheless, the model indicates the potential for a severe epidemic to develop, although it may take a substantial amount of time for the epidemic to develop.

Based on the findings of Renaud *et al.* [17], Detels *et al.* [4], and Autran [2], which suggest that potent antiretroviral therapy use by HIV infected persons will have a substantial beneficial effect on arresting $CD4^+$ count decline, parameter assignments are as illustrated in Table 6, indicating, on average, complete arrest in $CD4^+$ count declination.

Based on the parameter assignments of Tables 1 through 6, the maximum real part of the eigenvalues is 0.0366, which is larger than when the transition rates to less severe stages are set to 0. Thus, the larger eigenvalue indicates a quicker rate of spread as compared to the previous example. Trajectories based on the parameter assignments of Tables 1 through 6 are illustrated in Fig. 2.

Table 6. Transitions rates among stages.

$\gamma_f(1,2) = \gamma_m(1,2) = \gamma_f(2,1) = \gamma_m(2,1) = 1/12,$
$\gamma_f(2,3) = \gamma_m(2,3) = \gamma_f(3,2) = \gamma_m(3,2) = 1/52.62,$
$\gamma_f(3,4) = \gamma_m(3,4) = \gamma_f(4,3) = \gamma_m(4,3) = 1/62.89$

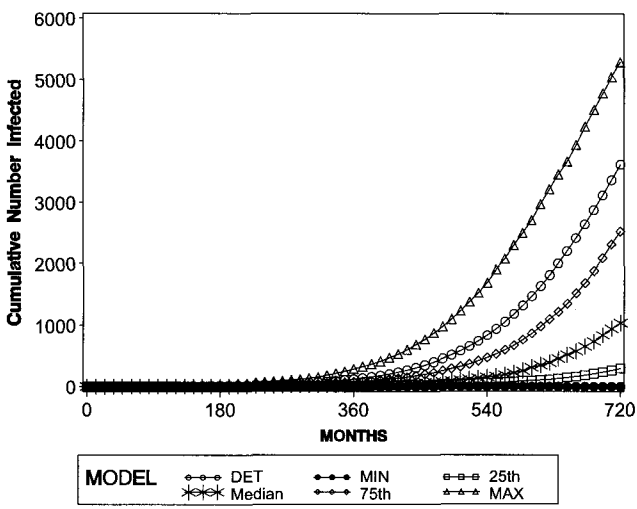


Fig. 2. HIV/AIDS treatments implemented.

Simultaneous comparison of Fig. 2 with Fig. 1 illustrates when HIV/AIDS treatments are taken into consideration, the epidemic will have a quicker rate of spread; therefore, resulting in an increase in infections of susceptibles. All corresponding trajectories of Fig. 2 compared to Fig. 1, with the exception of the minimum trajectory, indicate a substantial increase in the cumulative number of new infectives. Direct comparison of the deterministic solution at 720 months indicates approximately 1500 cumulative infected coupled females for Fig. 1 and 3800 cumulative infected coupled females for Fig. 2.

To estimate the probability of infection, the fraction of those realizations of the process that contained no secondary infectives was calculated as a function of time. Figure 3 contains the plot of the probability of extinction as a function of time for both systems. Similar to our simultaneous comparison of Fig. 1 and Fig. 2, simultaneous comparison of the extinction probabilities as a function of time provide empirical evidence that the system incorporating HIV treatments has a lower level of extinction; therefore, increase infections are evident in the system with HIV treatment.

While the need for HIV/AIDS treatments is not in question, the Jacobian methodology demonstrates that there is a potential detrimental impact for the susceptible population. By prolonging the life-span of infectives, the infectives have more opportunities to infect susceptibles, which,

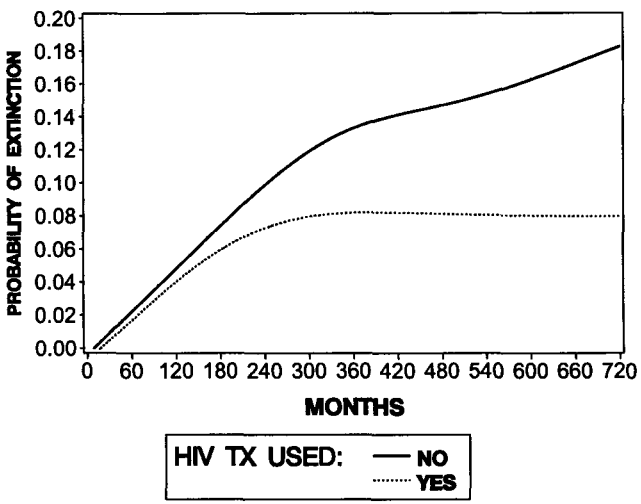


Fig. 3. Comparison of extinction probability over duration of simulations.

on average, will result in a quicker rate of spread for the epidemic. With recent published success of HIV/AIDS treatment protocols [9], it is urgent to inform and educate the public to the HIV/AIDS heterosexual epidemic. Infectives experiencing positive results, might question their medical status and the need to adopt stringent safe-sex practices and honest communication of health status. Thus, improvement in partnership selectivity is needed. Improvement in communication skills is targeted to help one to learn about a partner's prior sexual behavior and level of risk, such information will presumably lead to safer sexual behaviors such as abstaining from sex with high risk partners and screening of partners prior to engaging in sexual activity or a committed relationship [20], corresponding to both sets of β-parameters in the Mode–Sleeman model. Threshold investigations, based on these increased acceptance parameters, may serve as benchmarks for where a random-mixing population must achieve to limit the severity of the epidemic.

In the discussed experiments, selective skills for choice of marital partners would need to be greatly increased in order to return the system to the rate of progression seen prior to the introduction of HIV/AIDS treatments. The β-parameters for couple formations would need to be 1.25. This represents that a susceptible would choose a susceptible marital partner 78% of the time over an infected potential partner in stage 1 of the disease, 92% of time over an infected in stage 2 of the disease, 98% of time over an infected in stage 3 of the disease, and 99% of the time over an infected in stage 4 of the disease. In order to achieve a stable system resilient to infection spread, further inflation of the β-parameters is required.

8. Discussion and Summary

This stochastic model considers a semi-Markov process based on competing risks for males and females. The multiple facets of the heterosexual population incorporated by our stochastic model are couple formation, couple dissolution, recruitments, death, selectivity of partners for couple formation and extra-marital contact, progression through disease severity, the effects of HIV/AIDS treatments, and infection through sexual contacts. By allowing the time increment to become small, the stochastic model gives rise to a system of embedded differential equations. A stability analysis performed by linearizing the embedded differential equation around the infection-free equilibrium and determining when the maximum real part of the eigenvalues crosses zero gives a threshold condition dependent on the

multidimensional parameter setting. The solution of the system of differential equations is a linear combination of the exponential of each eigenvalue; therefore, the magnitude of the maximum real part of the eigenvalues indicate the rate of disease progression, with more positive values indicating a faster rate of spread. For statistical simplicity, our attention was focused on a heterosexual population with quick rate of couple formation and long duration of partnership. However, the model referenced in this paper are capable of handling many other scenarios, such as the inclusion of covariates such as age and race. Refer to Mode and Sleeman [14] for an extensive discussion on these issues.

As our experiments illustrated, when the effect of HIV/AIDS treatments is ignored the disease impact and progression is less then when the effect of HIV/AIDS treatments is incorporated. In order to achieve a system comparable to the disease impact prior to the implementation of HIV/AIDS treatment, selection skills must be greatly increased.

One fear is that infected individuals may question their health status when experiencing beneficial HIV therapy; therefore, it is urgent susceptible individuals screen their partners. It is not realistic to assume screening would constitute medical testing, but screening could consists of questions concerning sexual promiscuity and risky drug use behavior. Susceptible individuals need to posses the skills to recognize potentially risky partners. While the life-span of infected individuals, as well as their health status, continues to improve, the spread of the epidemic may continue to rise, if HIV education and prevention efforts are not advanced.

Recent investigations [1, 3] have also considered the issue of the effect of HIV treatments on infection spread and have had similar findings as discussed in this article. The models discussed in these papers are simpler and do not account for the various facets of the population as discussed here. Blower *et al.* [3] and Quinn *et al.* [16] have discussed the positive effect of HAART therapy has on reducing viral load, which in turn also reduces infectivity. This does provide evidence of an interaction of the effect of HAART therapy and the probability of infection. The model described here can accommodate this feature of the disease behavior, but accurate archival parameter estimates of infection rates incorporating this feature are needed. Presently, more research is being performed in this area; therefore, accuracy of these initial findings still remain uncertain. Thus, for now, the findings described by the experiments discussed in this article can be considered as the worst-case scenario. If HAART therapy, does reduce infection rates, then the estimates described here maybe slightly inflated. Regardless of the effect HAART therapy has on infection rates, the coupling of increased

educational efforts and HIV treatments will have an overall positive effect on the disease severity, with a higher probability for disease extinction.

In closing it should also be mentioned that stability of the disease-free equilibrium could also be attained if it were possible to lower the probabilities that a susceptible becomes infected per sexual contact with an infected person. On the practical side, one on the widely used methods for achieving this reduction would be the use of condoms and other protective devices. The Jacobian methodology described in this paper could be used to find regions of the parameter space such that the disease free system were stable, but, due to space limitations, the results of computer experiments designed to find such regions will not be reported here.

References

[1] R. M. Anderson, S. M. Gupta and R. M. May, Potential of community-wide chemotherapy and immunotherapy to control the spread of HIV-1, *Nature* **350** (1991) 356–359.

[2] B. Autran, Effect on antiretroviral therapy on immune reconstitution, *Antiviral Therapy* **4**(Supplemental 3) (1999) 3–6.

[3] S. M. Blower, H. B. Gershengorn and R. M. Grant, A tale of two futures: HIV and antiviral therapy in San Francisco, *Science* **287** (2000) 650–654.

[4] R. Detels, A. Munoz, G. McFarlane, *et al.*, Effectiveness of potent antiretroviral therapy on time to AIDS and deaths in men with known HIV infection duration, *J. Amer. Medical Assoc.* **280** (1998) 1497–1503.

[5] K. Dietz and K. P. Hadeler, Epidemiological models for sexually transmitted diseases, *J. Math. Biol.* **26** (1988) 1–25.

[6] L. Gilman and A. J. Rose, *APL — An Interactive Approach*, 2nd edn. (Wiley and Sons, New York, 1976), pp. 219–233.

[7] C. M. Gray, J. M. Schapiro, M. A. Winters and T. C. Merigan, Changes in CD4(+) and CD8(+) T Cell subsets in response to highly active antiretroviral therapy in HIV type 1 infected patients with prior protease inhibitors experience, *AIDS Res. Human Retroviruses* **14** (1998) 561–569.

[8] B. Hirschel and M. Opraqvil, The year in review: Antiretroviral treatments, *AIDS* **13**(Supplement A) (1999) S177–S188.

[9] R. S. Hogg, M. V. O'Shaughnessy, N. Gataric, *et al.*, Decline in death from AIDS due to new antiretrovirals, *Lancet* **349** (1999) 1443–1445.

[10] J. M. Hyman and J. Li, Disease transmission models with biased partnership selection, *Appl. Numer. Math.* **24** (1997) 379–392.

[11] J. M. Hyman, J. Li and E. A. Stanley, The differential infectivity and staged progression models for the transmission of HIV, *Math. Biosci.* **155** (1999) 77–109.

[12] I. M. Longini, W. S. Clark, G. A. Satten, R. H. Byers and J. M. Karon, Staged Markov models based on CD4+ T-Lymphocytes for the natural history of HIV infection, in *Models for Infectious Human Diseases — Their*

Structure and Relation to Data, eds. V. Isham and G. Medley (Cambridge University Press, 1996), pp. 439–459.

[13] C. J. Mode and C. K. Sleeman, A new design of stochastic partnership models of epidemics of sexually transmitted diseases with stages, *Math. Biosci.* **156** (1999) 95–122.

[14] C. J. Mode and C. K. Sleeman, *Stochastic Processes in Epidemiology: HIV/AIDS, Other Infectious Diseases, and Computers* (World Scientific, Singapore, 2000), pp. 358–362, 404–410.

[15] MMWR, Sexual risk behavior of STD clinic patients before and after Earvin "Magic" Johnson's HIV-infection announcement — Maryland, 1991–1992, **42** (1993) 45–48.

[16] T. C. Quinn, M. J. Wawer, N. Sewankambo, D. Serwadda, C. Li, F. Wabwire-Mangen, *et al.*, Viral load and heterosexual transmission of human immunodeficiency virus type 1 — Rakai project study group, *New England J. Med.* **342** (2000) 921–929.

[17] M. Renaud, C. Katalana, A. Mallet, *et al.*, Determinants of paradoxical CD4 cell reconstitution after protease inhibitor — containing antiretroviral regimen, *AIDS* **13** (1999) 669–676.

[18] C. K. Sleeman and C. J. Mode, A computer exploration of some properties of non-linear stochastic partnership models for sexually transmitted diseases with stages, *Math. Biosci.* **156** (1999) 123–145.

[19] M. R. Spiegel, *Applied Differential Equations*, 3rd. edn. (Englewood Cliffs, NJ: Prentice Hall Inc., 1981), pp. 441–443.

[20] D. J. Whitaker, K. S. Miller, D. C. May and M. L. Levin, Teenage Partners' Communication about sexual risk and Condom Use: The importance of Parent-Teenager Discussion, *Family Planning Perspectives* **31** (1999) 117–121.

This page is intentionally left blank

CHAPTER 4

MODELING AND IDENTIFICATION OF THE DYNAMICS OF THE MF-INFLUENCED FREE-RADICAL TRANSFORMATIONS IN LIPID-MODELING SUBSTANCES AND LIPIDS

J. BENTSMAN, I. V. DARDYNSKAIA, O. SHADYRO, G. PELLEGRINETTI, R. BLAUWKAMP and G. GLOUSHONOK

This work presents two mathematical models explicitly reflecting the magnetic-field-induced transitions in biologically significant processes: oxidation of n-hexane and linolenic acid, and describes the methodology used in obtaining the models. For the n-hexane oxidation, the range of the magnetic field strength is found (0.05–0.3 T) with the trend indicating a significant magnetic-field-induced change in the reaction rates (up to 50% at 0.2 T). For the linolenic acid oxidation, a pronounced magnetic-field-induced change in the rate of malonaldehyde (MDA) production is found (at 0.1 T). The equations describing the effects of the magnetic field on the photoinduced free radical reaction of oxidation involving a lipid-modeling substance, hexane, and a fatty acid, linolenic acid, are obtained on the basis of chemical kinetics and data from batch experiments. The magnetic-field-induced changes in n-hexane and linolenic acid oxidation are validated (the latter only for diene conjugates) using the identification technique based on the real time input-output data in separately conducted flow-through experiments.

Keywords: Mathematical modeling; stochastic H_∞ identification; magnetic-field-induced transitions; lipid-modeling substance; oxidation; hexane; fatty acid.

0. Introduction

In recent years there appeared studies, based on the experiments at the cellular level as well as on animals, which report that magnetic fields can interact with, and produce changes in, biological structures. One of the major candidate biophysical mechanisms for MF effects arises through the influence of MF on free radical reactions, which always play an important role in the processes of damaging of the living organism, therefore, any changes in the rates of these reactions are likely to alter their damaging effects.

The proposed biophysical mechanisms for MF effects on biological structures can be briefly described as follows [23, 26]. When a spin-correlated radical pair is formed, the radical electron spins can be either parallel (T-triplet state) or antiparallel (S-singlet state). Pauli Exclusion Principle, however, postulates that electrons cannot bond if their spins are parallel, and therefore chemical bond formation between two reacting radicals requires that they be in the singlet, rather than the triplet, state. Confining the discussion to the triplet state formed free radical pair, either spin-orbit or hyperfine interactions can induce the electron transition from triplet to singlet state. This transition is referred to as *intersystem crossing* or T-S interconversion (S-T conversion is also possible) [13, 28]. If after T-S conversion a single state spin-correlated radical pair remains in the cage (a space sufficiently small to ensure that the probability of recombination of the original spin-correlated pair is much higher than that of the reaction of each of the radicals of this pair with the other molecules) [23], the pairwise elimination of the singlet state radicals by recombination is likely to take place. This recombination (geminate or cage recombination) competes with the escape of the radicals from the cage into the "volume" (a space outside of the cage were free radical reactions could take place) and subsequent formation of products different from those of cage recombination. The magnetic field effect (MFE) consists in engaging the electrons through their magnetic moments and altering the intersystem crossing rate. This magnetic field action, therefore, may either speed up or slow down the rate of free-radical recombination, depending on the precise field value, thereby influencing the ratio of cage to escape reaction yields and rendering the overall chemical reaction field sensitive [1].

Recently a number of experimental studies on the influence of the magnetic field on the chemical reactions of free radicals and on processes involving triplet excited molecules in solutions has been carried out [14, 17], and physically justified models of these phenomena have been proposed [25]. However, the relation between the magnetic field strength and the reaction rate (including qualitative change in terms of rate increase or decrease) has not been established for many such reactions in the ranges of field strength frequently encountered by humans, and the review of reported results reveals a significant knowledge gap in this area. There are also only a few well-documented examples involving biological systems. McLaucham and Steiner [13], Harkins and Grissom [11], and Grissom [8] show that an applied external magnetic field can alter the rate of T-S interconversion in radicals, and potentially affect enzyme activity. Chignel and Sik, [3] showed

that application of a static external magnetic field (3350G) during UV-irradiation reduced the time of 50% ketoprofen-sensitized photohemolysis of human erytrocytes.

The effect of the magnetic field on biologically relevant chemical reactions can be looked at as taking place at four time scales, corresponding to the process stages in the descending order of the process speed. The fastest time scale corresponds to the physical stage, where the energy absorption and redistribution process, which leads to excitation and/or ionization of molecules, takes place. The second, physico-chemical, stage corresponds to the formation of charges or neutral radicals and/or ions. The third stage corresponds to the formation of stable chemical products. The fourth stage is the formation of biological effects as the result of the previous three stages. Mathematical models which describe the time evolution of the products of the reactions and their biological effects provide a framework that clearly sorts out the entire behavior of the magnetic-field-induced transitions in biological systems. These models are usually constructed on the basis of physico-chemical laws governing the behavior of process variables such as concentrations of the reactions, and identification methods using direct input-output measurements collected during real time experiments where the process transients or response to random or sinusoidal excitation can be observed with the sufficient time-resolution. Such detailed model verification, for example, has been carried out for the bistable laser-induced dimerization of sulfonyl chloride with composition affecting optical density in [7] via high time resolution optical density measurements.

The time-resolved techniques for studying MF effects on chemical reactions have been utilized in [24] in a flow-through experiment where photochemical quantum yield dependence on magnetic field has been studied using optical density changes and in [10] where the evolution of the intermediates in the MF-influenced reaction has been monitored in real time via optical detection.

Due to the fact that mathematical model can never entirely capture the reality, and therefore always "undermodels" the true process, it is important to supply a model with a "quality tag", i.e. the quantified degree of undermodeling, often referred to as modeling uncertainty, or unmodelled dynamics [9, 12]. The recently developed robust H_∞ identification methods [9, 12, 19, 20, 27] are suitable for this task, since they generate models along with the H_∞ norm bounds on unmodelled dynamics. The H_∞ identification can also yield a solution (the best possible under the circumstances) for a significant plant/model mismatch. This is especially important for biologically

relevant chemical reactions which are known to have high level of dynamic uncertainty and yield noisy experimental data. Under such conditions traditional identification methods are known to break down if the model is not chosen sufficiently close to the *a priori* unknown plant dynamics. Therefore, it looks attractive to enhance the modeling methodology with H_∞ identification methods in an attempt to obtain the models for EMF-influenced biologically significant reactions known to have high dynamic uncertainty level and assess the model quality in terms of the explicit bound on the modeling uncertainty.

The mathematical modeling of MF-sensitive free-radical reactions is still in infancy. The physico-chemical stage has been studied in the work of R. Z. Sagdeev (1977) [23] where the influence of a constant magnetic field on the reactions involving free radicals and triplet molecules in solutions has been examined and the theoretical principles of the influence of the magnetic field on the recombination processes of free radicals (taking into account the spin effect) and the quenching of triplets have been described. Two models, diffusion and exponential ones, were considered for phenomenological description of the recombination of the radicals. In Bachelor *et al.* (1993) [1], using a time resolved experiment it was shown that dynamics of evolution of the free radical cloud lies between the dynamics of the diffusion model and that of the exponential one. Some features of the MF-influenced chemical reaction dynamics are presented in [10] on the basis of the time-resolved experiments. The first fully developed explicit mathematical description of the dynamics of an MF-sensitive chemical reaction is given by the present authors in [2].

The stage of the formation of the biologically significant products of magnetic-field-induced transitions is investigated to a much lesser extent. Therefore, it is of great interest to investigate the influence of the external magnetic field on the free radical reactions of oxidation that could take place in biological lipid structures and to build the models of the dynamics of these reactions with the refinement and validation of these models on the basis of the experimental data. The description of the MF-sensitive reaction in [2] represents also the first experimentally validated and refined mathematical model of the MF-induced transitions of a biologically significant process, oxidation of a lipid-modeling substance, hexane. The journal publication [2], however, is focused only on the mathematical model and does not contain detailed experimental data and description of the chemical reaction kinetics.

Building on the material in [2] and including experimental data and chemical kinetics equations, the present chapter describes the development

and validation of mathematical models of the MF-induced transitions in oxidation of hexane (Sec. 3) and another biologically significant substance — linolenic acid (Sec. 4), the latter belonging to the class of fatty acids.

1. Objectives and Motivation

The objectives of this work were:

(1) To select and experimentally verify the conditions under which the free radical processes that take place in biological structures are most likely to be sensitive to the influence of the external magnetic fields.
(2) Under these conditions, to experimentally examine the magnetic-field-induced changes in these processes in the range of MF strength from zero through the values where pronounced effects of MF exposure are most likely to be found.
(3) To develop predictive mathematical models with the explicit dependence on the magnetic field strength describing the temporal evolution (dynamics) of these processes. Refine and validate these models via processing the experimental data by robust (H_∞) and standard identification algorithms as well as through the computation of the model trajectories and their comparison to the experimental process temporal evolution data.

In order to approach the stated objectives a model describing the effect of magnetic field on the free radical reactions of oxidation in lipids and lipid-modeling substances (substances which can undergo free radical transformations similar to lipids) was chosen.

Lipids are the essential cell compounds which are generally regarded to be the most sensitive to the influence of free radicals. Among them phospholipids, containing a significant amount of unsaturated fatty acid residues, are the most vulnerable to a free radical attack. Free radical processes that take place in the hydrophobic part of lipids lead to their peroxidation, which is considered a prevalent feature of the free radical inflicted cellular injury [21].

The peroxidation of polyunsaturated fatty acids usually involves three operationally defined processes: initiation, propagation and termination. These processes can be described by the following reactions:

$$\mathrm{r-r} \rightarrow 2\ {}^{\bullet}\mathrm{r} \tag{1.1}$$

$${}^{\bullet}\mathrm{r} + \mathrm{RH} \rightarrow {}^{\bullet}\mathrm{R} + \mathrm{rH} \tag{1.2}$$

$${}^{\bullet}\mathrm{R} + \mathrm{O_2} \rightarrow \mathrm{ROO}^{\bullet} \tag{1.3}$$

$$ROO^{\bullet} + RH \rightarrow ROOH + R \tag{1.4}$$

$$^{\bullet}R + {}^{\bullet}R \rightarrow R - R \tag{1.5}$$

$$^{\bullet}R + ROO^{\bullet} \rightarrow ROOR \tag{1.6}$$

$$2ROO^{\bullet} \rightarrow ROOR + O_2. \tag{1.7}$$

As seen from this scheme, the process of lipid oxidation occurs in a form of a chain reaction, with reaction (1.2) being the reaction of the chain initiation.

Lipid peroxidation may be initiated by any primary free radical that has sufficient reactivity to extract hydrogen atom from a reactive methylene group of an unsaturated fatty acid. In our experiment the photosensitized reactions were selected to provide the initiation of free radical reaction of oxidation in lipids. This type of initiation requires molecular oxygen, a sensitizing dye, or pigment, and exciting light. In the case when the formation of free radicals in lipids takes place in the presence of exciting light and sensitizing agent in the form of carbonyl compound, the formation of a geminate radical pair in the triplet state initiation phase could take place. Thus, the formation of geminate radical pairs in the photosensitized reactions in lipids could induce the sensitivity of the free radical processes that take place in them to the external magnetic field.

2. Framework for Fitting Mathematical Models to the Experimental Data

Two types of experiments, batch and flow-through, were conducted to obtain the experimental data and to reveal the MF effects on lipids (linolenic acid) and lipid modeling substances (hexane). A separate experiment, the reaction of the photoinduced transformation of iso-propanol in the presence of acetone, was conducted to localize the influence of magnetic field in system model. This reaction indicated that the dependence of the reactions of oxidation on magnetic field could be adequately reflected through the functional dependence of the intersystem crossing rate coefficient in the differential equations of the reaction dynamics on the magnetic field strength.

The mathematical modeling framework for the batch experiment has been confined to nonlinear continuous time ordinary differential equations (ODE's), describing the second, physico-chemical, stage of the formation of free radicals under UV irradiation and the third stage, that of the formation of stable chemical products. The approach taken was that of a "gray box" modeling, namely, the ODE's were obtained from the fundamental physico-chemical relations and parametrized by a set of constants which

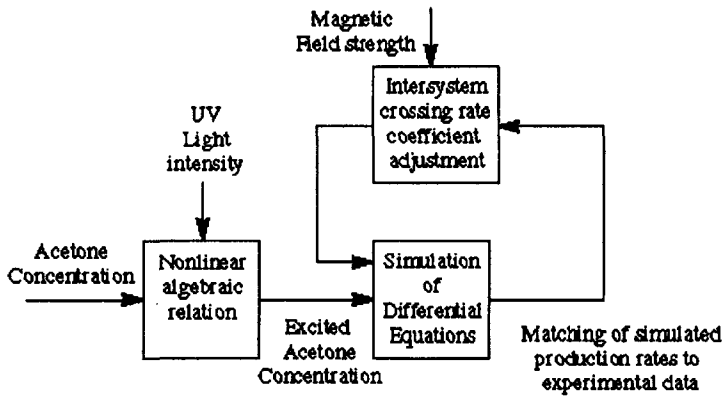


Fig. 1. Conceptual diagram of mathematical modeling using batch experiment data.

were adjusted to match the experimental data and to reflect the effects of magnetic field on the formation of stable chemical products.

A conceptual diagram of mathematical modeling using the batch experiment data is presented in Fig. 1.

As seen from this figure, the batch experiment data were used as follows. For a given experimental acetone concentration and UV intensity, the excited acetone concentration was computed using a nonlinear relation which was also included in the model, the rates of the production of stable chemical reaction products were measured as functions of magnetic field strength, and the corresponding model-generated rates, computed through model decomposition and ODE solvers, were matched to the experimental ones via the adjustment of the intersystem crossing rate coefficient in the model and very minor changes in the other coefficients, wherever necessary.

While it is of interest to carry out batch experiments to exhaustively cover the whole range of possible acetone concentrations, this is prohibitively time consuming, therefore the flow-through experiments have been pursued were the input switches between the maximal and the minimal acetone concentrations, and the mixing inside the reaction vessel makes acetone concentrations sweep through the whole range of interest.

The quality of mixing could be assessed by how well the identified mixing dynamics between input and output acetone concentrations matches perfect mixing dynamics which is known to be of the first order. The mathematical modeling framework for the flow-through experiment has been confined to the affine discrete time ODE's, i.e. linear discrete time ODE's

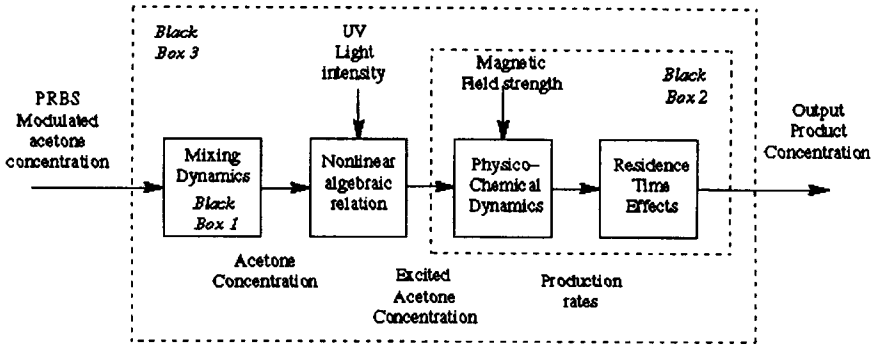


Fig. 2. Conceptual diagram of mathematical modeling using flow through experiment data.

with a constant offset, with the approach being that of a "black box" modeling. The values of an offset and model parameters would depend upon the particular identification procedure. A conceptual diagram of mathematical modeling using the flow-through experiment data is presented in Fig. 2.

As seen from this figure the flow-through experiment data were used as follows. Keeping the input flow rate into and out of the reaction vessel constant, the input acetone concentration was modulated as a pseudo-random binary sequence (PRBS). The fluid was stirred in the reaction vessel under UV and MF irradiation, the exit flow concentrations of acetone and reaction products were measured, and the excited acetone concentration was computed using a known nonlinear relation. Then, in the absence of the magnetic field and under fixed nonzero value of magnetic field strength, the input-output models were identified which related (i) PRBS input to the concentration of acetone in the exit flow ("black box 1"), (ii) PRBS input to the concentration of the reaction products in the exit flow ("black box 3"), and (iii) the computed excited acetone concentration to that of the reaction products ("black box 2").

3. Modeling and Identification of the MF-Influenced Oxidation of Hexane

3.1. *Experimental part*

The investigation was performed using a stirred-tank reaction vessel which had a volume of 900 µl. This apparatus is shown in Fig. 3.

A valve at the outlet of the vessel maintains a constant flow rate of the exit stream at 30 µl/min. Three regulating valves are placed at the input

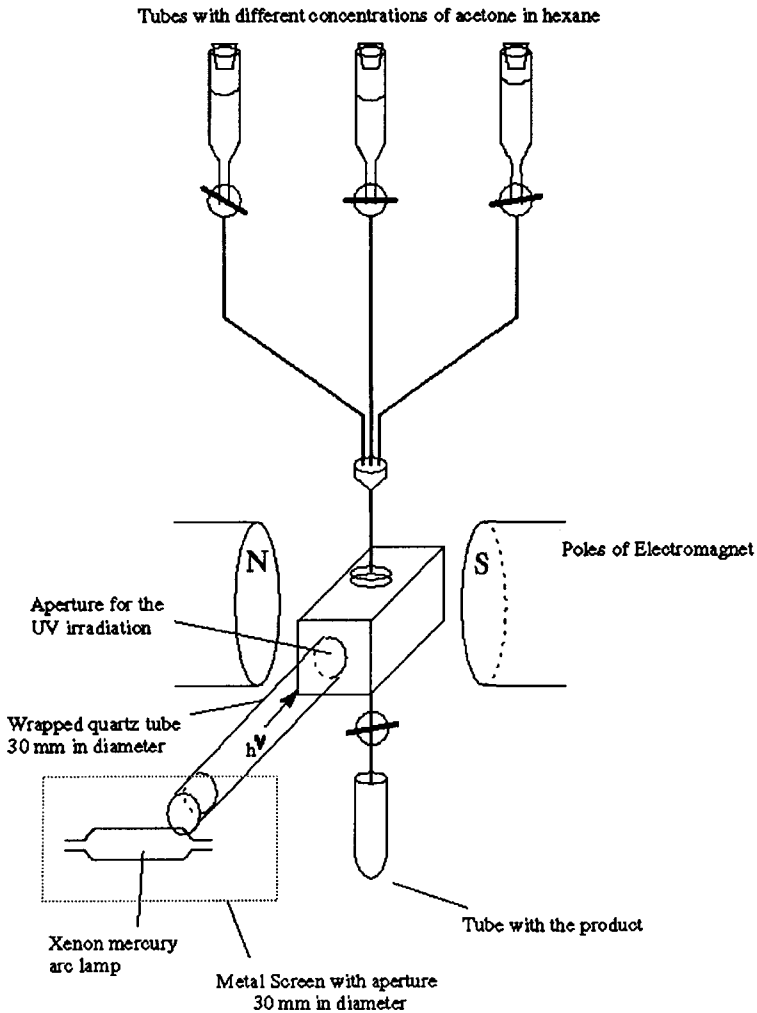


Fig. 3. Apparatus for the flow through experiment.

to the reaction chamber to permit the selection of n-hexane with various concentrations of acetone in the inlet stream. This stream is formed by liquid flows resulting from switching among vessels that contain prepared solutions with acetone concentrations of 6.0×10^{-3} mol/l, 8.0×10^{-2} mol/l, and 1.0 mol/l. The reaction vessel is held continuously under ultraviolet light with 260–280 nm wavelength to provide a constant energy source for the oxidation reaction. The reaction vessel is placed within the coils of an electromagnet, which can provide varying magnetic fields. The exit stream

is sampled and analyzed using gas-chromatographic equipment to evaluate the concentration of acetone and reaction products, ketones and alcohols.

In order to investigate the dynamics of the reaction, experiments were performed which varied the input concentration while sampling and analyzing the exit stream every three minutes. In these experiments, the input concentration of the acetone was modulated as a pseudo-random binary sequence (PRBS), alternating at random sample times between 6.0×10^{-3} mol/l and 1.0 mol/l. Specifically, the reaction was first brought to steady-state conditions with an input concentration of 8.0×10^{-2} mol/l. Then, according to randomly preselected time samples, the input concentration was switched between the minimum and the maximum concentrations, while the exit stream was analyzed at every sampling instant. These concentrations were chosen because steady state data indicated that the steady state concentration of the reaction products was a nearly linear function of the logarithm of acetone concentration for acetone concentrations between 3.0×10^{-2} mol/l and 2.0×10^{-1} mol/l. On a logarithmic scale, the 8.0×10^{-2} mol/l concentration lies almost exactly between the minimum and maximum concentrations. This value has been chosen as the logarithmic average for the magnitude of the PRBS sequence. The preselection of pseudo-random sample times for the alternation of concentrations was done in such a way that the concentration of acetone in the reaction vessel would stay in the "near-linear" range, and so that it would be suitably "rich" in frequency content; i.e. the spacing between concentration alternations varied sufficiently. The curves describing accumulation (concentration growth) of the reaction products as a function of the UV-exposure duration were observed to be linear for each of the products for almost the entire duration of the process. The nonhomogenuity of the UV irradiation was evaluated and found not to noticeably affect the outcome of the experiment, therefore in the model it is neglected.

To investigate the influence of the magnetic field on the photo-induced free-radical transformations of n-hexane the commercial chemical reagents were utilized. The purity of these chemicals was verified by chromatographic method. N-hexane that contains diluted oxygen in tube was placed in a plastic rack between the poles of ERS-220 electromagnet and exposed to UV-irradiation at room temperature in the presence of acetone with stable concentration 10^{-2} mol/l as the sensitizer.

The experiment was conducted as follows. Samples with the identical initial composition were irradiated at the time intervals, sequentially increasing by the increment of 5 minutes; i.e. the first sample was irradiated during 5 min. and then analyzed, the second identical sample was

irradiated during the time interval of 10 minutes and then analyzed and so on. The irradiation was performed either in the presence or absence of applied static magnetic field. The initial magnetic field strength was set at 0.4 T and for each of the subsequent experiments with magnetic field the field strength was sequentially reduced by 0.05 T until the whole range was exhausted. The analysis of molecular products of the reaction of photoinduced oxidation of hexane was performed by the method of the gas-liquid chromatography (GLC) with the flame ionizing detection.

3.2. *Reaction scheme and differential equations, describing the process of photo-induced oxidation of hexane*

The main products of the photolysis of *n*-hexane in the presence of acetone were alcohols (hexanol-2 and hexanol-3) and ketones (hexanon-2 and hexanon-3). The formation of these products occurs according to the following scheme:

$$\mathrm{Ac} \xrightarrow{h\nu} {}^3[\mathrm{Ac}]^* \tag{3.1}$$

$$\begin{matrix} {}^3[\mathrm{Ac}]^* \\ \uparrow \quad \uparrow \end{matrix} + \mathrm{C_6H_{14}} \rightarrow \begin{matrix} [{}^\bullet\mathrm{AcH} \cdots {}^\bullet\mathrm{C_6H_{13}}] \\ \uparrow \quad \downarrow \end{matrix} \tag{3.2}$$

$$\begin{matrix} [{}^\bullet\mathrm{AcH} \cdots {}^\bullet\mathrm{C_6H_{13}}] \\ \uparrow \quad \downarrow \end{matrix} \xrightarrow{\mathrm{isc}} [{}^\bullet\mathrm{AcH} \cdots {}^\bullet\mathrm{C_6H_{13}}] \tag{3.3}$$

$$\begin{matrix} [{}^\bullet\mathrm{AcH} \cdots {}^\bullet\mathrm{C_6H_{13}}] \\ \uparrow \quad \uparrow \end{matrix} \rightarrow \text{nonradical products} \tag{3.4}$$

$$[{}^\bullet\mathrm{AcH} \cdots {}^\bullet\mathrm{C_6H_{13}}] \rightarrow {}^\bullet\mathrm{AcH} + {}^\bullet\mathrm{C_6H_{13}} \tag{3.5}$$

$${}^\bullet\mathrm{C_6H_{13}} + \mathrm{O_2} \rightarrow \mathrm{C_6H_{13}OO^\bullet} \tag{3.6}$$

$$\mathrm{C_6H_{13}OO^\bullet} + \mathrm{C_6H_{14}} \rightarrow \mathrm{C_6H_{13}OOH} + {}^\bullet\mathrm{C_6H_{13}} \tag{3.7}$$

$$2\mathrm{C_6H_{13}OO^\bullet} \rightarrow \underset{\text{(ketones)}}{\mathrm{C_6H_{12}O}} + \underset{\text{(alcohols)}}{\mathrm{C_6H_{13}OH}} + \mathrm{O_2} \tag{3.8}$$

$$2\mathrm{C_6H_{13}OO^\bullet} \rightarrow \underset{\text{(ketones)}}{2\mathrm{C_6H_{12}O}} + \mathrm{H_2O_2} \tag{3.9}$$

$$2\mathrm{C_6H_{13}} \rightarrow \text{nonradical products} \tag{3.10}$$

$$2\mathrm{C_6H_{13}OO^\bullet} + \mathrm{AcH} \rightarrow \mathrm{C_6H_{13}OOH} + \mathrm{Ac} \tag{3.11}$$

$$2{}^\bullet\mathrm{AcH} \rightarrow \mathrm{Ac} + (\mathrm{CH_3})_2\mathrm{CHOH} \tag{3.12}$$

$C_6H_{13} = CH_3CH_2CH_2{}^{\bullet}CHCH_2CH_3$, or $CH_3CH_2CH_2CH_2{}^{\bullet}CHCH_3$
$Ac = CH_3COCH_3$ $\quad {}^{\bullet}AcH = (CH_3)_2{}^{\bullet}COH$

The curves describing accumulation (concentration growth) of hexanol-2, hexanol-3, hexanon-2, and hexanon-3 as a function of the UV-exposure duration were observed to be linear for each of the products for almost the entire duration of the experiment, as shown in Fig. 4.

This fact indicates that these substances are the initial products of the photolysis. The numerical values of the slopes of the linear regions of the curves in Fig. 4 give the corresponding concentration growth rates for each of the products.

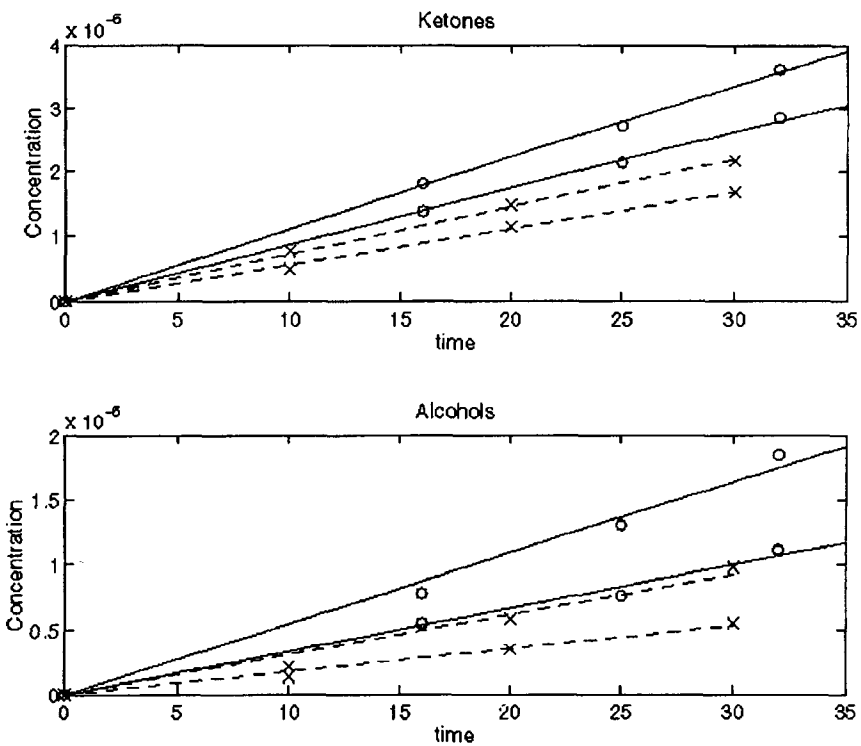


Fig. 4. Dependence of concentration growth on the UV exposure duration. The slopes of these lines are corresponding concentration growth rates in $B = 0.0\,\mathrm{T}$ (solid lines) and $B = 0.2\,\mathrm{T}$ (dashed lines).

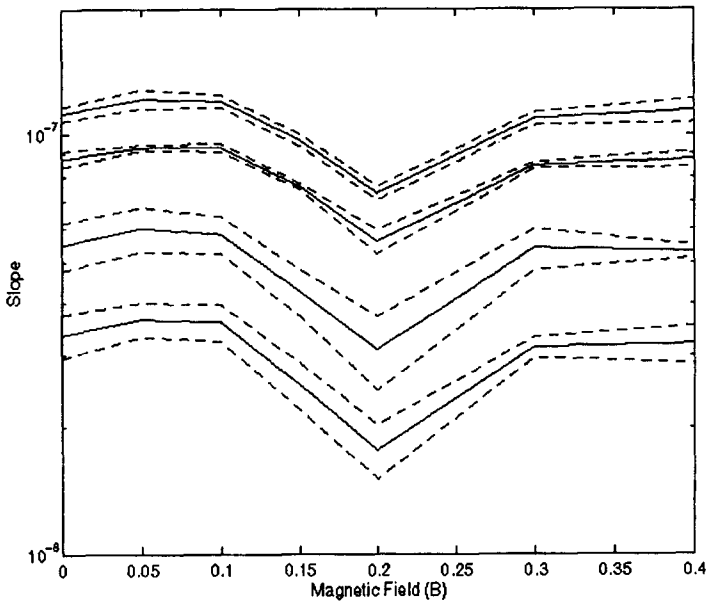


Fig. 5. Slope data for output production of C_3, C_2, C_3OH, C_2OH in hexane reactions versus magnetic field strength (in Tesla). Concentrations are plotted on a log scale, to show relative changes in magnitude.

The results of the experimental studies of the concentration growth rates dependence of the *n*-hexane oxidation products on magnetic field strength are summarized in Fig. 5.

As illustrated in this figure, the application of the magnetic field in the range of 0–0.1 T causes a slight increase in the growth rates, followed by a pronounced growth rates decrease in the range of 0.1–0.2 T. The most significant decrease of the rate of product formation is seen to occur at 0.2 T. Thereafter, the growth rates increase in the interval 0.2–0.3 T, reaching a plateau between 0.3–0.4 T. No experiments were conducted for the field strengths higher than 0.4 T. Thus, the batch experiments show that the process of photo-induced oxidation of hexane is sensitive to the magnetic field. The observed variation of the product growth rates could be explained by the dependence of the free-radical recombination in the "cage" on the magnetic field strength.

The system of differential equations of the photo-induced hexane oxidation obtained on the basis of methods of competitive kinetics is given below and is further referred to as System 3.1.

$dx_1/dt = k_1 u - 8.4 k_2 x_1$	k_1 — quantum yield of $^3Ac^*$ u — dose-rate of the UV-irradiation x_1 — concentration of the excited acetone molecules
$dx_2/dt = 8.4 k_2 x_1 - k_3 x_2 - k_5 x_2$	x_2 — concentration of the triplet radical pairs $[^\bullet AcH \cdots {}^\bullet C_6H_{13}]$
$dx_3/dt = k_3 x_2 - k_4 x_3$	x_3 — concentration of singlet radical pairs $[^\bullet AcH \cdots {}^\bullet C_6H_{13}]$
$dx_4/dt = k_5 x_2 + 8.4 k_7 x_6 - k_6 x_4 \cdot 10^{-2} - 2k_{10} x_4^2$	x_4 — concentration of $^\bullet C_6H_{13}$ radicals
$dx_5/dt = k_5 x_2 - 2k_{12} x_5^2 - k_{11} x_6 x_5$	x_5 — concentration of $^\bullet AcH$ radicals
$dx_6/dt = k_6 x_4 \cdot 10^{-2} - 2(k_8 + k_9) x_6^2 - k_{11} x_6 x_5$	x_6 — concentration of $C_6H_{13}OO^\bullet$ radicals
$dx_7/dt = 8.4 k_7 x_6 + k_{11} x_6 x_5$	x_7 — concentration of $C_6H_{13}OOH$
$dx_8/dt = k_9 x_6^2 + 2k_9 x_6^2$	x_8 — concentration of ketones
$dx_9/dt = k_8 x_6^2$	x_9 — concentration of alcohols

System 3.1.

The last three equations do not include the terms in the right hand side that describe the saturation, or equilibrium, of the last three products, a long term behavior observed experimentally after about thirty minutes. Analysis of the above equations reveals that after the ultraviolet light excitation is turned on, the concentrations x_1, x_2, and x_3 evolve very fast and quickly settle into steady states, which appear in the other slower equations as constant values. The next three equations, governing the production of x_4, x_5, and x_6 involve a more complicated combination of fast and slow dynamics, but these variables also eventually settle into stable steady state values, with a constant concentration of products in the solution versus time. Finally, the last three equations depend only on the previous quantities. The derivatives of the concentrations, therefore, approach positive constants within a short time period. This is consistent with the almost linear concentration growth behavior observed in the batch experiments.

These equations also provide the basis for the understanding of dynamics of the flow-through test with stirring. In this test the solution undergoes the ultraviolet light irradiation during specific fixed time interval. The concentrations of products x_7, x_8, and x_9 increase in the presence of the excited

acetone, and when the solution passes through the test setup, and out of the influence of the light, the reactions stop, and the concentrations remain at their final values. These final values depend on the rates of the production (the slopes in the batch test) which themselves might depend on magnetic field, the acetone concentration, and the time the solution remains in the influence of the light. In the tests with MF and without it, the input and output flow rates are identical and constant resulting in the same time of light irradiation, the irradiation parameters are identical, and the time patterns of the change in the input acetone concentration are identical as well, therefore the only variation from the test with MF to the test without it will be in the rates of production. Consequently, any differences in the final concentrations of products will be directly related to the changes in the slopes of their production. This implies that the detection of the MF effects can be carried out via comparison of the input-output relation identified from the data of the test with MF and that obtained from zero MF test data. If this comparison yields relations that match the ratios of slopes of the concentration growth in the corresponding batch tests, *then the presence or absence of the MF effects is validated by two experiments with completely different data generation and processing methodologies.*

3.3. *Localization of the influence of magnetic field in the system description*

The results of the batch experiments are presented in Fig. 6.

This figure shows the dependence of the concentration growth rate on the value of magnetic field that turns out to be rather pronounced. The differential equations presented above must be modified to include this effect. Since the dynamics of the influence of magnetic field on the free radical cloud is extremely fast in comparison to that of the formation of stable chemical products, it was decided to introduce the MF effects into the model in the form of the direct dependence of certain reaction parameters (coefficients) on the value of the magnetic field strength. Mathematical models of MF-influenced chemical reactions, however, contain a large number of coefficients, and at the outset it is not clear which of them should reflect the MF effects. Therefore, as the first step in model development it was decided to determine parameters which should reflect the MF influence on the overall reaction by experimentally finding the reaction stage most sensitive to the MF irradiation and subsequently introducing the functional dependence on the magnetic field into the coefficients in this stage. Although the available theory suggests that the stage of the competition

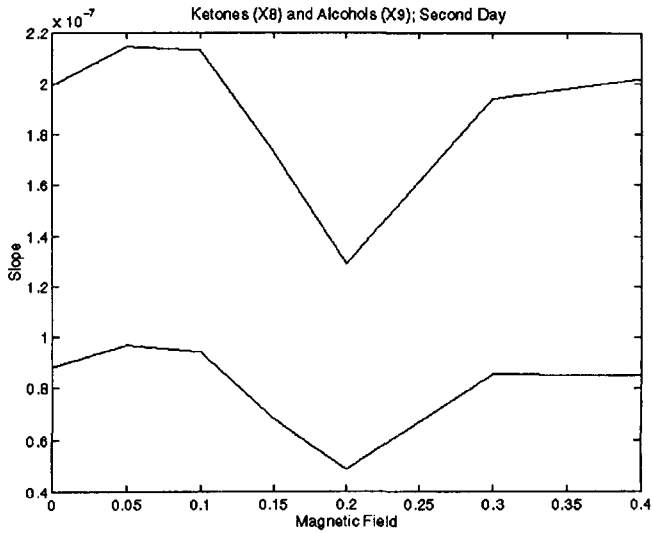


Fig. 6. Slope data for output production of ketones (upper curve) and alcohols in hexane reactions versus magnetic field.

between the "cage" recombination of free radicals with the formation of the "cage" products and the escape of the radicals from the "cage" to the "volume" with subsequent formation of "non-cage" products is the most sensitive to magnetic field, the sensitivity in this stage strongly depends on the method of free radical initiation and should be ascertained for each particular initiation type. Such verification could be most easily done by selecting a test reaction with the initiation type of interest where the "cage" products could be easily detected and analyzed. In the present work all free radical reactions have been initiated by ultraviolet light irradiation with acetone as the sensitizing agent. For this reaction initiation type, the reaction of the photoinduced transformation of iso-propanol in the presence of acetone was chosen as the test reaction with pinacol being the "cage" product. The reasoning behind such choice is as follows. Irradiation of the iso-propanol by UV in the presence of acetone causes the formation of a radical pair which consists of two (CH3)2COH radicals. These two radicals could interact with each other in the reactions of two different types, the reaction of recombination and the reaction of disproportionation. The process of recombination of these two radicals leads to the formation of pinacol. Although, besides "cage", the formation of pinacol could also take

place in the "volume", the experiments on radiolyses of the aqueous solutions of iso-propanol, where two similar radicals of (CH3)2COH are formed, clearly show that in the case when these two radicals escape to the "volume" they usually (10:1) undergo disproportionation with the resulting formation of acetone and iso-propanol [18]. Therefore, pinacol which is the product of photoinduced transformation of iso-propanol in the presence of acetone could be formed mainly as the result of the reaction of recombination of two radicals in the "cage". This product could be easily analyzed during the reaction of photoinduced transformation under the influence of the magnetic field of various strengths. Thus, the data on the rate of the formation of pinacol under the influence of the magnetic field, obtained from this experiment, could point exactly at the field-dependent stage in the photochemically-induced free radical transformations of organic compounds in the presence of acetone as the sensitizing agent and thereby single out the parameter (coefficient) in the mathematical description of the reactions of this type which needs to reflect MF-induced changes in the reaction rate. The process of the photolysis of iso-propanol in the presence of acetone could be described by the following scheme:

$$[\mathrm{Ac}] \xrightarrow{h\nu} {}^3[\mathrm{Ac}]^* \tag{3.13}$$

$$\begin{array}{ccc} {}^3[\mathrm{Ac}]^* & + & (\mathrm{CH_3})_2\mathrm{CHOH} \\ \uparrow & & \uparrow \end{array} \rightarrow \begin{array}{c} \uparrow \quad\quad \uparrow \\ [{}^{\bullet}\mathrm{AcH}\cdots(\mathrm{CH_3})_2{}^{\bullet}\mathrm{COH}] \\ \downarrow \quad\quad \uparrow \end{array} \tag{3.14}$$

$$[{}^{\bullet}\mathrm{AcH}\cdots(\mathrm{CH_3})_2{}^{\bullet}\mathrm{COH}] \xrightarrow{\mathrm{isc}} [{}^{\bullet}\mathrm{AcH}\cdots(\mathrm{CH_3})_2{}^{\bullet}\mathrm{COH}] \tag{3.15}$$

$$[{}^{\bullet}\mathrm{AcH}\cdots(\mathrm{CH_3})_2{}^{\bullet}\mathrm{COH}] \rightarrow \underset{\text{(pinacol)}}{(\mathrm{CH_3})_2\mathrm{C(OH)C(OH)(CH_3)_2}} \tag{3.16}$$

$$[{}^{\bullet}\mathrm{AcH}\cdots(\mathrm{CH_3})_2{}^{\bullet}\mathrm{COH}] \rightarrow {}^{\bullet}\mathrm{AcH} + (\mathrm{CH_3})_2{}^{\bullet}\mathrm{COH} \tag{3.17}$$

$$\underset{\text{(acetone)}}{\mathrm{Ac}} + \underset{\text{(iso-propanol)}}{(\mathrm{CH_3})_2\mathrm{CHOH}} \tag{3.18}$$

$$\mathrm{Ac} = \mathrm{CH_3COCH_3}, \quad {}^{\bullet}\mathrm{AcH} = (\mathrm{CH_3})_2{}^{\bullet}\mathrm{COH}$$

The results of our experimental studies of the dependence of the concentration growth rate of pinacol on magnetic field strength showed that the application of the magnetic field in the range of 0–0.15 T causes an increase in the pinacol growth rate, followed by a pronounced growth rate decrease in the range of 0.3 T. Thereafter, the growth rate increases in the interval 0.3–0.4 T. Thus, the batch experiment shows that magnetic

field with the strength between 0.0–0.4 T could influence the rate of the formation of the products of the recombination of two radicals in the "cage", and that the rate of intersystem crossing between triplet and singlet spin-correlated states in the reactions of photosensitized transformations of organic compounds in the presence of acetone could be field-dependent. This observation, therefore, indicates that the dependence of these reactions on magnetic field could be adequately reflected through the functional dependence of the intersystem crossing rate coefficient in the differential equations of the reaction dynamics on the magnetic field strength.

Although, as indicated in [15], excited singlet acetone can also be involved in hydrogen abstraction reactions, the conclusion of [15] (items 3 and 4) indicates that the efficiency of intramolecular triplet reactions is noticeably higher than that of the singlet hydrogen abstraction and, correspondingly, the efficiency of product formation in the singlet reaction is generally low ($< 25\%$) and may become negligibly small in some cases. For these reasons we consider that the inclusion of reactivity of excited singlet state acetone into the kinetic scheme corresponding to the model presented is not critical for capturing the EMF-induced effects by the model, and, therefore, this reactivity is omitted.

3.4. *Development of the model with dependence on magnetic field*

From the experiment with iso-propanol, it follows that the equation which reflects the change in the concentration of singlet radical pairs x_3, and the "inter-system crossing" constant k_3 which multiplies x_2 in this equation, are the most likely to be affected by a change in magnetic field. By changing k_3 and repeatedly solving numerically differential equations given above it might be possible to find the values of k_3 which yield very close matching between the empirical slopes obtained from the experimental data in the whole range of MF values and those obtained from the numerical solution. This procedure then would generate a function k_3 versus B, thereby explicitly introducing the effect of the magnetic field into the differential equations, and tailoring the behavior of the equations to the results of the experiments.

The complexity and nonlinearity of the differential equations, especially the large difference in the time constants for coupled equations makes this procedure difficult. The following steps were followed to obtain the steady state slopes of the outputs. The first six equations were solved for the equilibrium point with nominal parameter values, i.e. values obtained from

the literature and numerous prior experiments conducted by the authors. These values are:

$$k_1 \cdot u = 5.95 \cdot 10^{-7}; \quad k_2 = 2 \cdot 10^6; \quad k_4 = 2 \cdot 10^9; \quad k_5 = 4 \cdot 10^5;$$
$$k_6 = 10^9; \quad k_7 = 0.53; \quad k_8 = 1.5 \cdot 10^7; \quad k_9 = 8 \cdot 10^6; \quad k_{10} = 1.1 \cdot 10^9;$$
$$k_{11} = 2 \cdot 10^9; \quad k_{12} = 2 \cdot 10^9.$$

The values of K2, K3 are taken from [29]; K4, K6, K11, K12 — from [16, p. 572]; K5 — from [6]; K7 — from [5, p. 25]; K8, K9 — from [22, p. 25]; K10 — from [16, 30, p. 380].

This equilibrium point gave constant rates (slopes of the solution curves) of the generation of the output products in the remaining three equations. Then, by varying k_3, and thereby changing the value of the equilibrium point, a range of possible slopes for the last three equations was obtained. This range turned out to be very narrow for an order of magnitude variation in k_3 ($10^4 < k_3 < 10^6$). To maximize this range, the other model parameters were varied over a range of 10–20%. The constants then were increased or decreased in such a way as to provide a larger range of the output slopes for the same variations in k_3. The new values for the constants are:

$$k_7 = 0.39, \quad k_8 = 2.2 \cdot 10^7, \quad k_9 = 1.4 \cdot 10^7,$$
$$k_{11} = 10^9, \quad k_{12} = 10^9, \quad \text{the rest unchanged.}$$

By changing k_3 in the range 10^4–10^7 in System 3.1 these new values of constants permit a numerical matching of the entire range of experimentally obtained slopes. This permits the model to cover the ratios 1 through 1.5418 for ketones and 1 through 1.8105 for alcohols, encompassing the production rates under no magnetic field as well as all the MF affected rates. This matching of the slopes yields the set of (k_3, B) pairs (see Table 1).

Table 1. Relation between the values of intersystem crossing rate coefficient and MF strength obtained by matching numerical solutions to experimental data.

$B =$	$k_3 =$
0	$9.5026 \cdot 10^4$
$5 \cdot 10^{-2}$	$2.9583 \cdot 10^4$
$1 \cdot 10^{-1}$	$4.1707 \cdot 10^4$
$1.5 \cdot 10^{-1}$	$3.2894 \cdot 10^5$
$2 \cdot 10^{-1}$	$1.6766 \cdot 10^6$
$3 \cdot 10^{-1}$	$1.2045 \cdot 10^5$
$4 \cdot 10^{-1}$	$1.0494 \cdot 10^5$

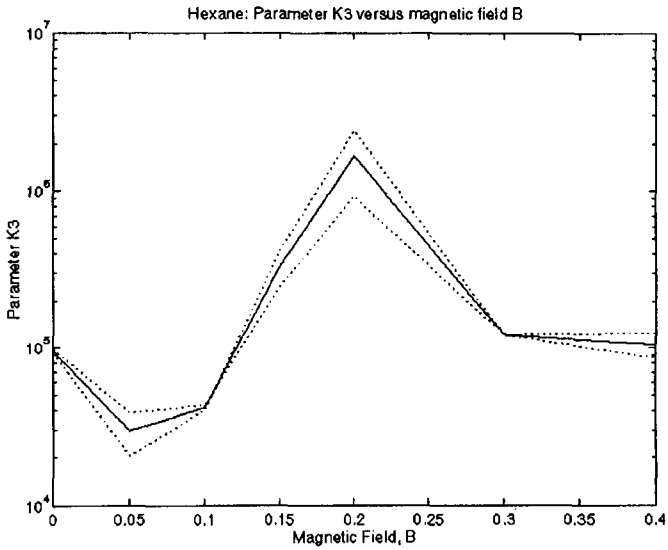


Fig. 7. Nonlinear hexane model development. Relationships (dotted curves) between parameter k_3 and magnetic field, obtained by matching the slopes for ketones and alcohols versus magnetic field data to the same slope values obtained from simulation via adjusting parameter k_3. The solid curve shows the average.

The function k_3 versus B shown in Fig. 7 is obtained by interpolating between the values of this set.

Thus, the complete mathematical model of the hexane oxidation irradiated by the magnetic field in the range 0–0.4 T is given by System 3.1 along with the graphical dependence of k_3 on magnetic field strength shown in Fig. 7. The accuracy of this model in terms of representing the changes in the final product concentration growth rates as a function of magnetic field is demonstrated in Fig. 8.

3.5. *Procedures for identification of the reaction dynamics under MF influence using the flow-through experimental data*

Reference [7] indicates that flow-through experiments provide a good setting for studying sensitivity of chemical reactions to magnetic field. The purpose of the flow-through experiment with hexane carried out in this work is to use robust system identification methods [9, 10, 12, 19, 20] to independently assess and quantify the influence of the magnetic field on the

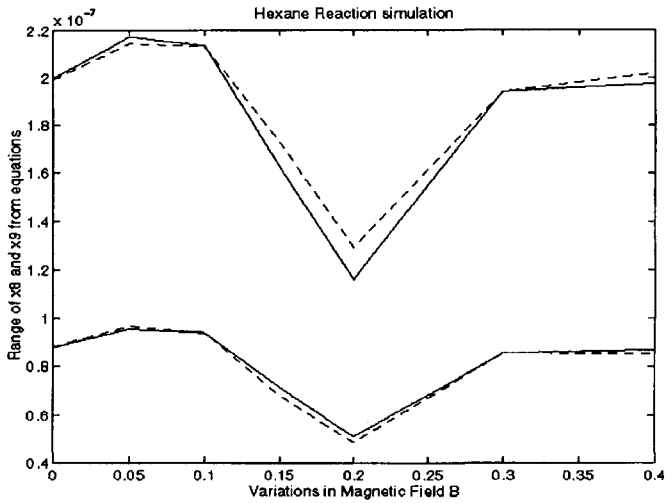


Fig. 8. Simulation of slopes for output production of ketones (upper curve) and alcohols in hexane reactions versus magnetic field. Actual slope data from experiments is indicated by the dotted lines.

flow-through process with mixing for a broad range of the oxidizer concentrations at the value of magnetic field strength $B = 0.2\,\mathrm{T}$ where the batch process displayed the highest sensitivity and thereby partially validate the mathematical model developed on the basis of the batch experiments at the point with the highest MF effect.

In the flow-through test, the input acetone concentration is modulated, resulting in the experimental output acetone concentration given in Fig. 9, the computed output excited acetone concentration given in Fig. 10, and the output product concentration with magnetic field on ($B = 0.2\,\mathrm{T}$) and off, given in Figs. 11 and 12. The excited acetone concentration is computed from the measured output acetone by interpolating the data in Table 2. This interpolation is shown in Fig. 13.

The measured acetone concentration data do not lend themselves easily to a convenient identification of a linear model due to a nonlinear relation to the excited acetone concentration. For a plot of the log of acetone concentration versus excited acetone concentration (Fig. 13), a region close to linear is demonstrated for acetone concentrations between 10^{-2} and $5 \cdot 10^{-2}$, but the actual acetone concentration range lies between $2 \cdot 10^{-2}\,\mathrm{mol/l}$ and $2 \cdot 10^{-1}\,\mathrm{mol/l}$, producing the excited acetone levels which are nonlinearly scaled. For this reason, black box 2 in Fig. 2 to be identified will have

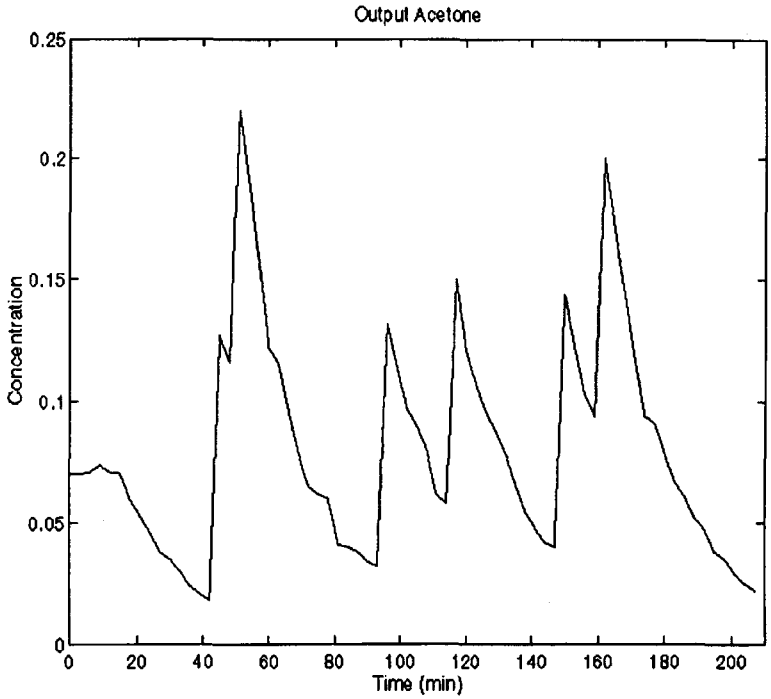


Fig. 9. Output acetone concentration in the flow through experiments.

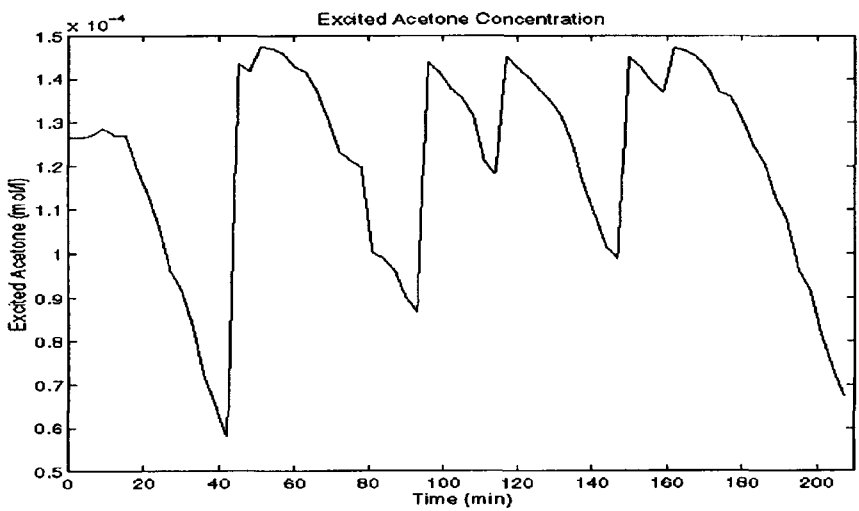


Fig. 10. Estimated excited acetone concentration, computed from the output acetone concentration.

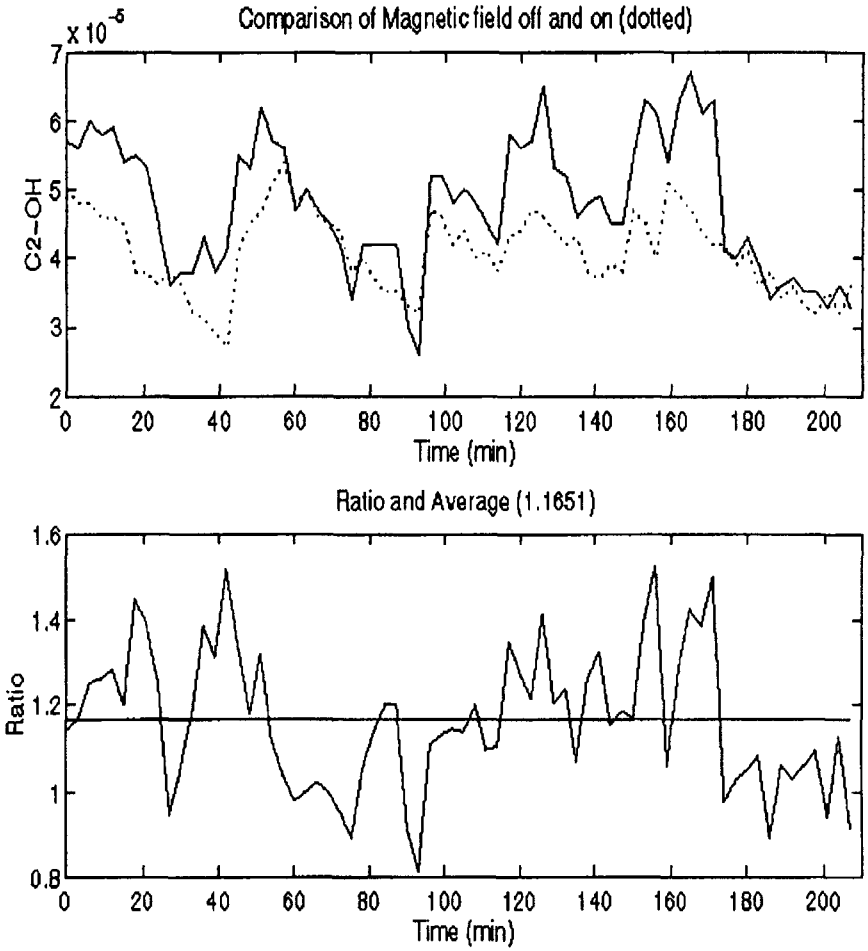


Fig. 11. Time domain data of C_2OH flow-through hexane experiment comparing the cases of magnetic field off, $B = 0.0\,\mathrm{T}$ (solid line) and magnetic field on, $B = 0.2\,\mathrm{T}$ (dotted line). The second plot shows the ratio of the two concentrations as a function of time, and the overall average.

the estimated excited acetone concentration as the input and experimental product (C_2OH and C_3OH) concentration as the output. The linear models of black boxes in Fig. 2, given below by the identified discrete time transfer functions, can be viewed as dynamic relations between input and output obtained via averaging input/output data by statistical and/or frequency domain methods over a broad range of the input acetone concentration values. Black box 1 represents the mixing dynamics of acetone, black

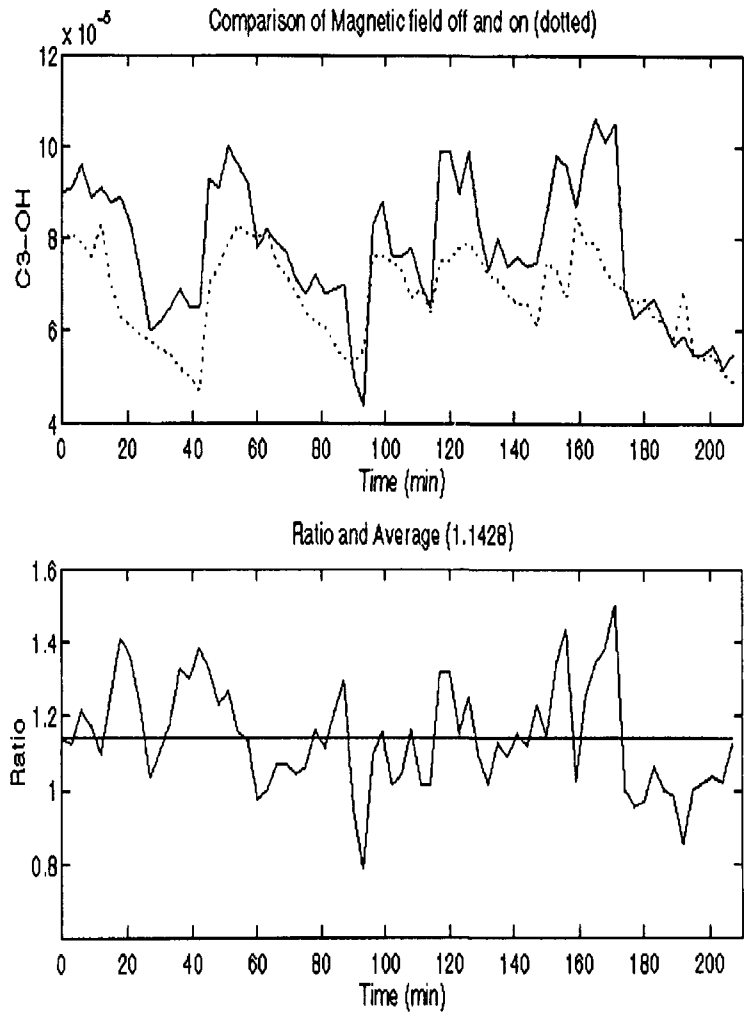


Fig. 12. Time domain data of C3OH flow through hexane experiment comparing the cases of magnetic field off, $B = 0.0\,\mathrm{T}$ (solid line) and magnetic field on, $B = 0.2\,\mathrm{T}$ (dotted line). The second plot shows the ratio of the two concentrations as a function of time, and the overall average.

box 2 — the relation between computed excited acetone concentration and final reaction products, and black box 3 — the transfer function of the whole system including both mixing and reaction dynamics.

Decomposition of box 3 into boxes 1 and 2 does not introduce a significant approximation error due to the fact that the reaction reaches steady

Table 2. Relation between the values of experimental acetone concentration data and computed (estimated) excited acetone concentration.

Acetone:	Excited Acetone	Acetone:	Excited Acetone
10^{-4}	$4.1384 \cdot 10^{-7}$	$3 \cdot 10^{-2}$	$8.3285 \cdot 10^{-5}$
10^{-3}	$4.0349 \cdot 10^{-6}$	$4 \cdot 10^{-2}$	$9.8863 \cdot 10^{-5}$
$2 \cdot 10^{-3}$	$7.9516 \cdot 10^{-6}$	$5 \cdot 10^{-2}$	$1.1067 \cdot 10^{-4}$
$3 \cdot 10^{-3}$	$1.1765 \cdot 10^{-5}$	$6 \cdot 10^{-2}$	$1.1964 \cdot 10^{-4}$
$4 \cdot 10^{-3}$	$1.5460 \cdot 10^{-5}$	$7 \cdot 10^{-2}$	$1.2643 \cdot 10^{-4}$
$5 \cdot 10^{-3}$	$1.9066 \cdot 10^{-5}$	$8 \cdot 10^{-2}$	$1.3160e \cdot 10^{-4}$
$6 \cdot 10^{-3}$	$2.2584 \cdot 10^{-5}$	$9 \cdot 10^{-2}$	$1.3550 \cdot 10^{-4}$
$7 \cdot 10^{-3}$	$2.5998 \cdot 10^{-5}$	10^{-1}	$1.3847 \cdot 10^{-4}$
$8 \cdot 10^{-3}$	$2.9309 \cdot 10^{-5}$	$2 \cdot 10^{-1}$	$1.4721 \cdot 10^{-4}$
$9 \cdot 10^{-3}$	$3.2546 \cdot 10^{-5}$	$3 \cdot 10^{-1}$	$1.477 \cdot 10^{-4}$
$1 \cdot 10^{-2}$	$3.5679 \cdot 10^{-5}$	$4 \cdot 10^{-1}$	$1.477 \cdot 10^{-4}$
$2 \cdot 10^{-2}$	$6.2756 \cdot 10^{-5}$	2	$1.4780 \cdot 10^{-4}$

concentration growth rates in the time scale the order of magnitude faster than that of mixing. Thus, by using the exit stream concentration of the acetone as an input, instead of the concentration of the selected input vial for a particular sample time, the mixing dynamics of the acetone can be effectively eliminated. This permits the identification to focus on the reaction dynamics, which is the primary interest.

Due to the fixed residence time of the flow-through experiment, the ratios to be compared to the constant slope (concentration growth rate) ratios of the batch experiment are those of the output steady state concentrations. The flow-through steady state product concentrations can be obtained from the input/output data as the final values of the unit step responses of the identified linear models with an offset. These final values are, in fact, the steady state gains (i.e. the frequency responses for zero frequency: $\omega = 0$) of these models plus an offset, computed by summing the coefficients of the transfer functions and adding the offset value. For example, for a transfer function of the form:

$$H(z) = [B_0]/[A_0 z + A_1]$$

a steady state gain is $B_0/(A_0 + A_1)$. Here z is understood as either one step advance in time domain: $x(t)z = x(t+1)$, or the complex argument $z = re^{j\omega}$ of z-transform.

Two different forms of models were obtained as a result of the identification procedure. The first one corresponds to a linear model (zero offset). The form of the data fit is

$$A(z)y(t) = B(z)u(t-1).$$

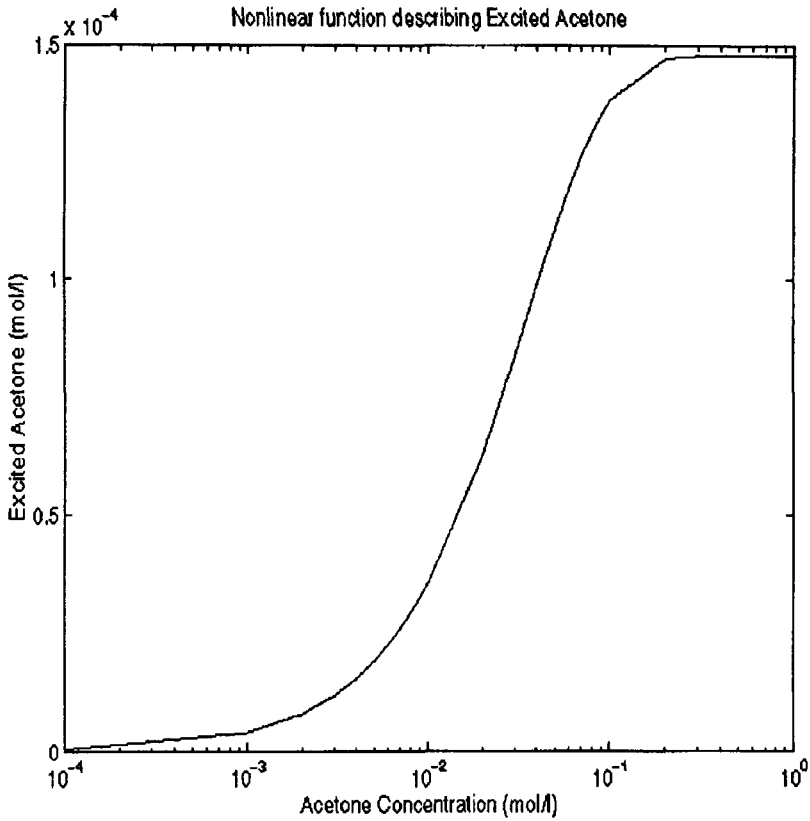


Fig. 13. Plot of function relating excited acetone concentration to acetone concentration.

The second one considers an offset in the data, so that the form of the data fit is

$$A(z)(y(t) - y_0) = B(z)(u(t-1) - u_0).$$

The physical significance of this offset becomes clear after examining the original non-linear input/output relation. Identification reveals that the mixing dynamics between input and output acetone concentrations is nearly linear and first order, yielding $A(z) = A_0 z + A_1$ and $B(z) = B_0$. This is demonstrated by the accuracy of the data fit of the simulated acetone concentration at the output of the model to the experimental data. On the other hand, the relationship between output acetone concentration and

C_2OH concentration and C_3OH concentration is complex. Since this relation is modeled by a linear system, the best fit to the data would include an offset y_0 from the origin. This is demonstrated by a decrease in the residual noise variance for the models with offset. It is interesting to note, though, that both model forms yield reasonably close parameter values, with similar time constants and gains.

Three identification methods have been applied to the experimental data: least squares (LS), empirical transfer function estimate (ETFE), both using MATLAB system identification toolbox, and stochastic H_∞ identification method of [19, 20]. It is interesting to point out that the initial set of data contained a one time step mismatch between the input and the output data, which could be viewed as the plant/model mismatch introduced by the measurements. This mismatch resulted in the failure of the standard LS method to produce any results, while the latter two methods yielded acceptable models. In the subsequent data sets this mismatch was removed, and the LS algorithm started converging to a set of parameter values. Only the results of the third method, however, will be presented below since it yields the process model with the smallest error bound on the unmodelled dynamics.

The latter identification algorithm attempts to estimate the model closest in the H_8 norm to the plant by minimizing the maximal plant/model mismatch as estimated by the standard empirical transfer function estimate (ETFE) of the transfer function from the input sequence $(u(kT))$ to the output error $(e(kT))$. As indicated in [19, 20], by establishing a connection between this minimax problem and a sequence of weighted least squares (WLS) problems, the estimate itself is computed via an iterative sequence of weighted least squares estimates. The weighting filter in this sequence is updated to asymptotically satisfy the H_∞ minimization criterion. The convergence of this procedure under relatively mild assumptions has been proven and computationally supported in [19, 20]. Convergence of the parameter estimates in the present paper is very similar to that of [19, 20].

In the present work, following the approach of [19, 20], iterations of a least-squares minimization algorithm with frequency weighting were carried out, adjusting the weights until the peak of the error frequency spectrum is made as low as possible. The error of the model with an offset for $A_0 = 1$ is given as a one step ahead prediction error by: $e(kT) = y(kT) + A_1[y((k-1)T) - y_0] - B_0[u((k-2)T) - u_0] - y_0$.

The error frequency spectrum is represented by the empirical transfer function estimate from u to e, denoted by $\mathrm{ETFE}(e, u, M)$, where M is the

number of frequency samples. The magnitude plot of this sampled transfer function from u to e is the function to be minimized in the weighted least squares sense through the selection of the model parameters. It represents the errors which result from modeling uncertainty, rather than random noise injected into the process or measurements, since it shows the frequency domain correlation with the input.

The minimization of the magnitude peak of ETFE(e, u, M), denoted further as L_1, is carried out through the iteratively weighted least squares (IWLS) procedure of [19, 20], as follows. Let at k'th iteration $\theta_k \in \mathbf{R}^m$ and $q_k \in \mathbf{R}^M$ denote vectors of model parameters and weights, respectively, $\theta_{k,i}$, $i = 1, \ldots, m$, and $q_{k,i}$, $i = 1, \ldots, M$, denote their i'th components, and M samples of ETFE(e, u, M) be arranged into M-vector with the complex-valued components, denoted further as $\mathrm{ETFE}(e, u, M)_k : \mathbf{R}^m \to \mathbf{C}^M$. Fix a step size $\alpha \in (0, 1]$ and a weighting exponent β,

(i) initialize counter k and weighting vector q_k as $k = 1$ and

$$q_1 = \left[\frac{1}{\sqrt{M}} \frac{1}{\sqrt{M}} \cdots \frac{1}{\sqrt{M}}\right]^T, \text{ respectively;}$$

repeat:

(ii) find θ_k (i.e. model coefficients B_0 and A_1 and offsets u_0 and y_0 in the present work) that minimize $\|q_k \bullet \mathrm{ETFE}(e, u, M)_k\|_2^2$, where $(\bullet)$ denotes componentwise multiplication (this step represents WLS minimization carried out in the present work by means of Levenberg-Marquardt algorithm);

(iii) update each component of the weighting vector:

$$\tilde{q}_{k+1,i} = \sqrt{q_{k,i}^2\left((1-\alpha) + \alpha\left|\mathrm{ETFE}(e, u, M)_{k,i}\right|^{\beta}\right)}, \quad i = 1, \ldots, M,$$

(iv) normalize weights: $q_{k+1} = \tilde{q}_{k+1}/\|\tilde{q}_{k+1}\|_2$;

(v) increment k;

until $k > k_{\mathrm{MAX}}$ or $\|\mathrm{ETFE}(e, u, M)_k\|_\infty - \sum_{i=1}^M q_{k,i}^2 |\mathrm{ETFE}(e, u, M)_{k,i}| / \sum_{i=1}^M q_{k,i}^2 < \varepsilon$, $\varepsilon > 0$; the values of k_{MAX} and ε in the stopping criteria above are usually obvious from the convergence behavior of the estimates. The $\|f\|_\infty$ norm is the maximum absolute value over the components of f, $\|q\|_2$ is the Euclidean norm of vector q.

Thus, IWLS approach generates a sequence of weights q_k such that a sequence of WLS solutions

$$\theta_k \in \arg\min_{\theta \in R^m} \|q_k \bullet \mathrm{ETFE}(e, u, M)_k\|_2^2$$

converges to the H_∞ solution

$$\theta^* \in \arg\min_{\theta \in R^m} \|\mathrm{ETFE}(e, u, M)\|_\infty.$$

3.6. *Identification results*

The H_∞ identification method was used to identify black box 1 (mixing dynamics from PRBS input acetone concentration to exit stream acetone concentration) and black box 3 (the overall system dynamics from PRBS input acetone concentration to the final product concentrations). Fixing $A_0 = 1$, the identification procedure produced the polynomials $B(z) = B_0 = \text{const.}$ and $A(z) = A_0 z + A_1$ and the offsets y_0 and u_0 as well as the measures of the identification accuracy given by the bound on the unmodelled dynamics, denoted below by L_1. The results of the computations are collected in Table 3.

The relatively small size of the bound on the unmodelled dynamics given in Table 3 indicates that the quality of the identified models is relatively high. The relations between output acetone and final product concentrations were determined taking into account the known nonlinear relation between acetone and excited acetone concentrations. The transfer functions (black box 2 in Fig. 2) are then identified from the excited acetone concentration to the output. Since the dynamics of the oxidation process is extremely fast in comparison to the sampling rate of the experimental data collection, the relation between excited acetone concentration and reaction products could be well described by a constant.

3.7. *Validation of the nonlinear mathematical model and the region of model validity*

The ratios of the output concentrations generated by the linear models computed on the basis of the flow-through experiment data can be compared

Table 3. H_∞ identification results. Input: PRBS acetone concentration, outputs: concentrations of output acetone and alcohols C_2OH and C_3OH.

$B = 0.0$	$B = 0.2$
Acetone, $L_1 = 0.36$ $B_0 = 11.3210$, $A_1 = -0.8421$ $y_0 = 6.4580$, $u_0 = 0.08$	Acetone, $L_1 = 0.15$ $B_0 = 10.2407$, $A_1 = -0.8405$ $y_0 = 6.2620$, $u_0 = 0.08$
C_2OH, $L_1 = 0.03$ $B_0 = 0.1841$, $A_1 = -0.8075$ $y_0 = 0.4480$, $u_0 = 0.08$	C_2OH, $L_1 = 0.012$ $B_0 = 0.1115$, $A_1 = -0.8767$ $y_0 = 0.3882$, $u_0 = 0.08$
C_3OH, $L_1 = 0.039$ $B_0 = 0.2978$, $A_1 = -0.8172$ $y_0 = 0.7373$, $u_0 = 0.08$	C_3OH, $L_1 = 0.016$ $B_0 = 0.1844$, $A_1 = -0.8853$ $y_0 = 0.6383$, $u_0 = 0.08$

to the corresponding ratios of the concentration growth rates generated by the full nonlinear model computed on the basis of the batch experiment data to verify the relative change in the reaction rates corresponding to the change in magnetic field. These ratios for C_2OH and C_3OH production for the cases of MF strength $B = 0$ and $B = 0.2$ are computed using the output concentration of the model, y. Given an input, u, the steady state output of the models obtained via H_∞ identification method and presented in Table 3 can be computed by the formula:

$$y = [B_0/(1 + A_1)](u - u_0) + y_0.$$

Due to the nonlinearity of the actual process, the following ratios produced by the linear models are evaluated at the minimum, mean, and maximum input values, $u_0 = 0.006, 0.08, 1.0$, respectively.

The ratios produced by the H_∞ identification are as follows:

$$C_2OH: \quad \frac{\frac{0.1841}{(1-0.8075)}(u - 0.08) + 0.448}{\frac{0.1115}{(1-0.8767)}(u - 0.08) + 0.388} = 1.13 \text{ to } 1.41,$$

$$C_3OH: \quad \frac{\frac{0.2978}{(1-0.8172)}(u - 0.08) + 0.737}{\frac{0.1844}{(1-0.8853)}(u - 0.08) + 0.638} = 1.14 \text{ to } 1.32.$$

The error bounds of the H_∞ identification are sufficiently small to ensure high confidence in the identification results. The above ratios are also consistent with the batch experiment data and therefore they indeed validate the nonlinear reaction model of the photosensitized free radical hexane oxidation under the influence of magnetic field developed on the basis of the batch experiment. As shown in Figs. 6–13, the experimentally supported range of model validity is as follows: (a) magnetic field strength: 0.0 Tesla–0.4 Tesla, (b) excited acetone concentration: 10^{-3} mol/l–$5 \cdot 10^{-1}$ mol/l, (c) cumulative alcohols concentration: $4 \cdot 10^{-8}$ mol/l–$2 \cdot 10^{-4}$ mol/l.

4. Modeling and Identification of the MF-Influenced Oxidation of Linolenic Acid

4.1. *Experimental part*

The experimental study of the magnetic field influence on the photo-induced free-radical oxidation of lipids was carried out using the linolenic acid supplied by Sigma Chemical Co. For the exposure studies the solution of the linolenic acid in chloroform (5.263 mg/ml) in quantity necessary for the preparation of the experiment was placed into a tube, and CHCl3 was

dried by rotary evaporation in argon at room temperature. The substance remaining in the tube was diluted by 0.1 mol Na3PO4 prepared using the bidistilled water. After the one hour dilution the prepared solution of the linolenic acid was used in the experiments.

Three batch experiments were performed. In the first batch experiment the $0.5 \cdot 10^{-2}$ mol/l concentration of linolenic acid was utilized. The acetone at the final concentration of $5 \cdot 10^{-2}$ mol/l was added to linolenic acid contained in the test tubes immediately before irradiation. The reaction vessel was placed within the coils of an electromagnet, capable of providing varying magnetic fields and was held continuously under ultraviolet light with 260–280 nm wavelength to provide a constant energy source for the oxidation reaction. The exit stream was sampled and analyzed. In the second batch experiment the linolenic acid with the concentration of 10^{-2} mol/l was utilized. The experimental procedure was identical to that of the first experiment. In the third batch experiment the linolenic acid and the acetone with the concentrations of 10^{-2} mol/l and 10^{-1} mol/l respectively, were utilized. The experimental procedure differed from that of the first experiment only in the irradiation duration and the final value of the MF strength set here at 30 min. and 0.2 T, respectively.

4.2. *Reaction scheme and differential equations, describing the process of photo-induced oxidation of linolenic acid*

The effect of the magnetic field on the photo-induced peroxidation of linolenic acid was studied by measuring the concentration of malonaldehyde (MDA) and the diene conjugates (DC). The formation of these products occurs according to the following scheme:

$$\begin{array}{lcll} \mathrm{Ac} & \xrightarrow{h\nu} & {}^3[\mathrm{Ac}]^* & (4.1) \\ & & \uparrow \quad \uparrow & \\ {}^3[\mathrm{Ac}]^* + \mathrm{LH} & \rightarrow & [{}^\bullet\mathrm{L}_1 \ldots {}^\bullet\mathrm{AcH}] & (4.2) \\ \uparrow \quad \uparrow & & \uparrow \quad \uparrow & \\ [{}^\bullet\mathrm{L}_1 \ldots {}^\bullet\mathrm{AcH}] & \rightarrow & [{}^\bullet\mathrm{L}_2 \ldots {}^\bullet\mathrm{AcH}] & (4.3) \\ \uparrow \quad \uparrow & & \uparrow \quad \downarrow & \\ [{}^\bullet\mathrm{L}_2 \ldots {}^\bullet\mathrm{AcH}] & \xrightarrow{\mathrm{isc}} & [{}^\bullet\mathrm{L}_2 \ldots {}^\bullet\mathrm{AcH}] & (4.4) \\ \uparrow \quad \downarrow & & & \\ [{}^\bullet\mathrm{L}_2 \ldots {}^\bullet\mathrm{AcH}] & \rightarrow & \text{nonradical products} & (4.5) \\ \uparrow \quad \uparrow & & & \\ [{}^\bullet\mathrm{L}_2 \ldots {}^\bullet\mathrm{AcH}] & \rightarrow & {}^\bullet\mathrm{L}_2 + {}^\bullet\mathrm{AcH} & (4.6) \end{array}$$

$$^{\bullet}L_2 + O_2 \rightarrow L_2OO^{\bullet} \quad (4.7)$$

$$L_2OO^{\bullet} + LH \rightarrow L_2OOH + {}^{\bullet}L_1 \quad (4.8)$$

$$L_2OO^{\bullet} \rightarrow {}^{\bullet}L_3 \quad (4.9)$$

$$^{\bullet}L_3 + LH \rightarrow L_3H + {}^{\bullet}L_1 \quad (4.10)$$

$$L_3H \xrightarrow{H_2O\,O_2,\,t} \text{MDA} + \text{other products} \quad (4.11)$$

$$2\,L_2OO^{\bullet} \rightarrow L_2OOL_2 + O_2 \quad (4.12)$$

$$^{\bullet}L_2 + L_2OO^{\bullet} \rightarrow L_2OOL_2 \quad (4.13)$$

$$2\,{}^{\bullet}AcH \rightarrow Ac + (CH_3)_2CHOH \quad (4.14)$$

$$^{\bullet}L_1 \rightarrow {}^{\bullet}L_2 \quad (4.15)$$

LH — linolenic acid. $^{\bullet}L_1$ — the radical of the abstraction of a hydrogen atom from a reactive methylene group of the linolenic acid.

$^{\bullet}L_2$ — an alkyl radical of the diene conjugate. $L_2OO^{\bullet}$ — a peroxyl radical of the diene conjugate. $^{\bullet}L_3$ — a cyclic peroxide radical. L_3H — cyclic endoperoxide. L_2OOH — hydroperoxides of the linolenic acid diene conlugates. L_2OOL_2 — peroxides of the diene conjugates.

In the reaction scheme acetone in its triplet state ^{3}[Ac] abstracts a hydrogen atom from fatty acid to generate a triplet radical pair, Eq. (4.2). The latter can follow two pathways: (1) intersystem crossing (ISC) to a singlet radical pair which recombines to give non-radical products, Eq. (4.5), or (2) separation, Eq. (4.6) — to produce radicals that are available for reaction with other molecules. The escaped radicals could react with oxygen to form peroxyl radicals ($LOO^{\bullet}$) and the propagation of lipid peroxidation proceeds by a well-known mechanism.

As shown in the scheme, during the photoinduced oxidation of linolenic acid the process of peroxidation takes place. In our experiments the yield of this peroxidation was measured by quantifying the diene conjugation (for the third batch and flow-through experiments) and malonaldehyde (for all batch experiments).

The results of the batch experiments on the dependence of the linolenic acid oxidation products concentration growth rates on magnetic field strength are summarized in Figs. 14–16. Figure 14 shows the change in the MDA concentration growth rates in the first batch experiment as a function of magnetic field strength.

The data from two test days of the first experiment are plotted by dashed lines with data points of each day distinguished by (o) and (*) symbols, and with the average function drawn with a solid line. As

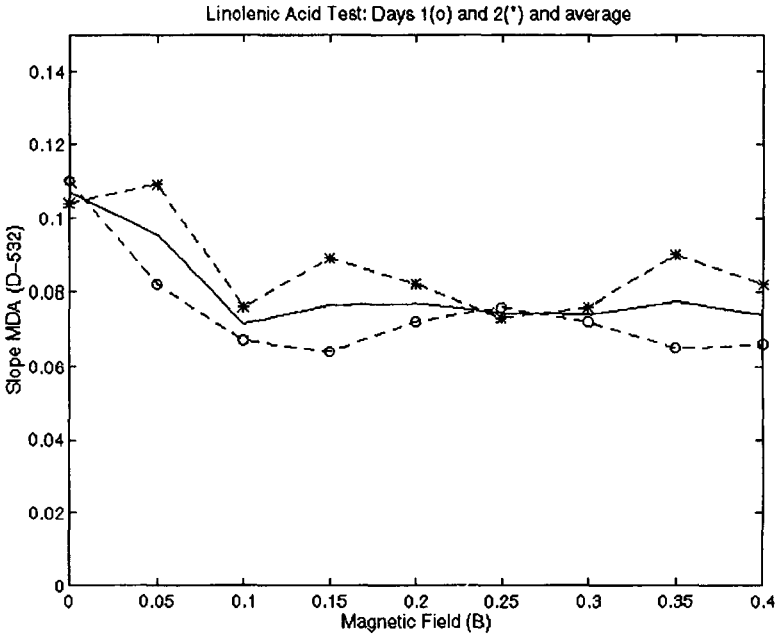


Fig. 14. The values of MDA concentration growth rate in linolenic acid oxidation versus magnetic field; day 1, day 2 and the average of both days. First batch experiment: $C_{LA} = 0.5 * 10^{-2}$ mol/l, uv = 20 min, $C_{AC} = 5 * 10^{-2}$ mol/l.

illustrated in this figure, the application of the magnetic field causes a growth rates decrease in the range of 0.0–0.1 T. Thereafter, the growth rates slightly increase in the interval 0.1–0.25 T, reaching a plateau between 0.25 T and 0.4 T. No experiments were conducted for the field strengths higher than 0.4 T.

The results of the second batch experiment are shown in Fig. 15.

As illustrated in this figure, the application of the magnetic field causes an MDA concentration growth rates increase in the range of 0.0–0.15 T. Thereafter, a slight decrease is observed in the interval 0.15–0.20 T, followed by a significant decrease in the range of 0.25–0.35 T and a plateau between 0.35 and 0.4 T. No experiments were conducted for the field strengths higher than 0.4 T.

The results of the third batch experiment are shown in Fig. 16.

The upper plot shows that the application of the magnetic field causes a slight decrease of diene conjugates production rate in the range of 0.0–0.06 T, followed by a slight rates increase in the range of 0.06–0.1 T, reaching

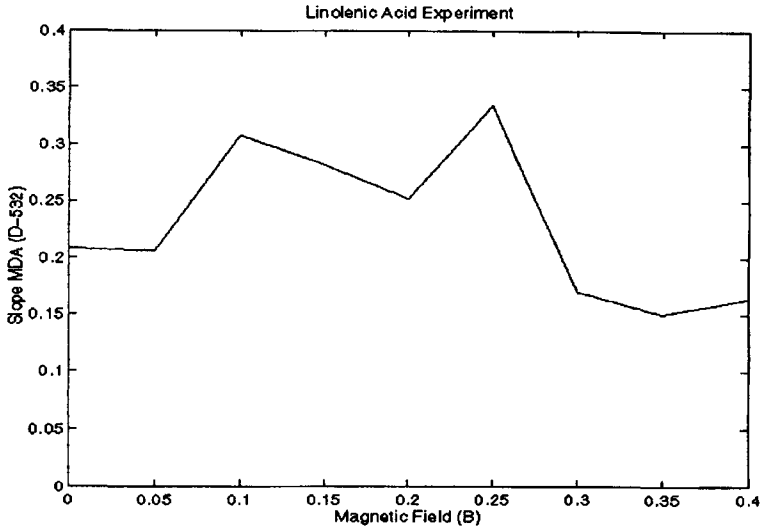


Fig. 15. The values of MDA concentration growth rate in linolenic acid oxidation versus magnetic field. Second batch experiment: $C_{LA} = 10^{-2}$ mol/l, uv = 20 min, $C_{AC} = 5 * 10^{-2}$ mol/l.

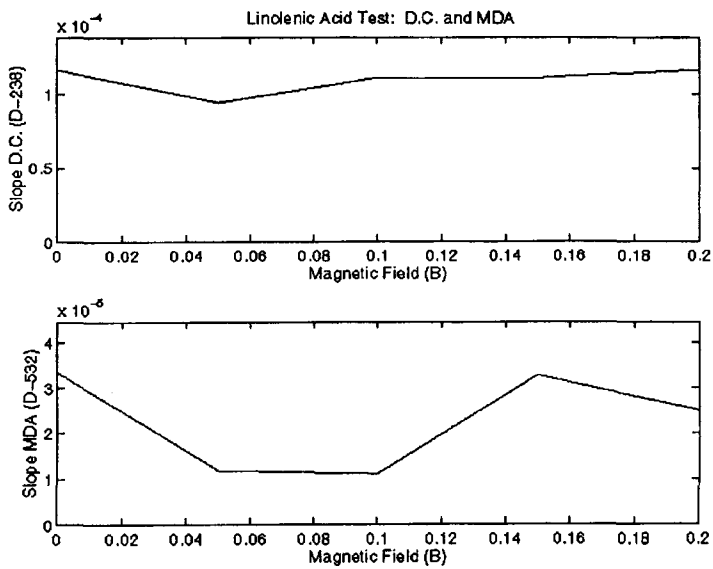


Fig. 16. Concentration growth rates of diene conjugates (upper plot) and MDA in linolenic acid oxidation versus magnetic field. Third batch experiment: $C_{LA} = 10^{-2}$ mol/l, uv = 30 min, $C_{AC} = 0.1$ mol/l.

the plateau between 0.1–0.2 T. The lower plot shows that the application of the magnetic field causes a significant growth rates decrease of MDA production in the range of 0.0–0.06 T with the plateau reached between 0.06 and 0.1 T. Thereafter, the MDA production rates increase in the interval 0.1–0.16 T and then slightly decrease in the range 0.16–0.2 T. No experiments were conducted for the field strengths higher then 0.2 T.

The results of the flow-through experiment with magnetic field off (solid curve) and on (dotted curve) for diene conjugates are shown in Fig. 17.

The lower plot indicates that the average change in the diene conjugates growth rates caused by magnetic field for the concentration of linolenic acid $0.5 \cdot 10^{-2}$mol/l and magnetic field strength 0.17 T is very small: 1.78%; i.e. within the measurement error, so that further flow-through experiments should be conducted to analyze MDA growth rate change.

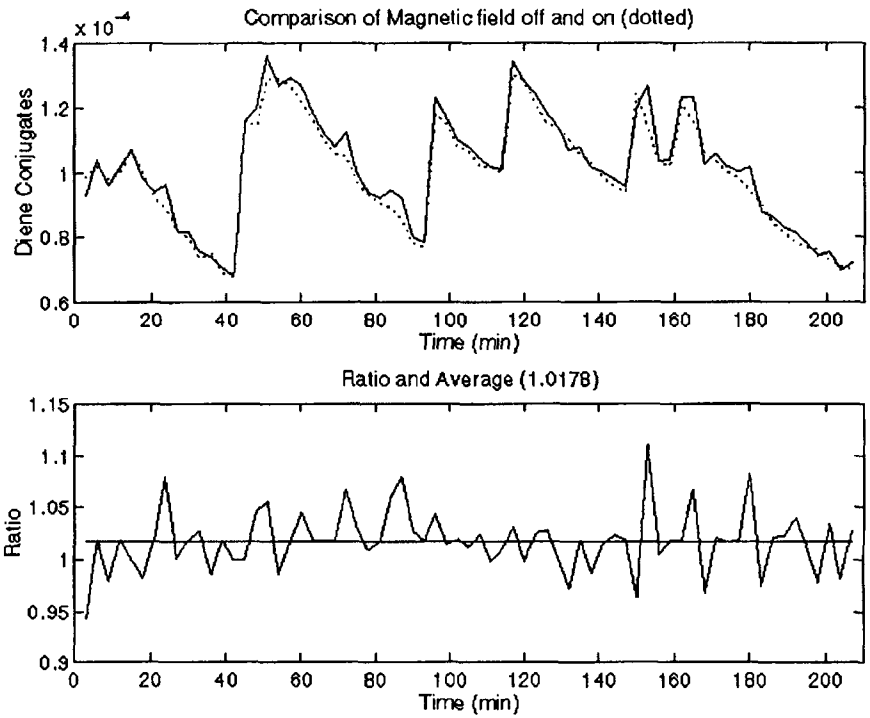


Fig. 17. Time domain data of diene conjugate concentration in the flow through linolenic acid experiment comparing the cases of magnetic field off ($B = 0.0$ T) (solid line) and magnetic field on ($B = 0.17$ T) (dotted line). The second plot shows the ratio of the two concentrations as a function of time, and the overall average ratio.

Our findings for the common trend displayed by the batch experiments can be explained by the following mechanism: application of an external magnetic field during UV-irradiation of a solution of linolenic acid increases the rate of intersystem crossing of triplet radical pairs in several intervals of the field strength as shown in Figs. 14 and 16. This allows fewer radical pairs to dissociate, resulting in lower average concentration of fatty acid free radicals in the solution and lower formation of malonaldehyde. We conjecture that in the heterogeneous structures the processes of the oxidation under the influence of the magnetic field are significantly more complicated then in the homogeneous structures, and they could depend not only on the strength of the magnetic field but also on the concentrations of substrate and sensitizer as well as the duration of the UV-irradiation.

The system of differential equations of the photo-induced oxidation of linolenic acid obtained on the basis of methods of competitive kinetics is given below and is further referred to as System 4.1.

$dx_1/dt = k_1u_1 - k_2x_1x_2$	k_1 — quantum yield ^{3}[Ac]*
$dx_2/dt = -k_2x_1x_2$	x_1 — concentration of excited acetone
$\quad -(k_8x_9 + k_{10}x_{12})x_2$	
$dx_3/dt = k_2x_1x_2 - k_3x_3$	x_2 — concentration of lipid LH
$dx_4/dt = k_3x_3 - (k_4 + k_6)x_4$	x_3 — concentration of triplet radical pairs
	$\uparrow \quad \uparrow$
	[•L$_1$ ... •AcH]
$dx_5/dt = k_4x_4 - k_5x_5$	x_4 — concentration of triplet radical pairs
	$\uparrow \quad \uparrow$
	[•L$_2$ ... •AcH]
$dx_6/dt = k_6x_4 + k_{15}x_{11}$	x_5 — concentration of singlet radical pairs
$\quad - k_7x_6x_8 - k_{13}x_6x_9$	
	$\uparrow \quad \uparrow$
	[•L$_2$ ... •AcH]
$dx_7/dt = k_6x_4 - 2k_{14}x_7^2$	x_6 — concentration of •L$_2$ radicals
$dx_8/dt = k_{12}x_9^2 - k_7x_6x_8$	x_7 — concentration of •AcH radicals
$dx_9/dt = k_7x_6x_8 - (k_8x_2 + k_9$	x_8 — concentration of O_2
$\quad + 2k_{12}x_9 + k_{13}x_6)x_9$	
$dx_{10}/dt = k_8x_9x_2$	x_9 — concentration of L$_2$OO• radicals
$dx_{11}/dt = k_8x_9x_2 + k_{10}x_{12}x_2$	x_{10} — concentration of (L$_2$OOH)
$\quad - k_{15}x_{11}$	
$dx_{12}/dt = k_9x_9 - k_{10}x_{12}x_2$	x_{11} — concentration of •L$_1$ radicals

$dx_{13}/dt = k_{10}x_{12}x_2 - k_{11}x_{13}$ x_{12} — concentration of cyclic peroxide •L_3 radicals

$dx_{14}/dt = k_{11}x_{13}$ x_{13} — concentration of cyclic peroxide L_3H

$dx_{15}/dt = k_{12}x_9^2 + k_{13}x_6x_9$ x_{14} — concentration of MDA

$dx_{16}/dt = k_8x_9x_2 + k_{12}x_9^2 + k_{13}x_6x_9$ x_{15} — concentration of peroxides of diene conjugates.

$y_1 = x_{14}$

$y_2 = x_{10} + x_{15}$ — the measured concentration of diene conjugates in solution.

$k_1u_1 = 5.95 \cdot 10^{-7}$, $k_2 = 2 \cdot 10^6$, $k_3 = 10^{10}$, $k_4 = 10^5$, $k_5 = 10^9$, $k_6 = 4 \cdot 10^5$, $k_7 = 9 \cdot 10^6$, $k_8 = 30$; $k_9 = 10^8$, $k_{10} = 1$, $k_{11} = 10$, $k_{12} = 3 \cdot 10^7$, $k_{13} = 5 \cdot 10^7$, $k_{14} = 2 \cdot 10^9$, $k_{15} = 10^{10}$.

System 4.1.

Analysis of the above equations reveals that the dynamics of the equations is more complicated than that of the case with hexane. The variables x_1, x_2, and x_3 evolve very fast after the ultraviolet light excitation is turned on and settle into steady state values which depend on the other, slower states, especially on x_9 and x_{12}. The equations of System 4.1 do not lend themselves to easy solution by the methods employed earlier. Still, these states will eventually settle into the constant values of concentration versus time. As before, the variables of interest (x_{14} and x_{16}) depend only on the other states, therefore the derivatives of the concentrations (the concentration growth rates) can be shown to quickly converge to positive constants. This is consistent with the almost linear concentration growth observed in the batch experiments.

For the flow through test configuration, these equations are still valid. The test provides only a specific, finite time for the solution to be under the ultraviolet light. The product concentrations increase at some constant rate while they are in the presence of the excited acetone, and when the solution passes through the test setup, and out of the influence of the light, the reactions will stop, and the concentrations will remain at their final values. These final values depend on the rates of the production (the slopes in the batch test) which themselves might depend on magnetic field, the acetone concentration, and the time the solution remains in the influence of the light. In the tests with MF and without it, the input and output flow rates are identical and constant resulting in the same time of light irradiation, the irradiation parameters are identical, and the time patterns of the

change in the input acetone concentration are identical as well, therefore the only variation from the test with MF to the test without it will be in the rates of production. Consequently, any differences in the final concentrations of products will be directly related to the changes in the slopes of their production. This implies that the detection of the MF effects can be carried out via comparison of the input-output relation identified from the data of the test with MF and that obtained from zero MF test data.

4.3. *Identification of the reaction dynamics under MF influence using the flow-through experimental data*

The model form for fitting the experimental data was selected as:

$$y(t) = B(q)/F(q)\{u(t-1) - du\} + C(q)/D(q)e(t) + dy.$$

Three identification methods have been employed in this part of the work: least squares (LS), empirical transfer function estimate (ETFE), both using MATLAB system identification toolbox, and H_∞ identification method described in Sec. 3.5.

The identification process demonstrated the following very important feature of the H_∞ identification method. Discrete models of real processes usually include a one-step measurement delay, which is routinely incorporated into the identification model structure by defining input to be $u(t-1)$, as in the model above. The process variables were sampled in real time but analyzed off-line, with the analysis results recorded with no measurement delay. This specificity of the data record for the linolenic acid did not become clear until after the identification had been carried out, and one step input delay was included into the standard model structure above. The results of the LS identification are presented in Table 4 and Figs. 18 and 19 for model with one step input delay, and Figs. 22 and 24 for input with no delay. The corresponding H_∞ identification results are given in Table 5 and Figs. 20 and 21 for model with one step input delay, and Figs. 23 and 25 for input with no delay.

As seen in Figs. 18 and 19, the LS identification algorithms could not identify the plant model under one step delay data/model mismatch for the experiments with magnetic field on and off. The mismatch resulted in the high variance of the residual noise shown in Table 4 and in the large errors in the estimated frequency response from the input signal to the error signal. The H_∞ identification algorithm, however, managed to adequately identify the plant parameters, and modeled the dynamics faithfully, as seen in Figs. 20 and 21, in spite of the large bound L_1 on modeling uncertainty

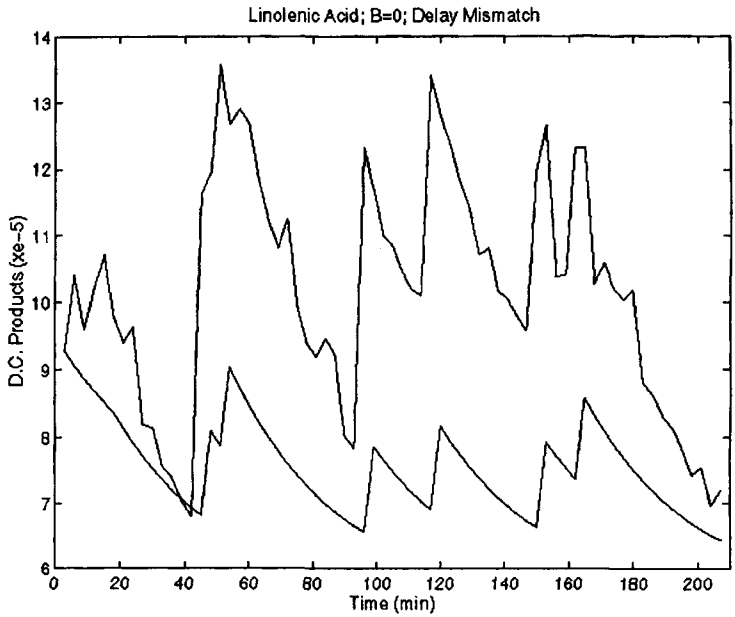


Fig. 18. Linear linolenic acid model identification using output error least squares identification, including a delay mismatch. Measured and simulated (lower curve) diene conjugates concentration versus time. Magnetic field off ($B = 0.0\,\mathrm{T}$).

for the case with measurement delay given in Table 5. Thus, for a one step input delay data/model mismatch, which clearly represents a large unmodeled dynamic perturbation, the H_∞ identification method shows a true strength in yielding a reasonably good model identification, while the LS method fails to do so.

For the case of no delay both methods yield similar results as seen in Figs. 22–25. For this case, the relatively small size of the residual noise variance and the bound on the unmodeled dynamics, given in Tables 4 and 5, respectively, indicate that the quality of the identified models is relatively high.

The relations between output acetone and final product concentration were determined taking into account the nonlinear relation between acetone and excited acetone concentrations given in Table 2 and interpolated in Fig. 13. The transfer functions (black box 2 in Fig. 2) are then identified from the excited acetone concentration to the output concentration of diene conjugates. Since the dynamics of the oxidation process is extremely fast in comparison to the sampling rate of the experimental data collection,

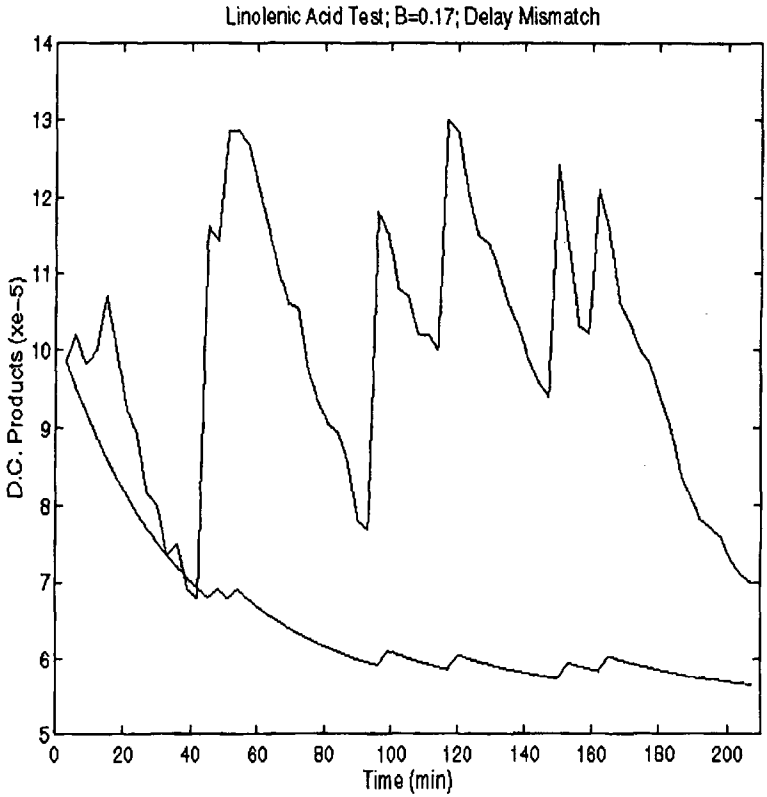


Fig. 19. Linear linolenic acid model identification using output error least squares identification, including a delay mismatch. Measured and simulated (lower curve) diene conjugates concentration versus time. Magnetic field on ($B = 0.17\,\mathrm{T}$).

the relation between excited acetone concentration and reaction products could be well described by a constant. This is indeed confirmed by the ETFE estimate given in Table 6 and presented in Figs. 26 and 27.

As seen in these figures, the ETFE estimates have almost flat magnitude spectra and phase spectra close to zero both with magnetic field on and off. This means that the model structure given earlier is simply a constant plus an offset, whose value depends on that of the magnetic field. Figures 28 and 29 show that by including the nonlinear function of the input into the modeling and using it to reduce input/output nonlinearity, the model output captures the shape of the nonlinear process behavior at the maximal concentration values (the tops of the curves), unlike linear fit in Figs. 22–25.

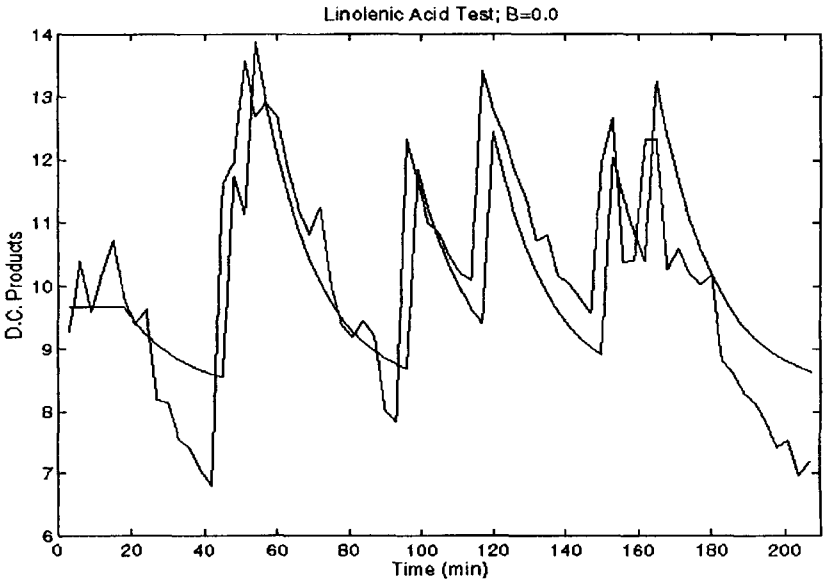


Fig. 20. Linear linolenic acid model identification using output error H-infinity identification, including a delay mismatch. Measured and simulated (smoother curve) diene conjugates concentration versus time. Magnetic field off ($B = 0.0\,\mathrm{T}$).

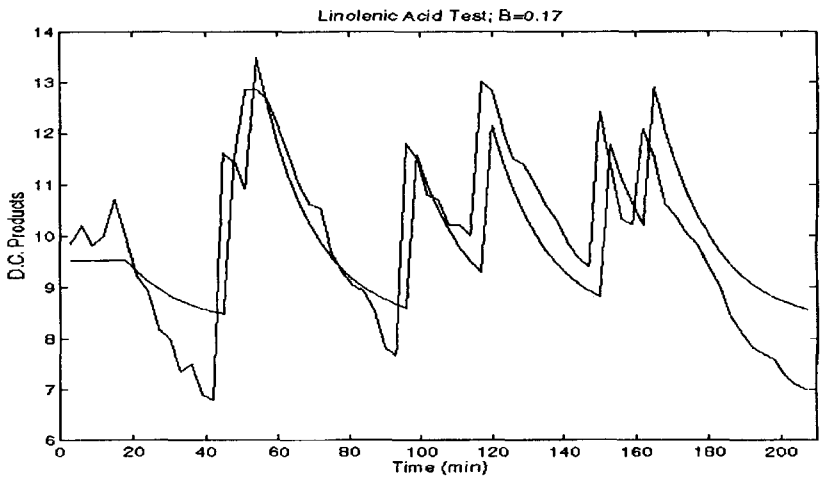


Fig. 21. Linear linolenic acid model identification using output error H-infinity identification, including a delay mismatch. Measured and simulated (smoother curve) diene conjugates concentration versus time. Magnetic field on ($B = 0.17\,\mathrm{T}$).

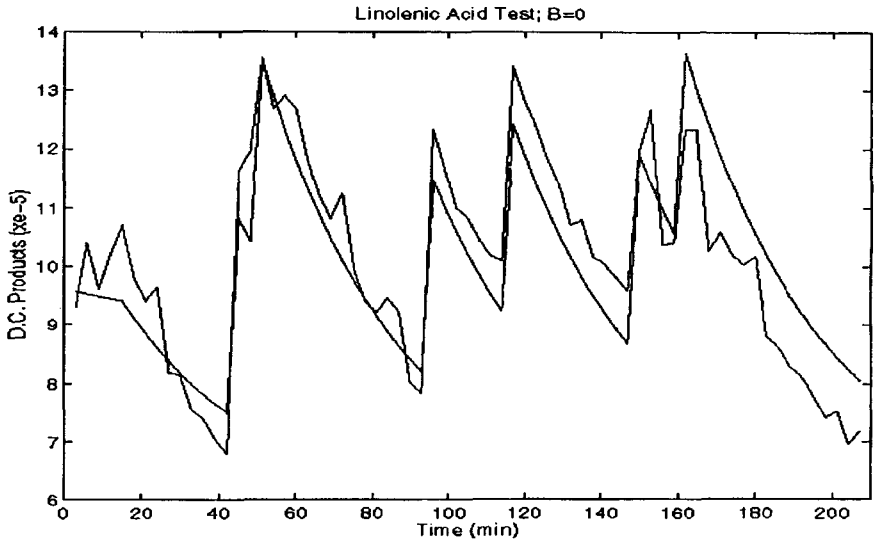


Fig. 22. Linear linolenic acid model identification using output error least squares identification. Measured and simulated (smoother curve) diene conjugates concentration versus time. Magnetic field off ($B = 0.0\,\mathrm{T}$).

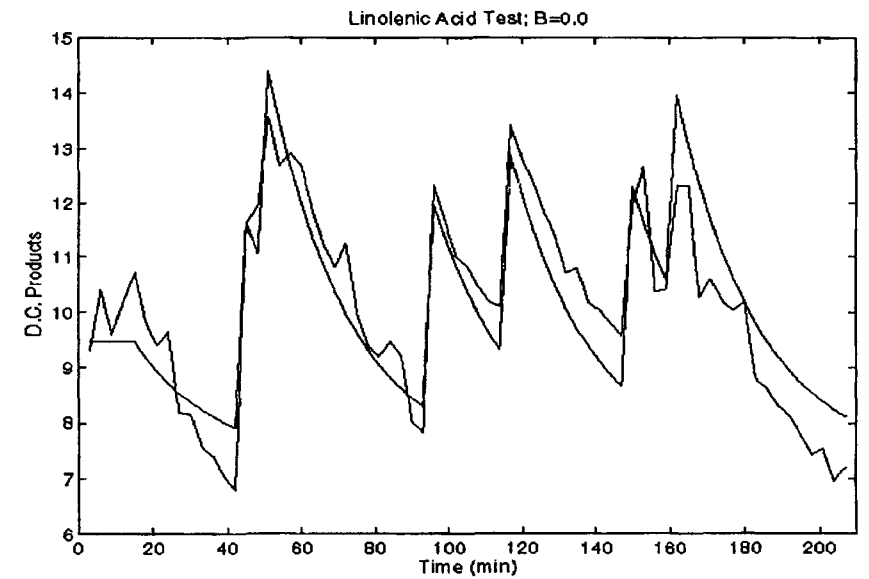


Fig. 23. Linear linolenic acid model identification using output error H-infinity identification. Measured and simulated (smoother curve) diene conjugates concentration versus time. Magnetic field off ($B = 0.0\,\mathrm{T}$).

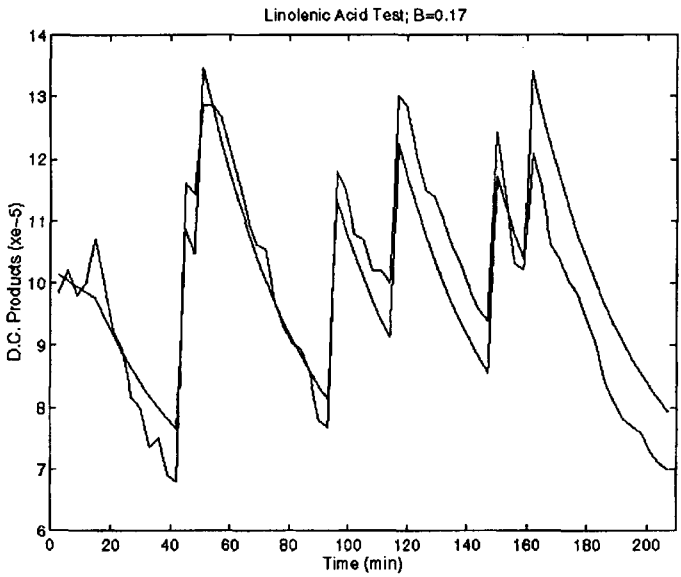


Fig. 24. Linear linolenic acid model identification using output error least squares identification. Measured and simulated (smoother curve) diene conjugates concentration versus time. Magnetic field on ($B = 0.17\,\mathrm{T}$).

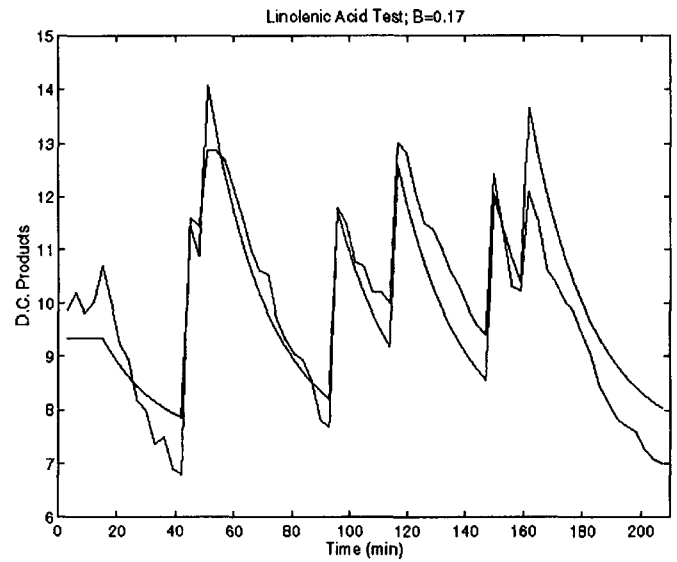


Fig. 25. Linear linolenic acid model identification using output error H-infinity identification. Measured and simulated (smoother curve) diene conjugates concentration versus time. Magnetic field on ($B = 0.17\,\mathrm{T}$).

Identification results

Table 4. Least squares identification results. Input: PRBS acetone concentration, outputs.

	No magnetic field $B = 0.0$	Magnetic field on $B = 0.17$
No measurement delay	DCvar = 0.3017 $B(z) = 3.484\,0$ $F(z) = 1 - 0.9206$ $dy = 5.5$, $du = 0$	DC_Bvar = 0.1662 $B(z) = 3.3938\,0$ $F(z) = 1 - 0.9198$ $dy = 5.5$, $du = 0$
With measurement delay	DCvar = 1.3210 $B(z) = 0\,1.3855$ $F(z) = 1 - 0.9128$ $dy = 5.5$, $du = 0$	DCBvar = 1.1280 $B(z) = 0\,0.2201$ $F(z) = 1 - 0.9150$ $dy = 5.5$, $du = 0$

Table 5. H_∞ identification results. Input: PRBS acetone concentration, outputs: concentrations of diene conjugates.

	No magnetic field $B = 0.0$	Magnetic field on $B = 0.17$
No measurement delay	$L_1 = 0.4087$ $F(z) = 1 - 0.8702$ $B(z) = 3.8297\,0$ $dy = 9.4615$, $du = 0.08$	$L_1 = 0.3158$ $F(z) = 1 - 0.8677$ $B(z) = 3.6804\,0$ $dy = 9.3350$, $du = 0.08$
With measurement delay (mismatch)	$L_1 = 2.9834$ $F(z) = 1 - 0.8242$ $B(z) = 0\,3.2440$ $dy = 9.6765$, $du = 0.08$	$L_1 = 2.8432$ $F(z) = 1 - 0.8226$ $B(z) = 0\,3.0601$ $dy = 9.5394$, $du = 0.08$

Table 6. Identified gains for ETFE in linolenic acid. Input: PRBS acetone concentration, outputs: concentrations of diene conjugates.

$B = 0.0$	$B = 0.17$
DC Gain = 63351 $dy = 10.13$, $du = 1.20 \cdot 10^{-4}$	DC Gain = 62524 $dy = 9.95$, $du = 1.20 \cdot 10^{-4}$

4.4. *Sensitivity of the concentration growth rates to magnetic field strength in the batch and flow-through experiments*

The comparison of the trial with no magnetic field and that with the magnetic field turned on is carried out by computing the ratio of the corresponding slopes of concentration change versus time. The slopes are computed at either a batch data point, or at a point that matches the magnetic field strength of the flow-through experiment.

For the latter case, the magnetic field is an interpolation at $B = 0.17$. These ratios are given below.

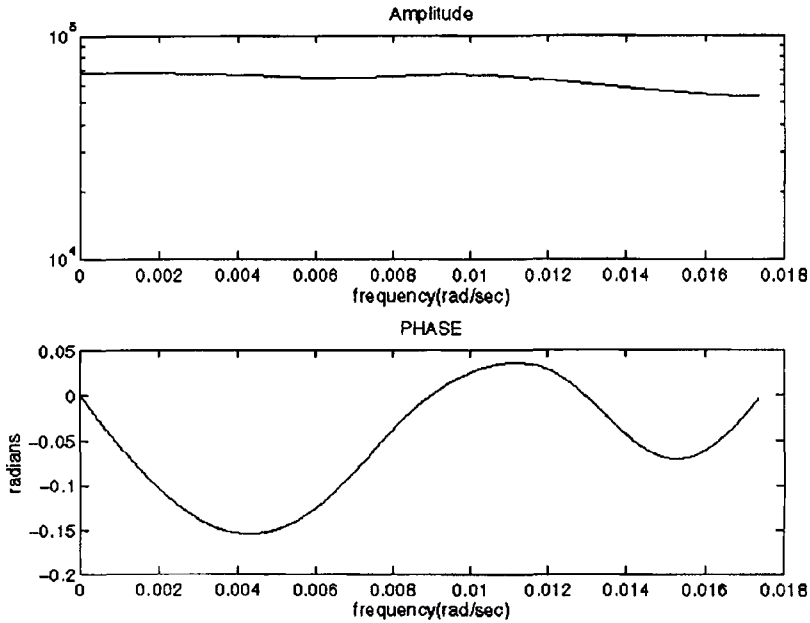


Fig. 26. Empirical transfer function estimate from the input of estimated excited acetone concentration to the output of diene conjugates concentration in linolenic acid reactions, for zero magnetic field. The nearly flat magnitude response indicates a constant relationship between the input and the output.

For the first set of batch experiment data for diene conjugates (cf. Fig. 14),

$$\frac{\mathrm{V[DC]}(B = 0.0)}{\mathrm{V[DC]}(B = 0.17)} = 1.395.$$

For the second set of batch experiment data for MDA (cf. Fig. 15),

$$\frac{\mathrm{V[MDA]}(B = 0.0)}{\mathrm{V[MDA]}(B = 0.25)} = 0.6228.$$

For the third set of batch experiment data for diene conjugates and MDA (cf. Fig. 16),

$$\frac{\mathrm{V[DC]}(B = 0.0)}{\mathrm{V[DC]}(B = 0.17)} = 1.019,$$

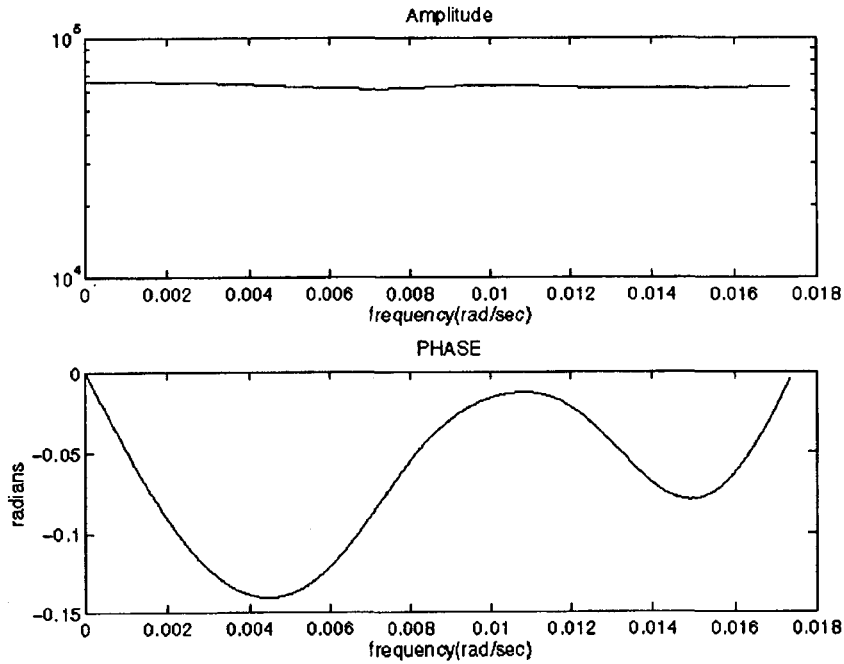


Fig. 27. Empirical transfer function estimate from the input of estimated excited acetone concentration to the output of diene conjugates concentration in linolenic acid reactions, with magnetic field ($B = 0.17\,\mathrm{T}$). The nearly flat magnitude response indicates a constant relationship between the input and the output.

$$\frac{\mathrm{V[MDA]}(B = 0.0)}{\mathrm{V[MDA]}(B = 0.17)} = 1.188.$$

If the ratios are computed at B = 0.1 a more dramatic effect is observed:

$$\frac{\mathrm{V[DC]}(B = 0.0)}{\mathrm{V[DC]}(B = 0.1)} = 1.05,$$
$$\frac{\mathrm{V[MDA]}(B = 0.0)}{\mathrm{V[MDA]}(B = 0.1)} = 3.00.$$

For the flow through test data for diene conjugates, the sensitivity of the oxidation to magnetic field is given by the ratio of the steady state concentration values between the test with no magnetic field and the test with the magnetic field set at $B = 0.17$.

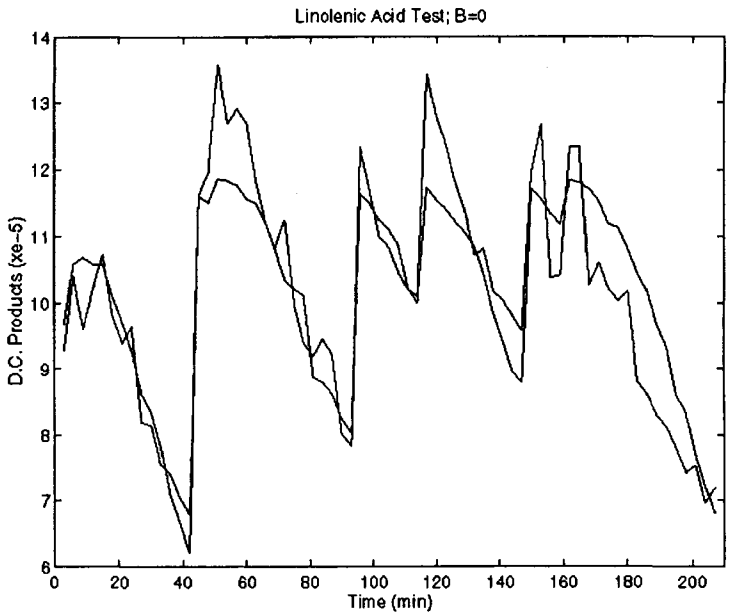


Fig. 28. Empirical transfer function of the constant gain of diene conjugate concentration from the input of estimated excited acetone concentration in linolenic acid reactions. Measured and simulated (lower curve) diene conjugates concentration versus time. Magnetic field off ($B = 0.0\,\mathrm{T}$).

From the Least Squares Identification of Table 4 ($B = 0.0$ and $B = 0.17$):

$$\frac{\frac{3.484}{(1-0.9206)}(u) + 5.5}{\frac{3.3938}{(1-0.9198)}(u) + 5.5} : \text{Ratio} = 1.01 \text{ to } 1.03.$$

From the H_∞ identification of Table 5 ($B = 0.0$ and $B = 0.17$):

$$\frac{\frac{3.8297}{(1-0.8702)}(u - 0.08) + 9.462}{\frac{3.6804}{(1-0.8677}(u - 0.08) + 9.316} : \text{Ratio} = 1.02 \text{ to } 1.05.$$

From the empirical transfer function estimate of Table 6 ($B = 0.0$ and $B = 0.17$):

$$\frac{63351(u - 1.20 \cdot 10^{-4}) + 10.13}{62524(u - 1.20 \cdot 10^{-4}) + 9.95} : \text{Ratio} = 1.01 \text{ to } 1.02.$$

As seen from the above ratios, there is a strong consistency between all three identification methods with the error bounds sufficiently small to

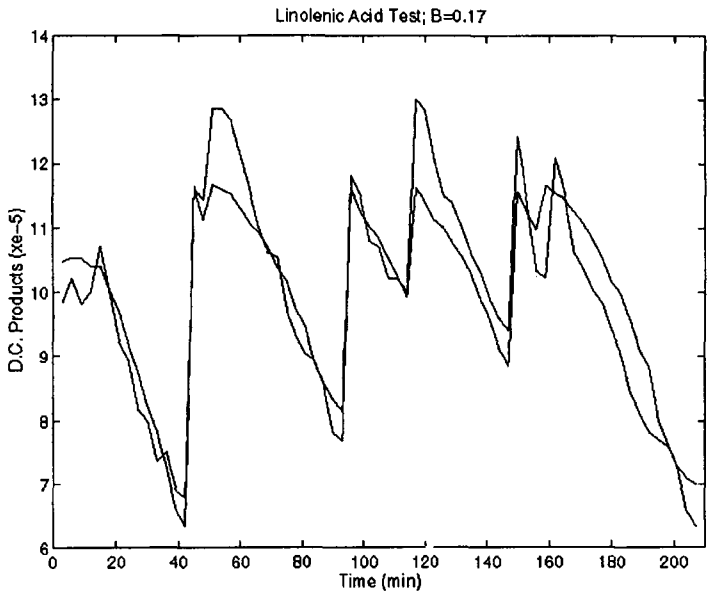


Fig. 29. Empirical transfer function of the constant gain of diene conjugate concentration from the input of estimated excited acetone concentration in linolenic acid reactions. Measured and simulated (lower curve) diene conjugates concentration versus time. Magnetic field on ($B = 0.17\,\mathrm{T}$).

ensure high confidence in the identification results. The above ratios are also consistent with the third batch experiment data indicating that for the chosen acid concentration the sensitivity of the production of diene conjugates to MF irradiation in the photosensitized free radical linolenic acid peroxidation is small and that MDA production sensitivity to MF exposure should be investigated.

4.5. *Development of nonlinear equation constants and dependence on magnetic field*

Finally, the differential equations (presented above) governing the production of the MDA and diene conjugates must be expanded to include the effects of the magnetic field on the dynamics. From analysis of the equations, it looks plausible that the equation which describes the production of the state x_5 and the "inter-system crossing" constant k_4 (which multiplies x_4 in these equations), are the most likely to be affected by the change in

magnetic field. It also looks plausible that System 4.1 is capable of supporting the experimentally observed behavior and yielding the constant slope of the production of x_{14} (MDA) and x_{16} (diene conjugates). By equating the empirical slopes obtained from the experimental data with the slopes obtained from the differential equations, it is possible to describe k_4 as a function of magnetic field, thereby introducing the effect of the magnetic field into the differential equations, and tailoring the behavior of the equations to be consistent with the results of the experiments.

Conclusion

The results of the experiment demonstrate a pronounced dependence of the oxidation of hexane on the strength of the MF irradiation. The methods of control theory permit obtaining a model with capability to predict the effects of MF influence on oxidation of lipid modeling substances and fatty acids. Establishing whether there is any link between the effects reported here and a danger from MF irradiation to humans and animals requires further investigation.

5. Problems

(1) Simulate system 3.1. Investigate the nature of system dynamics. Look for fast convergence onto slow manifold in the state space.
(2) Relate the simulation results to the experimental data given in Sec. 3.
(3) Relate equations of chemical kinetics for hexane oxidation to system of differential equations 3.1.
(4) Using discrete models in Sec. 3, generate input/output data sequences and carry out Least Squares as well as H_∞ identification.
(5) Repeat problems 1–4 for system 4.1.

Acknowledgment

This research was supported by the Electric Power Research Institute and the National Science Foundation grants CMS 20041 and CMS 0324630. Aik Hong Lee is gratefully acknowledged for scanning the original plots and inserting them into the manuscript.

References

[1] S. N. Batchelor, C. W. M. Kay, K. A. McLauchlan and I. A. Shkrob, Time-resolved and modulation methods in the study of the effects of magnetic

fields on the yields of free-radical reactions, *J. Phys. Chem.* **97** (1993) 13250–13258.

[2] J. Bentsman, I. V. Dardynskaia, O. Shadyro, G. Pellegrinetti, R. Blaukamp and G. Gloushonok, Mathematical modeling and stochastic H_∞ identification of the dynamics of the MF-influenced oxidation of hexane, *Math. Biosc.* **169** (2001) 129–151.

[3] C. F. Chignell and R. H. Sik, Magnetic field effects on the photohemolysis of human erythrocytes by ketoprofen and photoporphyrin, *Photochem. and Photobiol.* **62** (1995) 205–207.

[4] E. T. Denisov, Reaction rate constants of homolitical liquid-phase reactions. Moscow: Nauka, 1971 (in Russian).

[5] N. M. Emmanuel, G. E. Zaikov and Z. K. Maizus, Role of the system composition in radical-chain reactions of oxidation of the organic compounds. Moscow: Nauka, 1973 (in Russian).

[6] C. H. Evans, J. C. Sciano and K. U. Ingold, Influence of micellar size on the decay of triplet-derived radical pairs in micelles, *J. Amer. Chem. Soc.* **114** (1992) 140–146.

[7] J. Fakhfakh and J. Bentsman, Experiment with vibrational control of a laser-Illuminated thermochemical system, Transactions of ASME, *J. Dyn. Syst., Meas. Contr.* **112** (1990) 42–47.

[8] C. B. Grissom, Magnetic field effects in biology: a survey of possible mechanisms with emphasis on radical-pair recombination, *Chem. Rev.* **95** (1995) 3–24.

[9] G. Gu and P. P. Khargonekar, Linear and nonlinear algorithms for identification in H_8 with error bounds, IEEE *Trans. Auto. Contr.* **37** (1992) 953–963.

[10] C. A. Hamilton, J. P. Hewitt and K. A. McLauchlan, High resolution studies of the effects of magnetic fields on chemical reactions, *Molecular Phys.* **65** (1988) 423–438.

[11] T. T. Harkins and C. B. Grissom, Magnetic field effects on B_{12} ethanolamine ammonia lyase: evidence for a radical mechanism, *Science* **263** (1994) 958–961.

[12] A. J. Helmicki, C. A. Jacobson and C. N. Nett, Control oriented system identification: a worst case/deterministic approach in H_8, IEEE *Trans. Auto. Contr.* **36** (1991) 1163–1176.

[13] K. A. McLauchlan and U. E. Steiner, The spin-correlated radical pair as a reaction intermediate, *Molecular Radiation Physics* **73** (1991) 241–263.

[14] M. E. Michel-Beyerle, R. Haberkorn, W. Bube, E. Steffens, H. Schroder, H. J. Neusser and E. W. Schlag, Magnetic field modulation of geminate recombination of radical.

[15] W. M. Nau, Pathways for photochemical hydrogen abstraction by n, $\Pi*$-exited states, *Ber. Bunsen-Ges. Phys. Chem.* **102** (1998) 476–485.

[16] D. C. Nonhebel and J. K. Walton (ed) *Free-Radical Chemistry. Structure and Mechanism* (Cambridge University Press, Cambridge), 1974.

[17] N. Periasamy and H. Linschitz, Cage escape and spin rephasing of triplet ion-radical pairs: temperature-viscosity and magnetic field effects

in photoreduction of fluorenone by DABCO, *Chem. Phys. Lett.* **64** (1979) 281–285.

[18] A. K. Picaev, The Modern Radiation Chemistry. Radiolyzis of gases and liquids, Nauka, Moscow, 1986, pp. 362–363.

[19] S. Rangan and W. Ren, Stochastic H_∞ identification: an iteratively weighted least squares algorithm, *Proceedings of the 33rd IEEE Conference on Decision and Control*, Lake Buena Vista, FL, Dec. 14–16, 1994, pp. 3374–3379.

[20] S. Rangan and W. Ren, Stochastic H_8Identification: An Iteratively Weighted Least Squares Algorithm, Memorandum No. UCB/ERL M94/25, (College of Engineering, University of California, Berkeley), 1994.

[21] C. Rice-Evans and R. Burdon, Free radical lipid interactions and their pathological consequences, *Prog. Lipid Research* **32** (1993) 77–110.

[22] V. A. Roginski, Phenol oxidants. Reactivity properties and efficiency. Moscow: Nauka, 1973 (in Russian).

[23] R. Z. Sagdeev, K. M. Salikhov and Y. M. Molin, The influence on the magnetic field on processes involving radicals and triplet molecules in solutions, *Russian Chem. Rev.* **46** (1977) 569–601.

[24] W. Schlenker and U. E. Steiner, An efficient continuous flow technique for investigating the magnetic field dependence of photochemical quantum yields, *Ber. Bunsenges. Phys. Chem.* **89** (1985) 1041–1049.

[25] H. Staerk and K. Razi Naqvi, Magnetic field effects on spin rephasing in a photochemically produced radical pair investigated by a double-pulse technique using a nitrogen laser, *Chem. Phys. Lett.* **50** (1977) 386–388.

[26] H. Staerk and K. Razi Naqvi, Magnetic field effects on spin rephasing in a photochemically produced radical pair investigated by a double-pulse technique using a nitrogen laser, *Chem. Phys. Lett.* **50** (1977) 386–388.

[27] J. Tse, J. Bentsman and N. Miller, Minimax long range parameter estimation, *Proceedings of the 33rd IEEE Conference on Decision and Control*, Lake Buena Vista, FL, Dec. 14–16, 1994, pp. 277–282.

[28] N. J. Turro and B. Kraeutler, Magnetic field and magnetic isotope effects in organic photochemical reactions. A novel probe of reaction mechanisms and methods for enrichment of magnetic isotopes, *Ac. Chem. Res.* **13** (1980) 369–377.

[29] Von cinem Autoronkollectiv. Einfuhrung in die Photochemie. Berlin: VEB Deutcher Verlag der Wissenschaften, 1976.

This page is intentionally left blank

CHAPTER 5

COMPUTER SIMULATION OF SELF REORGANIZATION IN BIOLOGICAL CELLS*

DONALD GREENSPAN

In this paper we describe supercomputer simulations for the self reorganization of tissue which has been separated into endoderm, mesoderm, and ectoderm cells.

Keywords: Morphogenesis; self reorganization; endoderm; mesoderm; ectoderm.

1. Biological, Physical and Computational Preliminaries

1.1. *Introduction*

Steinberg [4] describes several interesting biological experiments in morphogenesis, that is, in the self reorganization of biological cells. For example, Holtfreter showed that embryonic tissue, consisting of distinct endoderm, mesoderm, and ectoderm layers, when separated out, could recombine into tissue with normal endoderm, mesoderm, and ectoderm layers. (See Fig. 1.) As another example, in an experiment by Wilson, cells and cell clusters obtained by squeezing a sponge through a fine silk cloth could reunite and aggregates could reconstruct themselves into functional sponges.

In this paper we will concentrate on a computer simulation of the Holtfreter experiment.

1.2. *Classical molecular mechanics*

For purposes of intuition, it will be important to review, first, how molecules interact. Within a larger body, molecules interact only locally, that is, with their nearest neighbors. This interaction is of the following nature [1]. If two molecules are pushed together they repel, if pulled apart they attract,

*Material in this paper has been adapted from Chapter 6 of PARTICLE MODELING, Birkhauser, Boston, 1997, by Donald Greenspan, and reprint permission has been granted by Springer Science and Business Media.

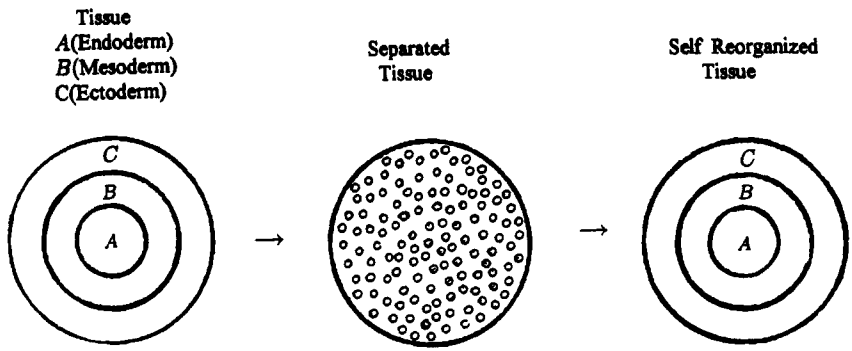


Fig. 1. The Holtfreter experiment.

and mutual repulsion is of a greater order of magnitude than is mutual attraction. Mathematically, this behavior is often formulated as follows. The magnitude F of the force $\vec{F}$ between two molecules which are locally r units apart is of the form

$$F = -\frac{G}{r^p} + \frac{H}{r^q}, \tag{1.1}$$

where, typically, $G > 0$, $H > 0$, $q > p > 6$. The negative term in (1.1) is the attraction term and the positive term is the repulsion term.

The major problem in simulating any physical body is that there are too many component molecules to incorporate into the model. The classical mathematical approach is to replace the large, but finite, number of molecules by an infinite set of points. In doing so, the rich physics of molecular interaction is lost because every point has an infinite number of neighbors which are arbitrarily close. A viable computer alternative is to replace the large number of *molecules* by a much smaller number of *particles* and then to readjust the parameters in (1.1). This is the engineering methodology called the *lumped mass* approach and it is this approach which we will follow.

1.3. *The computer algorithm*

The general idea outlined above will be implemented in the following constructive fashion. Consider N particles, P_i, $i = 1, 2, 3, \ldots, N$. For $\Delta t > 0$, let $t_k = k\Delta t$, $k = 0, 1, 2, 3, \ldots$. For each of $i = 1, 2, 3, \ldots, N$, let m_i denote the adhesive measure of P_i, and, in two dimensions, let P_i at t_k be located at $\vec{r}_{i,k} = (x_{i,k}, y_{i,k})$, have velocity $\vec{v}_{i,k} = (v_{i,k,x}, v_{i,k,y})$, and have acceleration $\vec{a}_{i,k} = (a_{i,k,x}, a_{i,k,y})$. Let position, velocity and acceleration be related

by the recursion formulas [2]:

$$\vec{v}_{i,\frac{1}{2}} = \vec{v}_{i,0} + \frac{1}{2}(\Delta t)\vec{a}_{i,0}, \quad \text{(Starter)} \tag{1.2}$$

$$\vec{v}_{i,k+\frac{1}{2}} = \vec{v}_{i,k-\frac{1}{2}} + (\Delta t)\vec{a}_{i,k}, \quad k = 1,2,3,\ldots \tag{1.3}$$

$$\vec{r}_{i,k+1} = \vec{r}_{i,k} + (\Delta t)\vec{v}_{i,k+\frac{1}{2}}, \quad k = 0,1,2,3,\ldots. \tag{1.4}$$

Formulas (1.2)–(1.4) are the popular *leap frog* formulas, which are computationally most convenient when the number of particle N is very large. At t_k, let the force acting on P_i be $\vec{F}_{i,k}$. We relate force and acceleration by the dynamical equation

$$\vec{F}_{i,k} = m_i\vec{a}_{i,k}. \tag{1.5}$$

As soon as the precise structure of $\vec{F}_{i,k}$ is given, the motion of each P_i will be determined explicitly and recursively by (1.2)–(1.5) from given initial data. The force $\vec{F}_{i,k}$ is now described as follows. Let $\vec{r}_{ij,k}$ be the vector from P_i to P_j at time t_k, so that $r_{ij,k}$, the distance between the two particles, is given by $r_{ij,k} = \|\vec{r}_{i,k} - \vec{r}_{j,k}\|$. Then the force $\vec{F}_{ij,k}$ on P_i exerted by P_j at time t_k is assumed to be

$$\vec{F}_{i,k} = \left(\left(-\frac{G_{ij}}{(r_{ij,k})^p} + \frac{H_{ij}}{(r_{ij,k})^q}\right)\frac{\vec{r}_{ji,k}}{r_{ij,k}}\right),$$

in complete analogy with (1.1). The total force $\vec{F}_{i,k}$ on P_i due to all other particles different from P_i is defined by

$$\vec{F}_{i,k} = \sum_{\substack{j=1 \\ j\neq i}}^{N}\left(\left(-\frac{G_{ij}}{(r_{ij,k})^p} + \frac{H_{ij}}{(r_{ij,k})^q}\right)\frac{\vec{r}_{ji,k}}{r_{ij,k}}\right). \tag{1.6}$$

Note finally that the introduction of an additional parameter D is essential to assure that particle interactions are local. We will require that whenever $r_{ij,k} > D$, then (1.6) must be replaced by

$$\vec{F}_{i,k} = \vec{0} \quad (r_{ij,k} > D). \tag{1.7}$$

2. Supercomputer Examples

2.1. *A morphogenesis simulation*

A large number of examples were run on a CRAY YMP/8. We will describe one in detail in this section and then discuss others in the next section.

Since, in general, particles do not adhere when in a gaseous state and are rigid when in a solid state, self reorganization can occur only in a liquid

or near-liquid state. Relative to this observation, previous calculations [2] allow us now to fix the parameters as follows: $\Delta t = 0.0001$, $p = 3$, $q = 5$, $G_{ij} = H_{ij} = 5m_i m_j$, $D = 2.2$. For, then, if P_i is to be a liquid particle, the speed v_i of P_i has been deduced for various adhesive measures m_i [2]. In particular;

$$m_i = 2000 \text{ implies } 100 \leq v_i \leq 170 \tag{2.1}$$

$$m_i = 4000 \text{ implies } 90 \leq v_i \leq 160 \tag{2.2}$$

$$m_i = 10000 \text{ implies } 50 \leq v_i \leq 80. \tag{2.3}$$

Let us now examine a particular example. Consider a square region in the XY plane whose vertices are $(-16,-16)$, $(-16,16)$, $(16,16)$, $(16,-16)$. On this region construct a triangular grid of 1072 points using the recursion formulas

$$\begin{aligned} x(1) &= -15.5, \quad y(1) = -16.0 \\ x(i+1) &= x(i) + 1.0, \quad y(i+1) = -16.0, \quad i = 1, 31 \\ x(33) &= -16.0, \quad y(33) = -15.0 \\ x(i+1) &= x(i) + 1.0, \quad y(i+1) = -15.0, \quad i = 33, 64 \\ x(i) &= x(i-65), \quad y(i) = y(i-65) + 2.0, \quad i = 66, 1072. \end{aligned}$$

This point set is shown in Fig. 2.

We now fix a set A which consists of 38 particles each with adhesive measure 10000, a set B of 266 particles each with adhesive measure 4000, and a set C of 768 particles each with adhesive measure 2000. The particles are distributed at the 1072 points shown in Fig. 2, with no two particles at the same point. A particle at the point $(x(i), y(i))$ is denoted by P_i. In Fig. 3, the A particles, which have the largest adhesive measures, are denoted by circles; the particles of set B, which have the intermediate adhesive measures, are denoted by quadrilaterals; and the particles of set C which have the smallest adhesive measures are denoted by triangles.

Next a velocity is assigned to each particle. In agreement with (2.1)–(2.3), each A particle is assigned a speed of 60 while each of the B and C particles is assigned a speed of 150. The XY direction and the corresponding $(\pm)$ signs of the velocity vectors are determined at random, and the resulting velocity is shown in Fig. 3 as a vector emanating from each particle's center. For a complete listing of all the initial data, see Greenspan [3].

The motion of the system is now determined by (1.2)–(1.7). However, in order to keep the particles within the square while they are in motion,

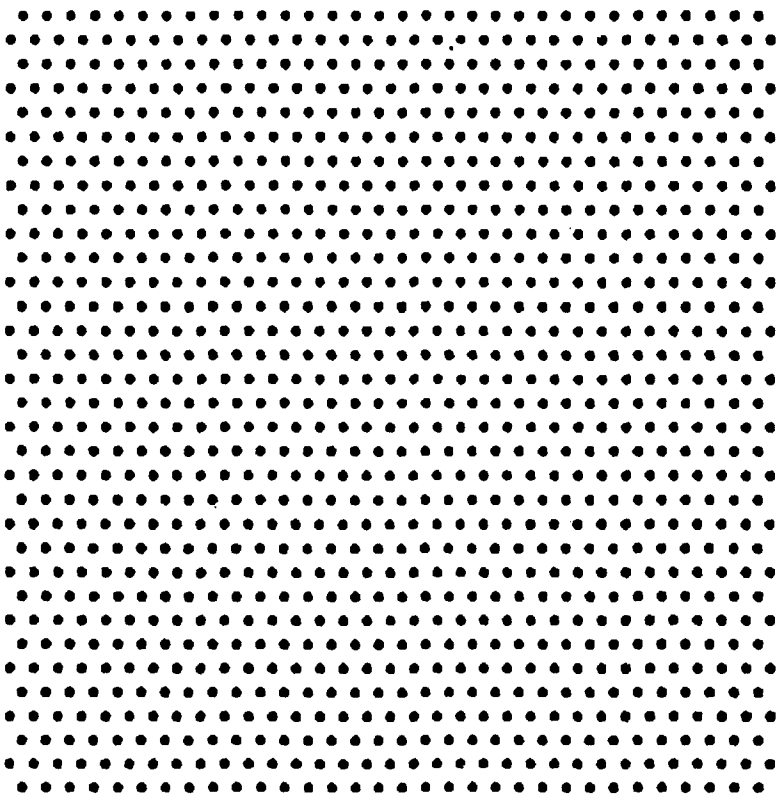


Fig. 2. 1072 points in a 32 cm by 32 cm square.

the following reflection rules are applied:

(a) if $x_i > 16$, reset $x_i \to 32.0 - x_i$, $v_{x,i} \to -0.99 v_{x,i}$, $v_{y,i} \to 0.99 v_{y,i}$
(b) if $x_i < -16$, reset $x_i \to -32.0 - x_i$, $v_{x,i} \to -0.99 v_{x,i}$, $v_{y,i} \to 0.99 v_{y,i}$
(c) if $y_i > 16$, reset $y_i \to 32.0 - y_i$, $v_{x,i} \to 0.99 v_{x,i}$, $v_{y,i} \to -0.99 v_{y,i}$
(d) if $y_i < -16$, reset $y_i \to -32.0 - y_i$, $v_{x,i} \to 0.99 v_{x,i}$, $v_{y,i} \to -0.99 v_{y,i}$

The small velocity damping in rules (a)–(d) insures numerical stability when using the time step $\Delta t = 0.0001$.

The resulting self reorganization is shown in Figs. 4–13. Figures 4–9 show the self reorganization of the A cells at $T = 1.5$, 9.0, 16.5, 24.0, 31.5, 39.0. Notice that these cells first reorganize into small groups which then converge to form a central core. The self reorganization of the B cells at

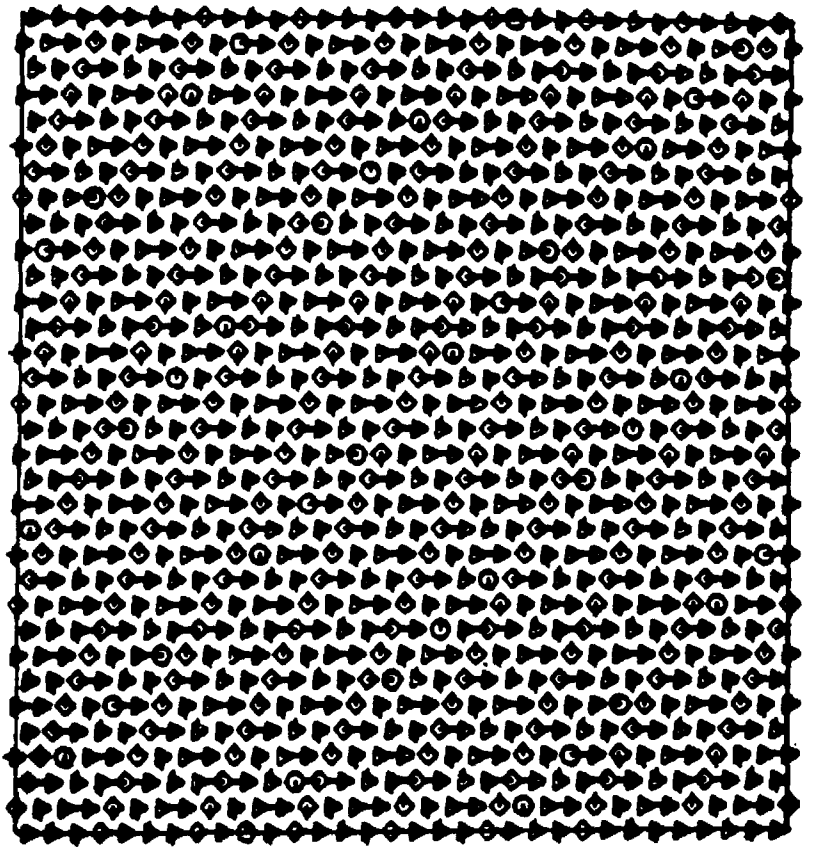


Fig. 3. The initial data.

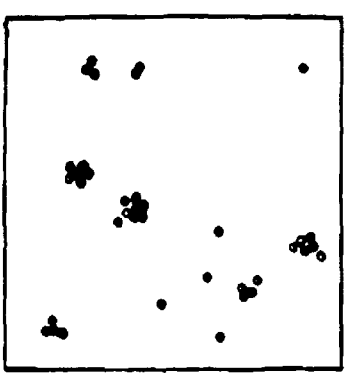


Fig. 4. $T = 1.5$.

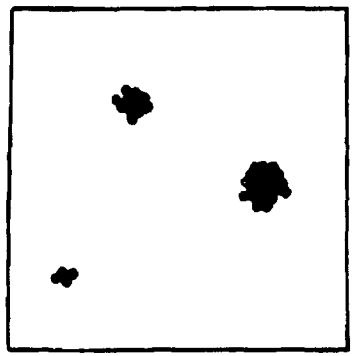


Fig. 5. $T = 9.0$.

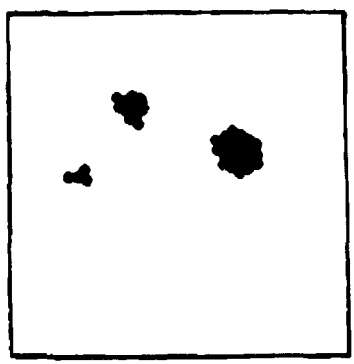


Fig. 6. $T = 16.5$.

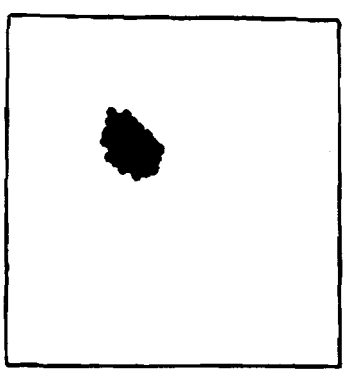


Fig. 7. $T = 24.0$.

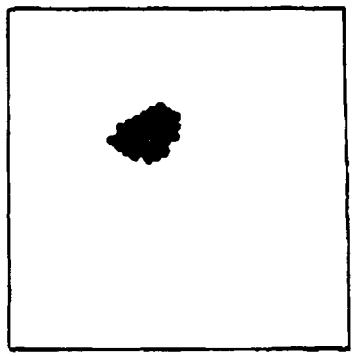


Fig. 8. $T = 31.5$.

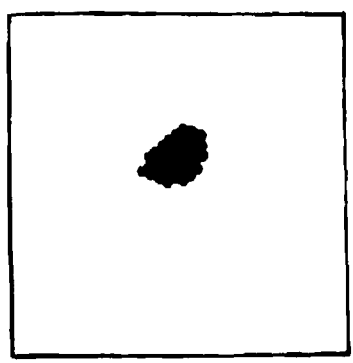


Fig. 9. $T = 39.0$.

Fig. 10. $T = 24.0$.

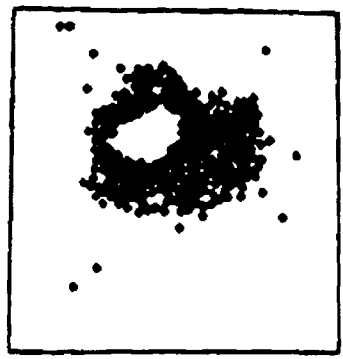


Fig. 11. $T = 31.5$.

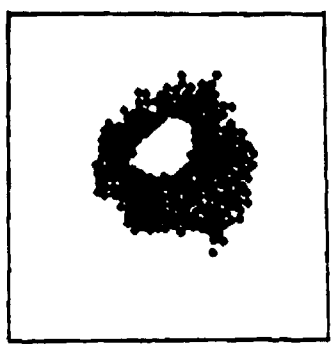


Fig. 12. $T = 39.0$.

the respective times $T = 24.0$, 31.5, 39.0 is shown in Figs. 10–12. Figure 13 shows the triple self reorganization of the A, B, and C sets at time $T = 39.0$. The exceptionally slow self reorganization of the sets B and C after the A particles formed into the core was accelerated by setting the damping factor 0.99 to 0.9 in rules (a)–(d) after $T = 24.0$.

With regard to computer time on the CRAY, 1000 time steps require 48 seconds of cpu time.

2.2. *Other examples*

If one varies parameters in Sec. 2.1 by no more that 5%, results completely analogous to those shown in Figs. 4–13.

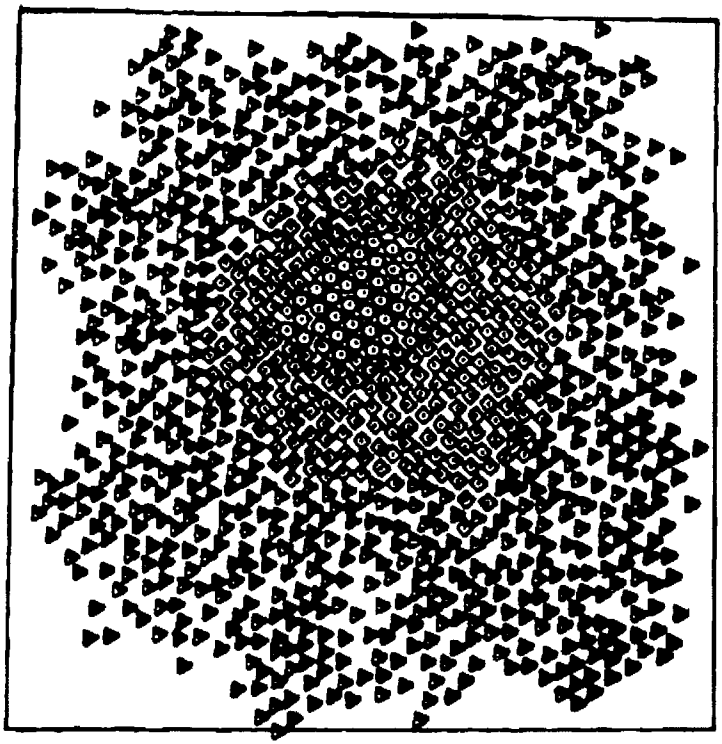


Fig. 13. $T = 39.0$.

The need for damping in rules (a)–(d) follows because the time step $\Delta t = 0.0001$ is relatively large. This protocol proves to be *economically* practical for the simulations. The damping rules however can be discarded if one wishes to use a time step $\Delta t = 0.00001$ or smaller.

Next note that a "close" choice of the m_i parameters, like $m_i = 10000$ for the A set and $m_i = 9500$ for the B set, results in exceptionally slow self-reorganization.

If in rules (a)–(d) the damping factor 0.99 is replaced by 0.9 from the start, then trapping often results, and, in particular, a particle from the set B can often be found in the interior of the set A after the set A has formed a core. The reason is that there follows an excessive loss of system kinetic energy, which yields premature solidification in the core.

If all other parameters are unchanged, the calculations are unstable for $D > 3.0$.

Finally note that the concept of temperature for *molecules* does yield a specific formula for temperature calculation. No such formula exists for *particles*. It would be of interest to develop such a formula, for it would then allow the determination of the temperature range in which morphogenesis can result.

References

[1] R. P. Feynman, R. B. Leighton and M. Sands, *The Feynman Lectures on Physics*, Vol I (Addison-Wesley, Reading, 1963).

[2] D. Greenspan, *Arithmetic Applied Mathematics* (Pergamon, Oxford, 1980).

[3] D. Greenspan, Particle simulation of biological sorting, TR 254, Math. Dept. (Univ. Texas at Arlington, 1988).

[4] M. S. Steinberg, Reconstruction of tissues by dissociated cells, in *Models for Cell Rearrangement*, ed. G. D. Mostow (Yale University Press, New Haven, 1975), pp. 82–99.

This page is intentionally left blank

CHAPTER 6

MODELLING BIOLOGICAL GEL CONTRACTION BY CELLS: CONSEQUENCES OF CELL TRACTION FORCES DISTRIBUTION AND INITIAL STRESS

S. RAMTANI

Université Paris 13 — IUT de Saint-Denis, Laboratoire des Propriétés Mécaniques et Thermodynamiques des Matériaux, CNRS-UPR9001, Institut Galilée, Université Paris 13, 99, av. J.B. Clément, 93430 Villetaneuse, France
salah.ramtani@lpmtm.univ-paris13.fr

Models based on the Murray–Oster continuum framework have been applied to a variety of biological settings in order to investigate the morphogenesis of living tissues. Collagen-matrix contracted by fibroblast model is an important example and a suitable way to study reciprocal geometric and mechanical interactions that regulate wound contraction of connective tissue cells. This contraction, which is due to cell traction forces, is essential in wound healing and pathological contractures. In the present contribution, where thin disk sample geometry is considered, an attempt is made to investigate the effect of initial stress upon the kinematics of contraction. This aspect is probably source of novel insight into the roles of key biological parameters in determining the biomechanical properties of contracted biological gel. Our hope is that this contribution will find a logical sound and contribute to gain a greater understanding of wound contraction mechanism.

1. Introduction

The mechanical interactions developed by motile cells with fibers in the surrounding extracellular matrix is essential to cell behavior and plays a major role in soft tissues and tissue-equivalent reconstituted gels, and thus to many biomedical and tissue engineering problems [1, 4, 8, 14, 16–19, 24, 27–28, 33]. The ability of cells to organize collagen fibrils is fundamental to a variety of processes found in angiogenesis, embryogenesis, fibrosis, scar formation, and wound healing [3–6, 26–28, 30–34]. It has been suggested that fibroblast reorganize the collagen lattice either as a result of isometric tension applied to the collagen fibrils [1–23, 25, 27–29, 32]. Although the

mechanism of wound contraction, which is a clinically important biological process, is not completely understood. In fact what is well known is based on the two following theories: (a) the one advanced by Ehrlich [8–9] suggest that wound contraction results from migrating fibroblasts that move trough and rearrange connective tissue in granulating wounds. The activity of the fibroblasts on the connective tissue is sufficient to cause centripetal movement of the skin margin. Simultaneous formation of collagen cross-links maintains the dimensions of the wound as it decreases its surface area over time, (b) the second theory, which carries a recent study proposed by [13], has been suggested by Gabbiani *et al.* [10] who originally described myofibroblasts; these highly specialized contractile fibroblasts found in granulating wounds are attached to one another by *cell-cell connection* and to the extracellular matrix (ECM). Thus, they are capable of contracting synchronously to generate the centripetal force of wound contraction [10, 12]. Proponents of the later theory suggest that collagen has very little to do with wound contraction [15]. Based on experimental observations, Murray *et al.* [22] had proposed a continuum model for mesenchymal morphogenesis which take into account the interaction between cells and ECM and which has been extensively used [1, 20–21, 23, 29–32]. Despite the fact that above studies have permitted to more understand the cell-ECM interactions mechanisms, the exact form of these forces and their relative distribution is still an open question [20]. Moreover, it has been shown that excessive and permanent contractile forces are characteristic of abnormal healing responses such as keloid scarring and other fibrocontractive diseases [23, 27]. Thus, there is a clear need to study the contraction mechanism which reflects the macroscopic manifestation of the intrinsic and local cell-matrix fiber mechanical interaction [3–4, 9].

In the present work, which is based on the theory proposed by Murray and Oster [22], the interactive processes of cell migration and matrix deformation are derived from mass conservation equations for cell and extra-cellular matrix (ECM) which are coupled to the mechanical force balance for the tissue-equivalent composite. The above model is revisited with the new assumption dealing with: (a) the *centripetal* character of the cell traction force and, (b) the effect of the *initial stress* due probably to the cell-cell interactions. The mixed system of parabolic-hyperbolic-elliptic partial differential equations, obtained after spatial rescaling together with the ordinary differential derived from the zero stress condition at the free boundary, are solved numerically by the use of finite difference method. From our numerical investigation, it is clearly shown that the initial stress

has a predominant role in the mechanism of contraction. This aspect, which is often omitted, is probably source of novel insight into the roles of key biological parameters in determining the biomechanical properties of contracted biological gel. Our hope is that this contribution will find a logical sound and contribute to gain a greater understanding of wound contraction mechanism.

2. The Mechanocellular Model

Whereas in biochemical models cells respond in a programmed manner to chemical concentrations, in biomechanical models they participate directly in the dynamics of pattern formation and react actively and passively to mechanical forces. The basis of most biomechanical models of pattern formation lies largely in experimental observations of the effect of cell traction on artificial substrates [1, 7, 13–16].

From theoretical view point, Murray and Oster [22] proposed a mechanocellular model for pattern formation based on the following assumptions:

(a) cell's migration occurs through the fibrous network of extracellular matrix;
(b) cell's motility induces large traction responsible in part of the extracellular matrix deformation and;
(c) this deformation and adhesion gradient influence the direction of the movement of cells (haptotaxis).

According to this theory, the basic variables are cell density $n(M,t)$ and ECM density $\rho(M,t)$; these are locally averaged species variables that depends on space M and time t. The mechanical consequences of the cell-ECM traction forces, the intrinsic response of the tissue-composite and the external resistance to tissue movement due to fibrous attachments to underlying tissues, are encapsulated by a force-balance equation that governs the tissue displacement, $\mathbf{u}(M,t)$. Specifically, the model equations have the following forms:

- For the mass conservation of the local cells concentration,

$$\frac{\partial n}{\partial t} + \operatorname{div}(\mathbf{J}) = P(n,\rho) \tag{1}$$

where $n(M,t)$ is the number of cells per unit volume, $\mathbf{J}$ is the flux of cell per unit area and $P(n,t)$ is the mitotic rate process. The cell flux

vector is given by

$$\mathbf{J} = n\frac{\partial \mathbf{u}}{\partial t} - D_n \, \mathbf{grad}(n) \tag{2}$$

where D_n is a diffusion coefficient which represents short-range effect in random dispersal. Each of the three terms of Eq. (2) reflects respectively, the convection term characteristic of the ECM deformation and the cell's diffusion by migration.

In order to account for the mitosis process which is viewed as a logistic process, we used the well known relation for the mitotic rate $P(M,t) = kn(N-n)$ where k is a growth rate and N is a maximum cell density.

- For the mass conservation for the local ECM concentration,

$$\frac{\partial \rho}{\partial t} + \operatorname{div}\left(\rho \frac{\partial \mathbf{u}}{\partial t}\right) = B(n,\rho) \tag{3}$$

where the rate of ECM secretion and degradation by fibroblasts $B(n,\rho)$ has been neglected in order to reflect the fact that the rate of ECM remodeling takes place on a relatively long time scale compared with the proliferative phase.

- For the local mechanical equilibrium, with body and inertial forces neglected,

$$\operatorname{div}(\boldsymbol{\sigma}) = \mathbf{0} \tag{4}$$

where the stress tensor for the composite material can be written as

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}^p + \boldsymbol{\sigma}^a + \boldsymbol{\sigma}^0 \overline{\mathbf{grad}(\mathbf{u})} \tag{5}$$

and in which $\boldsymbol{\sigma}^p$, $\boldsymbol{\sigma}^a$ and $\boldsymbol{\sigma}^0$ are the passive, the active and the initial stress tensors, respectively. The superimposed bar represents the transpose of the displacement gradient tensor.

We consider the behavior of the ECM as a linear, isotropic, compressible viscoelastic solid. Then, we shall write the ECM stress tensor $\boldsymbol{\sigma}^p$ as follows:

$$\boldsymbol{\sigma}^p = \mu_1 \frac{\partial}{\partial t}\varepsilon + \mu_2 \frac{\partial}{\partial t}\operatorname{tr}[\varepsilon]\,\mathbf{I} + \frac{E}{(1+\nu)}\left[\varepsilon + \frac{\nu}{1-2\nu}\operatorname{tr}[\varepsilon]\,\mathbf{I}\right] \tag{6}$$

where ε is the linearized strain tensor, $\mathbf{I}$ is the identity tensor, μ_1 and μ_2 are related to the shear and bulk viscosities, E is the Young's modulus and ν is the Poisson's ratio.

Although this area has received significant attention, the origin of the forces exerted by the cells is poorly understood, but stress fibers, i.e. aligned

microfilaments, have been seen in the cytoplasm of these cultured cells and depend on adhesion between cell surface receptors and binding sites on collagen fibers [2, 5]. To account for the short-range of the active cell-matrix interaction contribution, Murray and Oster [22] considered the stress σ^a_{ij} as a negative pressure proportional to the product ρn. When a fibroblast is embedded and cultured in a collagen gel, for example, the collagen fibers near the cell align in a radial pattern [28], apparently also in response to an applied tension. Taking advantage of previous works, the centripetal direction of the internal traction force is chosen as follows

$$\boldsymbol{\sigma}^a = \frac{\tau_0 \rho}{1+\lambda n^2} n \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{7}$$

where τ_0 and λ are positive constants. Moreover, λ defines the saturation cell density; i.e. cell motion is restricted by contact inhibition, whereas τ_0 is the constant value of the initial traction parameter.

By the use of the relations (5)–(7) and with the following simplification $\boldsymbol{\sigma}^0 = \sigma^0 \mathbf{I}$, the equation of motion (4) can be rewritten more explicitly in polar coordinates

$$\begin{aligned} &(\mu_1+\mu_2)\frac{\partial}{\partial t}\left(\frac{\partial^2 u}{\partial r^2}+\frac{1}{r}\frac{\partial u}{\partial r}-\frac{u}{r^2}\right) \\ &\quad+\left[\frac{E(1-\nu)}{(1+\nu)(1-2\nu)}+\sigma^0\right]\left(\frac{\partial^2 u}{\partial r^2}+\frac{1}{r}\frac{\partial u}{\partial r}-\frac{u}{r^2}\right) \\ &\quad+\frac{\partial}{\partial r}\left(\tau_0\frac{\rho n}{1+\lambda n^2}\right)+\frac{1}{r}\frac{\tau^0 \rho n}{1+\lambda n^2}=0. \end{aligned} \tag{8}$$

The symmetry properties associated with the hypothesis of free boundary condition at the moving surface gives respectively

$$\left\{\frac{\partial n}{\partial r}=0, \quad \frac{\partial \rho}{\partial r}=0, \quad u=0\,; \quad r=0\right. \tag{9}$$

$$\begin{aligned} &(\mu_1+\mu_2)\frac{\partial^2 u}{\partial r \partial t}+\frac{\mu_2}{r}\frac{\partial u}{\partial t}+\frac{E(1-\nu)}{(1+\nu)(1-2\nu)}\left[\frac{\partial u}{\partial r}+\frac{\nu}{1-\nu}\frac{u}{r}\right] \\ &\quad+\sigma^0\frac{\partial u}{\partial r}+\tau_0\frac{\rho n}{1+\lambda n^2}=0 \end{aligned} \tag{10}$$

at the current moving boundary position $S(t)$.

We finally assume that local cell and ECM concentrations are initially distributed according to the uniform and normal laws, respectively. The

initial conditions are therefore,

$$n(r,0) = n_0, \quad \rho(r,0) = \frac{\rho_0}{\sqrt{\pi}} e^{-\frac{r^2}{2S^2(t)}}, \quad u(r,0) = 0. \tag{11}$$

In order to solve the set of equations we have defined dimensionless variables as follow

$$\begin{aligned} &n^* = \frac{n}{N}, \quad \rho^* = \frac{\rho}{\rho_0}, \quad u^* = \frac{u}{a}, \quad r^* = \frac{r}{a}, \quad t^* = \frac{t}{T}, \\ &D^* = \frac{DT}{a^2}, \quad k^* = kNT, \quad \lambda^* = \lambda N^2 \\ &\mu_1^* = \frac{\mu_1 + \mu_2}{ET}\nu_1, \quad \mu_2^* = \frac{\mu_2}{ET}\nu_1, \quad \tau_0^* = \tau_0 \frac{\rho_0 N}{E}\nu_1, \quad p^\bullet = \sigma^0 \frac{\nu_1}{E} \\ &\nu_1 = \frac{(1+\nu)(1-2\nu)}{(1-\nu)}, \quad \nu_2 = \frac{\nu}{1-\nu} \end{aligned} \tag{12}$$

where a is the initial radius of the disk sample, T is a characteristic time, that is a factor scale.

The governing equations, boundary and initial conditions are written in the new relative frame $\xi = \frac{r^\bullet}{S(t)}$ which fixes the boundary at $\xi = 1$ for all time. The set of equations is then

$$\begin{aligned} &-\frac{\xi}{S}\frac{dS}{dt}\frac{\partial n}{\partial \xi} + \left(\frac{\partial n}{\partial t}\right)_\xi + \frac{n}{S}\frac{\partial}{\partial \xi}\left(-\frac{\xi}{S}\frac{dS}{dt}\frac{\partial u}{\partial \xi} + \mathrm{v}\right) \\ &\quad + \frac{1}{S}\left(\frac{\partial n}{\partial \xi} + \frac{n}{\xi}\right)\left(-\frac{\xi}{S}\frac{dS}{dt}\frac{\partial u}{\partial \xi} + \mathrm{v}\right) \\ &\quad - D\left(\frac{1}{S^2}\frac{\partial^2 n}{\partial \xi^2} + \frac{1}{S\xi}\frac{\partial n}{\partial \xi}\right) + kn(n-1) = 0 \end{aligned} \tag{13}$$

$$\begin{aligned} &-\frac{\xi}{S}\frac{dS}{dt}\frac{\partial \rho}{\partial \xi} + \left(\frac{\partial \rho}{\partial t}\right)_\xi + \frac{\rho}{S}\frac{\partial}{\partial \xi}\left(-\frac{\xi}{S}\frac{dS}{dt}\frac{\partial u}{\partial \xi} + \mathrm{v}\right) \\ &\quad + \frac{1}{S}\left(\frac{\partial \rho}{\partial \xi} + \frac{\rho}{\xi}\right)\left(-\frac{\xi}{S}\frac{dS}{dt}\frac{\partial u}{\partial \xi} + \mathrm{v}\right) = 0 \end{aligned} \tag{14}$$

$$\begin{aligned} &\left(\frac{\partial^2 u}{\partial \xi^2} + \frac{1}{\xi}\frac{\partial u}{\partial \xi} - \frac{u}{\xi^2}\right)(1+p) + \mu_1 \frac{\partial^2}{\partial \xi^2}\left(-\frac{\xi}{S}\frac{dS}{dt}\frac{\partial u}{\partial \xi} + \mathrm{v}\right) \\ &\quad + \frac{\mu_1}{\xi}\frac{\partial}{\partial \xi}\left(-\frac{\xi}{S}\frac{dS}{dt}\frac{\partial u}{\partial \xi} + \mathrm{v}\right) - \frac{\mu_1}{\xi^2}\left(-\frac{\xi}{S}\frac{dS}{dt}\frac{\partial u}{\partial \xi} + \mathrm{v}\right) \\ &\quad + S\frac{\partial(\tau_0 n\rho/1+\lambda n^2)}{\partial \xi} + S\frac{\tau_0 n\rho}{1+\lambda n^2} = 0 \end{aligned} \tag{15}$$

where v is the cell-matrix composite velocity.

The initial conditions and boundary conditions are therefore in the relative frame:

$$n(\xi, t_0) = n_0, \quad \rho(\xi, t_0) = \frac{\rho_0}{\sqrt{\pi}} e^{-\frac{\xi^2}{2}}, \quad u(\xi, t_0) = 0 \tag{16}$$

$$\left(\frac{\partial n}{\partial \xi}\right)_{\xi=0} = 0, \quad \left(\frac{\partial \rho}{\partial \xi}\right)_{\xi=0} = 0, \quad u(\xi = 0, t) = 0 \tag{17}$$

$$\left(\frac{\partial n}{\partial \xi}\right)_{\xi=1} = 0, \quad \left(\frac{\partial \rho}{\partial \xi}\right)_{\xi=1} = 0, \quad u(\xi = 1, t) = S(t) - 1, \mathrm{v}(t) = \frac{dS}{dt}. \tag{18}$$

3. Model Predictions and Discussion

Then, this transformed nonlinear governing moving boundary value problem is reduced to a differential algebraic system of equations. This system is solved applying centered finite difference approximation to derivatives and a *Newton–Raphson* method.

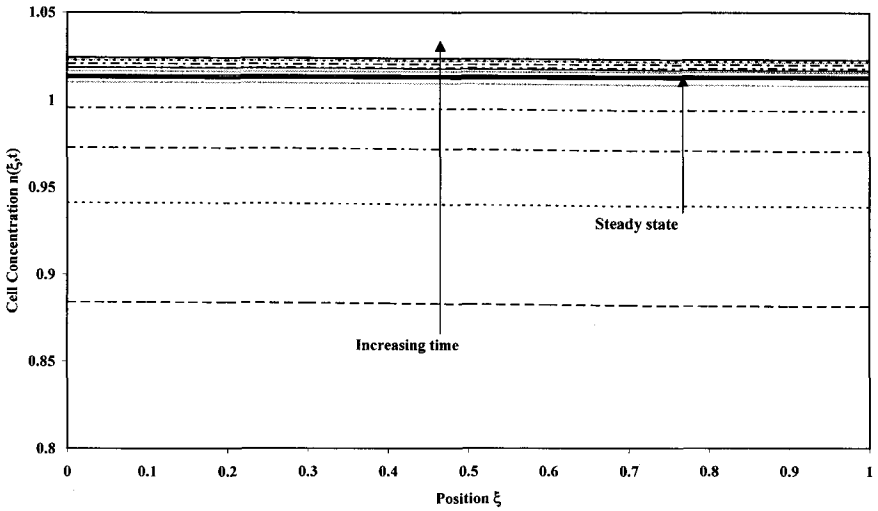
3.1. *Uniaxial cell traction force effect*

We simulate the gel contraction for an inhomogeneous initial distribution of ECM density over the gel. First, an attempt is made to compare the uniaxial cell traction force hypothesis ($\sigma_{\theta\theta} = 0$) with the spherical one ($\sigma_{rr} = \sigma_{\theta\theta}$) in the case where there is no initial stress. For the chosen set of parameters values taken back from Barocas *et al.* [20] and given in Table 1, we represents in the relative frame (0, ξ) and at successive time steps:

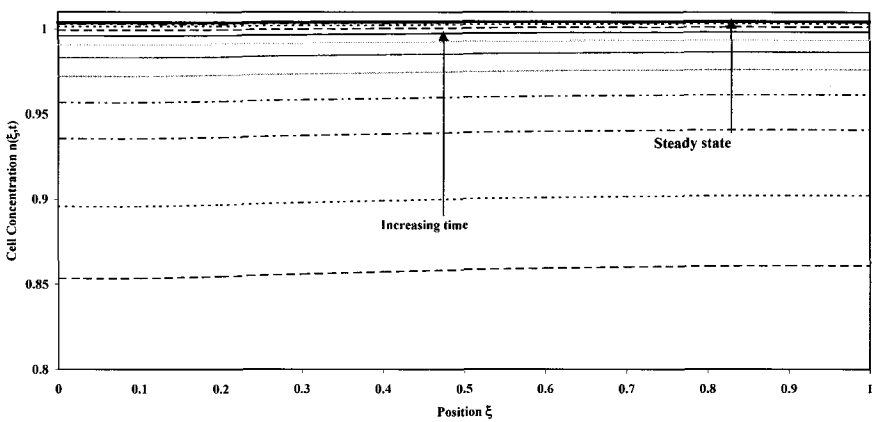
(a) the simulated evolution of the cells concentrations (Fig. 1). It is clearly shown that the results are sensibly the same. However, one can note that for the spherical hypothesis the steady value (bold line) is not the maximal one contrarily to the centripetal hypothesis;
(b) the simulated evolution of the apparent ECM concentrations is given Fig. 2. One notes different distributions between the two hypotheses and observes a reduction of the density on the left part of the curve 2b ($\xi < 0.5$);

Table 1. Model's parameters.

Dimensionless Parameters	n_0	ρ_0	τ_0	D	k	λ	ν_2	μ_1	μ_2
	0.80	1.00	1.00	0.10	1.00	0.80	0.96	1.0	0.1



(a)



(b)

Fig. 1. Simulated gel contraction for an homogeneous initial cell distribution. Cell density along the gel radius is plotted as a function of time without initial stress. (a) Spherical active stress hypothesis, (b) Uniaxial active stress hypothesis.

(c) the simulated evolution of the local volumetric dilatation illustrated in Fig. 3. The decrease of the ECM concentration is certainly due in part to the increase of the local volume as shown in Fig. 3(b). This result is in connection with the results shown in Fig. 4.

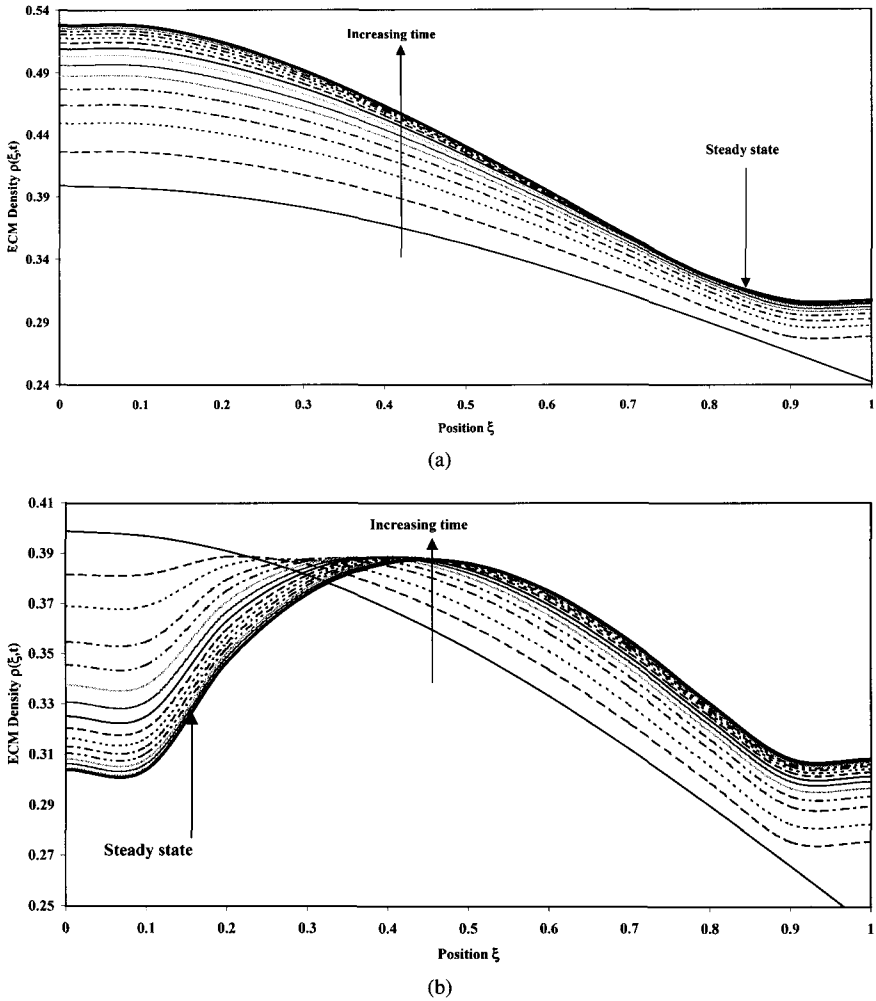
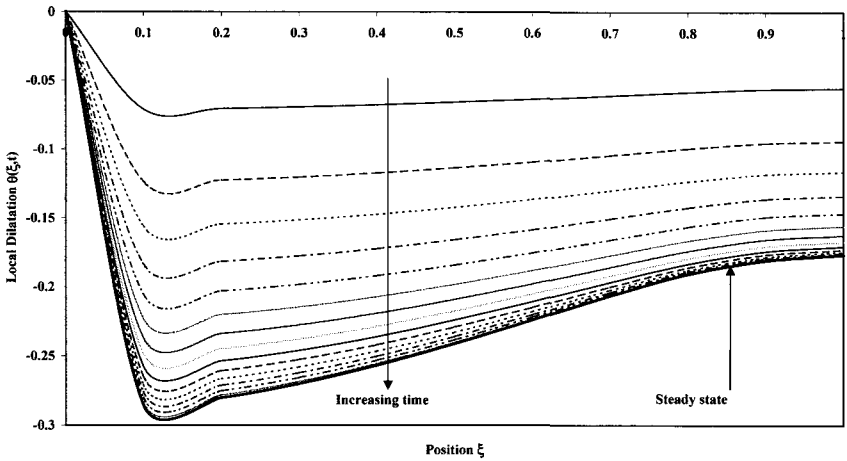
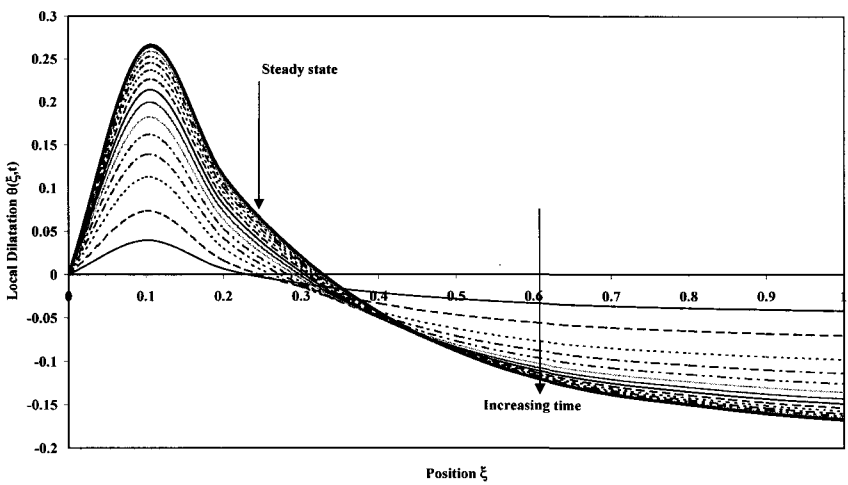


Fig. 2. Simulated gel contraction for an inhomogeneous initial gel distribution without initial stress. Gel density along the gel radius is plotted as a function of time. (a) Spherical active stress hypothesis, (b) Uniaxial active stress hypothesis.

Effectively, it exists one area under tension even though the sample is globally under compression. Until now the measure of the boundary displacement attracted all the attention of experimental investigators. Maybe it is necessary to look at the displacement of the interior points.

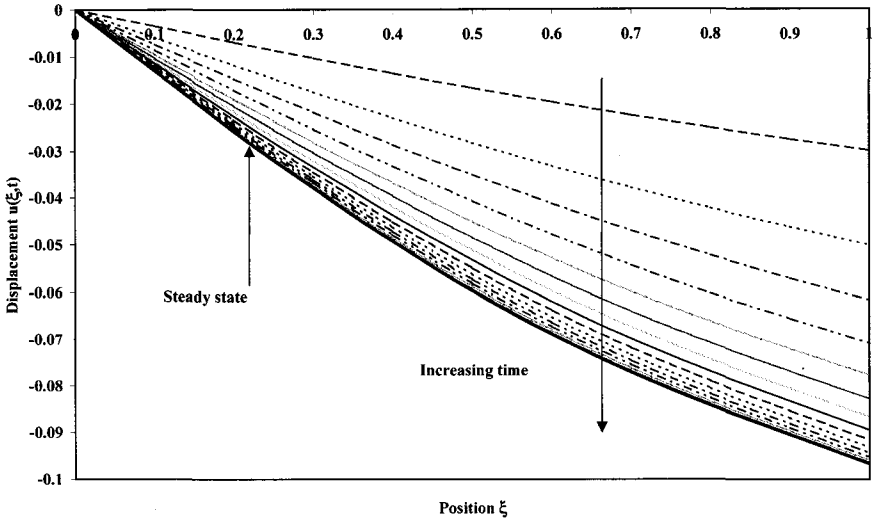


(a)

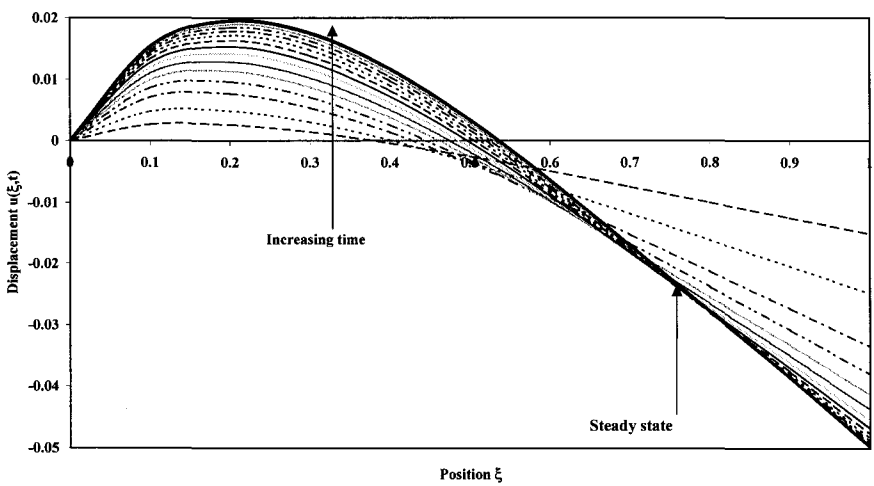


(b)

Fig. 3. Simulated gel contraction for an homogeneous initial cell distribution and an inhomogeneous initial gel distribution without initial stress. Local dilatation along the gel radius is plotted as a function of time. (a) Spherical active stress hypothesis, (b) Uniaxial active stress hypothesis.



(a)



(b)

Fig. 4. Simulated gel contraction for an homogeneous initial cell distribution and an inhomogeneous initial gel distribution without initial stress. Displacement of each point with time is plotted as a function of the gel radius. (a) Spherical active stress hypothesis, (b) Uniaxial active stress hypothesis.

3.2. *Initial stress effect*

With the hypothesis of an uniaxial cell traction force, we first represents in the relative frame $(0, \xi)$ the effect of increasing initial stress upon the steady values of:

(a) the cell concentration which is considerably increasing up to 40% of the initial concentration (Fig. 5),
(b) the ECM density which exhibits a highly nonlinear distribution as shown in Fig. 6,
(c) the local volumetric dilatation which is also increasing as shown in Fig. 7.

Second, we illustrates during the time period $(0,\ 3T)$ the evolution of both the boundary and midpoint displacements given in Figs. 8 and 9, respectively. In Fig. 8, we observe a significant increasing of the boundary as a function of the initial stress. It is shown in Fig. 9 that the behavior of midpoint is first compacted (*negative displacement*) before changing and to be in extension (*positive displacement*). Note that the apparition of the extension is delayed by the increase of the initial stress.

We have examined the mechanical interactions of cells with tissue-equivalent gels revisiting the well known monophasic theory to describe the biomechanics of the gel contraction. In particular, consequences of the cell

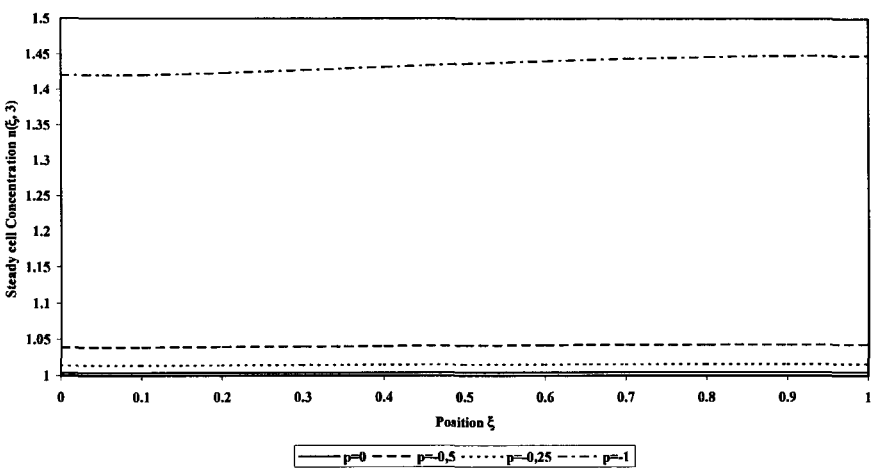


Fig. 5. Simulated gel contraction for an homogeneous initial cell distribution and an inhomogeneous initial gel distribution. Steady cell density along the gel radius is plotted as a function of initial stress.

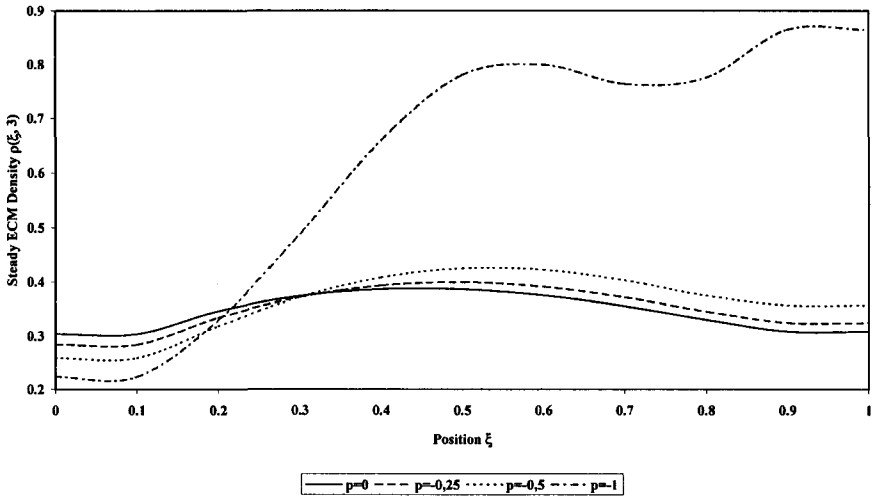


Fig. 6. Simulated gel contraction for an homogeneous initial cell distribution and an inhomogeneous initial gel distribution. Steady gel density along the gel radius is plotted as a function of initial stress.

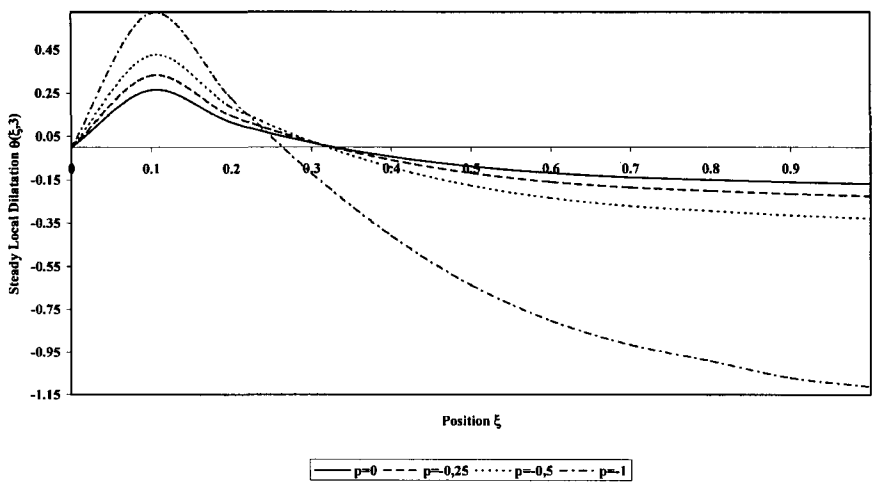


Fig. 7. Simulated gel contraction for an homogeneous initial cell distribution and an inhomogeneous initial gel distribution. Steady local dilatation along the gel radius is plotted as a function of initial stress.

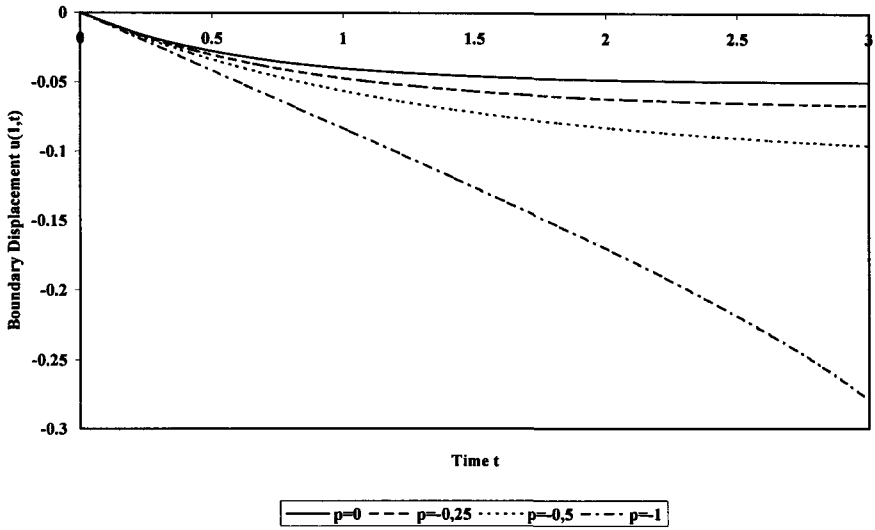


Fig. 8. Simulated gel contraction for an homogeneous initial cell distribution and an inhomogeneous initial gel distribution. Steady free boundary displacement along the gel radius is plotted as a function of initial stress.

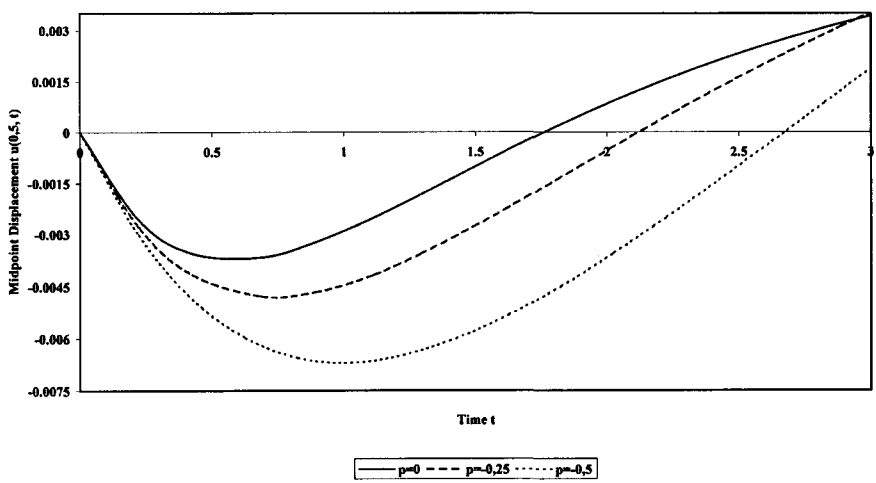


Fig. 9. Simulated gel contraction for an homogeneous initial cell distribution and an inhomogeneous initial gel distribution. Steady midpoint displacement along the gel radius is plotted as a function of initial stress.

traction forces distribution and the initial stress effects are investigated. In this study, it is clearly shown that these two aspects plays a predominant role during the contraction process. In particular, the initial stress effect which is often omitted can affect the manner in which cells restructure the surrounding collagen network and this aspect is central to the modeling of such biomaterials.

Nomenclature

- n Local cell concentration
- ρ Local ECM concentration
- $\mathbf{u}$ Displacement vector for cell/matrix composite
- u Radial displacement for cell/matrix composite
- $\mathbf{J}$ Net cell flux vector due to active migration
- D_n Cell motility coefficient
- k Logistic growth rate constant
- N Maximum cell concentration
- $\boldsymbol{\sigma}$ Total stress tensor for cell/ECM composite
- $\boldsymbol{\sigma}^p$ Stress tensor for ECM
- $\boldsymbol{\sigma}^a$ Stress tensor associated with the active traction stress
- ε Small strain tensor for cell/ECM composite
- θ Local dilatation of cell/ECM composite
- E Young's modulus
- ν Poisson's ratio
- μ_1 Shear viscosity
- μ_2 Bulk viscosity
- $\mathbf{I}$ Identity tensor
- τ^0 Traction parameter
- λ Contact inhibition parameter
- a Radius of the disk sample
- $S(t)$ Position of disk boundary
- r Radial coordinate
- ξ Relative frame
- T Time scale factor
- t Time
- $\mathbf{v}$ Velocity of cell/ECM composite

References

[1] V. H. Barocas, A. G. Moon and R. T. Tranquillo, The fibroblast-populated collagen microsphere assay of cell traction force: Part 2 measurement of the cell traction parameter, *J. Biomech. Eng.* **117** (1995) 161.

[2] V. H. Barocas and R. T. Tranquillo, An anisotropic biphasic theory of tissue-equivalent mechanics: The interplay among cell traction, fibrillar network deformation, fibril orientation and cell contact guidance, *J. Biomech. Eng.* **119** (1997) 137.

[3] V. H. Barocas and R. T. Tranquillo, Biphasic theory and *In Vitro* assays of cell-fibril mechanical interactions in tissue-equivalents gels, in *Cell Mechanics and Cellular Engineering*, eds. V. C. Mow, F. Guilak, T. Son-Tay and Hochmuth (Springer-Verlag, 1994).

[4] E. Bell, B. Ivarsson and C. Merrill, Production of a tissue-like structure by contraction of collagen lattices by human fibroblasts of different proliferative potential *in vitro*, *Proc. Natl. Acad. Sci.* **76** (1979) 1274.

[5] K. Burridge, Are stress fibers contractile?, *Nature* **294** (1981) 691.

[6] R. A. Clark and P. M. Henson, *The Molecular and Cellular Biology of Wound Repair* (Plenum, New York, 1988).

[7] P. Delvoye, P. Wiliquet, J. L. Levêque and B. V. Lapière, Measurement of mechanical forces generated by skin fibroblasts embedded in three-dimensional collagen gel, *J. Invest. Dermatol.* **97** (1991) 898.

[8] H. P. Ehrlich, The role of connective tisue matrix in wound healing, *Prog. Clin. Biol. Res.* **266** (1988) 243.

[9] H. P. Ehrlich and J. B. M. Rajaratnam, Cell locomotion forces versus cell contraction forces for collagen lattice contraction: An *in vitro* model of wound contraction, *Tissue Cell.* **22** (1990) 407.

[10] G. Gabbiani, G. B. Ryan and G. Majno, Presence of modified fibroblast in granulation tissue and their possible role in wound contraction, *Experientia* **27** (1971) 549.

[11] C. G. Galbraith and M. P. Sheetz, Forces on adhesive contacts effect cell function, *Curr. Opin. Cell Biol.* **10** (1998) 566.

[12] F. Grinnel, Fibroblast, myofibroblasts, and wound contraction, *J. Cell Biol.* **124** (1994) 401.

[13] F. Grinnel, Fibroblast-collagen-matrix contraction: Growth factor signalling and mechanical loadings, *Trends Cell Biol.* **10** (2000) 362.

[14] A. K. Harris, Tissue culture cells on deformable substrata: Biomechanical implications, *J. Biotech. Engin.* **106** (1984) 19.

[15] H. L. Joseph, F. J. Roisen, G. L. Anderson, J. H. Baker, L. J. Weiner and G. R. Tobin, Inhibition of wound contraction with locally injected lathyrogenic drugs, *Amer. J. Surg.* **174** (1997) 347.

[16] M. S. Kolodney and R. B. Wysolmerski, Isometric contraction by fibroblasts and endothelial cell in tissue culture: A quantitative study, *J. Cell Biol.* **117** (1992) 73.

[17] J. A. Madri and M. A. Pratt, Endothelial cell-matrix interactions: *in vivo* models of angiogenesis, *J. Histochem. Cytochem.* **34** (1986) 85.

[18] M. H. McGrath and R. H. Simon, Wound geometry and the kinetics of wound contraction, *Plast. Reconst. Surg.* **72** (1983) 66.

[19] D. Montandon, G. Gabbiani, G. B. Ryan and G. Majno, The contractile fibroblast: Its relevance in plastic surgery, *Plast. Reconstr. Surg.* **52** (1973) 286.

[20] A. G. Moon and R. T. Tranquillo, Fibroblast-populated collagen microsphere assay of cell traction force: Part 1 continuum model, *AIChE J.* **39** (1993) 163.

[21] V. Moulin, G. Castilloux, A. Jean, D. R. Garrel, F. A. Auger and L. Germain, *In vitro* models to study wound healing fibroblasts, *Burns.* **22** (1996) 359.

[22] J. D. Murray and G. F. Oster, Cell traction models for generating pattern and form in morphogenesis, *J. Math. Biol.* **19** (1984) 265.

[23] L. Olsen, J. A. Sherrat and P. K. A. Maini, Mechanochemical model for adult dermal wound contraction and the permanence of the contracted tissue displacement profile, *J. Theor. Biol.* **177** (1995) 113.

[24] P. Rompré, F. A. Auger, L. Germain *et al.*, Influence of initial collagen and cellular concentrations on the final surface area of dermal and skin equivalents, A Box-Behnken Analysis. In vitro Cell, *Dev. Biol.* **26** (1990) 983.

[25] R. Rudolph, Contraction and the control of contraction, *World J. Surg.* **4** (1980) 279.

[26] R. Rudolph, J. Van de Berg and P. H. Ehrlich, Wound contraction and scar contracture, in *Wound Healing. Biochemical and Clinical Aspects*, eds. I. K. Cohen, R. F. Diegelmann and W. J. Lindblad (1992), p. 96.

[27] O. Skalli and G. Gabbiani, The biology of the myofibroblast relationship to wound contraction and fibrocontractive diseases, in *The Molecular and Cellular Biology of Wound Repair*, eds. R. A. F. Clarck and P. M. Henson (Plenum Press, New York, 1988), p. 373.

[28] D. Stopak and A. K. Harris, Connective tissue morphogenesis by fibroblast traction, *Dev. Biol.* **90** (1982) 383.

[29] P. Tracqui, D. E. Woodward, C. Cruywagen, J. Cook and J. D. Murray, A mechanical model for fibroblast-driven wound healing, *J. Biol. Syst.* **3** (1995) 1075.

[30] L. Tranqui and P. Tacqui, Mechanical signalling and angiogenesis. The integration of cell-extracellular matrix coupling, *C. R. Acad. Sci. Life Sci.* **323** (2000) 31.

[31] R. T. Tranquillo and J. D. Murray, Continuum model of fibroblast-driven wound contraction: Inflammation-mediation, *J. Theor. Biol.* **158** (1992) 135.

[32] R. T. Tranquillo, M. A. Durrani and A. G. Moon, Tissue engineering science: Consequences of cell traction force, *Cytotechnology* **10** (1992) 225.

[33] J. P. Trinkaus, *From Cells into Organs.* 2nd edn. (Englewood Cliffs, Prentice-Hall, 1984).

[34] I. V. Yannas, J. F. Burke, D. P. Orgill and E. M. Skrabut, Wound tissue can utilize a polymeric template to synthesize a functional extension of skin, *Science* **215** (1982) 174.

This page is intentionally left blank

CHAPTER 7

PERISTALTIC TRANSPORT OF PHYSIOLOGICAL FLUIDS

J. C. MISRA* and S. K. PANDEY
Department of Mathematics, Indian Institute of Technology
Kharagpur-721302, India
**jcm@maths.iitkgp.ernet.in*

Physiological fluids in human or subhuman primates are, in general, pumped by the continuous periodic muscular oscillations of the ducts through which the fluids pass. These oscillations are supposed to be caused by the progressive transverse contraction waves that propagate along the walls of the duct. **True peristalsis** is usually defined as a coordinated reaction in which a wave of contraction is preceded by a wave of relaxation. Some electrochemical reactions are held responsible for this phenomenon. In fact, it is a reflex process. The swallowing of food through the oesophagus, the movement of chyme through the small intestine, the colonic transport in the large intestine, the passage of urine from the kidneys to the urinary bladder through the ureters, the spermatic flows in the ductus efferentes of the male reproductive tract, the vas deferens and the cervical canal, and the movement of ovum in the fallopian tube are all based upon the mechanism of peristaltic transport. The vasomotion of some blood vessels, e.g. venules and arterioles and the motion in the lymphatic vessels have also been found to be of peristaltic nature. Even some worms move peristaltically. Moreover, biomechanical pumps are fabricated to save blood or similar fluids from any possible contamination arising out of the contact with the pump machinery while pumping the fluid.

The peristaltic motion experienced in physiological flows is classified into different categories, a few of them being: (i) *rush peristalsis*, (ii) *anti-peristalsis* and (iii) *mass peristalsis.*

Rush peristalsis is the ordinary peristalsis found in different physiological transportations. This term is mainly associated with the flow in the small intestine.

Anti-peristalsis is the same peristalsis but it acts in the opposite direction. For example, in the oesophagus it moves in the oral direction. It is present in man in the second and the third parts of duodenum.

Mass peristalsis is found in the large intestine and is analogous to the rush peristalsis in the small intestine. Indeed, it is the main movement of the large intestine.

In order to have a proper understanding of the peristaltic transport in physiological systems, it is felt that we should have some information about the relevant matters. It is with this end in view, we first discuss briefly a few phenomena and some physiological systems associated with peristalsis.

1. Phenomena Associated with Peristalsis

Two very important fluid dynamical phenomena inherent in peristalsis are: (i) **reflux** and (ii) **trapping**.

Reflux. There are two contradicting definitions prevailing from the beginning of the investigation on peristaltic motion of physiological fluids. One was propagated by Fung and Yih [20] and the other by Shapiro *et al.* [57]. In fact, they meant two different phenomena. Shapiro *et al.* associated their definition with the backward migration of bacteria from the bladder to the kidneys. According to them, it refers to the presence of fluid particles that move, on the average, in the direction opposite to the net flow. The backward migration takes place near the walls. It was also experimentally verified by Weinberg *et al.* [72]. According to Fung and Yih, it is the average mean flow reversal near the axis of the duct. A similarity with vesicoureteral reflux was expected with this definition. Shapiro *et al.* maintained that in order to examine the retrograde motion of fluid particles, Eulerian time-mean velocity should be taken into consideration whereas Fung and Yih stressed on Lagrangian displacement of fluid particles. In the light of this controversy reflux, hereafter, will be denoted by $reflux^1$ for the definition of Shapiro *et al.* and $reflux^2$ for the definition of Fung and Yih as these two still perpetuate.

Trapping. Shapiro *et al.* [57] held that at high flow rates and large occlusions there is a region of closed streamlines in the wave frame and thus some fluid is found trapped within a wave of propagation. The trapped fluid mass is found to move with the mean speed equal to that of the wave.

2. Physiological Systems Associated with Peristalsis

2.1. *Digestive system*

The human digestive canal (cf. Fig. 1) is a long muscular duct comprising mouth, tongue, pharynx, oesophagus, stomach, small intestine, large intestine, rectum and anal canal.

2.2. *Oesophagus*

It is a long muscular tube that commences at the neck opposite to the long border of cricoid cartilage and extends from the lower end of the pharynx to the cardiac orifice of the stomach. The cardiac sphincter regulates the proximal end of the stomach and the one which guards the distal end is known as pyloric sphincter. Small intestine follows the pyloric sphincter. It is about 76 mm long in an adult human being and is subdivided into duodenum, jejunum and ileum (see Fig. 1). Large intestine joins the lower end of the small intestine at the ileocolic sphincter. The last part in which the large intestine opens is rectum together with anal canal.

Swallowing (or **deglutition**) takes place in three stages: (i) *first, i.e. buccal,* (ii) *second, i.e. pharyngeal,* and (iii) *third, i.e. oesophageal.* The first one is voluntary but the remaining two are controlled reflexly.

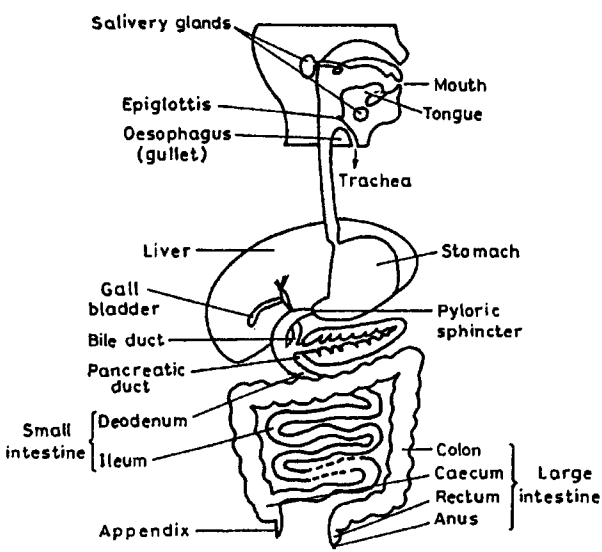


Fig. 1. Human digestive system.

Buccal deglutition: The food after mastication is rolled into a bolus. Owing to the contraction of mylohyoid muscles upward and backward movements start in the tongue which throws the bolus into the pharynx.

Pharyngeal deglutition: The soft palate is elevated and the nasal cavity is closed. There is a rise in the larynx together with the hyoid bone. The vocal cords are adducted and respiration is inhibited for a moment. Then there is an elevation in the epiglottis, which takes the bolus away from the laryngeal opening. The pharynx reopens and gathers the bolus.

Oesophageal deglutition: The larynx retrieves its normal position and the bolus is propelled into the oesophagus by the contraction of cricopharyngeus muscle. Peristaltic wave begins to propagate down the oesophagus carrying the bolus to the lower end where it is squeezed out into cardiac sphincter. The rate of propagation of the peristaltic wave is 20–40 mm per second. The vagus and the local plexus control the peristaltic movement. The cardiac sphincter relaxes within 2 sec of the swallowing. Gravity has little role in this process as the rate of progress along the oesophagus is not affected by posture, whether spine or erect (cf. Bosma [5]).

2.3. *Stomach*

In empty state two kinds of movements are seen in the stomach:

(i) **Tonus rhythm** — Rhythmic variations of tone occur at the rate of about 3 per min.
(ii) **Hunger contraction** — At intervals a series of strong contractions, called hunger contractions, takes place for about 30 sec. The entire stomach is involved in it.

Two different types of movements are observed in the two halves — *pylorus* and *fundus*, after taking food.

(i) **Fundus movement:** Here is tonic contraction but no peristalsis. A constant pressure is maintained upon the contents and these parts of the stomach send out more and more food into the pylorus, which in the mean time churns and pushes the food mass into the duodenum.
(ii) **Pylorus movement:** This part exhibits movements like peristalsis. They are waves of constriction. The waves activate near the incisura angularis and move towards the pylorus slowly. They become stronger as they proceed but almost die out normally near the pyloric sphincter and never continue up to duodenum. Each such progressive wave varies rhythmically in intensity. The recurrence of the wave is seen at the

rate of 3–4 per min. In addition, three or more waves may exist on the pylorus at the same time. Sometimes the pyloric sphincter does not open and the food mass suffers from a backward reflection in an axial stream. After some time when such a wave becomes sufficiently strong, the sphincter is opened and as a result, a part of the gastric contents is propelled out into the duodenum. The sphincter closes itself immediately and the peristaltic process continues until it opens again.

Vomiting. It is the act of forcible expulsion of the stomach contents through the mouth. In the beginning a feeling of nausea is experienced, followed by excess salivation. Glottis becomes closed and the nasopharynx is also shut off by elevating the soft palate. The stomach, the cardiac sphincter and the oesophagus relax and then there is a rise in intra-abdominal pressure.

It is a reflex process. A vomiting centre is situated in the medulla and is closely related to the vagus nucleus. Certain drugs (Apomorphine, etc.), toxins (such as those of uraemia) and increased intravascular pressure (as in the cases of brain tumour, asphysia, meningitis, etc.) directly stimulate this centre. It can be stimulated reflexly in various ways. The afferent impulses may arise in the throat, stomach, intestine, uterus, heart and from other viscera. The efficient impulses — both excitatory and inhibitory are carried in the vagus. The cause of vomiting is gastric irritation and its purpose is to drive out the irritant from the stomach (cf. Borison and Wang [4]).

2.4. *Small intestine*

The movement in the human small intestine are of four kinds: (i) *segmentation*, (ii) *peristalsis*, (iii) *anti-peristalsis* and (iv) *pendular movement.*

Segmentation: These are local constrictions followed by immediate relaxation. The constriction occurs at the site of maximum distension. In animals the group of constrictions succeeds at the rate of 20–30 per min. The rate is slower in man. The frequency which is 17 per min in duodenum and 12 per min in ileum is inversely proportional to the distance from the stomach. These most fundamental movements of the intestine are myogenic in nature and are independent of all nerves. Their functions include proper admixture of food with the digestive juices, helping absorption by bringing the mucus membrane into closer contact with food and increasing vascular and lymphatic circulation through the wall of the gut.

Peristalsis: Peristalsis in small intestines is called "*The law of intestine*" or "*Myenteric reflex*". The presence of food acts as the normal stimulus causing relaxation below and constriction above the food-bolus. As the wave travels downwards the food is moved in a spiral manner and the direction of rotation is anti-clockwise. The length of bowel traversed in making a copalate spiral is about 30 cm on average.

Peristalsis with two different speeds, are observed in the small intestine. It depends on nervous and chemical agents.

A special manifestation of peristaltic movement in the ileum is called gastro-ilial reflex. This is a brisk peristalsis set up in the ileum after meal reflexly, although peristalsis is generally very sluggish in the last part of ileum. The purpose is to drive out the ilial contents into caecum creating space for fresh supply.

Anti-peristalsis: It moves in the oral direction and is present in man in the second and third parts of duodenum only. Weak anti-peristalsis too takes place in the terminal part of the ileum and in this way restrains a rapid passage of the ilial contents into caecum. In the duodenum it helps through admixture, and also causes duodenal regurgitation into the stomach.

Pendular movement: It is side to side movement of individual loops of the intestine as a consequence of the rush of the food material through the lumen. It is absolutely a passive movement.

Large intestine:

It has four types of motion: (i) *rhythmic variations of tone*, (ii) *peristalsis*, (iii) *mass peristalsis* and (iv) *anti-peristalsis*.

Rhythmic variations of tone: This occurs throughout the large intestine but not always and is not at all concerned with propulsion; it rather maintains adequate circulation through the wall and helps in the absorption of water.

Peristalsis: It is not the same as rush peristalsis seen in the small intestine. It is a weak peristalsis alternately shortening and elongating in the transverse colon.

Mass peristalsis: This is the chief movement of large intestine governed by gastrocolic reflex. It occurs twice or thrice a day and after meal and during defaecation.

Anti-peristalsis: In man it is rarely seen but is well marked in animals such as cats.

2.5. *Ureter*

The two nearly 300 mm long muscular ducts joining the kidneys to the bladder are known as ureters (cf. Fig. 2). They are located in the extraperitonial tissue behind the peritoneum to which they closely adhere. The upper aspect of the ureter lies in the abdomen while its continuing lower part is in the pelvis. The only known function of the ureter is that it collects urine from the kidneys and squeezes it out to the bladder at the ureterovesical junction against a pressure gradient. This ureterovesical junction functions as a one-way valve and refrains fluid from going back into the ureters from the bladder. At rest it is totally collapsed and gets activated when needed to function. The fluid is passed peristaltically with almost full occlusion of the duct. The diastolic phase is found to be twice as long as the systolic phase. The cross-section is almost circular when it is fully distended while it adopts a star-like shape with flat quasi-two-dimensional lobes when contracted. Several waves of length ranging from 10 mm to 150 mm per min have been observed experimentally. The largest diameter is found to be 5 mm. In a normal ureter, the composition of urine is unchanged, but in a diseased state abnormal elements such as red or white cells or tumor cells may be present. The propulsion of fluid in a ureter is easily found to be primarily due to peristaltic motion (cf. Weinberg [71]).

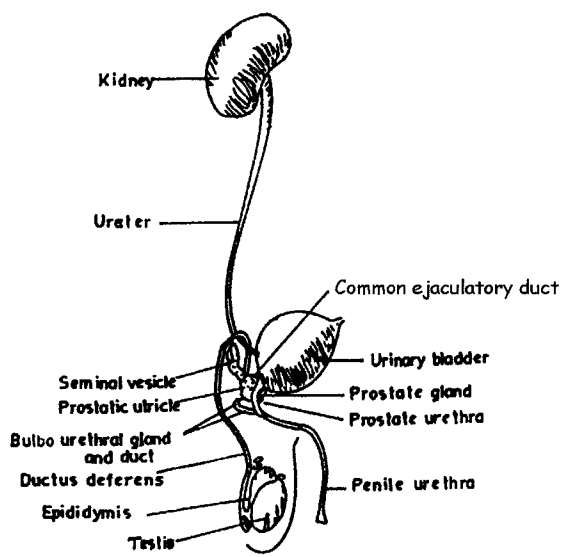


Fig. 2. Male urogenital organs.

2.6. *Vas deferens*

The vas deferens (or ductus deferens) is a long duct originating at the testis near epididymis and joins the seminal vesicle to form the common ejaculatory duct. It is the main duct through which seminal fluid passes. Its average length in an adult human is 45 cm.

2.7. *Experimental investigations on peristalsis*

Latham's [31] experiment (cf. Fig. 3) included a test duct of clear flexible polyvinylchloride with a wall thickness of 0.05 in. It was confined, in a 180° arc between a steel band and a stationary back plate formed in a semicircle of 16 in radius, such that the tube became approximately rectangular in shape, height 2.5 in, and a mean width of about 0.3 in. The ends of the test duct, outside the semicircular arc of flattening, joined vertical reservoirs to maintain the fluid at a constant elevation. For adjusting the pressure rise between the reservoirs and measuring mean flow, a control valve was placed

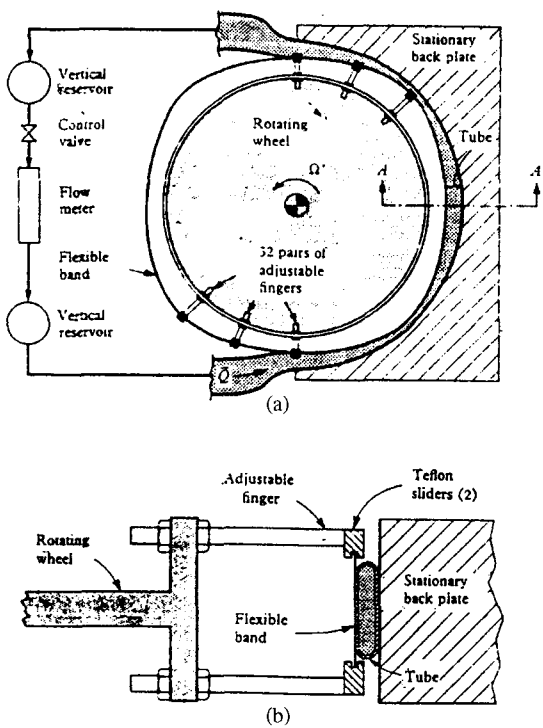


Fig. 3. Latham's apparatus of quasi-two-dimensional experiment (a) plane view, (b) section.

in the network. 32 pairs of adjustable fingers were mounted on the rotating wheel. Arrangements were made, by adjusting the fingers, for the propagation of an integral number of waves of approximately sinusoidal nature, having an amplitude of one-third of the half width of the rectangular duct. The motion of the fluid was approximately two-dimensional.

In order to attain different viscosity levels, mixtures of either glycerin and water or corn syrup and water were used. The main features of this experiment were that the wave speed, the pressure rise and the viscosities could be adjusted. For Reynolds number, $R < 0.2$, no significant difference was observed. However, for $R > 0.2$, pumping performance was found to be degraded. For $R = 38$, pumping drastically deteriorated.

Although it had some drawbacks such as (i) the wave was in only one wall and (ii) the whole duct was curved in a semicircle, the results of that experiment were generally in good agreement with the theoretical investigation of Shapiro [56].

A more refined experimental investigation was carried out by Weinberg *et al.* [72] on an improved apparatus for Reynolds number ranging from the inertia-free limit to values in which inertial effects were significant. Various mixtures of glycerin and water were used as the working fluid in order to obtain the range of viscosity necessary for wide variations of Reynolds number.

The pumping duct, rectangular in cross-section, was bounded by a rigid semi-circular back wall, a flexible moving wall in which longitudinal waves of transverse displacement were driven by roller cams, and two transparent cover plates. The rectangular duct was 10 in high, with a mean width of 0.50 in, giving a mean aspect ratio of 20. It was laid out on a semi-circle of radius 17.24 in. Exactly three wavelengths were fitted within the arc length of 54.0 in, so that the wavelength was equal to 18 in and the ratio of the half-width of the channel to the wavelength was 0.014 (cf. Fig. 4).

The dimensionless time-mean flow was measured as a function of dimensionless pressure rise per wavelength for very small values of R and with three different amplitude ratios ($\phi = 0.4$, 0.7 and 0.9).

No effects of Reynolds number in the range $R = 0.024$ to 0.034 were observed. The slight difference from the theory of Shapiro *et al.* [57] was attributed to the rectangular cross-section of the experimental pumping duct and end wall effects. Reflux[1] was experimentally confirmed by inserting dye near the wall. The phenomenon of trapping as predicted by Shapiro *et al.* [57] was also verified. The following conclusions were made:

(i) The inertia-free theory is valid up to $R \cong 1$. (ii) The phenomenon of reflux is determined by the Lagrangian time-mean velocity rather than by

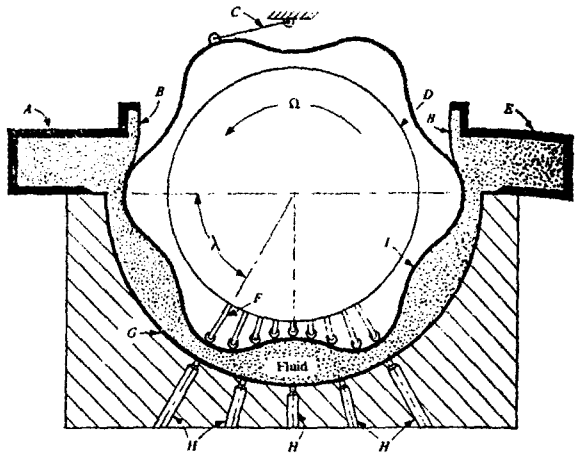


Fig. 4. Apparatus of Weinberg's experiment. A, upstream transition chamber to reservoir above; E, downstream transition chamber to reservoir above; B, spring steel flaps for sealing; C, cable to restrain rotational motion of moving wall I; D, cam rotor; F, radially adjustable arms and roller cams; G, semi-circular back wall; H, pressure-taps; I, flexible moving wall.

the Eulerian time-mean velocity. (iii) The second order expansion in R is valid up to $R \cong 10$.

Yin and Fung [73] were of the opinion that since the experimental verification of Shapiro's model [56] had some drawbacks like consideration of the vibration of only one wall while the mathematical model included vibrations of both the wall, the comparison was not quite satisfactory. In order to match the experiment with the mathematical model, they extended and modified the theoretical analysis of Fung and Yih [20] by imposing vibrations in only one wall. They also tried to rectify the three dimensional effects that arise owing to finite width-to-height ratio. The experimental results matched very closely with the theoretical predictions. The difference between the two was attributed partly to the experimental error and partly to the perturbation technique used in achieving the solution. They also performed experimental verification for reflux.

Brown and Hung [8] who conducted an experiment as well as a numerical investigation claimed that their numerical solution agreed closely with experimental flow visualization and concluded that (i) the transport effectiveness is markedly reduced for pumping against a mild adverse pressure drop, and (ii) increasing the wave amplitude leads to the development of

traveling vortices within the core region of the peristaltic flow. An experimental investigation of the peristaltic flow of a mixture of fluid and solid particles was also carried out by Hung and Brown [26].

3. Theoretical Studies on Peristaltic Transport

Various studies on peristaltic flows, Newtonian as well as non-Newtonian, have been carried out by different investigators with varied considerations. We give here a brief review of the same in a systematic manner.

3.1. *Newtonian flows*

As mentioned earlier, the investigation of peristalsis from a mechanical point of view was launched with a crude experiment by Latham [31] who examined the problem analytically too. The results of that experiment were generally in good agreement with the theoretical investigation of Shapiro [56].

Although investigations similar to that of peristalsis were reported earlier by considering varying breadth along the length such as that according to cosine law (without any reference to peristalsis), a theoretical investigation truly for the peristaltic motion was carried out by Burns and Parkes [10], who studied the flow of a Newtonian fluid through a pipe and a channel by considering sinusoidal variations in the walls along the length. Two cases were examined in particular, viz, (i) peristaltic motion with no pressure gradient and (ii) flow under pressure along a pipe or a channel with fixed walls and sinusoidally varying cross-section. Perturbation solutions, in powers of the ratio of the amplitude of the variation in the pipe radius or channel breadth to the mean radius or the breadth respectively, were given for the stream function.

A contemporary investigation was reported by Shapiro [56] for two-dimensional peristaltic pumping under the two conditions: (i) the appropriate Reynolds number is so small that the flow may be considered inertia-free, and (ii) the length of the peristaltic wave is very long compared to the width of the tube. The small Reynolds number approximation was endorsed by Jaffrin [28]. He further extended the analysis by considering higher order terms to include cases where Reynolds number is higher. An exact solution having a parabolic velocity profile of Poiseuille flow was presented under the said assumptions, which made the flow steady in the wave frame. They also discussed the reflux phenomenon in detail. This was

followed by a further investigation by Barton and Raynor [1] using large wavelength approximation for intestinal flow. They analyzed in a greater detail the case for small Reynolds number.

Fung and Yih [20] formulated a mathematical model in order to study peristaltic pumping using perturbation technique by applying Fourier series expansion used by Taylor [68] for arbitrary Reynolds number but small amplitude ratio (i.e. the ratio between the amplitude of the wave and the width of the channel). The channel was supposed to be of constant width and infinite length. Apart from the application of their model to the flow of blood in arterioles and venules their investigation was focused on the situation where there is some obstruction in the ureter or in the ureter bladder junction. Consequently, dilation of the ureter takes place at the site of the observation and the amplitude of the peristaltic waves become relatively small. In this context, there arises a question whether pumping takes place or not. The answer was found to be in the affirmative when the pressure gradient is less than a critical value depending on the situation. Whenever it exceeds the critical value, a flow reversal (or reflux) is observed. The corresponding analysis for the two-dimensional flow was extended for the axisymmetric case by Yin and Fung [73] for practical applications to biological problems. The two results are qualitatively similar but quantitatively different.

An important investigation for ureteral flow was put forward by Shapiro *et al.* [57] who solved the problem of the flow of a Newtonian fluid through a circular cylindrical tube as well as through a channel by considering the propagation of an integral number of sinusoidal waves of arbitrary amplitude along the walls of the tube/channel of infinite length under the assumption of very small Reynolds number. They derived mathematical expressions for mechanical efficiency of pumping, for the phenomena such as reflux and trapping and also limits for reflux and trapping. A separate expression of the reflux function for small amplitude was also presented by them. Their results including that of reflux were experimentally verified through Latham's experiment [31]. These few studies laid the real foundation stone for the investigation of peristaltic pumping.

Chow [17] generalized the solution put forward by Fung and Yih [20] by considering axisymmetrical geometry with initially non-stationary flow. In fact, it was also more general than that presented by Yin and Fung [73]. The solution given in the form of a power series expansion, dealt with two cases, viz. (i) when the amplitude-radius ratio of the pipe and Reynolds number

are small but the radius-wavelength ratio is unrestricted, and (ii) the radius-wavelength ratio is small but the other two quantities are unrestricted.

Lykoudis and Roos [35] pointed out that the shape of the ureter during peristalsis is not sinusoidal. In the light of this they solved the problem for arbitrary wave shapes and determined the minimum and maximum pressures in a tube the displacements of whose wall vary according to a power-law in the axial direction. The existence of *reflux*[2] with adverse pressure gradient was, however, ignored by them. Manton [37] extended their approach to investigate some general properties of peristalsis. In his asymptotic expansion he accounted for the inertial and viscous effects to an extent greater than that considered by Lykoudis and Roos [35]. These authors determined expressions for the relationship between the mean pressure gradient and the volume flux. A necessary and sufficient condition for the occurrence of trapping was also obtained. They found that reflux occurs whenever there is an adverse mean pressure gradient, independent of the shape of wave. An estimate of the amount of reflux was also derived. Mahrenholtz *et al.* [36] examined the influence of wave form on peristaltic transport of a Newtonian fluid for high Reynolds number in a highly occluded channel.

A study on ureteral peristalsis was made by Griffiths [21] by considering the ureter as a collapsible muscular tube. The tube was supposed to be non-uniform and of finite length and was subjected to non-uniform external pressure. They observed that peristaltic pumping occurs effectively for low flow-rate and pressure. At higher mean flow rates the peristaltic contractions of the ureter may even obstruct the flow of urine. Li and Brasseur [34] made an attempt to explore the pressure shear rate (at the wall) distribution of a Newtonian flow. The conventional sinusoidal wave equation was improved by considering the position of the wall as a function of the minimum radius of the tube, which vibrates in only one direction. The amplitude, which is equal to the radius of the tube minus the minimum tube radius is, however, to be adjusted whenever the degree of contraction of the tube is varied. They studied the difference between integral number and non-integral number of waves propagating along a tube of finite length. Single bolus transport in oesophagus was also discussed. An active membrane model for peristaltic pumping with periodic activation waves has recently been reported by Carew and Pedley [12]. The predictions of their analysis on phase-lag in wall constriction with respect to peak activation wave, lumen occlusion due to thickening of the lumen material with

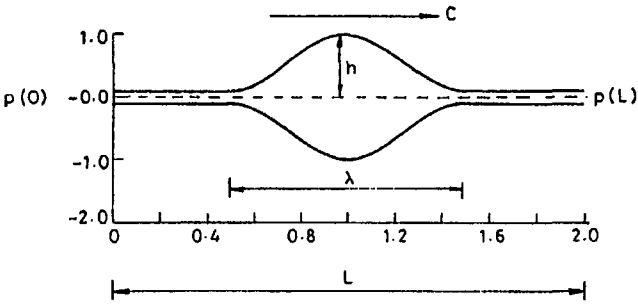
smooth muscle, and the general bolus shape were reported to be in qualitative agreement with experimental observation.

3.2. *Non-Newtonian flows*

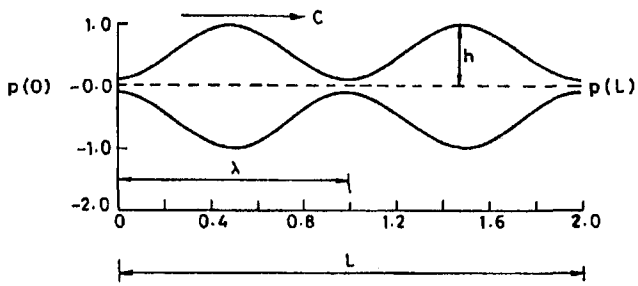
As mentioned above, several authors considered the fluid to behave like a Newtonian fluid for physiological peristalsis including the flow of blood in arterioles. But such a model has only restricted application. Casson [13] derived a semiempirical equation for the flow behavior of varnishes and printing inks by assuming the presence of interparticle forces and disruptive stresses in chain like flocculus. Scott Blair [55] opined that the same description was applicable for blood flow too. He used the available experimental data for human blood and plotted a graph of square root of strain-rate against square root of shear stress, which showed a remarkable linearity with a nonzero value for the intercept on the stress-axis. Experimental data available for animal blood too were reported to conform to this observation. Merrill *et al.* [38] adopted two different methods to obtain data for two different suspensions of red cells in plasma. All the four sets of data displayed linear graphs having positive stress intercepts. Moreover, this model was found to be satisfactory over a large range of shear rates. Charm and Kurland [14, 15] demonstrated that by using Casson's equation, it is possible to extrapolate blood viscometry information obtained at shear-ranges of 5 to 200 $\sec^{-1}$ to shear-rates of 10000 to 100000 $\sec^{-1}$ with less than five percent error.

Although viscoelastic behavior of blood as well as that of the blood vessel wall was adequately taken care of in several investigations (cf. Bohme and Friedrich [3]; Misra and Patra [47]; Imaeda and Goodman [27]), in view of the experimental observations mentioned above, the Casson fluid model of blood seems to bear the potential to explore some important aspects of blood flow through small vessels.

Like blood since other physiological fluids also are mostly of non-Newtonian nature, it is worthwhile to study the dynamics of such fluids by taking their non-Newtonian behavior into consideration. Patel *et al.* [50] found that human faeces is a non-Newtonian power low fluid. Further, Han and Bernett [25] pointed out that bronchial mucus behaves like a non-Newtonian fluid. Raju and Devanathan [53] reported a theoretical investigation for blood flow by considering blood as a non-Newtonian power-law fluid. They employed the perturbation technique used by Chow [17] to solve the problem of the flow in a cylindrical tube with a sinusoidal wave of small amplitude. A similar problem was later considered by Devi



(a) Propagation of single wave of contraction.



(b) Propagation of train-waves of contraction against a pressure difference of $p(L)-p(0)$ along the length of the oesophagus.

Fig. 5. Schematic representation of the peristaltic transport through the oesophagus. c is the velocity of the wave, λ the wavelength, L the oesophageal length, and h the position of the activated wall from the center line.

and Devanathan [18] where the fluid was taken to be micropolar. For viscoelastic liquids, the solution of the problem was presented by Bohme and Friedrich [3]. They also discussed mechanical efficiency of pumping for such liquids. An analysis of this problem for Casson fluid model applicable to blood flow was carried out by Srivastava and Srivastava [63], by considering a peripheral layer of a Newtonian fluid.

Misra and Pandey [44] investigated the transport of a food-bolus through the oesophagus by developing a mathematical model. The oesophagus was treated as a circular tube of finite length and the transport of the masticated food-grains was taken to be governed by a power-law, where the power-law index was supposed to vary, depending on the kind of the food material. (This consideration was based upon the experimental data reported by Patel *et al.* [50] for a similar case.) The peristaltic transport

was supposed to take place axisymmetrically where a single wave was considered to propagate along the wall. The wall of the oesophagus is supposed to be brought under the influence of a periodic transverse contraction wave, owing to which the passage is first shortened by way of contraction of muscles. Then its path is retracted so that its original position is attained. This process continues until the propellant (food material) is completely squeezed out. Misra and Pandey [44] represented such a motion by an equation of form

$$h(z,t) = a - 0.5\phi\left\{1 + cos\frac{2\pi}{\lambda}(z - ct)\right\},$$

where z denotes the axial distance, t the time variable, a the radius of the stationary tube, ϕ the amplitude of the wave, λ the wavelength, c the wave speed and h the radial displacement of the wave from the centreline (cf. Fig. 5).

The wall equation of the tube was taken to fit the natural oesophageal wall contraction that did not involve the expansion beyond the stationary boundary. The spatial as well as the temporal dependence of pressure was studied for a fixed time-averaged flow-rate in the laboratory frame of reference. Comparison was made between the effects of a single wave transport and the propagation of train-waves (cf. Fig. 5) with an integral number of waves in the train. On the basis of the study it has been concluded that in the single wave propagation, there is a forward flow within the wave, while beyond the wave in the oesophagus, there is a tendency of retrograde motion and that if there is some fluid within the duct apart from that in the wave, the average flow will be in the opposite direction. It has been reported that in the train-wave case, the flow is everywhere positive except at the junction of two waves where the rate of backward flow is very high.

Basing upon the observations of the study, Misra and Pandey [44] made a conjecture that it is easier to swallow a pseudoplastic fluid than a dilatant fluid. They also remarked that in the single wave case, the occlusion of the duct should be sufficient to overcome the tendency of retrograde flow in the other region of the oesophagus and that the oesophagus undergoes total occlusion while transporting a single food-bolus.

3.3. *Non-stationary initial flows*

Chow [17] studied the problem of a non-stationary initial flow for the axisymmetric case. Srivastava and Srivastava [61, 62] studied the problem by incorporating an initial flow induced by an arbitrary periodic pressure

gradient for axisymmetric flow in order to model the flow through pulmonary arteries, arterioles, venules and other microvessels. The interaction between peristaltic flow and pressure driven motion was examined by Pozrikidis [51] for molecular connective transport.

3.4. *Two-phase flows*

As mentioned earlier, the investigation on peristaltic transport of a mixture of fluid and solid particles was initiated by Hung and Brown [26] who conducted an experiment for the channel flow of a single bolus. They found that for a neutrally buoyant particle propelled along the axis of the channel by a single bolus, the net particle displacement can be either positive or negative. The instantaneous force acting upon the particle and the resulting particle trajectory are sensitive to the Reynolds number of flow. The net forward movement of the particle increases slightly with the increase in particle size but decreases rapidly as the gap-width of the bolus increases. A reduction in wave amplitude along with an increase in wave speed may lead to a retrograde particle motion. Further, when the centre of the particle is off the longitudinal axis, the particle will undergo rotation as well as translation. Lateral migration of the particles was found to occur in the curvilinear flow region of the bolus leading to a reduction in the net longitudinal transport. The applications included transport of such a mixture for various technological purposes. A theoretical investigation was attempted by Kaimal [30] for the peristaltic motion of a fluid, in which rigid particles are uniformly distributed through an axisymmetric tube of arbitrary wave shape for low Reynolds numbers. He concluded that the presence of particles does not disturb the flow-field. His study included reflux and trapping. This model too was mainly meant for engineering applications. Srivastava and Srivastava [64] used Drew's model [19] and solved a two-dimensional problem of peristaltic transport of a mixture of a Newtonian fluid and small spherical solid particles by neglecting inter-molecular forces for arbitrary Reynolds number and considering the ratio between the amplitude of the wave and the width of the channel to be small. *Reflux*[2] was also dealt with by them. The analysis was aimed at providing a model for chyme flow in the small intestine, spermatic fluid in the cervical canal and the flow of diseased fluid in arterioles.

Peristaltic pumping induced by a sinusoidal travelling wave of moderate amplitude was investigated by Misra and Pandey [41] in the axisymmetrical case for a Newtonian viscous incompressibe fluid mixed with rigid spherical particles of identical size. They employed a continuum mechanics approach,

where the equations governing the conservation of mass and conservation of linear momentum for the fluid and the solid particle phases were taken as follows:

For the Fluid Phase

$$(1-C)\rho_f\left[\frac{\partial}{\partial t}+v_f\frac{\partial}{\partial r}+u_f\frac{\partial}{\partial z}\right]v_f$$
$$=-(1-C)\frac{\partial p}{\partial r}+(1-C)u_s(C)\left[\frac{\partial^2}{\partial r^2}+\frac{\partial^2}{\partial z^2}+\frac{1}{r}\frac{\partial}{\partial r}-\frac{1}{r}\right]v_f+CS(v_p-v_f),$$

$$(1-C)\rho_f\left[\frac{\partial}{\partial t}+v_f\frac{\partial}{\partial r}+u_f\frac{\partial}{\partial z}\right]u_f$$
$$=-(1-C)\frac{\partial p}{\partial z}+(1-C)u_s(C)\left[\frac{\partial^2}{\partial r^2}+\frac{\partial^2}{\partial z^2}+\frac{1}{r}\frac{\partial}{\partial r}\right]u_f+CS(u_p-u_f),$$

$$\frac{\partial}{\partial r}(1-C)v_f+\frac{\partial}{\partial z}(1-C)u_f+\frac{1}{r}(1-C)v_f=0$$

For the Particulate Phase

$$C\rho_p\left[\frac{\partial}{\partial t}+v_p\frac{\partial}{\partial r}+u_p\frac{\partial}{\partial z}\right]v_p=-C\frac{\partial p}{\partial r}+CS(v_f-v_p),$$
$$C\rho_p\left[\frac{\partial}{\partial t}+v_p\frac{\partial}{\partial r}+u_p\frac{\partial}{\partial z}\right]u_p=-C\frac{\partial p}{\partial z}+CS(u_f-u_p),$$
$$\frac{\partial}{\partial r}Cv_p+\frac{\partial}{\partial z}Cu_p+\frac{1}{r}Cv_p=0.$$

In the equations given above, z represents the direction of the wave propagation, whereas r stands for the radial coordinate, (u_f, v_f) denote the axial and radial velocity components of the fluid phase, and (u_p, v_p) those of the particulate phase; ρ_f, ρ_p, $(1-C)\rho_f$ and $C\rho_p$ are, respectively, the actual densities of the materials consisting of the fluid and the solid particle phases, the fluid-phase density, and particle-phase density, C being the volume fraction of the particles in the mixture, p is the pressure, $\mu_s(C)$ is the particle-fluid mixture viscosity and S the drag coefficient of the interaction for the force exerted by one phase on the other. The expression of the drag coefficient was selected as (cf. [67])

$$S=\frac{9}{2}\frac{\mu_0}{a^2}\lambda'(C),$$

where $\lambda'(C) = \frac{4+3[8C-3C^2]^{1/2}+3C}{[2-3C]^2}$, μ_0 being the fluid viscosity, and a the radius of each solid particle suspended in the fluid. The above expression for the drag coefficient bears the potential to account for the finite particulate fractional volume through the function $\lambda'(C)$. For the viscosity of the suspension, the following empirical relation (cf. [15]) was used.

$$\mu_s(C) = \mu_0 \frac{1}{1 - qC},$$

where $q = 0.070 \exp[2.49C + \frac{1107}{T} \exp(1 - 1.69C)]$, in which T represents the absolute temperature (°K).

Charm and Kurland [15] asserted that the formula gives reasonable accuracy for values of C up to $C = 0.6$. The no-slip and impermeability conditions constituted the boundary conditions of the problem discussed by Misra and Pandey [41].

They used a perturbation technique, choosing the amplitude ratio (wave amplitude/tube radius) as the perturbation parameter. The analysis was carried out by duly accounting for the nonlinear convective acceleration terms and the no-slip condition on the wavy wall. The governing equations were developed up to the second order of the amplitude ratio. It was shown that the zeroth order terms yield the Poiseuille flow, while the first order terms give the Orr-Sommerfeld equation. In the absence of the pressure gradient and the wall motion, the mean flows (for the mixture of the fluid and the solid particles) and the mean pressure gradient (averaged over time) were found to be proportional to the square of the amplitude ratio. On the basis of this study they made the following conclusions:

(i) The mean flow induced by the peristaltic motion is proportional to the square of the amplitude ratio and depends on the mean pressure gradient induced by the peristaltic motion.
(ii) At a certain critical value of the pressure gradient, the reversal of flow takes place, which is favored by the presence of particles.
(iii) The mean flow in the axisymmetric case may exhibit the reversal of flow at the boundaries also.

As an illustration of the applicability of their analytical work, they investigated the peristaltic flow through the ureter, by using the necessary data reported by Orkins [49], Bergman [2], Weinberg [71], Boyarsky [6] and Griffiths [21].

3.5. *Two-layer flows*

Flows in certain physiological processes like the vasomotion of some blood vessels, motion in ductus efferentes of the male reproductive tract, transport of spermatozoa in the cervical canal, movement of chyme in the gastrointestinal tract, involve flow of a mucus layer adhered to the innermost surface of the walls of the ducts (cf. Guyton [24]). It is observed that the viscosity of the fluid in the peripheral region is different from that in the core region. Bugliarello and Sevilla [9] as well as Cockelet [16] showed by carrying out experiments that for blood flowing through small vessels, there is a peripheral layer of plasma, which is a Newtonian fluid, and a core region which is non-Newtonian, that can be regarded as a suspension of erythrocytes. Taking this fact into consideration, Shukla *et al.* [58] tried to include a peripheral layer of different viscosity in peristaltic flows through tubes and channels using Stokes approximation. They applied the tube solution to intestinal flows and the channel solution to the flows in the ductus efferentes of the male reproductive tract. Shukla and Gupta [59] further extended this analysis to incorporate power-law nature of the fluid in order to apply to blood flow problems. Both of these studies, however, ignored the conservation of mass in separate layers. Though quantitatively the flow rate might not have been affected to a very large extent, the shape of the interface was wrongly deduced. Brasseur *et al.* [7] pointed out this mistake and presented a correct solution where the interface was considered as a streamline in the steady wave frame. The mechanical efficiency and also the phenomena of trapping and reflux were also elucidated for channel flow. An extension of this study to axially symmetric case was carried out by Rao and Usha [52]. Peristaltic transport of a biological fluid in a pipe of elliptic cross-section was studied by Usha and Rao [70].

An analytical study of the two-dimensional flow of a power-law fluid with a peripheral layer was conducted by Misra and Pandey [43]. By using large wavelength approximations, the solution was obtained in the form of a stream function from which the shape of the interface was determined. A relation between the flow-rate and the pressure difference was established and using that relation, analytical expressions for the maximum pressure and the maximum flow-rate were derived. They also deduced the expressions for the mechanical efficiency of pumping, the trapping limit and the reflux limit. The study reveals that the flow increases with an increase in the flow behavior index or with an increase in the peripheral layer viscosity. They concluded that in the case of peristaltic pumping of physiological fluids for which the viscosity of the peripheral layer is usually less than that of

the core region, a thinner peripheral layer whose viscosity is considerably large (but does not exceed the viscosity of the core region) bears the potential to enhance the flow-rate. This is in coherence with the observation of Brasseur *et al.* [7]. Misra and Pandey [43] made an observation that the maximum pressure difference for physiological power-law fluids is less than that for Newtonian fluids. They conjectured that the maximum pressure difference increases indefinitely for a more viscous peripheral layer when the occlusion is large enough and further that in the case of total occlusion, it may not be possible to check the flow by applying a finite pressure difference, however large. They proclaimed that the peristaltic pumping is more efficient when the physiological power-law fluid has a thinner but a more viscous peripheral layer and is subjected to large occlusions and also that the pumping efficiency of physiological power-law fluids is less than that of Newtonian fluids.

Misra and Pandey [45] developed a mathematical model with an aim to study the pulsatile flow of chyme through the small intestine treated as a long cylindrical intestinal duct under the influence of a mucus layer existing adjacent to the inner surface of the duct. The chyme was taken to be propelled by the sinusoidal motion of the wall. The wall motion is due to some electrochemical reactions that take place within the human body. Both the chyme and the mucus are treated as power-law fluids having different viscosity. Small Reynolds numbers and inertia-free flows have been investigated. This is in coherence with the observation made by Han and Barnett [25] that mucus layer is non-Newtonian by nature. Small Reynolds number and inertia-free flows have been investigated with particular emphasis, because of the observation made by Jaffrin and Shapiro [29], Buthand [11] and Jaffrin [28] who on extending the inertia-free flow of Shapiro *et al.* [57] to next higher order terms for Reynolds number and wave number, found that their results were in good agreement with the zeroth order solution when Reynolds number and wave number are small. Misra and Pandey [45] concluded that the peripheral layer thickness is less uniform when it is less viscous and that a reduction in the value of the flow behavior index makes the peripheral layer thicker while waxing and thinner while waning. They pointed out that unlike two-layered axisymmetric Newtonian flows, reflux does not take place in the entire domain in the corresponding flow of power-law fluids.

It was pointed out by Misra and Ghosh [40] that blood flow in the micro-vessels of the lung may be described as a channel flow, where as that in arterioles and venules as an axisymmetric flow. The peristaltic flow of

blood in small vessels was investigated by Misra and Pandey [46] through the development of a mathematical model in which blood was treated as a two-layer fluid where the core region was described by Casson model and the peripheral region was taken to be Newtonian viscous. Wave frame steady solutions for channel flow as well as axisymmetric flow were presented by them. Consideration of mass conservation has been made separately in the two layers. It has been shown that the higher the viscosity of the peripheral layer, the greater is the flow rate. The study indicates further that (i) a thinner peripheral layer enhances the flow rate, whereas the flow-rate reduces when the yield stress increases, (ii) the flow-rate in the case of a single layer is higher than the two-layer flow-rate when the peripheral layer is more viscous than the core layer and (iii) the flow-rate in the case of axisymmetric flow is greater than that of channel flow under identical conditions.

4. Flows through Tubes of Non-Uniform Cross-Section

Lee and Fung [32] studied flow in small blood vessels of non-uniform cross-section, considering the flow to be of peristaltic nature. Peristaltic transport as well as mixing of chyme in small intestine was investigated by Lew *et al.* [33].

Considering the non-uniform geometry of viscometric capillary tubes and blood vessels, Manton [37] examined the peristaltic flow of a Newtonian fluid through an axisymmetric tube whose radius varies slowly in the axial direction and whose wall is subjected to arbitrary wave propagation. Application of Stokes approximation and use of a perturbation technique were made for performing the analysis. Gupta and Seshadri [23] presented a solution of peristaltic pumping of Newtonian fluids in channels and tubes of non-uniform cross-section with a particular reference to the spermatic flow in the vas deferens. They concluded that peristalsis is responsible for one-third of the total flow in the vas deferens. Similar solutions of peristaltic flows in non-uniform tubes were reported by Rath [54], Srivastava and Srivastava [61].

Misra and Pandey [42] studied the nonlinear peristaltic flow of a Newtonian viscous incompressible fluid through a tapered tube, where the wave propagating along the wall of the tube is sinusoidal and the initial flow is Hagen-Poiseuille. The derived analytical expressions were computed to have an in-depth study of an important physiological problem, viz. spermatic flow through the vas deferens, in which the peristaltic motion is quite

dominant. Their theoretical prediction for the flux-rate was found to be in good agreement with the experimentally measured values reported by Guha *et al.* [22] for rhesus monkeys.

5. Numerical Investigations

A finite element approach was adopted by Tong and Vawter [69] to analyze peristaltic pumping, by considering that both the wavelength and the wave-amplitude have a strong influence on the flow-field. They studied the reflux phenomenon for short wavelengths, as well as for longer wavelengths. Their method for solving peristaltic flow problems was subsequently extended by Nergin *et al.* [48].

Computational investigations of two-dimensional non-linear peristaltic flows under the assumption of finite wall-wave curvature and Reynolds number were carried out by Brown and Hung [8]. They used orthogonal curvilinear coordinates and employed an implicit finite-difference technique for solving the problem. The same problem was also studied by them experimentally. It was concluded that (i) the inertia-free theory is valid up to Reynolds number of the order of 1, and (ii) the second order expansion in Reynolds number is valid up to Reynolds number of the order of 10.

Takabatake and Ayukawa [65] used upwind SOR method to solve two-dimensional peristaltic motion with moderate Reynolds number and compared their results with those achieved by applying perturbation techniques. It was found that the validity of the perturbation solutions given by Jaffrin [28] and Zien and Ostrach [74] are confined within a range narrower than that they had predicted. It was concluded that the *reflux*[1] phenomenon in the flow changes the whole situation according to Reynolds number. They also claimed to find a good agreement of their computational results with experimental results.

Takabatake *et al.* [66] adopted an upwind finite difference technique to replace the channel cross-section of Takabatake and Ayukawa [65] by a circular one. They inferred that much greater peristaltic mixing and transport occur in a circular tube than that in a plane channel. Their discussion included mechanical efficiency of pumping, reflux and trapping. They also pointed out the term left out in the calculations for the mechanical efficiency in the case of a circular cylindrical tube and concluded that the efficiency is more in this case than that in the case of the channel.

A numerical simulation of the peristaltic reflex of a small bowel was presented by Miftakhov and Wingate [39].

References

[1] G. Barton and S. Raynor, Peristaltic flow in tubes, *Bull. Math. Biophy.* **30** (1968) 663–688.

[2] H. Bergman, *The Ureter* (Harpens & Row, New York, 1967).

[3] G. B. Bohme and R. Friedrish, Peristaltic flow of viscoelastic liquids, *J. Fluid Mech.* **128** (1983) 109–122.

[4] H. L. Borison and S. C. Wang, Physiology and pharmacology of vomiting, *Pharmacol. Rev.* **5** (1953) 193.

[5] J. F. Bosma, Deglutition: Pharyngeal stage, *Physiol. Rev.* **37** (1957) 275.

[6] S. Boyarsky (ed.), *Neuorgenic Bladder* (Williams and Wilkins Co., Baltimore, 1967).

[7] J. G. Brasseur, S. Corrsin and N. Q. Lu, The influence of a peripheral layer of different viscosity on peristaltic pumping with Newtonian fluids, *J. Fluid Mech.* **174** (1987) 495–519.

[8] T. D. Brown and T. K. Hung, Computational and experimental investigations of two dimensional non-linear peristaltic flows, *J. Fluid Mech.* **83** (1974) 249–272.

[9] G. Bugliarello and J. Sevilla, Velocity distribution and other characteristics of steady and pulsatile blood flow in fine glass tubes, *Biorheology* **7** (1970) 85–107.

[10] J. C. Burns and T. Parkes, Peristaltic motion, *J. Fluid Mech.* **29** (1967) 731–743.

[11] H. Buthaud, The influence of unsymmetry, wall slope and wall motion on peristaltic pumping at small Reynolds number, M. S. thesis, Department of Mechanics, The John Hopkins University (1971).

[12] E. O. Carew and T. J. Pedley, An active membrane model for peristaltic pumping: Part 1-Periodic activation waves in an infinite tube, *Trans. ASME J. Biomech. Engng.* **119** (1997) 66–76.

[13] N. Casson, Rheology of dispersive systems, ed. C. C. Mills, Pergamon, Oxford, **84** (1959).

[14] S. E. Charm and G. S. Kurland, Viscometry of human blood for shear rates of 0–100000 $\sec^{-1}$, *Nature* **206** (1965) 617.

[15] S. E. Charm and G. S. Kurland, *Blood Flow and Micro-Circulation* (John Wiley, New York, 1974).

[16] G. R. Cockelet, The rheology of human blood, in *Biomechanics, Its Foundation and Objectives*, eds. Y. C. Fung, N. Perrone and M. Anliker (Prentice-Hall, Englewood Cliffs, N. J., 1972), pp. 63–103.

[17] T. S. Chow, Peristaltic transport in a circular cylindrical pipe, *Trans. ASME, Ser. E. J. Appl. Mech.* **37** (1970) 901–905.

[18] G. Devi and R. Devanathan, Peristaltic motion of a micropolar fluid, *Proc. Indian Acad. Sci. A* **81** (1975) 149–163.

[19] D. A. Drew, Stability of a Stokes layer of a dusty gas, *Phys. Fluids* **22** (1979) 2081–2084.

[20] Y. C. Fung and C. S. Yih, Peristaltic transport, *Trans ASME, Ser. E. J. Appl. Mech.* **35** (1968) 669–675.

[21] D. J. Griffiths, Flow of urine through the ureter: A collapsible, muscular tube undergoing peristalsis, *J. Biomech. Engg.* **111** (1989) 206–211.

[22] S. K. Guha, H. Kaur and A. M. Ahmed, Mechanism of spermatic flow in the vas deferens, *Med. Biol. Engg.* **13** (1975) 518.

[23] B. B. Gupta and V. Seshadri, Peristaltic transport in non-uniform tubes, *J. Biomechanics* **9** (1982) 105–109.

[24] A. C. Guyton, *Textbook of Medical Physiology* (Saunders, Philadelphia, 1971).

[25] C. D. Han and B. Barnett, Measurements of the rheological properties of biological fluids, *Rheology of Biological Systems*, eds. Hanry L. Gabelnick and Mitchell Litt, Charles C. Thomas Publisher, **III** (1973).

[26] T. K. Hung and T. D. Brown, Solid-particle motion in two-dimensional peristaltic flows, *J. Fluid Mech.* **73** (1976) 77–96.

[27] K. Imaeda and F. O. Goodman, Analysis of non-linear pulsatile blood flow in arteries, *J. Biomechanics* **13** (1980) 1007–1022.

[28] M. Y. Jaffrin, Inertia and streamline curvature on peristaltic pumping, *Intl. J. Engng. Sci.* **11** (1973) 681–699.

[29] M. Y. Jaffrin and A. H. Shapiro, Peristaltic pumping, *Annual Reviews of Fluid Mechanics* **3** (1971) 13–36.

[30] M. R. Kaimal, Peristaltic pumping of Newtonian fluid with particles suspended in it at low Reynolds number under long wavelength approximations, *Trans. ASME, J. Appl. Mech.* **45** (1978) 32–36.

[31] T. W. Latham, Fluid motions in the peristaltic pump, M.S. thesis, M.I.T. (1966).

[32] J. S. Lee and Y. C. Fung, Flow in non-uniform small blood vessels, *Microvasc. Res.* **3** (1971) 272.

[33] H. S. Lew, Y. C. Fung and C. B. Lowenstein, Peristaltic carrying and mixing of chyme in small intestine, *J. Biomechanics* **4** (1971) 297.

[34] M. Li and J. G. Brasseur, Non-steady peristaltic transport in finite-length tubes, *J. Fluid Mech.* **248** (1993) 129–151.

[35] P. S. Lykoudis and R. Roos, The fluid mechanics of the ureter from a lubrication point of view, *J. Fluid Mech.* **43** (1970) 661–674.

[36] O. H. Mahrenholtz, M. G. Mank and R. U. Zimmermann, The influence of wave form on peristaltic transport, *Biorheology* **15** (1978) 501–510.

[37] M. J. Manton, Long wavelength peristaltic pumping at low Reynolds number, *J. Fluid Mech.* **68** (1975) 467–476.

[38] F. W. Merrill, A. M. Benis, E. R. Gilliland, T. K. Sherwood and E. W. Saltzman, Pressure-flow relations of human blood in hollow fibres at low flow rates, *J. Appl. Physiology* **20** (1965) 954–967.

[39] R. Miftakhov and D. Wingate, Numerical simulation of the peristaltic reflex of the small bowel, *Biorheology* **31** (1994) 309–325.

[40] J. C. Misra and S. K. Ghosh, A mathematical model for the study of blood flow through a channel with permeable walls, *Acta Mech.* **122** (1997) 137–153.

[41] J. C. Misra and S. K. Pandey, Peristaltic transport of a particle-fluid suspension in a cylindrical tube, *Computers Math. Applic.* **28** (1994) 131–145.

[42] J. C. Misra and S. K. Pandey, Peristaltic transport in a tapered tube, *Math. Comput. Modelling* **22** (1995) 137–151.

[43] J. C. Misra and S. K. Pandey, Peristaltic transport of a non-Newtonian fluid with a peripheral layer, *Internat. J. Engng. Sci.* **37** (1999) 1841–1858.

[44] J. C. Misra and S. K. Pandey, A mathematical model for oesophageal swallowing of a food bolus, *Math. Comput. Modelling* **33** (2001) 997–1009.

[45] J. C. Misra and S. K. Pandey, Peristaltic flow of a multilayered power-law fluid through a cylindrical tube, *Internat. J. Engg. Sci.* **39** (2001) 387–402.

[46] J. C. Misra and S. K. Pandey, Peristaltic transport of blood in small vessels: study of a mathematical model, *Computers Math. Applic.* **43** (2002) 1183–1193.

[47] J. C. Misra and M. K. Patra, A non-Newtonian fluid model for blood flow through arteries under stenotic conditions, *J. Biomech.* **26** (1992) 1129–1141.

[48] M. P. Nergin, W. J. Shack and T. J. Lardner, A note on peristaltic pumping, *Trans. ASME J. Appl. Mech.* **96** (1974) 520–521.

[49] L. A. Orkins, *Trauma in the Ureter: Pathogenesis and Management*, F. A. Davis Co., Philadelphia, Pa. (1964).

[50] P. D. Patel, B. F. Picologlou and P. S. Lykoudis, Biorheological aspects of colonic activity — Part II: Experimental investigation of the rheological behaviour of human faeces, *Biorheology* **10** (1973) 441–445.

[51] C. Pozrikidis, A study of peristaltic flow, *J. Fluid Mech.* **180** (1987) 521–527.

[52] A. R. Rao and S. Usha, Peristaltic transport of two immisible viscous fluids in a circular cylindrical tube, *J. Fluid Mech.* **298** (1995) 271–285.

[53] K. K. Raju and R. Devanathan, Peristaltic motion of a non-Newtonian fluid, *Rheol. Acta* **11** (1972) 170–178.

[54] H. J. Rath, Peristaltic flow through a lobe-shaped tube, *Intl. J. Mech. Sci.* **24** (1982) 359–367.

[55] G. W. Scott Blair, An equation for the flow of blood, plasma and serum through glass capillaries, *Nature* **183** (1959) 631.

[56] A. H. Shapiro, Pumping and retrograde diffusion in peristaltic waves, *Proc. Workshop in Ureteral Reflux in Children* (1967), pp. 109–126.

[57] A. H. Shapiro, M. Y. Jaffrin and S. L. Weinberg, Peristaltic pumping with long wavelengths at low Reynolds number, *J. Fluid Mech.* **37** (1969) 799–825.

[58] J. B. Shukla, R. S. Parihar, B. R. P. Rao and S. P. Gupta, Effects of peripheral layer viscosity on peristaltic transport of a bio-fluid. *J. Fluid Mech.* **97** (1980) 2225–2237.

[59] J. B. Shukla and S. P. Gupta, Peristaltic transport of a power-law fluid with variable consistency, *Trans. ASME J. Biomech. Engg.* **104** (1982) 182–186.

[60] L. M. Srivastava and V. P. Srivastava, Peristaltic transport of a two layered model of a physiological fluid, *J. Biomechanics* **15** (1982) 257.

[61] L. M. Srivastava and V. P. Srivastava, Peristaltic transport of a physiological fluid Part I — Flow in a non-uniform geometry, *Biorheology* **20** (1983) 153.

[62] L. M. Srivastava and V. P. Srivastava, Peristaltic transport of a physiological fluid, Part III — Applications, *Biorheology* **20** (1983) 179.

[63] L. M. Srivastava and V. P. Srivastava, Peristaltic transport of blood: Casson Model-II, *J. Biomechanics* **17** (1984) 821–829.

[64] L. M. Srivastava and V. P. Srivastava, Peristaltic transport of a particle-fluid suspension, *Trans. ASME J. Biomech. Engg.* **111** (1989) 157–165.

[65] S. Takabatake and K. Ayukawa, Numerical study of two-dimensional peristaltic flows, *J. Fluid Mech.* **122** (1982) 439–465.

[66] S. Takabatake, K. Ayukawa and A. Mori, Peristaltic pumping in circular cylindrical tubes: A numerical study of fluid transport and its efficiency, *J. Fluid Mech.* **193** (1988) 269–283.

[67] C. K. W. Tam, The drag on a cloud of spherical particles in low Reynolds number flow, *J. Fluid Mech.* **38** (1969) 537–546.

[68] G. I. Taylor, Analysis of the swimming of microscopic organisms, *Proc. Roy. Soc. London, A* **209** (1951) 447.

[69] P. Tong and O. Vawter, An analysis of peristaltic pumping, *Trans. ASME J. Appl. Mech.* **39** (1972) 857–862.

[70] S. Usha and A. R. Rao, Peristaltic transport of a bio-fluid in a pipe of elliptic cross-section, *J. Biomech.* **28** (1995) 45–52.

[71] S. R. Weinberg, Physiology of the ureter, in *The Ureter*, ed. H. Bergeman (Harper & Row, 1967), pp. 48–66.

[72] S. L. Weinberg, E. C. Eckstein and A. H. Shapiro, An experimental study of peristaltic pumping, *J. Fluid Mech.* **49** (1971) 461–479.

[73] F. Yin and Y. C. Fung, Peristaltic waves in a circular cylindrical tube, *Trans. ASME, Ser. E. J. Applied Mech.* **36** (1969) 93–112.

[74] T. Zien and S. Ostrach, A long wave approximation to peristaltic motion, *J. Biomech.* **3** (1970) 63–75.

This page is intentionally left blank

CHAPTER 8

MATHEMATICAL MODELLING OF DNA KNOTS AND LINKS

J. C. MISRA* and S. MUKHERJEE
Department of Mathematics, Indian Institute of Technology Kharagpur 721302, India
jcm@maths.iitkgp.ernet.in

A. K. DAS
Department of Biotechnology, Indian Institute of Technology Kharagpur 721302, India

1. Introduction

A defining moment for DNA research was the discovery of its structure half a century ago on 25 April 1953 by James Watson and Francis Crick [15] describing the entwined embrace of two strands of deoxyribonucleic acid (DNA). The structure of DNA is the foundation for understanding different physiological phenomenon like molecular damage and repair, replication and inheritance of genetic material, as well as the diversity and the evolution of species. One of the longstanding issues in molecular biology is the three-dimensional structure (shape) of proteins and deoxyribonucleic acid (DNA) in solution in the cell and the relationship between structure and function. Ordinarily, the structure of protein and DNA is determined by X-ray crystallography or electron microscopy. Because of the close packing needed for crystallization and the manipulation required to prepare a specimen for electron microscopy, these methods provide little direct evidence for molecular shape in solution.

The structure of DNA suggests its three dimensional arrangement as two very long curves that are intertwined millions of times, linked to other curves, and subjected to four or five successive orders of coiling to convert it into a compact form for information storage. Hence the arrangement of these two curves is called the duplex DNA, consisting of two

backbone strands wound about each other in right-handed helical fashion. DNA strand consists of sugar phosphate backbone with a nitrogenous base attached to each sugar. The four bases are adenine (A), guanine (G), cytosine (C) and thymine (T) (cf. [2, 15, 16]). The two strands are held together by hydrogen bonding between the bases with A always paired with T with two hydrogen bonds and G always paired with C with three hydrogen bonds. The bacterial DNA is usually circular. Although human DNA is linear, it is extremely long and tacked down to a protein scaffold at various points on the DNA. This periodic attachment endows human DNA with topological constraints similar to those for circular DNA. Although DNA is considered to be the master molecule of the body, actually protein is the working molecule. Hence these topological constraints of DNA can interfere with vital cellular processes such as replication and transcription due to different mode of interactions with proteins. Enzymes are usually proteins and are involved in these topological entanglement problems that arise through cellular metabolism and replication. In this case topoisomerases, which are enzymes that mediate the passage of one segment of DNA through an enzyme-bridged transient break in the backbone strands of another DNA segment, are responsible for unlinking the DNA. Other enzymes called recombinases break two DNA segments and interchange the ends, resulting in an exchange of genetic information. Tangle calculus has been successfully used to study recombinases. The topological approach to enzymology is an experimental protocol in which the descriptive and analytical powers of knot theory and tangle calculus are employed in an indirect effort to determine the enzyme mechanism and the structure of active enzyme-DNA complexes *in vitro* (in a test tube).

Due to the uniqueness of the bonding partner for each nucleotide, knowledge of the sequence along one backbone implies knowledge of the sequence along the other backbone. In the classic Watson–Crick double helix model for DNA [2], the ladder is twisted in a right-hand helical fashion, with an average and nearly constant pitch of approximately 10.5 base pairs per full helical twist. The local helical pitch of duplex DNA is a function of both the local base pair sequence and the cellular environment in which the DNA lives; if a DNA molecule is under stress, or is constrained to live on the surface of a protein, or is being acted upon by an enzyme, the helical pitch can change. The packing, twisting, and topological constraints all taken together mean that topological entanglement poses serious functional problems for DNA. This entanglement would interfere with, and be exacerbated by, the vital life processes of replication, transcription, and

recombination. For information retrieval and cell viability, some geometric and topological features must be introduced into the DNA. For example, the Watson–Crick helical twist of duplex DNA may require local unwinding in order to make room for a protein involved in transcription to attach to the DNA. The DNA sequence in the vicinity of a gene may need to be altered to include a promoter or repressor. During replication, the daughter duplex DNA molecules become entangled and must be disentangled in order for replication to proceed to completion. After the process is finished, the original DNA conformation must be restored. Some enzymes maintain proper geometry and topology by passing one strand of DNA through another by means of a transient enzyme-bridged break in one of the DNA strands. Other enzymes break the DNA apart and recombine the ends by exchanging them, a move performed by recombinases. Recently, it has been found that Topoisomerase III and IV also help in DNA recombination. The description and quantification of the three-dimensional structure of DNA and the changes in DNA structure due to the action of these enzymes have required serious use of geometry and topology in molecular biology. This use of mathematics as an analytic tool is particularly important because there is no experimental way to observe the dynamics of enzymatic action directly.

In the experimental study of DNA structure and enzyme mechanism, biologists developed the topological approach to enzymology, shown schematically in Figs. 1 and 2. In this approach, one performs experiments on circular substrate DNA molecules [7]. Cloning techniques to contain regions that a certain enzyme will recognize and act upon, genetically engineer these circular substrate molecules. The circular form of the substrate molecule traps an enzymatic topological signature in the form of DNA knots and links (catananes). These DNA knots and links of the reaction product DNA molecules are observed by gel electrophoresis and electron microscopy.

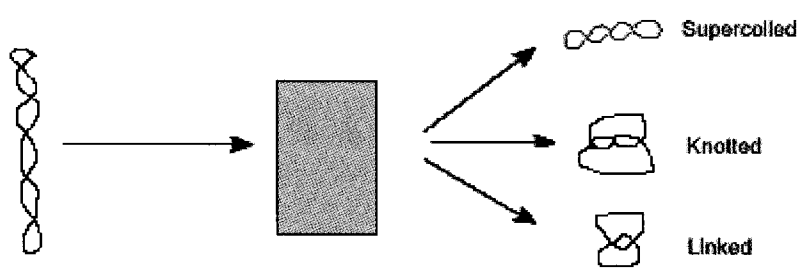


Fig. 1. Topological approach to enzymology.

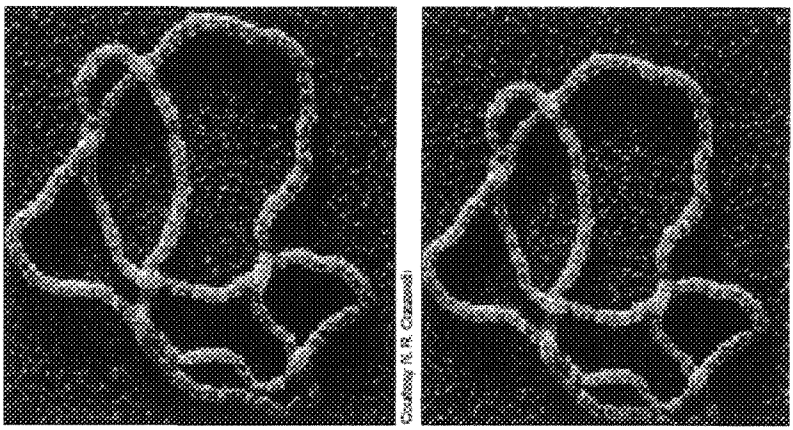


Fig. 2. (a) DNA (+) Whitehead link, (b) DNA knot 6_2^*.

By observing the changes in geometry (supercoiling) and topology (knotting and linking) in DNA caused by an enzyme, the enzyme mechanism can be described and quantized.

The topological approach to enzymology poses an interesting challenge for mathematicians as to how one can deduce enzyme mechanisms from the observed changes in DNA geometry and topology. This requires the construction of mathematical models for enzyme action and the use of these models to analyze the results of topological enzymology experiments. The entangled form of the product DNA knots and links contains information about the enzymes that made them. In addition to utility in the analysis of experimental results, the use of mathematical models forces all of the background assumptions about the biology to be laid out carefully. At this point they can be examined and dissected, and their influence on the biological conclusions drawn from experimental results can be determined.

Ernst and Sumners [7] were the first to introduce tangle model. They also used the model to analyze the Tn3 resolvase site-specific recombination system. It was proved mathematically that, in a processive recombination event, Tn3 resolvase binds to its unknotted, negatively supercoiled substrate (sites in direct repeat), fixes three negative supercoils, and each round of recombination introduces a positive crossing in the domain. It was also proved that, given biologically reasonable assumptions, this is the only possible explanation for the experimental data. In 2001 Darcy [4] modelled the Xer recombinase using 4-plat oriented equation. But since Xer is

non-processive the model gave an infinite number of solutions. The solutions of the model depended upon the initial assumptions that were made.

2. Mathematical Background

Site-specific recombination affects the topology of circular DNA substrates. These changes in topology can be characterized experimentally. Based on the experimental data, biological models for enzymatic mechanisms can be proposed. Only a mathematical treatment of this problem can give a definite answer. The fields of knot theory and low dimensional topology are needed to analyze site-specific recombination reactions.

2.1. *Topological tools for DNA analysis*

Fortunately for biological applications, most of the circular DNA falls into the mathematically well-understood family of 4-plats (cf. [1, 6, 8, 11]). This family consists of knot and link configurations produced by patterns of plectonemic supercoiling of pairs of strands about each other. All "small" knots and links are members of this family — more precisely, all prime knots with crossing number less than 8 and all prime (two-component) links with crossing number less than 7 are 4-plats. For *in vitro* binding of circular DNA with enzymes, we can consider the enzyme mechanism as a machine that transforms 4-plats into other 4-plats. We need a mathematical language for describing and computing these enzyme-mediated changes. In many enzyme-DNA reactions, a pair of sites that are distant on the substrate circle are juxtaposed in space and bound to the enzyme. The enzyme then performs its topological moves, and the DNA is then released. A mathematical language is also required to describe configurations of linear strings in a spatially confined region. This is accomplished by means of the mathematical concept of tangles, which were introduced into knot theory by Conway [2]. Tangle theory is knot theory done inside a 3-ball with the ends of the strings firmly glued down. A mathematical model for the study of enzymatic action on DNA knots and links was recently developed and analyzed by Misra *et al.* [13].

The family of tangles that can be converted to the trivial tangle by moving the endpoints of the strings on S^2 is the family of rational tangles [14]. Equivalently, a rational tangle is one in which the strings can be continuously deformed (leaving the endpoints fixed) entirely into the boundary 2-sphere of the 3-ball, with no string passing through itself or through another string.

Rational tangles form a homologous family of 2-string configurations in B^3 and like 4-plats, look like DNA configurations being built up out of a pattern of plectonemic supercoiling of pairs of strings. More specifically, enzymes are often globular in shape and are topologically equivalent to our unit-defining ball. Thus, in an enzymatic reaction the enzyme bound DNA forms a 2-string tangle. Since the amount of bound DNA is small, the enzyme-DNA tangle so formed admits projections with few nodes and therefore is very likely rational. For example, all locally unknotted 2-string tangles having less than five crossings are rational. There is a second, more natural argument for rationality of the enzyme-DNA tangle. In all cases studied intensively, DNA is bound to the surface of the protein. This means that the resulting protein-DNA tangle is rational, since any tangle whose strings can be continuously deformed into the boundary of the defining ball is automatically rational.

There is a classification scheme for rational tangles that is based on a standard form that is a minimal alternating diagram. The classifying vector for a rational tangle is an integer entry vector $(a_1, a_2, \ldots, a_n)$ of odd or even length, with all entries (except possibly the last) non-zero and having the same sign and with $|a_1| > 1$. The integers in the classifying vector represent the left-to-right (west-to-east) alternation of vertical and horizontal windings in the standard tangle diagram, always ending with horizontal windings on the east side of the diagram. Horizontal winding is the winding between strings in the top and bottom (north and south) positions; vertical winding is the winding between strings in the left and right (west and east) positions. By convention, positive integers correspond to horizontal plectonemic right-handed supercoils and vertical left-handed plectonemic supercoils; negative integers correspond to horizontal left-handed plectonemic supercoils and vertical right-handed plectonemic supercoils. Two rational tangles are of the same type if and only if they have identical classifying vectors. Due to the requirement that $|a_1| > 1$ in the classifying vector convention for rational tangles, the corresponding tangle projection must have at least two nodes. There are four rational tangles $\{(0); (0;0); (1); (-1)\}$ that are exceptions to this convention ($|a_1| = 0$ or 1).

Tangles can be used to build a model that will compute the topology of synaptic complex in a single recombination event, with knowledge of the topology of the substrate and product. In site-specific recombination on circular DNA substrate, two kinds of geometric manipulation of the DNA occur. The first is a global ambient isotopy, in which a pair of distant recombination sites are juxtaposed in space and the enzyme binds to the

molecule(s), forming the synaptic complex. Once synapsis is achieved, the next move is local and due entirely to enzyme action. Within the region occupied by the enzyme, the substrate is broken at each site, and the ends are recombined.

Within the region controlled by the enzyme, the enzyme breaks the DNA at each site and recombines the ends by exchanging them. Hence the enzyme itself can be modeled as a 3-ball. The synaptosome consisting of the enzyme and bound DNA forms a 2-string tangle.

2.2. *Definitions*

A **knot** is a simple closed curve embedded in 3-space. A **link** is a disjoint union of such simple closed curves (cf. [1, 11]). Two knots A and B are said to be equivalent if and only if A can be smoothly deformed into B, and we write $A \equiv B$. A knot that can be deformed to lie on a plane, with no crossings, is called a **trivial knot**, or the **unknot**. Likewise, a **trivial link** with two components consists of two circles that can be deformed to lie flat on a plane. The 2-dimensional representation of a 3-dimensional knot is known as the **projection** of a knot. The crossing number of a knot is the minimum number of crossings over all the projections of a knot. For example, the crossing number of a trefoil knot is 3 (see Fig. 3). The problem of deciding when two knots or links are equivalent is not easy. Many invariants of knots and links, both geometric and algebraic, have been developed throughout the years. Some examples of geometric invariants are the crossing number of a link, and the linking number of a link with two oriented components.

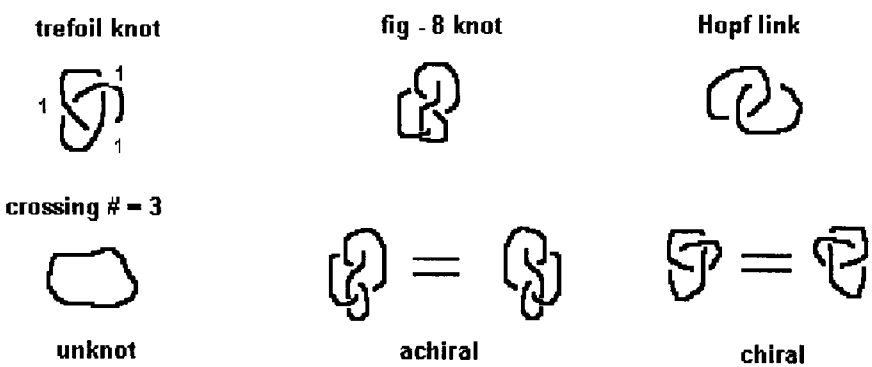


Fig. 3. Different types of knots.

The following definitions will lead to another knot invariant. Let K be a link with a fixed orientation. The link obtained after inverting the orientation of K is denoted by $(-K)$, and it is called the **inverse** of K. Likewise, the link obtained by reflection of K with respect to a plane, is called the **mirror image** of K and is denoted by K^*. If $K = (-K)$, then K is said to be **invertible**. K is **achiral** if $K = K^*$. If $K \neq K^*$, K is said to be **chiral**.

2.3. *2-string tangles*

A unit ball is considered in R^3. In the XY plane (see Fig. 4) the positive Y-axis is considered to point north, and the positive X-axis to point east. Let {NE, NW, SE, SW} be four fixed equatorial points of the unit ball. A **2-string tangle** can be thought of as two strands with end-points {NE, NW, SE, SW} together with the unit ball to that contains them. The basic definition is illustrated in Fig. 4.

As is the case with knots, tangles are also studied through their projections. A tangle diagram is the image of the 2-string tangle when it is projected onto the equatorial disc. Two tangle diagrams represent equivalent tangles if strands of the one can be deformed into strands of the other.

There are 3 types of tangles (see Fig. 5):

- **Rational Tangle:** a, a′) any rational tangle can be obtained from the trivial tangle shown in a) by moving the strands' ends on the boundary of the ball.
- **Locally knotted:** b) a locally knotted tangle contains a knotted strand.
- **Prime Tangle:** c) tangles which are not rational or locally knotted are said to be prime.

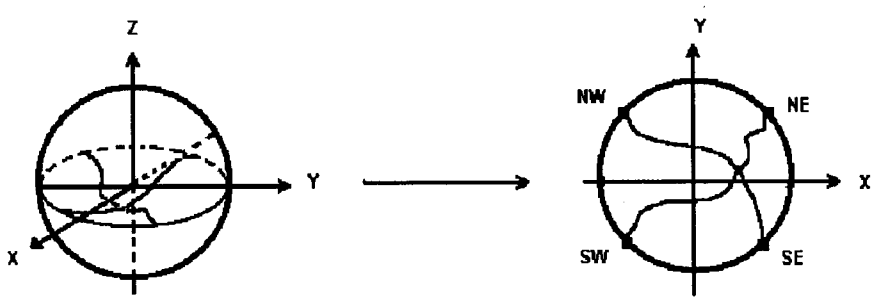


Fig. 4. Projection of a unit ball from R^3 to S^2.

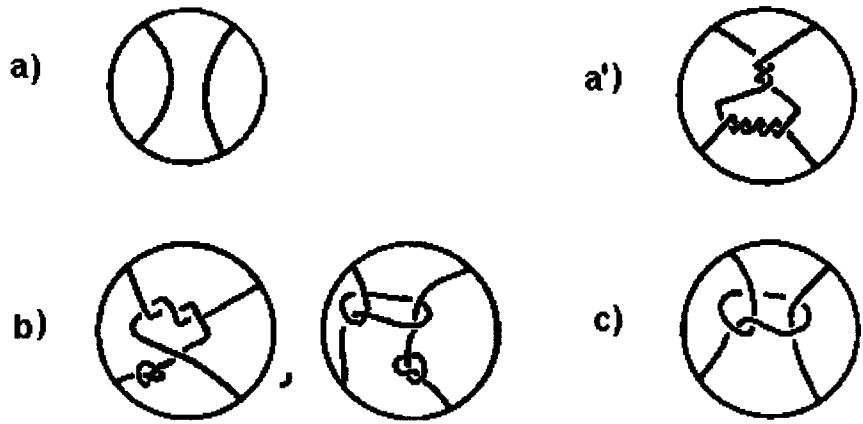


Fig. 5. Different types of tangles.

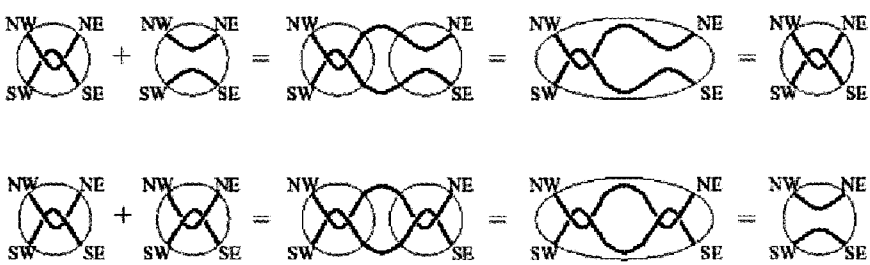


Fig. 6. Tangle addition.

Given two tangles A and B, the **tangle addition** $A + B$ is defined in the figure above (cf. [3, 9]), as in Fig. 6. The resulting object $A + B$ is obtained by gluing NE of A to NW of B, and SE of A to SW of B. One may note that the sum of two tangles is not always a tangle since the strands of $(A + B)$ can include a simple closed curve.

The figure given below (Fig. 7) is used to define two other tangle operations called **numerator** and **denominator**. Given a tangle A, $N(A)$ and $D(A)$ denote these operations, respectively, and they produce knots and 2-component links. $N(A + B)$ and $D(A + B)$ can be defined in a similar way. Note that if $A + B$ is not a 2-string tangle, the result of $N(A + B)$ or $D(A + B)$ can be a link of more than two components.

2.4. *Rational tangle*

Rational tangle is a tangle whose strands can be deformed to a trivial tangle by moving the ends of strands on the boundary [14]. Rational tangles admit

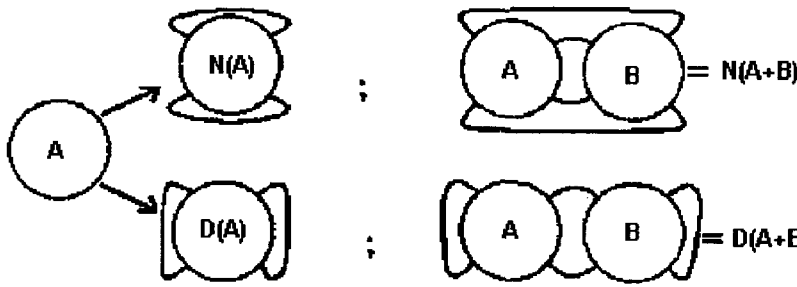


Fig. 7. Tangle operations.

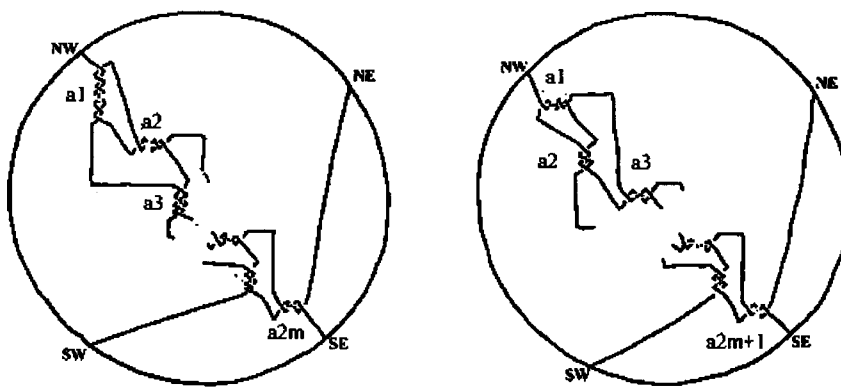


Fig. 8. Tangle surgery.

of a classification in which a unique standard vector with integer entries is associated with each equivalence class of rational tangles. Such a vector $(a_1, \ldots, a_m)$ must satisfy the following conditions:

(1) $a_i \neq 0$ when $0 < i < m$
(2) all entries are of the same sign
(3) a_1 is not equal to 1 or -1.

The tangle can be constructed from its associated vector as shown in the figure above (Fig. 8).

Four exceptional tangles are excluded by the convention; they can be visualized in Fig. 9 together with their standard vector.

For each equivalence class of rational tangles, denoted by A, the standard vector associated to it is called the Conway Symbol for A (cf. [2]). To

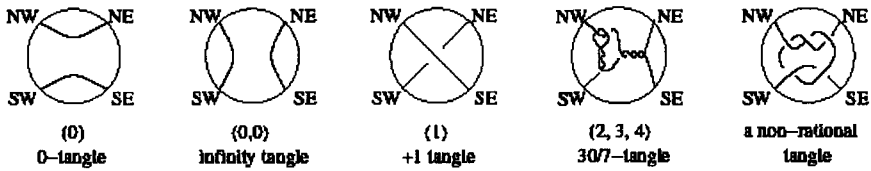


Fig. 9. Canonical form of rational tangles.

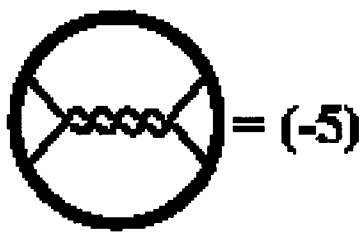


Fig. 10. Integral tangle.

each Conway symbol can be associated a unique extended rational number; $\frac{\beta}{\alpha} \in Q \cup \{\infty\}$

$$\beta/\alpha = a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{\cdots + \cfrac{1}{a_m}}}}.$$

A tangle is integral (shown in Fig. 10), if its canonical vector is of the form (z) for some integer z. It may be noted that integral tangles are in one-to-one correspondence with the integers, and that they are drawn as a row of horizontal twists (positive or negative).

2.5. *4-plats*

A 4-plat is a knot or a link that admits a representation consisting of a braid on 4 strings closed up as in Fig. 11. The classification of 4-plats shows that each 4-plat K is characterized by a vector $\langle c_1, c_2, \ldots, c_{2n+1} \rangle$ such that c_1

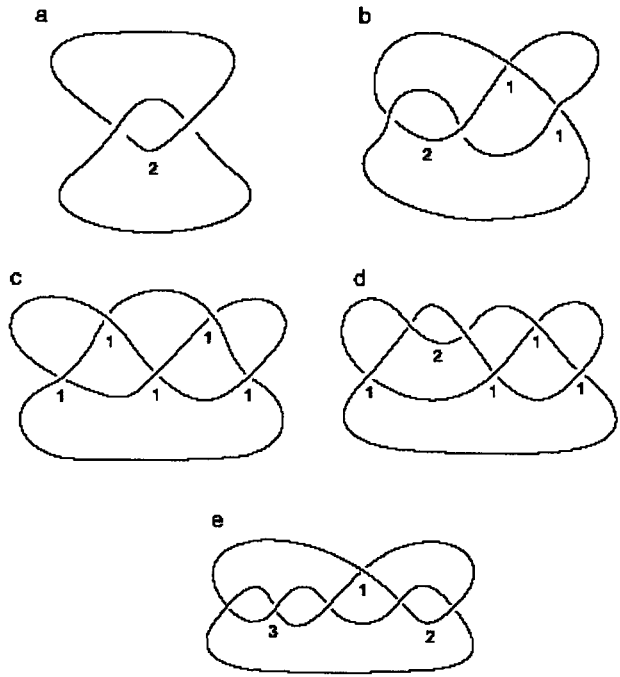


Fig. 11. Standard 4-plats.

and c_{2n+1} are different from zero. A rational number is assigned, by means of the following continued fraction calculation, to each vector:

$$\beta/\alpha = \cfrac{1}{c_1 + \cfrac{1}{c_2 + \cfrac{1}{\cdots}}}.$$

The link K is denoted by $b(\alpha, \beta)$ and is termed as the Conway notification for K.

2.6. *Classification of 4-plats*

$b(\alpha, \beta)$ and $b(\alpha', \beta')$ are equivalent and non-oriented links if and only if

$$\alpha = \alpha'; \quad \beta' \equiv \beta (\text{mod } \alpha).$$

See Figs. 12 and 13.

The numerator closure of the sum of two rational tangles is a rational knot or link (Fig. 14).

(1)

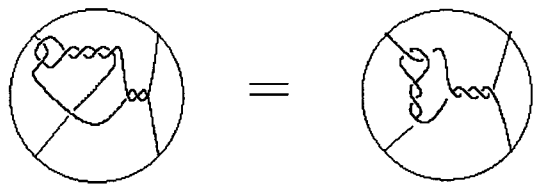


Fig. 12. Equivalent tangles.

(2)

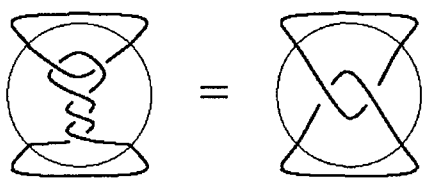
$$N((2,3,0)) = N\left(0 + \frac{1}{3 + \frac{1}{2}}\right) = N\left(\frac{2}{7}\right)$$

and $N((2)) = N\left(\frac{2}{1}\right) = N\left(\frac{2}{1}\right) = N((2))$

Thus $N((2,3,0)) = N\left(\frac{2}{7}\right) = N\left(\frac{2}{1}\right) = N((2))$,

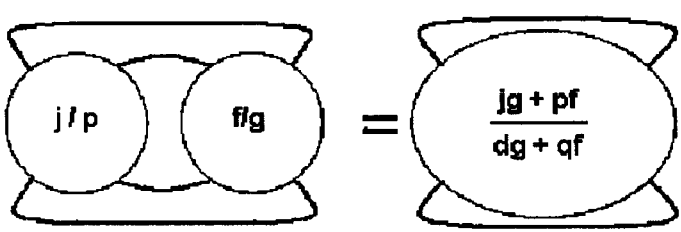
since $7 = 1 + 2(3)$.

(a)



(b)

Fig. 13. (a), (b) Rational knot/link equivalence.



where dp - qj = 1

Fig. 14. Numerator closure for sum of two tangles.

3. Biological Statement and Assumptions

The main goal when doing tangle analysis of experimental data arising from site-specific recombination reactions is to understand the enzymatic mechanism. The tangle model studies topological changes in DNA caused by the enzymes. The mechanism of recombinases involves local interaction of two DNA strands (Fig. 15).

One of the goals of the tangle model is to compute the topology of the synaptosome (enzyme + bound DNA), before and after the enzymatic action. In an attempt to translate an enzymatic action on DNA into the language of mathematical 2-string tangles, DNA molecule with its two recombination sites as an embedding of one or more circles in 3-space has been considered. Therefore the substrate, DNA is considered to be a knot or a link. Each circular DNA molecule is represented by the axis of its double-helix (a simple closed curve in R^3). A single event of recombination consists of two movements. One of them is a global movement where, by ambient isotopy of R^3, the recombination sites are juxtaposed inside a ball. The ball represents the enzyme, together with any accessory proteins that bind the DNA substrate and are required for recombination. The ball with the two strands of bound DNA represents, by definition, the local synaptic complex (or synaptosome). The second movement is a local movement in the interior of the ball where two strands are cut at the recombination sites, and then recombined. At this stage, the part of the knot or link that was left in the exterior of the ball remains fixed. Mathematically, the ball divides the space into two regions. Each region will be defined on the basis of its biological role.

A ball with two embedded strands is, by definition, a 2-string tangle. Therefore the enzyme with the accessory proteins and the bound DNA form a 2-string tangle, where the proteins form the ball that defines the tangle. Call this tangle E. Likewise, the recombination sites can be surrounded by a small ball in the interior of E. Let P be this tangle in the interior of

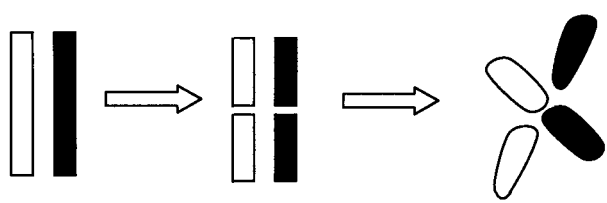


Fig. 15. Local interaction of two DNA strands.

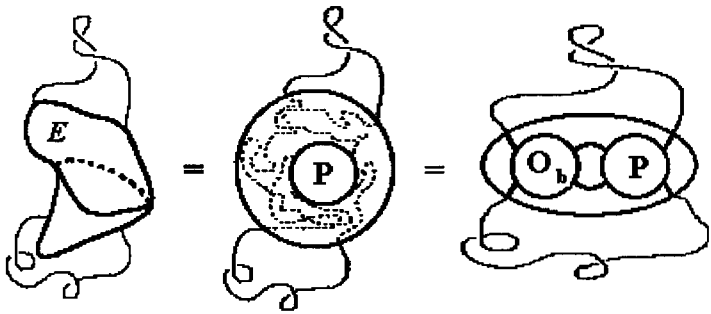


Fig. 16. Sketch of a recombination site.

E where the DNA is cut by the enzyme. This description is illustrated in Fig. 16.

4. Tangle Model Assumptions

In a site-specific recombination reaction, the recombinase and accessory proteins bind to the DNA. Enzyme and proteins are modelled as a ball; the circular DNA is modelled as a knot or link that intersects the ball in two strands. The synaptosome is a 2-string tangle called E. We can look upon E as the sum of two tangles, $O_b + P$.

Assumption 1. $E = O_b + P$, *where* O_b *contains the entire DNA that is bound to the enzyme or to the accessory proteins, except for the recombination sites that are contained in* P. *That is, the enzyme mechanism in a single recombination event is constant, independent of the geometry (supercoiling) and topology (knotting and catenation) of the substrate population. Moreover, recombination takes place entirely within the domain of the enzyme ball, and the substrate configuration outside the enzyme ball remains fixed while the strands are being broken and recombined inside and on the boundary of the enzyme.*

We assume that any two pre-recombination copies of the synaptosome are identical, meaning that we can by rotation and translation superimpose one copy on the other, with the congruence so achieved respecting the structure of both the protein and the DNA. We likewise assume that all of the copies of post-recombination synaptosome are identical.

Let O_f be the tangle formed by the ball $S^3 - E$ that contains the DNA not bound to the enzyme/accessory proteins complex. It may be noted that both topology and sequence of O_b and O_f remain unchanged upon

recombination. $\boldsymbol{O_f}$ contains all the relevant topological information from the free DNA. Assumption 1 allows to see the whole synaptic complex simply as:

$$\boldsymbol{N(O_f + O_b + P) = N(O_f + E)}.$$

One may note that both $\boldsymbol{O_b}$ and $\boldsymbol{O_f}$ remain unchanged upon recombination. Let $\boldsymbol{O}$ be the "outside tangle" defined by the following sum:

$$\boldsymbol{O = O_f + O_b}.$$

In the calculations one will usually refer to the tangle $\boldsymbol{O}$ instead of $\boldsymbol{O_b}$ and $\boldsymbol{O_f}$. If the substrate is a knot or link K_1, then the synaptic complex can be represented by a **substrate equation** of the form:

$$\boldsymbol{N(O + P) = K_1}.$$

Recombination occurs during the local movement, and strand exchange is restricted to the tangle P. This motivates the second mathematical assumption.

Assumption 2. *The recombinase action corresponds to a tangle surgery where the tangle P is changed by the tangle R.*

With this assumption, after one round of recombination leading to a knotted or linked product of type K_2, the parental tangle P is removed from the synaptosome and replaced by the recombinant tangle R. The outside tangle O remains unchanged. The post-recombination synaptic complex is represented by the **product equation**:

$$N(O + R) = K_2.$$

Therefore, one round of recombination action is translated to the following system with two tangle equations:

$$\left.\begin{aligned} N(O_f + O_b + P) = N(O + P) = K_1 \\ N(O_f + O_b + P) = N(O + P) = K_2 \end{aligned}\right\} \tag{1}$$

where $\{\boldsymbol{O_f}, \boldsymbol{O_b}, \boldsymbol{P}, \boldsymbol{R}\}$ are unknown. In general, two tangle equations on 4 unknowns are not enough to find a unique solution array $(\boldsymbol{O_f}, \boldsymbol{O_b}, \boldsymbol{P}, \boldsymbol{R})$, or even a finite number of solutions. Electron micrographs of the synaptic complex can sometimes characterize $\boldsymbol{O_f}$.

For unknotted substrates it can generally be deduced that $\boldsymbol{O_f}$ is rational; in particular, $\boldsymbol{O_f} = (\mathbf{0})$ and therefore, $\boldsymbol{O = O_f + O_b = O_b}$.

4.1. *Other substrates*

When the tangle model was used to study a resolvase system, the following assumption was crucial to unveil the enzymatic mechanism:

Assumption 3. *The recombination mechanism is constant, independent of the geometry (supercoiling) and topology (knotting and linking) of the substrate population.*

This means, in part, that the recombination is restricted to the interior of the ball, and that the substrate's configuration outside the ball remains fixed during this event. It also implies that both P and R are constant, they do not depend on the nature of neither substrate nor product of recombination, and they are characteristic of the enzyme. Any change in the substrate would be translated into a change in the tangle O (in particular a change in $\boldsymbol{O_f}$). It follows from Assumption 3 that the tangles $\{\boldsymbol{O_b}, \boldsymbol{P}, \boldsymbol{R}\}$ are constants reflecting enzyme binding and mechanism, while the tangle $\boldsymbol{O_f}$ reflects the variable geometry and topology of the substrates. In the case of enzymes with topological selectivity and specificity (e.g. Gin, Tn3 and Xer), given a fixed substrate $\boldsymbol{K_1}$ the tangles $\boldsymbol{O}$, $\boldsymbol{P}$ and R are constants uniquely determined by the enzyme. Furthermore, if one considers two experiments where a given enzyme acts on topologically different substrates, then two systems of equations appear in the tangle analysis. Assumption 3 allows taking P and R constant in both the systems. The tangle O will be denoted as $\boldsymbol{O^{(1)}}$ for experiment 1 and as $\boldsymbol{O^{(2)}}$ for experiment 2. In the cases of Gin and Xer the assumption of constant mechanism is supported by experimental data (Gin, Xer). On the other hand, there are some enzymes such as λ-int, mutant Gin and FLP that have no topological selectivity. In those cases, for a single substrate, the tangle O can vary. Assumption 3 in these cases only implies that P and R are constants. Thus the mechanism is not constant and it is not clear whether the enzymatic binding (characterized by the tangle O) changes from one substrate type to another.

5. Site-Specific Recombination

Site-specific recombination is one of the ways in which nature alters the genetic code of an organism, either by moving a block of DNA to another position on the molecule or by integrating a block of alien DNA into a host genome (cf. [2, 15]). One of the biological purposes of recombination is the regulation of gene expression in the cell, because it can alter the

relative position of the gene and its repressor and promoter sites on the genome. Site-specific recombination also plays a vital role in the life cycle of certain viruses, which utilize this process to insert viral DNA into the DNA of a host organism. An enzyme that mediates site-specific recombination on DNA is called a recombinase. A recombination site is a short segment of duplex DNA whose sequence is recognized by the recombinase. Site-specific recombination can occur when a pair of sites (on the same or on different DNA molecules) becomes juxtaposed in the presence of the recombinase. The pair of sites is aligned through enzyme manipulation or random thermal motion (or both), and both sites (and perhaps some contiguous DNA) are then bound by the enzyme. This stage of the reaction is called synapsis. We shall call this intermediate protein-DNA complex formed by the part of the substrate that is bound to the enzyme together with the enzyme itself the synaptosome and the entire DNA molecule(s) involved in synapsis (including the parts of the DNA molecule(s) not bound to the enzyme), together with the enzyme itself, the synaptic complex. It is our intent to deduce mathematically the path of the DNA in the black mass of the synaptosome, both before and after recombination.

After forming the synaptosome, a single recombination event occurs: the enzyme then performs two double-stranded breaks at the sites and recombines the ends by exchanging them in an enzyme-specific manner. The synaptosome then dissociates, and the enzyme releases the DNA. We call the pre-recombination unbound DNA molecule(s) the substrate and the post-recombination unbound DNA molecule(s) the product. During a single binding encounter between enzyme and DNA, the enzyme may mediate more than one recombination event; this is called processive recombination. On the other hand, the enzyme may perform recombination in multiple binding encounters with the DNA, which is called distributive recombination. Some site-specific recombination enzymes mediate both distributive and processive recombination.

Site-specific recombination involves topological changes in the substrate. In order to identify these topological changes, one chooses to perform experiments on circular DNA substrate. One must perform an experiment on a large number of circular molecules in order to obtain an observable amount of product. Using cloning techniques, one can synthesize circular duplex DNA molecules, which contain two copies of a recombination site. At each recombination site, the base pair sequence is in general not palindromic and hence it induces a local orientation on the substrate DNA circle. If these induced orientations from a pair of sites on a singular circular molecule

agree, this site configuration is called direct repeats (or head-to-tail), and if the induced orientations disagree, this site configuration is called inverted repeats (or head-to head). If the substrate is a single DNA circle with a single pair of directly repeated sites, the recombination product is a pair of DNA circles and can form a DNA link (or catanane). If the substrate is a pair of DNA circles with one site each, the product is a single DNA circle and can form a DNA knot (usually with direct repeats). In processive recombination on circular substrate with direct repeats, the products of an odd number of rounds of processive recombination are DNA links, and the products of an even number of rounds of processive recombination are DNA knots. If the substrate is a single DNA circle with inverted repeats, the product is a single DNA circle and can form a DNA knot. In all the figures where DNA is represented by a line drawing, duplex DNA is represented by a single line, and supercoiling is omitted.

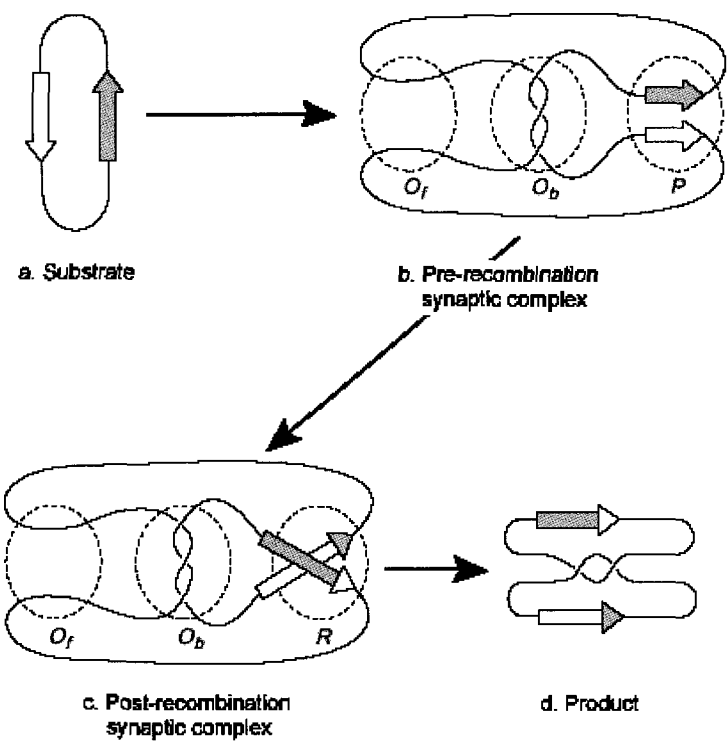


Fig. 17. A single recombination event: direct repeats.

The geometry and topology of circular DNA substrate are experimental control variables. The geometry and topology of the recombination reaction products are observable. *In vitro* experiments usually proceed as follows:

Circular substrate is prepared, with all of the substrate molecules representing the same knot type. The amount of supercoiling of the substrate molecules is also a control variable. The substrate molecules are reacted with a high concentration of purified enzyme, and the reaction products are fractionated by gel electrophoresis. Gel electrophoresis discriminates among DNA molecules on the basis of molecular weight; given that all molecules have the same molecular weight (as is the case in these topological enzymology experiments), electrophoresis discriminates on the basis of subtle differences in the geometry (supercoiling) and topology of the DNA molecules. Under the proper conditions gel velocity is (surprisingly) determined by the crossing number of the knot or the link, knots and links of the same crossing number migrate with the same gel velocities. After running the gel, the DNA molecules are removed from the gel and coated with Rec A protein. It is this new observation technique (Rec A-enhanced electron microscopy) that makes possible the detailed knot-theoretic analysis of reaction products. Rec A is an E. coli protein that binds to DNA and mediates general recombination in E. coli. The process of Rec A coating fattens, stiffens, and stretches (untwists) the DNA. This facilitates the unambiguous determination of crossings (nodes) in an electron micrograph of DNA.

6. Processive Recombination

Processive recombination must be incorporated to the tangle model without contradicting the assumption of constant mechanism [2, 15]. Since P is assumed to be changed by R upon one round of recombination, R will be assumed to go to $\boldsymbol{R} + \boldsymbol{R}$ after two rounds and so on and so forth. In this way processive recombination is modelled by tangle addition. Experimental data obtained from processive recombination adds equations to the system. These equations involve the same unknowns as before.

Assumption 4. *Processive recombination acts by tangle addition* (Fig. 18). The implication is that, after n rounds of processive recombination, the post-recombination synaptosome is $(\boldsymbol{O_b} + \boldsymbol{nR})$. This leads to a new equation for each round of recombination:

$$\boldsymbol{N(O + nR) = n\text{th round product}}$$

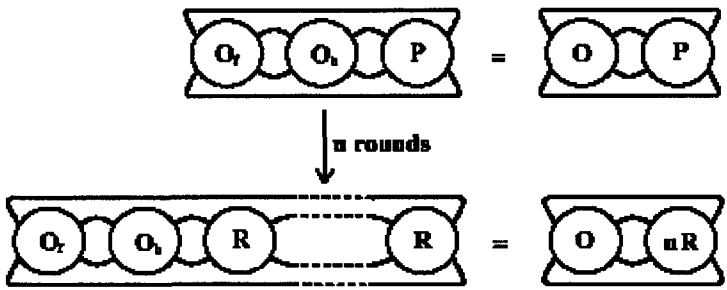


Fig. 18. Processive recombination.

In the tangle analysis of Tn3 resolution and in that of Gin inversion, data arising out of the first few (three or four) rounds of recombination are enough to find unique solutions to the tangle equations. In addition, these computations correctly predict the products of additional rounds of processive recombination.

7. Useful Facts and Theorems About Tangles

(1) Both $N(a/b)$ and $D(a/b)$ are 4-plats. The knot/link $N(a/b)$ is the 4-plat $S(a,-b)$. The knot/link $D(a/b)$ is the 4-plat $S(b,a)$.
(2) The tangle corresponding to $a_1\backslash b_1$ is the same as the tangle corresponding to $a_2\backslash b_2$ if and only if $a_1\backslash b_1 = a_2\backslash b_2$.
(3) $a_1\backslash b_1 + a_2\backslash b_2 \cong$ a rational tangle unless either $b_1 = \pm 1$ or $b_2 = \pm 1$.
(4) $a/b + t = (a + bt)/t$.
(5) $N(A + C) = N(C + A)$ where A and C are arbitrary tangles.
(6) $N(A+C) =$ 4-plat implies at least one of A and C is rational or locally knotted.
(7) $D(A + C) = D(A) \cong D(C)$.
(8) $N(AN(c_1, \ldots, c_n) + B) = N(A + BN(c_n, \ldots, c_1))$ where n is odd.

Theorem 1. *Let U and R be tangles such that $N(U + iR) =$ 4-plat for some $i \geq 2$, and $N(U + jR) \neq N(U + iR)$ for some j. Then R is a rational tangle. If $i \geq 3$, then R is an integral triangle.*

Proof. If R were locally knotted, then $N(U + iR)$, $i \geq 2$ would be composite. Since 4-plats are prime, R cannot be locally knotted. Suppose R is a prime tangle. By tangle properties $U + (i-1)R$ is rational or locally knotted and R prime implies that $(i-1)R$ prime and U must be ∞-tangle or locally knotted.

Now U cannot be a ∞-tangle. If U were the infinity tangle, then $N(U+iR) = D(iR) = D(R)\#\cdots\#D(R)$. Since 4-plats are prime, $D(R) =$ unknot. But $N(U + iR) = D(iR) =$ unknot $= D(jR) = N(U + jR)$, a contradiction. Thus if U is locally unknotted, R must be rational.

If $i \geq 3$, then R does not have parity ∞ since 4-plats have at most two components. If $i \geq 3$, U is locally unknotted, and R is not integral, then if U is not integral, $U+R$ and $(i-1)R$ are prime. But $N(U+iR) =$ 4-plat would then contradict the tangle property. If U is integral and if R is rational, then $N(U + iR) =$ 4-plat, $i \geq 3$, if and only if R is integral. Thus if U is locally unknotted and $i \geq 3$, then R must be integral.

Suppose U is locally knotted. Then if U' is the tangle formed from U by removing the local knot, then $N(U' + iR) =$ unknot, since 4-plats are prime. $N(U + jR) \neq N(U + iR)$ implies that $N(U' + jR) \neq N(U' + iR)$. Since the unknot is a 4-plat and U' is locally knotted, R is rational if $i \geq 2$ and integral if $i \geq 3$. □

Theorem 2. *If $N(U + P) =$ 4-plat and $N(U + R) =$ 4-plat where $P = a_1/b_1$, $R = a_2/b_2$, $a_1b_2 - a_2b_1 \neq \pm 1$, then U is either a rational tangle or ambient isotopic to a sum of two rational tangles.*

Proof. If $N(U + P) =$ 4-plat and $N(U + R) =$ 4-plat where $P = a_1/b_1$, $R = a_2/b_2$, $a_1b_2 - a_2b_1 \neq \pm 1$, then the cyclic surgery theorem implies that the double branched cover of the tangle U is a Seirfert fibered space. This means, U is ambient isotopic to a Montesinos tangle (Ernst [8]). □

Theorem 3. *Let U and R be tangles such that $N(U + iR) = K_i$ for $0 \leq i \leq 3$, where K_i's are 4-plats, and $\{K_1, K_2, K_3\}$ represent at least 2 different link or knot types. Then there is at most one solution for U and U is either rational or the sum of two rational tangles.*

Theorem 4. *Lat $E = t/w -$ tangle, $(w, t) = 1$ and $ay - bx = 1$. Then the following are equivalent for $|t| \geq 2$. For $t = \pm 1$, (2) and (3) are equivalent and imply (1):*

(1) $d_R\ (N(a/b), N(z/v)) \leq 1$.

(2) *If $w = \pm 1 \bmod t$, $N(z/v) = N((tb + w)a)/(-ty + w)x))$ or $N((-tx + (tk + w)a)/(-ty + (tk + w)b))$. Else $w \cong \pm 1 \bmod t$ and $N(z/v) = N((tp^2b + sa)/(-tp^2y - sx))$ or $N((-tp^2x + sa)/(-tp^2y + sb))$ where $s = tp(-q + pk) \pm 1$, $(p, q) = 1$, $p > 0$.*

(3) $N(a/b) = N(U+0)$ *and* $N(z/v) = N(U+t/w)$ *have the following solutions when* $|t| \geq 2$:

If $w \neq \pm 1 \bmod t$, *then* U *must be rational and* $U = a/(b+ka)$ *or* $a/(-x+ka)$.

If $w = \pm 1 \bmod t$, *then* U *must be ambient isotopic to a sum of at most two rational tangles and* $U = (U_1 + U_2)$ *o* $(h,0)$ *where* $U_1 = (-bja(d-kj))/(pb+a(pk-q))$ *or* $(xj+a(d-kj))/(-px+a(pk-q))$ *and* $U_2 = j/p$, $pd-qj = 1$ *and* $h = (-w \pm 1)/t$ *if* $(-w \pm 1) \in Z$. *If* $t = \pm 1$, *then the above list contains all solutions when* U *is ambient isotopic to a sum of rational tangles.*

Proof. $d_R(N(a/b), N(z/v)) \leq 1$ if and only if there exists a U such that $N(a/b) = N(U+0)$ and $N(z/v) = N(U+t/w)$. By Theorem 2, U is either a rational tangle or ambient isotopic to the sum of two rational tangles. If U is a rational tangle, $N(U+0) = N(a/b)$ implies by tangle fact that $U = a/(b+ka)$ or $a/(-x+ka)$ and $N(z/v) = N((tb+w)a)/(-ty+w)x))$ or $N((-tx+(tk+w)a)/(-ty+(tk+w)b))$. If U is ambient isotopic to the sum of two rational tangles, $U_1 + U_2$, then since $N(U+0)$ is a 4-plat, $U = (U_1+U_2)\,\mathrm{o}\,(h,0)$. Solving $N((U_1+j/p)\,\mathrm{o}\,(h,0)+0) = N(a/b)$ implies $U_1 = (-bja(d-kj))/(pb+a(pk-q))$ or $(xj+a(d-kj))/(-px+a(pk-q))$ and $U_2 = j/p$, $pd-qj = 1$. If $N((U_1+j/p)\,\mathrm{o}\,(h,0)+t/w) = N((U_1+j/p+t/(ht+w))$. If U_1 or U_2 are non-integral, $N((U_1+U_2+t/(ht+w))$ is a 4-plat if and only if $ht+w = \pm 1$, i.e. $w = \pm 1 \bmod t$ in which case $h = (-w \pm 1)/t$ if $(-w \pm 1) \in Z$. Again by tangle properties if $s = tp(-q+pk) \pm 1$, $N(U_1+U_2 \pm t) = N((tp^2b+sa)/(-tp^2y-sx)) = N((-tp^2x+sa)/(-tp^2y+sb))$. □

Theorem 5. *If* $N(U+f_1/g_1) = unknot$ *and* $N(U+f_2/g_2) = N(2z/1)$ *where* $f_1g_2 - f_2g_1 = \pm 1$, *then* U *is rational.*

Lemma 1. *If* $N(U_i+P) = unknot$, $i = 1,2$, *and* $U_1 \neq U_2$, *then* P *is rational.*

Theorem 6. *If* $N(U+0/1) = N(1/0)$ *and* $N(U+1/w) = N(2k/1)$, *then* U *is rational.*

Corollary 1. *Suppose* $bx-ay = 1$, $N(U+0/1) = N(a/b)$ *and* $N(U+t/w) = N(z/v)$ *where* $N(a/b)$ *and* $N(z/v)$ *are unoriented* 4*-plats. If* $w \neq \pm 1$ *or if* U *is rational, then* $t/w = (xz-av')/(bv'-yz-kt)$ *and* $U = a/(b+ka)$

or $t/w = (bz - av')/(xv' - yz - kt)$ *and* $U = a/(x + ka)$ *where* v' *is any integer such that* $v'v^{\pm 1} = 1 \bmod z$. *If* $w = \pm 1 \bmod t$, *then* t *divides* $z \curlyvee a$.

8. Model for the Tn3 Resolvase

Tn3 resolvase is a site-specific recombinase that reacts with certain circular duplex DNA substrate with directly repeated recombination sites. One begins with supercoiled unknotted DNA substrate and treats it with resolvase. The principal product of this reaction is known to be the DNA 4-plat. Resolvase is known to act dispersively in this situation to bind to the circular DNA, to mediate a single recombination event, and then to release the linked product. It is also known that resolvase and free (unbound) DNA links do not react. However, once in twenty encounters, resolvase acts processively — additional recombinant strand exchanges are promoted prior to the release of the product, with yield decreasing exponentially with increasing number of strand exchanges at a single binding encounter with the enzyme. Two successive rounds of processive recombination produce the DNA 4-plat $\langle 2; 1; 1\rangle$; three successive rounds of processive recombination produce the DNA 4-plat $\langle 1; 1; 1; 1; 1\rangle$, whose electron micrograph appears in Fig. 2(a); four successive rounds of recombination produce the DNA 4-plat $\langle 1; 2; 1; 1; 1\rangle$ whose electron micrograph appears in Fig. 2(b). The discovery of the DNA knot $\langle 1; 2; 1; 1; 1\rangle$ substantiated a model for Tn3 resolvase mechanism.

For resolvase, the electron micrograph of the synaptic complex reveals that $O_f = (0)$, since the DNA loops on the exterior of the synaptosome can be untwisted and are not entangled. This observation from the micrograph reduces the number of variables in the tangle model by one, leaving us with three variables $\{O_b; P; R\}$. One can prove that there are four possible tangle pairs $\{O_b; R\}$, which can produce the experimental results of the first two rounds of processive Tn3 recombination (cf. [1, 9, 11]). The third round of processive recombination is then used to discard three of these four pairs of extraneous solutions. The following theorems can be viewed as a mathematical proof of resolvase synaptic complex structure. The model proposed in [2] is the unique explanation for the first three observed products of processive Tn3 recombination, assuming that processive recombination acts by adding on copies of the recombinant tangle R. Mathematical analysis makes feasible the reconstruction of DNA topology from gel electrophoresis, avoiding the technically difficult electron microscopy of Rec A-enhanced DNA.

Theorem 1. *Suppose that tangles O_b; P; and R satisfy the following equations*:

(1) $N(O_b + P) = \langle 1 \rangle$ (*the unknot*),
(2) $N(O_b + R) = \langle 2 \rangle$ (*the Hopf link*),
(3) $N(O_b + R + R) = \langle 2; 1; 1 \rangle$ (*the figure* 8 *knot*).

Then $\{O_b; R\} = \{(-3; 0); (1)\}$, $\{(3; 0); (-1)\}$, $\{(-2; -3; -1); (1)\}$, *or* $\{(2; 3; 1); (-1)\}$:

In order to decide the biologically correct solution, we have to utilize more experimental evidence. The third round of processive resolvase recombination determines which of these four solutions is the correct one.

Theorem 2. *Suppose that tangles O_b; P; and R satisfy the following equations*:

(1) $N(O_b + P) = \langle 1 \rangle$ (*the unknot*),
(2) $N(O_b + R) = \langle 2 \rangle$ (*the Hopf link*),
(3) $N(O_b + R + R) = \langle 2; 1; 1 \rangle$ (*the figure* 8 *knot*),
(4) $N(O_b + R + R + R) = \langle 1; 1; 1; 1; 1 \rangle$ (*the* (+) *Whitehead link*).

Then $O_b = (-3; 0)$; $R = (1)$, *and* $N(O_b + R + R + R + R) = \langle 1; 2; 1; 1; 1 \rangle$.

The correct global topology of the first round of processive Tn3 recombination on the unknot is shown in Fig. 17. Moreover, the first three rounds of processive Tn3 recombination uniquely determine $N(O_b+R+R+R+R)$, the result of four rounds of recombination. It is the 4-plat knot $\langle 1; 2; 1; 1; 1 \rangle$, and this DNA knot has been observed (cf. Fig. 2(b)). We note that there is no information in either Theorem 1 or Theorem 2 about the parental tangle P. Since P appears in only one tangle equation (Eq. (i)), for each fixed rational tangle solutions for O_b there are infinitely many rational tangle solutions to the equation for P. Most biologists believe that $P = (0)$, and a biomathematical argument exists for this claim.

9. Model for the Xer Recombinase and Topoisomerases III and IV

Xer recombinase acting on an unknotted substrate produces only one product, the link $N(4/1)$. Thus there are only two equations involving 3 unknowns and hence they have an infinite number of solutions.

$$N(U_1 + P) = unknot \qquad N(U_1 + R) = N(4/1)$$

This enzyme cannot act processively. So there is no experiment that can be performed in order to reduce the infinite number of solutions to a finite number. However, we can make a list of all possible solutions and propose experiments to reduce this list to a smaller number of solutions. These can then be analyzed to decide as to which solutions are the most biologically relevant, using additional biological assumptions. For example, if P and R are biologically restricted to have at most 4 crossings, then the solutions become finite.

9.1. *Biological model for recombinases and topoisomerases*

- Initial Configuration.
- The accessory proteins fix three negative crossings in the domain. Xer binds to the two recombination sites.
- Idea: the proteins and the three negative crossings remain fixed.
- One round of recombination produces one negative crossing in the domain.
- After recombination the enzyme releases the molecule.

9.2. *Biological model (unknotted substrates)*

- Substrate = unknotted circular DNA with sites in direct repeat.
- $K_1 = b(1,1) = \langle 1 \rangle$ [where K's are 4-plats]
- Product = 4-crossings right-handed torus link with antiparallel sites.
- $K_1 = b(4,3) = \langle 1,2,1 \rangle$

9.3. *Biological model (catenated substrates)*

- Substrate = 6-crossings right-handed torus link with anti-parallel sites.
- $K_1 = b(6,5) = \langle 1,4,1 \rangle$
- Product = 7-crossings knot or link.

9.4. *Tangle equations for unknotted substrates*

(i) $N(O+P) = \langle 1 \rangle = b(1,1)$
(ii) $N(O+R) = \langle 1,2,1 \rangle = b(4,3)$
together with the assumptions:

(a) $P = (0)$;
(b) $R = (k)$, k non-zero integer;
(c) O is rational or sum of 2 rational tangles.

We solve for O and R.

9.5. *Tangle equations for catenated substrates*

(i) $N(O+P) = \langle 1,4,1\rangle = b(6,5)$
(ii) $N(O+R) = K_2 = 7$-crossings knot or link.

in which

(1) $P = (0)$;
(2) $R = (k)$, k being a non-zero integer;
(3) O is rational or sum of 2 rational tangles;
(4) K_2 is a 4-plat.

We solve for O and R.

9.6. *Problems*

- Xer recombination is not processive. The action on substrates with a single topology provides only two tangle equations.
- For known P rational and K 4-plat: $N(O+P) = K$ has infinitely many solutions for O.
- For known P rational, K_1 and K_2 4-plats: $N(O+P) = K_1$, $N(O+R) = K_2$ do not lead a unique solution.

In order to solve the tangle equations, we intend to make use of the minimum possible assumptions, with an aim to put forward the results in a realistic manner.

9.7. *Results*

- **UNKNOTTED SUBSTRATES: 1. When O is rational:**
- **The solutions to the tangle equations are:**
- $O = (-3,0)$ and $R = (-1)$
- $O = (-5,0)$ and $R = (+1)$
- $O = (1)$ and $R = (3)$
- $O = (-1)$ and $R = (5)$

The last two cases produce 4-crossing links with the wrong site alignment. These cases must be discarded.

- **When O is the sum of two rational non-integral tangles, there exist no solutions.**

Conclusions

- $N(O+P) = b(1,1)$
- $N(O+R) = b(4,3)$ with sites in anti-parallel $P = (0)$, $R = (k)$,
- O is rational or the sum of two rational tangles.
- The only solutions to the system are:

  $O = (-3,0)$, $R = (-1)$;
  $O = (-5,0)$, $R = (+1)$

Results

- **CATENATED SUBSTRATES:**
- **When O is rational, the solutions to the tangle equations are:**

(a) $O = (6)$ and $R = (+1)$, $K_2 = b(7,6)$
(b) $O = (6)$ and $R = (-13)$, $K_2 = b(7,1)$
(c) $O = (6,2,0)$ and $R = (-1)$, $K_2 = b(7,6)$
(d) $O = (-5,-1)$ and $R = (4)$, $K_2 = b(14,9)$
(e) $O = (-5,-1)$ and $R = (-1)$, $K_2 = b(11,9)$
(f) $O = (-5,-1,-2,0)$ and $R = (+1)$, $K_2 = b(11,9)$

The solutions (a)–(c) have to be discarded, since they correspond to torus knots, while the solution (d) is to be discarded because it corresponds to a link of parental genotype. The solutions (e) and (f) are the only acceptable ones since they correspond to twist knots of recombinant genotype.

When $O = X + A$ with X and A rational non-integral, the solutions to the tangle equations are:

(1) $X = (-4,0)$, $A = (-2,0)$ and $R = (+3)$, $K = b(18,13)$
(2) $X = (-4,0)$, $A = (-2,0)$ and $R = (-1)$, $K = b(14,9)$
(3) $X = (-3,0)$, $A = (-3,0)$, and $R = (-1)$, $K = b(15,11)$
(4) $X = (-3,0)$, $A = (-3,0)$, and $R = (+3)$, $K = b(21,13)$

The solutions (1) and (2) have to be discarded since they correspond to links. The solutions (3) and (4) are the only acceptable ones since they correspond to knots of recombinant genotype.

10. Modelling Conclusions

The tangle is modelled here, assuming that for a given enzyme, the tangles P and R are constant, independent of the topology of the substrate. We

made the tangle analysis of two recombination events mediated by recombinases. We showed that if $P = (0)$, R is integral and O is rational or the sum of two rational tangles, and K_2 is a 4-plat, then there are only three solutions that explain the observed products in both the experiments.

Recapitulating, the tangle model looks upon the circular DNA substrate and products as knots or links. The site-specific recombinase and its accessory proteins are seen as a ball that intersects the DNA knot or link in two strands. The interior of the ball is divided into two regions. One of them is restricted to strand exchange and corresponds to a parental tangle P. This tangle can be chosen to be $P = (0)$. P represents the only region in the synaptic complex that changes upon recombination. The region outside P but inside the ball, called $\boldsymbol{O_b}$, traps all the conformation that, together with the change from P to R, determines the topology of the recombination products. Finally, the region outside the ball, $\boldsymbol{O_f}$ detects the variation between substrates with different topology. The tangle model assumes that the synaptic complex can be expressed as $\boldsymbol{N(O+P) = K_0}$ where $\boldsymbol{O = O_f + O_b}$ is called the outside tangle. Recombination is modelled by a tangle surgery that replaces P by the recombinant tangle R, thus leading to a product equation $\boldsymbol{N(O+R) = K_1}$. The assumption of constant mechanism implies that P and R are constants uniquely determined by the enzyme. In the cases when there are both topological selectivity and specificity (e.g. Tn3 resolvase, Gin, Xer), the tangle $\boldsymbol{O}$ is also determined uniquely by both the enzyme and the topology of the substrate. If there is no topological selectivity (e.g. λ -Int, mutant Gin and FLP) then, for a fixed substrate, P and R are constants but $\boldsymbol{O}$ can vary. Furthermore, processive recombination is modelled by tangle addition. A recombination event that consists of n-rounds of processive recombination is translated into a system of $(n+1)$ equations with unknowns $\boldsymbol{O^{(i)}}$, P and R. The tangle $\boldsymbol{O^{(i)}}$ is allowed to change from one equation to another if and only if there is no topological selectivity. This introduces more unknowns to the system, and the analysis becomes much more difficult. It was seen that solutions for a system of three tangle equations with three unknowns could be found if the unknowns are rational tangles.

Acknowledgment

The authors are grateful to the Ministry of Human Resource Development, Government of India (New Delhi) for sponsoring the Research Project entitled, "Mathematical Modelling and Computer Graphics Simulation for the Analysis of DNA".

References

[1] C. C. Adams, *The Knot Book* (W.H. Freeman and Company, New York, 1994).

[2] J. H. Conway, An enumeration of knots and links and some of their related properties, *Computational Problems in Abstract Algebra* (Pergamon Press, Oxford, 1970), pp. 329–358.

[3] R. H. Crowell and R. H. Fox, *Introduction to Knot Theory*, 1st edn. (Springer-Verlag, New York Heidelberg Berlin, 1963).

[4] I. Darcy, Biological distances on DNA knots and links: Applications to XER recombination, *J. Knot Theory Ramifications* **10** (2001) 269–294.

[5] I. Darcy and D. Sumners, A strand passage metric for topoisomerase action, in *Knots* (World Scientific, 1997), pp. 267–278.

[6] I. Darcy and D. Sumners, Rational tangle distances on knots and links, *Math. Proc. Camb. Phil. Soc.* **128** (2000) 497–510.

[7] C. Ernst and D. W. Sumners, A calculus for rational tangles: Applications to DNA recombination, *Math. Proc. Camb. Phil. Soc.* **108** (1990) 489–515.

[8] C. Ernst, Tangle equations, *J. Knot Theory Ramifications* **5** (1996) 145–159.

[9] C. Ernst and D. W. Sumners, Solving tangle equations arising in a DNA recombination model, *Math. Proc. Camb. Phil. Soc.* **126** (1999) 23–36.

[10] J. R. Goldman and L. H. Kaufman, Rational tangles, *Adv. Appl. Math.* **3** (1997) 300–332.

[11] J. R. Goldman and L. Kaufman, Rational tangles, *Adv. Appl. Math.* **18** (1997) 300–332.

[12] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore and J. E. Darnell, *Molecular Cell Biology*, 4th edn. (W.H. Freeman and Company, New York, 1999).

[13] J. C. Misra, S. Mukherjee and A. K. Das, A mathematical model for enzymatic action on DNA knots and links, *Math. Comput. Modelling* **39** (2004) 1423–1430.

[14] D. W. Sumners, C. Ernst, S. J. Spengler and N. R. Cozzarelli, Analysis of the mechanism of DNA recombination using tangles, *Quart. Rev. Biophys.* **28** (1995) 253–313.

[15] J. Watson and F. Crick, Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid, *Nature* **171** (1953) 737.

[16] G. Zubay, W. W. Parson and D. E. Vance, *Principles of Biochemistry: Energy, Proteins, and Catalysis* (McGraw-Hill, 1995).

CHAPTER 9

USING MONODOMAIN COMPUTER MODELS FOR THE SIMULATION OF ELECTRIC FIELDS DURING EXCITATION SPREAD IN CARDIAC TISSUE

G. PLANK

Institut für Medizinische Physik und Biophysik,
Karl Franzens Universität Graz, Austria
gernot.plank@meduni-graz.at

Within the last decade computer models of cardiac excitation spread have become increasingly more realistic due to the interdisciplinary merging of techniques from biophysics, mathematics, cardiology and computer sciences. Computer models are considered as an indispensable complement to experimental and clinical studies. Both experimental as well as clinical methods for the determination of the cardiac activation sequence rely in most instances on the measurement of potentials outside the myocardium. For the interpretation of such measurement a profound understanding of the relationship between electrical processes in the tissue and the electric field caused by them outside the tissue is essential. Monodomain computer models are one of the most frequently used tools for the investigation of this relationship since they represent a balanced trade-off between level of detail and computational tractability. This paper summarizes theoretical basics necessary for the implementation of monodomain computer models for the simulation of the cardiac excitation spread and the concomitant electric field and reviews numerical techniques used for this purpose.

1. Introduction

The human heart is a mechanical pump with four chambers, two upper chambers, the atria, which act to fill the two lower main pumping chambers, the ventricles. Electrical signals propagate wavelike through the heart muscle to coordinate the mechanical contraction which guarantees appropriate blood circulation under normal conditions. Disturbances of the electrical signal conduction may deteriorate or impede the coordination of the mechanical contraction leading to discomfort or even life-threatening conditions. As a vital organ the diagnosis of malfunctions of the heart has been

an important issue since ever. Clinical examinations are based on potentials measured on the body surface from which conclusions are drawn on the electrical processes occurring within the heart. In the last decades a further method was established which permits potential measurements inside the heart by means of electrodes introduced with catheters. Beyond mere measurements this method also facilitates the modification of conduction pathways by delivering high-frequency impulses to the tissue. During *in vitro* experiments with heart preparations the situation is quite different. A wider gamut of methods, most of them not applicable *in vivo*, is applied, providing measurement data from inside the tissue as well. Nevertheless, clinical examinations of the cardiac activation sequence still rely almost exclusively on the measurement of potentials outside the tissue. Therefore, the understanding of the relation between the electrical processes occurring within the tissue and the concurrent electric field outside the tissue is fundamental for the interpretation of clinically recorded signals. Numerous studies addressed this question trying to elucidate this source-field relationship with both *in vitro* experiments and numerical studies with computer models.

The development of computer models is hampered by both structural as well as functional complexities of cardiac tissue. The tissue is composed of irregularly shaped and nonuniformly interconnected cells, surrounded by a fluid-filled space (interstitium) with embedded connective tissue and blood vessels. The cell borders are defined by an isolating membrane with highly nonlinear electrical properties. The local functional behavior of the tissue is determined by the cell membrane as the location of the electrical sources, however, the interaction of a membrane patch with adjacent tissue depends on the passive electrical properties determined by the tissue structure. All that constitutes evident difficulties to find an appropriate representation as electrical network.

Early computer models represented the three-dimensional cardiac structure as a one-dimensional cable and adopted membrane models appropriate for nerve membranes. Improved experimental techniques revealed more and more details of the basic mechanisms of cardiac electrical activity. Favored by the rapidly increasing availability of computational resources, these mechanisms were integrated subsequently into computer models to investigate their effects.

Despite considerable technical advances, insurmountable limitations of experimental measurements persist. For technical reasons the number of different parameters which can be measured simultaneously during experiments is limited. Since the cardiac excitation process is determined by many factors, the analysis of interactions between them is considerably

complicated by the fact that just a small number of them can be measured and the majority of them remain unknown.

In contrast to experimental measurements, the use of computer models allows a clear separation of effects caused by structure (connecting network) and function (membrane kinetics). These models, however, are based on simplifying assumptions on structure and function of the tissue and have to be considered as an approximation of the actual physical system. Therefore, computer models on its own are not particularly useful since the discrimination between model artefact and real physical behavior is impossible without experimental validation. As a complementary method to experimental measurement, however, they have proven to be an extremely powerful tool permitting insights which cannot be gained otherwise.

Computational resources available today are by far not sufficient to integrate all the available physiological knowledge into computer models. In fact, depending on the question being answered, a trade-off has to be made between the amount of details considered in the model and the tractability of simulations. If models are too simple, wrong or imprecise predictions will result, if too complex, computations will become intractable.

Computer models typically consist of two parts, one representing the tissue structure and one the functional behavior of the cell membrane. Regarding the choice the structural representation of the tissue two model types exist, the monodomain and the bidomain model. Both models can be coupled with a large variety of ionic models to describe the dynamic membrane behavior.

Since the excitation spread in cardiac tissue and the corresponding electric field are inextricably linked, in a strict sense both excitation spread and corresponding field have to be considered simultaneously to account for the feedback effect of the field on the electrical processes in the tissue. Whenever these effects are of interest, a mathematical treatment based on a bidomain formulation is more appropriate, since it allows to account for the different electrical properties in both compartments inside and outside the tissue.

Under the assumption, however, that these effects are negligible due to comparably small potentials in the extracellular medium, the excitation spread in the tissue and the extracellular potentials can be computed sequentially. Models based on these assumptions are generally referred to as monodomain models. All the techniques described in this work deal with that model type. Monodomain models are particularly suitable for the simulation of cardiac tissue surrounded by an extensive fluid medium (volume conductor). This reflects, for instance, the common experimental setup of a

small tissue preparation immersed in an extended fluid bath or in a clinical context, potential measurements in the blood-filled cavities of the heart.

The aim of this work is to provide the basic knowledge necessary for the setup of monodomain computer models for the simulation of the cardiac excitation spread and the corresponding extracellular potentials evoked by the sources within the tissue.

2. Physiological Background

This chapter is intended as a brief summary of basic concepts used in cardiac electrophysiology for readers unfamiliar with the subject. Characteristic properties of the cardiac tissue like its cellular structure, the basic ionic mechanisms responsible electric potential differences across the membrane and the response of the membrane to the application of stimulation currents will be elucidated. A more detailed introduction into membrane biophysics is found in [85].

2.1. *Properties of cardiac cells*

Cardiac muscle is composed of densely packed cells arranged into fibrous bundles [111]. A single cell is typically 30–100 μm long and 8–20 μm wide. Cells have an approximately cylindrical geometry with step-like irregularities at the cell ends. The cells are bounded by a thin plasma membrane ($\sim$75 Å thick) which separates the fluid-filled intracellular space from the interstitial fluid surrounding the cell. In general the membranes of adjacent cells are separated by narrow clefts, but at some points the membranes are connected via protein channels, called the nexus or gap junctions. The gap junctional membrane is localized mainly to the intercalated disks at the cell ends and, to a lesser extent, along the length of the cell. A typical cell is connected with ten neighboring cells [107]. As a consequence of the low conductance of the cell membrane and the spatial arrangement of gap junctions, the average conductance along the cell axes (longitudinal direction) is approximately ten times higher than in the perpendicular direction (transverse direction) [13].

The main function of the membrane is to control the passage of substances (ions and molecules) into and out of the cell. Its main constituent is lipid which is organized to form a lipid bilayer. The membrane lipid excludes passage of ions. Exchange of ions between intra- and extracellular spaces is only possible via channel proteins which are embedded in the bilayer matrix of the membrane.

Solute transport through the membrane is facilitated by passive and active mechanisms. Passive transport tends to equilibrate ionic concentrations inside and outside the cell and requires no expenditure of metabolic energy, whereas active transport is metabolically driven. Passive mechanisms include diffusion as a result of the concentration gradient (intra- and extracellular ionic concentrations are different), facilitated diffusion via carrier proteins (carrier-mediated transport) and through ion channels. Ion channels are specialized membrane proteins that allow the rapid movement of small ions like Na^+, K^+, Cl^- and Ca^{2+}. An ion carrying channel consists of an aqueous pore (through which ions may traverse the membrane), a selectivity filter (reflecting the selective permeability property of the channel, which allows only ions of a certain species to permeate the channel) and a control gate which may be classified as voltage- or as ligand-gated, depending on the nature of the mechanism that triggers the gate. Voltage-gated channels open or close their gates depending on the potential difference across the membrane; ligand-gated channels act depending on the concentration of different ions, neurotransmitters, hormones or drugs among other substances. To maintain intracellular ionic concentrations, active transport mechanisms are necessary to antagonize the effect of passive transport. Active transport is driven by so-called ion pumps which use metabolic energy to transport ions against the concentration gradient.

The unequal ionic concentration in the intra- versus the extracellular space gives rise to diffusion of ions along the concentration gradient, whereby the rate of diffusion depends on the difference in concentrations and the membrane permeability. Regardless of the mechanism, the movement of ions across the membrane constitutes a flow of electric current since ions are carrying charges. The membrane accumulates these charges due to its associated capacitance resulting in a potential difference across the membrane. This potential difference is associated with an electric field which exerts forces on all charged particles within the membrane. A steady-state is reached when the ion fluxes driven by diffusion and electric field forces are equal. The corresponding non-zero electrical potential difference is called resting membrane potential.

A quantitative description of diffusion is given by

$$\mathbf{j_d} = -D_p \nabla C_p \tag{1}$$

known as Fick's law, where C_p is the concentration of an ion of species p as a function of position and D_p is the corresponding diffusion constant. The

flux $\mathbf{j_d}$ is the number of ions passing per unit time through a cross section of unit area. The flux resulting from the electric field forces is given by

$$\mathbf{j_e} = -\frac{D_p F Z_p}{RT} C_p \nabla \Phi \tag{2}$$

where $-\nabla\Phi$ is the electric field, Z_p the valence of the ion species, F the Faraday constant, R the gas constant and T the absolute temperature.

The equilibrium potential difference for a given ion can be found from $\mathbf{j_d} + \mathbf{j_e} = 0$. Under these conditions we obtain from (1) and (2)

$$\nabla C_p = -\frac{F Z_p C_p}{RT} \nabla \Phi \tag{3}$$

If we assume that quantities vary in the direction ξ perpendicular to the membrane only we obtain

$$\frac{dC_p}{d\xi} = -\frac{F Z_p C_p}{RT} \frac{d\Phi}{d\xi} \tag{4}$$

Rearranging and integrating from the inner to the outer membrane surface yields

$$\int_i^e \frac{dC_p}{C_p} = -\frac{F Z_p}{RT} \int_i^e d\Phi \tag{5}$$

resulting in

$$V_m = \Phi_i - \Phi_e = \frac{RT}{F Z_p} \ln\left(\frac{[C_p]_e}{[C_p]_i}\right) \tag{6}$$

This is the Nernst potential for which the ion of type p is in equilibrium with its diffusion force. For example, the potential difference E_K necessary for the potassium to be in equilibrium is given by

$$E_K = \frac{RT}{F} \ln\left(\frac{[K^+]_e}{[K^+]_i}\right) \tag{7}$$

which is approximately $E_K \approx -88\,\text{mV}$ for $[K^+]_e = 5.4\,\text{mM}$ and $[K^+]_i = 145\,\text{mM}$. In general, biological membranes cannot be in equilibrium for all ions, since their Nernst potentials are different. The resting condition can be only characterized as a steady state ($\partial V_m/\partial t = 0$) which requires the total ionic flux to be zero. Under these conditions and the assumption of constant field strength within the membrane the resting potential in a two-ion system is given by Goldman's equation

$$\Phi_i - \Phi_e = \frac{RT}{F} \ln \frac{P_K [K^+]_e + P_{Na} [Na^+]_e}{P_K [K^+]_e + P_{Na} [Na^+]_i} \tag{8}$$

where P_K and P_{Na} are the permeabilities for potassium and sodium ions. At the resting membrane potential $V_{rest} \approx -84\,\text{mV}$ the membrane is much

more permeable for potassium than for sodium and P_{Na} in the Goldman equation may be neglected. This yields the Nernst equation and the resting potential of the membrane is close to the Nernst equilibrium potential for potassium.

2.2. *The action potential*

During each heart beat all cardiac cells cycles through a deflection of the transmembrane voltage $V_m = \Phi_i - \Phi_e$ which is in the most general case characterized by five distinct phases: the rapid upstroke (phase 0), the early repolarization (phase 1), the plateau (phase 2), the repolarization (phase 3) and the resting phase (phase 4). Such a cycle is called the action potential (see Fig. 1(a)).

An action potential is the response of the membrane to a current either from an external source (stimulation) or from an adjacent membrane in excited state. No action potential is elicited unless the depolarization reaches a specific level, called the threshold potential. Once the threshold potential is reached an action potential is triggered which is always identical (all-or-nothing). That is, the shape of the waveform of the action potential is independent of the initial depolarization. This property is known as excitability.

The initial part of phase 0, generally referred to as the foot of the action potential (see Fig. 1(b)), reflects the passive membrane behavior. If the depolarization is driven by adjacent tissue the time course of the foot is exponential. Once the threshold is reached, active membrane behavior is triggered. Beyond the threshold, a positive feedback between membrane permeability for sodium ions and transmembrane potential begins. The depolarization of the membrane increases the sodium permeability due to the opening of the Na^+ channels. As a consequence, the sodium current increases which leads to a further depolarization of the membrane and to a further increase of the sodium permeability which drives the membrane towards the Nernst potential for sodium E_{Na}. The resulting upstroke of V_m is extremely fast due to positive feedback mechanism with a duration of approximately 1 ms only (see Fig. 1(b)). During phase 1, a slight repolarization occurs reflecting a decreasing number of open Na^+ channels and the opening of K^+ channels. The plateau phase of the action potential is mainly sustained by calcium currents. Subsequently the closing of the Ca^{2+} channels and the opening of the delayed outward rectifier K^+ channels repolarizes the transmembrane gradually back to its resting state.

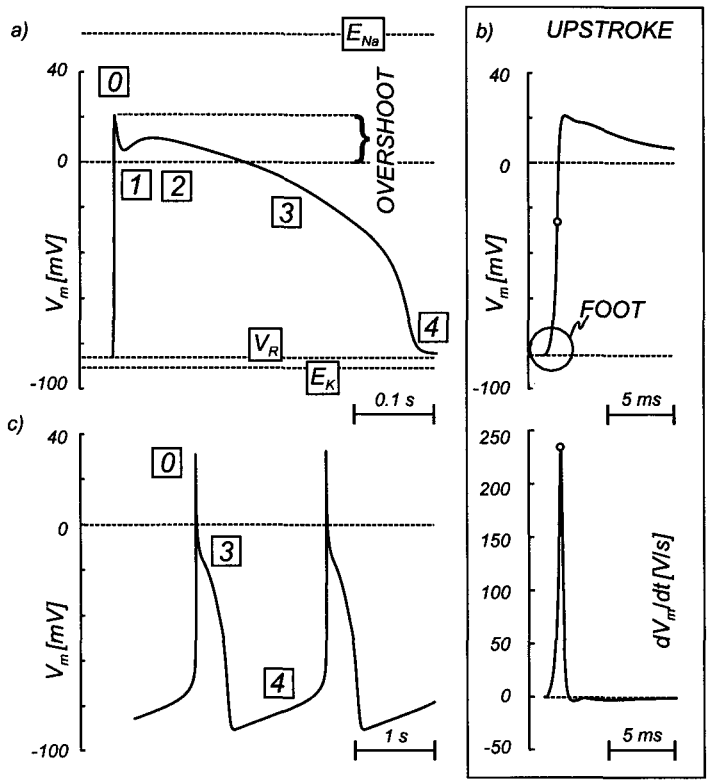


Fig. 1. (a) Phases of a ventricular action potential: Starting from the resting potential V_R which is close to the Nernst equilibrium potential for potassium E_K, the transmembrane voltage V_m cycles through an action potential. During the rapid upstroke (0) the membrane rapidly depolarizes towards the equilibrium potential for sodium E_{Na}, followed by early repolarization (1) and plateau (2). During repolarization (3) V_m returns gradually to its resting state (4). (b) Upstroke of the action potential: During the foot of the action the membrane behaves passively and the time course of V_m is exponential. Exceeding the threshold potential triggers active behavior, a positive feedback mechanisms causes a very fast upstroke of V_m. The derivative dV_m/dt of the transmembrane voltage demonstrates the short duration of the upstroke of only ≈ 1 ms. (c) Phases of a pacemaker action potential: A pacemaker action potential is characterized by the absence of a constant resting potential. The membrane depolarizes up to the threshold without any external stimulation and triggers an action potential automatically. Compared with (a) the phases (1) and (2) are not present.

The waveform of the action potential may differ depending on the cell type and not all the phases are observed in all types of cells. Particularly pacemaker cells like found in the sino-atrial node, in the atrioventricular node and in the His–Purkinje system show different waveforms

(see Fig. 1(c)). The salient property of these cells is the absence of a resting potential. During phase 4 pacemaker cells depolarize progressively permitting to reach the threshold without any external stimulation. This property is referred to as automaticity. Under normal conditions the cells of the sino-atrial node depolarize faster than all other pacemaker cells and thus are setting the pace of the heart beat. In case of malfunction of the sino-atrial node pacemaker cells of the atrio-ventricular node or the His–Purkinje system assume this duty.

Another important action potential characteristics is referred to as refractoriness. Once an action potential has been triggered, a subsequent depolarization will not elicit another action potential unless a certain minimum period of time (the absolute refractory period) has elapsed. For a subsequent time period (the relative refractory period), the threshold for the second depolarization is higher than normal. Under normal conditions, this prevents the action potential from returning to its origin since the excitation wave front would encounter tissue in refractory stake; Under pathological conditions, however, reentrant circuits way arise resulting in flutter or fibrillation of the heart.

3. Modelling the Membrane Kinetics

Based on the voltage clamp technique which allows the measurement of ionic currents during an action potential, Hodgkin and Huxley derived a quantitative model describing the cell membrane of a squid axon [49]. Although the model was developed for a nerve action potential, the mathematical formalism has been used in models for the cardiac action potential as well, even in contemporary models this formalism is virtually unchanged. The behavior of cardiac cells in different regions of the heart (pacemaker cells, atrial and ventricular cells) differs considerably and depends moreover on the species as well. Technical advances of the voltage clamp technique allowed to correct formulations used in older models and to identify of new currents. This led to the development of numerous models which account for more and more physiological details, described by a considerably increased number of state variables (the number of state variables increased from the Luo–Rudy phase I model to the phase II model from 9 to 30, a fact, which poses significant computational problems in large scale computations). Today specialized cardiac membrane models are available for the sino-atrial node [22], for Purkinje fibers [2, 24], for atrial [19, 62, 74, 96], and ventricular cells [5, 26, 28, 64–66].

A membrane model for the cardiac action potential describes the electrochemical events in a small membrane patch and the adjacent intracellular and interstitial media. The patch is assumed to be sufficiently small so that the diffusion in the adjacent media occurs essentially instantaneous, and at the same time sufficiently large that the membrane channels show their ensemble-averaged behavior and the probabilistic single channel behavior does not appear. Ionic current models in their most general form typically consist of three submodels: an electrical analog, a kinetic gating model, and a fluid compartment model.

The electrical analog connecting in parallel the ionic currents and the capacitive current is known as parallel-conductance model (see Fig. 2(d)). The total transmembrane current density $\tilde{i}_m$ [μA/cm^2] is given by

$$\tilde{i}_m = \tilde{c}_m \frac{\partial V_m}{\partial t} + \tilde{i}_{ion} = \tilde{c}_m \frac{\partial V_m}{\partial t} + \sum_p \tilde{i}_p \tag{9}$$

where $\tilde{i}_{ion}$ [μA/cm^2] is the sum of the ionic currents of ion species p and $\tilde{c}_m$ [μF/cm^2] is the specific membrane capacitance. Each parallel branch reflects the contribution of a partial current to the total transmembrane

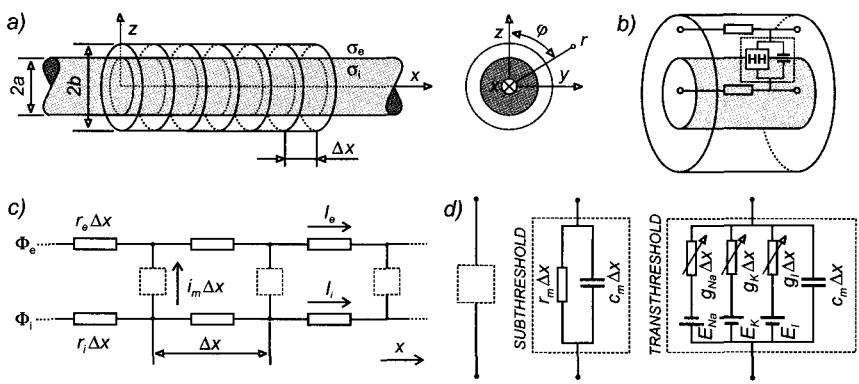


Fig. 2. (a) Physical model of a cylindrical fiber of radius a surrounded by a thin fluid film of radius b: Conductivity within the fiber is σ_i, respectively σ_e in the extracellular fluid. A coordinate system is chosen so that the axis x aligns with the cylinder axis. Cable models of cardiac fibers neglect potential variations with φ or r according to the core conductor assumptions. (b) Discrete cable element of length dx: Since cross-sectional potential variations are neglected the inner and outer cylinder may be replaced by a resistor. The membrane is modelled as a capacitance in parallel with an electric model of the resistive membrane properties of Hodgkin–Huxley type. (c) Linear core conductor model for restricted extracellular space. (d) Electical representation of a fiber element of length Δx under sub- and transthreshold conditions: The conductances g_{Na} and g_K are found from the Hodgkin–Huxley equations, E_{Na} and E_K are the equilibrium potentials for sodium and potassium, respectively.

current. The ionic currents branches are of the form of Ohm's law

$$\tilde{i}_p = \tilde{g}_p (V_m - E_p) \tag{10}$$

where $\tilde{g}_p$ [mS/cm^2] is the membrane conductance and E_p the Nernst equilibrium potential for the ion species p. The net driving force for ion species p is $(V_m - E_p)$ which is the deviation of the membrane potential from the equilibrium condition. Depending on the mechanism, the conductance $\tilde{g}_p$ may be represented as a simple constant or as a nonlinear term. Leakage or background currents accounting for nonspecific current flow are typically modelled with fixed conductances. Nonlinear currents in cardiac membrane models, however, are described by kinetic gating models based on the Hodgkin–Huxley formalism. According to this formalism, it is assumed that the conductance of each channel is determined by a number of independent subunits or gates, each having two possible states: open or closed. Ions may pass through the membrane via a particular channel if all subunits are in the appropriate state. Switching between open and closed state of a single channel is a stochastic process where instant and duration of the opening and closing processes random variables. The current derived from a large number of such channels corresponds to the macroscopically measured current. The state transition of the gates is governed by first-order kinetics. If y is the probability of a particular gate to be in the open state, the ensemble-averaged transient behavior of this type of gate is given by

$$\frac{dy}{dt} = \alpha(V_m)(1 - y) - \beta(V_m)y \tag{11}$$

where α and β are non-negative monotonic functions depending on V_m only.

To exemplify this formalism, the system of equations used by Hodgkin and Huxley in their first quantitative model of a nerve cell membrane is given (see Eq. (20)) [49]. Their model took into account the sodium current, responsible for the fast upstroke of the action potential, the potassium current for the repolarization of the membrane, and a leakage current. Currents were formulated as

$$\tilde{i}_{Na} = \tilde{g}_{Na}(V_m - E_{Na}) \tag{12}$$

$$\tilde{i}_K = \tilde{g}_K(V_m - E_K) \tag{13}$$

$$\tilde{i}_l = \tilde{g}_l(V_m - E_l) \tag{14}$$

where $\tilde{g}_l$ is a constant and $\tilde{g}_{Na}$ and $\tilde{g}_K$ depend on the gating variables m, h and n according to

$$\tilde{g}_{Na} = \hat{g}_{Na} m^3 h \tag{15}$$

$$\tilde{g}_K = \hat{g}_K n^3 \tag{16}$$

where E_{Na} and E_K are the Nernst equilibrium potentials and $\hat{g}_{Na}$ and $\hat{g}_K$ are the maximum conductances of sodium and potassium, respectively. Gating variable m represents the portion of open activation gates of a sodium channel and h the portion of open inactivation gates. In case of the potassium channel there is only one state variable determining the ion transport rate. The transient behavior of these gates is governed by

$$\frac{dm}{dt} = \alpha_m(V_m)(1-m) - \beta_m(V_m)m \tag{17}$$

$$\frac{dh}{dt} = \alpha_h(V_m)(1-h) - \beta_h(V_m)h \tag{18}$$

$$\frac{dn}{dt} = \alpha_n(V_m)(1-n) - \beta_n(V_m)n \tag{19}$$

Hence the total transmembrane ionic current of the Hodgkin–Huxley model is given by

$$\tilde{i}_{ion} = \hat{g}_{Na}m^3h(V_m - E_{Na}) + \hat{g}_K n^4(V_m - E_K) + \tilde{g}_l(V_m - E_l) \tag{20}$$

Contemporary membrane models like [19, 65, 74] include fluid compartment models. These models allow to account for variations of ionic concentrations in intracellular and interstitial compartments as a consequence of ionic fluxes by enforcing mass conservation with the fluxes between the compartments. A fluid compartment model leads to one or more first order, ordinary differential equations.

Thus a general membrane model comprises one equation imposing current conservation according to Kirchhoff's current law (9), some mass conservation equations describing the fluid compartment model, and some kinetic gating equations like given in (11) or (17)–(19). The common independent variable is time, the dependent variables are the transmembrane voltage, the ionic concentrations and the gating variables. All equations together constitute a system of first-order, nonlinear ordinary equations. Since analytic solutions are not known for this system, numerical methods have to be applied.

For mathematical methods presented in the following sections physiological details of the membrane models will not be considered, since they are not essential for the understanding of the basic concepts. Hence for the sake of simplicity the following abstractions are made. All model variables except the transmembrane potential are collectively referred to as membrane state variables and denoted by m_n with $n = 1, \ldots, N$ where N is the

number of state variables of a particular model. The mass conservation and gating equations are written as

$$\frac{dm_n}{dt} = f_n(m_n, V_m) \tag{21}$$

where $f_n(m_n, V_m)$ are nonlinear functions. The equation for the total transmembrane current is written as

$$\tilde{i}_m = \tilde{c}_m \frac{\partial V_m}{\partial t} + \tilde{i}_{ion}(m_n, V_m) \tag{22}$$

The sum of all ionic currents is given by the nonlinear function $\tilde{i}_{ion}(m_n, V_m)$ similar to (20), each single ionic current is given by an expression as in Eq. (12).

4. Modelling of Action Potential Propagation in Cardiac Tissue

Early studies of action potential propagation [48, 49] used the one-dimensional cable equation to describe the electrical behavior of a cylindrical nerve fiber. In contrast to nerve fibers, cardiac tissue is better characterized as a three-dimensional electrical network of complex geometry and discontinuous distribution of electrical parameters rather than a uniform continuous fiber. Experimental evidences [14, 125] suggested that cardiac tissue exhibits syncytial behavior which justified, to a certain extent, a homogenization of the discrete cellular structure into a uniformly continuous region. Based on the assumption of syncytial behavior, investigators began to apply the continuous cable theory to cardiac tissue considering the conduction along a representative fiber and compared the results with experimentally obtained data. Based on the one-dimensional cable theory models were extended to two and three dimensions to account for effects of anisotropy [13]. A detailed deduction of the mathematical description of the cable analysis may be found in various places like for instance in [37, 51, 54, 85], a summary of fundamental relations will be given here.

4.1. *Core conductor model*

4.1.1. *Electrical parameters of a cylindrical fiber*

For a uniform continuous cylindrical structure it is convenient to define its electrical parameters on a per unit length basis. The axial resistance per unit length r_i [$k\Omega$/cm] of the intracellular fluid (myoplasm) is defined as

the resistivity ρ_i [$k\Omega$ cm] divided by the cross sectional area. If we designate the radius of the cylinder with a, the resistance per unit length is given by

$$r_i = \frac{\rho_i}{a^2\pi} \tag{23}$$

and the conductivity per unit length g_i [mS cm] by

$$g_i = \sigma_i a^2 \pi \tag{24}$$

where σ_i is the conductivity of the myoplasm in [mS/cm].

The cylindrical membrane enclosing the myoplasm shows resistive as well as capacitive properties and might be characterized as a capacitance shunted with a leakage resistance. If we designate the specific resistance of the membrane with $\tilde{r}_m$ [$k\Omega\,\mathrm{cm}^2$], the specific conductance with $\tilde{g}_m$ [mS/cm^2] and the specific capacitance with $\tilde{c}_m$ [μF/cm^2], the per unit length quantities r_m [$k\Omega$ cm], g_m [mS/cm] and c_m [μF/cm] are given by

$$r_m = \tilde{r}_m / 2\pi a \tag{25}$$

$$g_m = \tilde{g}_m\, 2\pi a \tag{26}$$

$$c_m = \tilde{c}_m\, 2\pi a \tag{27}$$

4.1.2. *Electrical model of a single fiber*

A rigorous mathematical treatment of an infinite excitable fiber immersed in an extensive, homogeneous, conducting medium would require the solution of a three-dimensional field problem [12]. The potential field of such a fiber can be considered as quasi-static satisfying Laplace's equation in the external medium and the myoplasm. If we assume the membrane as infinitely thin Laplace's equation is satisfied everywhere and non-zero potential fields can only be explained with discontinuities across the membrane interface (that is, the sources of the fields are located on or within the membrane). If we designate the potential at an arbitrary point in the external medium in cylindrical coordinates (see Fig. 2(a)) with $\Phi_e(r, \varphi, x)$, then

$$\Delta\Phi_e(r, \varphi, x) = 0 \quad r \geq a \tag{28}$$

must be satisfied in the extracellular space and

$$\Delta\Phi_i(r, \varphi, x) = 0 \quad r \leq a \tag{29}$$

in the myoplasm.

Computer models based on these equations are rarely found, just a few studies are reported [58, 99, 124]. These models are based on a boundary

element approach which makes use of Green's theorem to simplify the three-dimensional problem given by Eqs. (28)–(29) to two dimensions.

Making various simplifying assumptions permits to reduce this three-dimensional problem to the essentially one-dimensional problem of a core conductor.

A formulation reduced by one dimension is obtained by assuming axial symmetry, that is $\partial/\partial\varphi = 0$ with φ denoting the azimuth angle. If the extracellular fluid is restricted as shown in Fig. 2(a) to a thin cylindrical fluid sheet of radius b it may be further assumed that all the involved quantities, extra- and intracellular potentials and currents, are a function of x only. This is equivalent to the assumption that at a given site x no radial potential gradients of Φ_i and Φ_e (i.e. no radial current flow) arise and consequently the current flow in both media must be confined to the axial direction only.

In the intracellular space the assumption of axial current flow seems to be well satisfied, since the diameter of a cardiac cell is small compared to its length. In the extracellular space the validity of this assumption depends on the relation of a and b, but also on the spatial distribution of the sources within the membrane. For values of $b < 1.5a$, however, the core conductor assumptions are well satisfied independent of the source distribution within the membrane [102, 118], for $b \gg a$ deviations from core conductor behavior will occur.

For instance, if we consider a bundle of parallel, tightly packed fibers instead of a single fiber, core conductor assumption might be well satisfied for a fiber near the center of the bundle, since the cross section available for interstitial current flow is comparable with the myoplasmic cross section. For fibers near the bundle surface, however, the interstitial space of the fibers is somewhat in closer contact with the surrounding fluid. This will give rise to radial current flow as well and, in consequence, the associated extracellular potentials will deviate from those expected from core conductor assumptions. In the context of monodomain computer models (see Sec. 4.2) it is common practice to neglect the extracellular resistance, since the extracellular potentials are small compared to the intracellular ones. Thus one forgive the ability to compute Φ_e directly from the model, however, the recovery of these potentials from the transmembrane current distribution is still possible (see Sec. 5).

Starting from the core conductor assumptions a single fiber can be represented as a discrete electrical network. Although the cable equations are based on a continuum, a representation as a repetitive network of finite-length Δx is equivalent for $\Delta x \to 0$. Potentials and currents are designated

with Φ_e and I_e along the extracellular path, and with Φ_i and I_i along the intracellular path, respectively (see Fig. 2(c)). Like illustrated in Sec. 3, the electrical behavior of the membrane depends on the transmembrane potential V_m. Two ranges are to be discriminated, a linear subthreshold range $V_m < V_{th}$, where the membrane is characterized as a passive RC-structure, and a nonlinear transthreshold range $V_m \geq V_{th}$, where a characterization based on non-linear kinetic models is required (see Fig. 2(d)).

4.1.3. *Cable equations*

The application of Kirchhoff's laws to the electric circuit analog of a core conductor (see Fig. 2(c)) leads to the cable equations. In the subsequent analysis we will continue to consider an infinitely long cylindrical continuous cable of radius a surrounded by an extracellular fluid cylinder of radius b, respectively its analog representation as an electric circuit like shown in Fig. 2(c).

According to Ohm's law the decrease in potentials Φ_i and Φ_e per unit length must be equal to the voltage drop caused by the axial currents I_i and I_e at the resistances r_i and r_e. Hence

$$\frac{\partial \Phi_e}{\partial x} = -I_e r_e \tag{30}$$

$$\frac{\partial \Phi_i}{\partial x} = -I_i r_i \tag{31}$$

From Kirchhoff's current law we conclude that the axial decrease in the intracellular current occurs as a consequence of the loss of current which enters the extracellular space by crossing the membrane. Expressed on a per unit length basis this yields

$$\frac{\partial I_i}{\partial x} = -i_m \tag{32}$$

The current leaving the intracellular space must appear in the extracellular space and sums up there to the extracellular current. A further increase may occur due to applied stimulation currents. If we express the stimulation current i_s as current per unit length as well, we obtain

$$\frac{\partial I_e}{\partial x} = i_m + i_s \tag{33}$$

The transmembrane voltage V_m is defined as the difference between intracellular and extracellular potential at the inner and outer surface of the

membrane. Since radial potential variations are neglected according to the core-conductor assumptions the transmembrane potential V_m is defined as

$$V_m = \Phi_i - \Phi_e \tag{34}$$

In absence of a stimulating current i_s the magnitudes of the intra- and extracellular axial currents are equal and $I_i = -I_e$ holds. Using this and Eqs. (30)–(31) allow the representation of the derivatives of Φ_i and Φ_e as a function of V_m. This yielding

$$\frac{\partial \Phi_i}{\partial x} = \frac{r_i}{r_i + r_e} \frac{\partial V_m}{\partial x} \tag{35}$$

$$\frac{\partial \Phi_e}{\partial x} = -\frac{r_e}{r_i + r_e} \frac{\partial V_m}{\partial x} \tag{36}$$

or if we just consider deflections from the resting values of V_m, Φ_i and Φ_e and designate them with v_m, ϕ_i and ϕ_e (these new quantities are equal to the original quantities aside from a constant) we obtain

$$\phi_i = \frac{r_i}{r_i + r_e} v_m \tag{37}$$

$$\phi_e = -\frac{r_e}{r_i + r_e} v_m \tag{38}$$

The deflections of Φ_i, Φ_e and V_m from their resting values are related by a simple voltage divider like expression.

The relation between the transmembrane current i_m and the transmembrane potential V_m can be found by subtracting (30) from (31)

$$\frac{\partial V_m}{\partial x} = -r_i I_i + r_e I_e \tag{39}$$

differentiating the result with respect to x

$$\frac{\partial^2 V_m}{\partial x^2} = -r_i \frac{\partial I_i}{\partial x} + r_e \frac{\partial I_e}{\partial x} \tag{40}$$

and substituting Eqs. (32)–(33) into (40)

$$\frac{\partial^2 V_m}{\partial x^2} = (r_i + r_e) i_m + r_e i_s \tag{41}$$

Equation (41) is valid under core conductor conditions, regardless whether the membrane is sub- or transthreshold. In absence of stimulating currents ($i_s = 0$), Eq. (41) shows that the transmembrane current i_m and the second spatial derivative of the membrane potential V_m are proportional.

A fiber immersed in an extended volume conductor represents a setup which is not conform with the core conductor assumption. If we assume the volume conductor as infinitely conductive with $r_e \approx 0$, no potential drops occur in the extracellular fluid and from $\Phi_e \approx 0$ follows $V_m \approx \Phi_i$. Hence (41) may be written as

$$i_m = \frac{1}{r_i} \frac{\partial \Phi_i^2}{\partial x^2} \tag{42}$$

The same result can be deduced by differentiating (31) with respect to x and substituting into (32). Since Eq. (42) was deduced from intracellular quantities only it is valid whether or not the core conductor assumptions are fulfilled in the extracellular domain. Equations (41) and (42) are considered as monodomain equations, since a single partial differential equation is used to describe the behavior of a fiber. In contrast to this, a bidomain description of a cylindrical fiber can be deduced from Eqs. (30) and (31). Differentiation of (30)–(31) and substitution of (32)–(33) yield

$$\frac{1}{r_i} \frac{\partial \Phi_i^2}{\partial x^2} = i_m + i_{si} \tag{43}$$

$$\frac{1}{r_e} \frac{\partial \Phi_e^2}{\partial x^2} = -i_m + i_{se} \tag{44}$$

where i_{si} and i_{se} are internally or externally applied stimulation currents. Note that Eqs. (41)–(44) are equally valid for ϕ_i, ϕ_e and v_m, since a spatial derivative is involved.

4.1.4. *Linear subthreshold conditions*

If the deflections of the transmembrane voltage from the resting potential are sufficiently small, the relationship of membrane current i_m and voltage V_m is given by a passive admittance. This subthreshold range where the membrane responds passively is referred to as linear or electrotonic. The electrical behavior of the membrane can be characterized as a capacity c_m in parallel with a resistance r_m. In contrast to transthreshold conditions, r_m is constant and does not depend neither on time nor on V_m.

Examination of the membrane behavior under electrotonic conditions is important for several reasons. The tissue ahead of a propagating action potential is characterized as electrotonic corresponding to the foot of the action potential in the temporal course of V_m (see Fig. 1(b)). Also for the study of electric stimulation subthreshold conditions are frequently used, since it is often considered as sufficiently accurate to determine whether

a stimulation pulse raises the membrane potential enough to reach the threshold voltage V_{th} or not. Furthermore, in experimental studies passive conditions has been used frequently to determine membrane parameters.

The transmembrane current under subthreshold conditions is given by

$$i_m = \frac{v_m}{r_m} + c_m \frac{\partial v_m}{\partial t} \tag{45}$$

where v_m denotes the excursions of the membrane potential from its resting value as defined before.

Substituting (45) in (41) yields

$$\frac{r_m}{r_i + r_e} \frac{\partial^2 v_m}{\partial x^2} - r_m c_m \frac{\partial v_m}{\partial t} - v_m = \frac{r_e r_m}{r_i + r_e} i_s \tag{46}$$

Characteristic properties of cable are identified as the time constant τ_m and the length constant λ which are defined as

$$\lambda = \left(\frac{r_m}{r_i + r_e} \right)^{1/2} \quad \text{and} \quad \tau_m = r_m c_m \tag{47}$$

Substituting (47) in (46) results in

$$\lambda^2 \frac{\partial^2 v_m}{\partial x^2} - \tau_m \frac{\partial v_m}{\partial t} - v_m = r_e \lambda^2 i_s \tag{48}$$

Assuming steady-state conditions ($\partial / \partial t = 0$) and a current injection of strength I_0 at $x = 0$, represented as a spatial delta function $i_s = I_0 \, \delta(x)$, is obtained the steady-state equation

$$\lambda^2 \frac{\partial^2 v_m}{\partial x^2} - v_m = r_e \lambda^2 I_0 \, \delta(x) \tag{49}$$

with the solution of the homogeneous form

$$v_m(x) = A \, e^{-x/\lambda} + B \, e^{x/\lambda} \tag{50}$$

where A and B are arbitrary constants. Imposing boundary conditions on A and B [85] to account for the effect of the stimulating current yields

$$v_m(x) = -\frac{r_e \lambda I_0}{2} \, e^{-x/\lambda}, \quad x \geq 0 \tag{51}$$

$$v_m(x) = -\frac{r_e \lambda I_0}{2} \, e^{x/\lambda}, \quad x \leq 0 \tag{52}$$

From inspection of (51) it can be concluded that the application of a stimulus current influences the transmembrane voltage, since v_m is different from zero at all sites x. The strongest influence of the stimulus occurs at the stimulus site itself and decreases exponentially with x. A positive current causes a more negative transmembrane potential (hyperpolarization),

whereas a negative current gives rise to an increase (depolarization) in v_m. The length constant λ represents a measure for the spatial extension of a disturbance as a consequence of a stimulating current, at $x = \lambda$ the change in v_m is $1/e$ of the magnitude at the stimulus site.

The general time-varying solution of Eq. (48) for $x \geq 0$ is given by

$$v_m(x,t) = -\frac{r_e \lambda I_0}{4} \left\{ e^{-x/\lambda} \left[1 - \operatorname{erf}\left(\frac{x}{2\lambda} \left(\frac{\tau_m}{t} \right)^{1/2} - \left(\frac{t}{\tau_m} \right)^{1/2} \right) \right] \right.$$
$$\left. - e^{x/\lambda} \left[1 - \operatorname{erf}\left(\frac{x}{2\lambda} \left(\frac{\tau_m}{t} \right)^{1/2} + \left(\frac{t}{\tau_m} \right)^{1/2} \right) \right] \right\} \tag{53}$$

where erf is the error function, defined by $\operatorname{erf}(y) = \int_0^y e^{-x^2} dx$. The solution for $x \leq 0$ is found from symmetry. A detailed elucidation of the solution steps is found in [85].

Since $\operatorname{erf}(\infty) = 1$ and $\operatorname{erf}(-\infty) = -1$, the spatial course of $v_m(x, t \to \infty)$ reduces the expression (53) to (51), valid under steady-state conditions. Using $\operatorname{erf}(-y) = -\operatorname{erf}(y)$ permits to derive the time course of v_m at the origin $x = 0$ from (53). This results in

$$V_m(x = 0, t) = -\frac{r_e \lambda I_0}{2} \operatorname{erf}\left(\sqrt{\frac{t}{\tau_m}} \right) \tag{54}$$

with a peak value of $-r_e \lambda I_0/2$ for $t = \infty$. The quantity $-r_e \lambda I_0$ is divided by two as a consequence of the tacit assumption of current sinks of strength $-I_0/2$ at $x = \pm\infty$, therefore half the current goes towards $-\infty$ and half towards $+\infty$.

Due to the presence of a capacitance, time is required to charge the membrane at a given distance x from the stimulus site to its steady-state value and τ_m is a measure for this effect (see Fig. 3(b)). For a given instant t, the spatial decay is exponential-like, with increasing t the spatial course $v_m(x)$ approaches the true exponential course of the steady-state solution in (51). This continuous decay of v_m with x is explained by the leakage resistance of the membrane with λ as a measure of this effect (see Fig. 3(a)).

4.1.5. *The propagating action potential*

Once a sufficiently large membrane patch is depolarized beyond the threshold voltage and active behavior is triggered, the membrane undergoes a change in transmembrane potential referred to as action potential. The rise of the potential relative to adjacent regions, where no potential changes occurred, leads to current flow between the active site and the surrounding

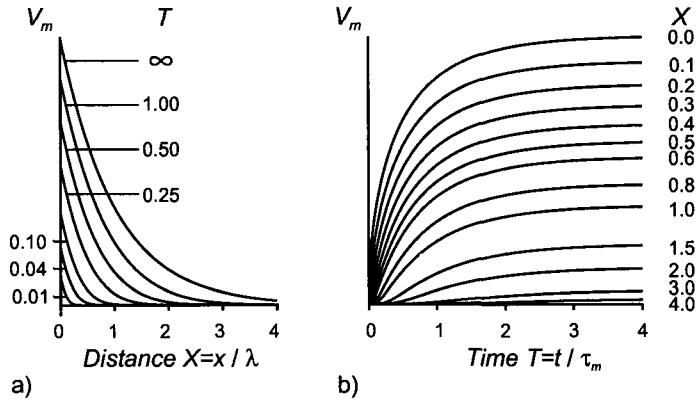


Fig. 3. Distribution of transmembrane voltage V_m of a passive cell membrane in response to the onset of a constant current, applied extracellularly at the origin $x = 0$. (a) shows the spatial distribution of V_m at different times, and (b) the time course of the potential at different distances along the fiber. Time T is expressed as multiples of the time constant τ, the distance X along the fiber as multiples of the length constant λ.

inactive regions. These currents, known as local circuit currents, depolarize the vicinity of the active site up to the threshold, so that the adjacent inactive tissue becomes active as well. Stimulation is only required to initiate this process, once it is triggered, these changes in membrane potential propagate through the tissue in a self-sustained manner. This is referred to as action potential propagation and is associated with the conditions $i_s = 0$ and $I_i = -I_e$. Consequently, (41) specializes to

$$\frac{1}{r_i} \frac{\partial^2 V_m}{\partial x^2} = i_m \tag{55}$$

where the extracellular resistance r_e is assumed to be zero. It is a known fact that the propagation of the action potential along a uniform fiber takes place without distortion and damping. This is only possible since cardiac tissue is an active medium capable of storing metabolic energy. The property of undistorted propagation is mathematically expressed by

$$V_m(x,t) = V_m(x - \theta t) \tag{56}$$

or equivalently as a differential equation, obtained by differentiating (56) twice using the chain rule

$$\frac{\partial^2 V_m}{\partial x^2} = \frac{1}{\theta^2} \frac{\partial^2 V_m}{\partial t^2} \tag{57}$$

Equation (57) is referred to as wave equation, where θ is the propagation velocity of the action potential.

For the foot of the action potential RC-behavior of the membrane can be assumed. Then the transmembrane current i_m is given by (45) and Eq. (55) can be rewritten using (47) to obtain

$$\left(\frac{\lambda}{\theta}\right)^2 \frac{\partial^2 V_m}{\partial t^2} - \tau_m \frac{\partial V_m}{\partial t} - V_m = 0 \tag{58}$$

to describe the foot of the action potential. The solution of this equation is the sum of two exponentials with the time constants

$$\frac{1}{\tau_{F_{1,2}}} = \frac{\tau_m \Theta^2}{2\lambda^2} \pm \frac{\Theta}{2\lambda}\sqrt{\frac{\tau_m^2 \Theta^2}{\lambda^2} + 4} \tag{59}$$

In most cases $\tau_m^2 \Theta^2/\lambda^2 \gg 4$ is well satisfied and the foot of the action potential is a monoexponential process, characterized by the time constant τ_F given by

$$\tau_F = \frac{(\lambda/\Theta)^2}{\tau_m} \tag{60}$$

Under transthreshold conditions the membrane behavior is nonlinear and more complex membrane models are required. Ionic membrane models usually describe the transmembrane current on a per unit area basis. Thus for the linkage of Eq. (55) with a kinetic model it is convenient to express the transmembrane current on a per unit area basis rather than on a per unit length basis. Both quantities, $\tilde{i}_m$ and i_m, are related through the cylindrical geometry by

$$i_m = 2\pi a \tilde{i}_m \tag{61}$$

This permits to rewrite (55)

$$\tilde{i}_m = \frac{a}{2\rho_i} \frac{\partial^2 V_m}{\partial x^2} \tag{62}$$

If the membrane is capable of conducting an action potential at constant velocity, Eq. (57) may be used to obtain

$$\tilde{i}_m = \frac{a}{2\rho_i \theta^2} \frac{\partial^2 V_m}{\partial t^2} \tag{63}$$

If the time course $V_m(t)$ at a given site x is known, the time course of $\tilde{i}_m(t)$ is determined as well. In general, the relation between $V_m(t)$ and $\tilde{i}_m(t)$ is complicated and there will not be a simple relation between them. As an example, the total transmembrane current as formulated in the Hodgkin–Huxley

model is stated here:

$$\tilde{i}_m = \tilde{c}_m \frac{\partial V_m}{\partial t} + \hat{g}_{Na} m^3 h (V_m - E_{Na}) + \hat{g}_K n^4 (V_m - E_K) + \tilde{g}_l (V_m - E_l) \tag{64}$$

From Eqs. (62)–(64) several conclusions can be drawn. The transmembrane current is determined by two factors, the electrical load seen by the membrane and the intrinsic properties of the membrane. The electric load is imposed by the properties of the conducting region which relates the current that crosses the membrane with the second spatial derivative of V_m. Equation (62) states how the potentials in the neighborhood of a certain patch and the resistive coupling affect the local transmembrane current. The intrinsic properties of the membrane like its capacitance and its time and voltage dependent permeability for ions of different species reflect the local dynamic behavior of the membrane, comprised by Eq. (64).

From inspection of (63), however, an important result can be deduced without having to solve the equation explicitly. One pair of functions $V_m(t)$ and $\tilde{i}_m(t)$ satisfying (63) will continue to be a solution if

$$\frac{a}{2\rho_i \theta^2} = \frac{1}{K} \tag{65}$$

where K is constant. Thus the conduction velocity is given by [50]

$$\theta = \sqrt{\frac{aK}{2\rho_i}} \tag{66}$$

The values for $\tilde{i}_m$ in Eqs. (62)–(64) must be equal. Equating two of them, (64) to (62) or (63), allows to solve for a propagating action potential. In general an analytical solution is not possible and numerical methods have to be applied. Equating (63) and (64) permits to solve for V_m as a function of time only. This method was originally used by Hodgkin and Huxley. They guessed a value θ and stepped through the solution as a function of time. For an incorrect guess of θ the solution was found to diverge, but with a correct θ the time course of the action potential was found. Modern computer methods, however, are based on equating (62) to the intrinsic membrane current given by (64) which allows to find solutions for V_m as a function of space and time.

4.1.6. *Finite length cables*

Up to this point, all the analysis was based on the assumption of an infinitely long cable, real cables, however, are of finite length. This

discrepancy gives reason to expect a distinct behavior. Differences in behavior between finite and infinite cables should be examined in terms of the input impedance Z_{in}, defined as

$$Z_{in} = \frac{v_m}{I_i} \tag{67}$$

and evaluated at the stimulus site, by simply comparing their input impedances under steady state conditions. For this purpose we will regard a semi-infinite cable with one end at $x = 0$ and the other end at $x = +\infty$. The same equations as for the infinite cable are valid apart from the minor difference that the voltage at $x = 0$ for the semi-infinite cable, given by

$$v_m(x) = -r_e \lambda I_0 \, e^{-x/\lambda} \tag{68}$$

is twice the voltage of the infinite cable. Insight is gained by regarding the infinite cable as a shunt of two semi-infinite cables (therefore the infinite cable has half the impedance of the semi-infinite cable), or equivalently, by taking into account that no current will flow in the $-x$ direction in the semi-infinite case. Consequently, the entire current I_0 will flow in the $+x$-direction instead of $I_0/2$ in the infinite case.

A stimulation current i_s is injected at $x = 0$, for $x \geq 0^+$, however, $i_s = 0$ holds and the axial currents are related by $I_e = -I_i$. Using this the intracellular current I_i is found from (39). This gives

$$I_i = -\frac{1}{r_i + r_e} \frac{\partial V_m}{\partial x} = -\frac{r_e \lambda I_0}{(r_i + r_e)\lambda} \, e^{-x/\lambda} \tag{69}$$

Division of (68) by (69) according to (67) permits to express the input impedance $Z_{in} = Z_0$ of the semi-infinite cable at $x = 0^+$ as

$$Z_0 = \left. \frac{-r_e \lambda I_0 \, e^{-x/\lambda}}{\frac{-r_e \lambda I_0}{(r_i + r_e)\lambda} \, e^{-x/\lambda}} \right|_{x=0^+} = \lambda (r_i + r_e) \tag{70}$$

or

$$Z_0 = \sqrt{(r_i + r_e) r_m} \tag{71}$$

To find Z_{in} in general for a cable of finite length $x = L$, terminated with an arbitrary load impedance Z_L at $x = L$, the homogeneous expressions

$$v_m(x) = A \, e^{-x/\lambda} + B \, e^{x/\lambda} \tag{72}$$

and

$$I_i = \frac{1}{Z_0} \left(A \, e^{-x/\lambda} - B \, e^{x/\lambda} \right) \tag{73}$$

will be used. Appropriate values for A and B are found by means of the reflection coefficient γ, a factor relating the terminal impedance Z_L with

the impedance of the infinite cable Z_0. The impedance at a certain site x is given by

$$Z(x) = Z_0\left(\frac{A\,e^{-x/\lambda} + B\,e^{x/\lambda}}{A\,e^{-x/\lambda} - B\,e^{x/\lambda}}\right) \tag{74}$$

At $x = 0$ the input impedance is found from (74) with

$$Z_{in} = Z_0\left(\frac{A+B}{A-B}\right) \tag{75}$$

and the terminal impedance Z_L at $x = L$ with

$$Z_L = Z_0\left(\frac{A\,e^{-L/\lambda} + B\,e^{L/\lambda}}{A\,e^{-L/\lambda} - B\,e^{L/\lambda}}\right) \tag{76}$$

The reflection coefficient γ, a term adopted from the theory of travelling electromagnetic waves due to similarities of the mathematical representation, at the cable end is defined as

$$\gamma(L) = \frac{A\,e^{-L/\lambda}}{B\,e^{L/\lambda}} = \frac{Z_L + Z_0}{Z_L - Z_0} \tag{77}$$

If the cable is terminated with $Z_L = Z_0$, the cable is equivalent to an infinite cable and no "reflections" will occur ($\gamma = \infty$). A reflection coefficient of $\gamma = \pm 1$ corresponding to a load impedance of $Z_L = \infty$, 0 represents a maximum discontinuity and everything will be "reflected". If we rewrite (76) by using (77) is obtained the input impedance as a function of the reflection coefficient

$$Z_{in} = Z_0\left(\frac{\gamma\,e^{2L/\lambda} + 1}{\gamma\,e^{2L/\lambda} - 1}\right) \tag{78}$$

A finite cable with a sealed end is considered as a cable terminated with an open circuit, that is $Z_L = \infty$ and $\gamma = 1$. For such a cable the input impedance is given by

$$Z_{in} = Z_0\left(\frac{e^{2L/\lambda} + 1}{e^{2L/\lambda} - 1}\right) = Z_0 \coth\left(\frac{L}{\lambda}\right) \tag{79}$$

Other special cable solutions like for cables terminated with a short circuit or with a finite impedance are found in [21, 37]. Regarding (79), for $L = 3\lambda$ the ratio of $Z_{in}/Z_0 = 1.01$, i.e. the deviation of the input impedance compared to a cable of infinite length is about 1%. For cables with $L < \lambda$ substantial deviations in behavior compared to infinite cables are to be expected. For instance $Z_{in}/Z_0 = 10$ for a cable with $L = 0.1\lambda$. This is of particular importance regarding numerical cable simulations. The majority

of studies considers the tissue terminated with sealed ends, although actually the behavior of infinite tissue is of interest. Hence, to avoid boundary effects, a certain minimum fiber length should be simulated. The observation site where simulation results are examined, should maintain a minimum distance of several λ to the cable ends. In multidimensional tissue λ depends on the direction of propagation (λ transverse to the fibers is usually much smaller than along the fibers). This should be taken into account to find an optimum tissue size for a simulation.

4.2. *Monodomain models*

4.2.1. *One-dimensional fiber*

The one-dimensional cable equation is given by

$$\frac{1}{r_i + r_e} \frac{\partial V_m^2}{\partial x^2} = i_m + \frac{r_i}{r_i + r_e} i_s \tag{80}$$

where injection of the stimulation current in the intracellular space is assumed contrary to (41). Rewriting in terms of conductivities and with the transmembrane current density $\tilde{i}_m$ yields

$$\frac{g_i g_e}{g_i + g_e} \frac{\partial V_m^2}{\partial x^2} = 2\pi a \tilde{i}_m + \frac{g_e}{g_i + g_e} i_s \tag{81}$$

Conductivities g_i and g_e are given on a per unit length base. That is, $g_i = \sigma_i A_i$ and $g_e = \sigma_e A_e$ where $A_i = a^2\pi$ and $A_e = (b^2 - a^2)\pi$ (see Fig. 2).

Although it is not essential for monodomain models to refer g_i and g_e to the entire cross-section (and it is not common practice to do it neither), it may be convenient particularly with regard to bidomain models and the comparison of used parameters. Bidomain models are based on the idea of two continuous interpenetrating intra- and extracellular domains which are separated everywhere by a membrane of unspecified topology [38, 45]. Therefore it is required to refer all quantities to the entire cross section $A_t = A_i + A_e$ of a discrete element which entails an adjustment of the intrinsic conductivities. If we rewrite (81) in terms of intrinsic conductivities σ_i and σ_e we obtain

$$\frac{\sigma_i A_i \, \sigma_e A_e}{\sigma_i A_i + \sigma_e A_e} \frac{\partial V_m^2}{\partial z^2} = 2\pi a \tilde{i}_m + \frac{\sigma_e A_e}{\sigma_i A_i + \sigma_e A_e} i_s \tag{82}$$

We define the fractions f_i and f_e as

$$f_i = \frac{A_i}{A_t} \quad \text{and} \quad f_e = \frac{A_e}{A_t} \tag{83}$$

and spread out the conductances of the respective domains of the entire volume to obtain the adjusted intrinsic conductivities $\bar{\sigma}_i$ and $\bar{\sigma}_e$ as

$$\bar{\sigma}_i = \sigma_i f_i \quad \text{and} \quad \bar{\sigma}_e = \sigma_e f_e \tag{84}$$

generally referred to as effective intra- and extracellular conductivity. Substituting (83) into (82) yields

$$\frac{\sigma_i f_i \, \sigma_e f_e}{\sigma_i f_i + \sigma_e f_e} \frac{\partial V_m^2}{\partial x^2} = \frac{2\pi a}{A_t} \tilde{i}_m + \frac{\sigma_e f_e}{\sigma_i f_i + \sigma_e f_e} \frac{1}{A_t} i_s \tag{85}$$

If we further define the surface-to-volume ratio β as the ratio of the total membrane area to the total volume we obtain for a cylindrical fiber element of length δx with the cross sectional fraction f_i

$$\beta = \frac{2\pi a \delta x}{A_t \, \delta x} = \frac{2\pi a \, \delta x}{A_i / f_i \, \delta x} = \frac{2\pi a \, \delta x}{a^2 \pi \, \delta x / f_i} = \frac{2 f_i}{a} \tag{86}$$

This permits to rewrite (85) as

$$\frac{\bar{\sigma}_i \, \bar{\sigma}_e}{\bar{\sigma}_i + \bar{\sigma}_e} \frac{\partial V_m^2}{\partial x^2} = \beta \tilde{i}_m + \frac{\bar{\sigma}_e}{\bar{\sigma}_i + \bar{\sigma}_e} I_s \tag{87}$$

where $I_s = i_s / A_t$ is identified as a stimulus current per unit volume. A common assumption is that the fiber lies in an infinite, homogeneous conductive bath. Under these conditions the extracellular space can be assumed to be grounded ($\bar{\sigma}_e \approx \infty$) and (87) simplifies to

$$\bar{\sigma}_i \frac{\partial V_m^2}{\partial x^2} = \beta \tilde{i}_m + I_s \tag{88}$$

This is a monodomain representation of a one-dimensional fiber in which a single partial differential equation describes the current flow in the intracellular space.

4.2.2. *Multi-dimensional tissue*

The monodomain model can be extended to two and three dimensions. Multi-dimensional models allow to account for the anisotropy of cardiac tissue which gives rise to a faster conduction velocity along than across the fibers. A general form of the monodomain model is

$$\nabla \cdot (\mathbf{D_i} \nabla V_m) = \beta \tilde{i}_m - I_s \tag{89}$$

where $\mathbf{D_i}$ is the intracellular conductivity tensor in [mS/cm] and I_s a current source per unit volume in (μA/cm^3).

If we continue to consider cardiac tissue as a set of parallel fibers it is mathematically convenient to choose a coordinate system whose axes align with the principal axes of $\mathbf{D_i}$. For instance, for tissue with straight

non-rotated fiber, a Cartesian coordinate system is defined such that the axis x aligns with the fiber direction. Then the conductivity tensor is a diagonal matrix $\mathbf{D_i} = \mathrm{diag}(\sigma_{iL}, \sigma_{iT}, \sigma_{iT})$ with the conductivity σ_{iL} along and σ_{iT} across the fiber. Under these conditions the discretized Eq. (89) may be interpreted as a network of resistors on a regular grid [47] like shown in Fig. 4 for a two-dimensional monodomain model. An expression [10, 79] typically used for such a model is

$$\sigma_{iL}\frac{\partial^2 V_m}{\partial x^2} + \sigma_{iT}\frac{\partial^2 V_m}{\partial y^2} = \beta\tilde{i}_m + I_s = \beta\tilde{i}_{ion} + \beta\tilde{c}_m\frac{\partial V_m}{\partial t} + I_s \tag{90}$$

If the fibers are curved or rotated, $\mathbf{D_i}$ is a function of space and the coordinate axes cannot be chosen to diagonalize $\mathbf{D_i}$ everywhere. For instance, in the three-dimensional ventricular tissue the fiber orientation rotates slowly with tissue depth from the epicardium to the endocardium. Under these conditions the diffusion term in the most general case two-dimensional $\mathbf{D_i}(x, y)$ is represented as

$$\begin{aligned}\nabla\cdot(\mathbf{D_i}\nabla f) &= \frac{\partial}{\partial x}\left(\sigma_{xx}\frac{\partial f}{\partial x} + \sigma_{xy}\frac{\partial f}{\partial y}\right) + \frac{\partial}{\partial y}\left(\sigma_{xy}\frac{\partial f}{\partial x} + \sigma_{yy}\frac{\partial f}{\partial y}\right)\\ &= \sigma_{xx}\frac{\partial^2 f}{\partial x^2} + 2\sigma_{xy}\frac{\partial^2 f}{\partial x\partial y} + \sigma_{yy}\frac{\partial^2 f}{\partial y^2} + \left(\frac{\partial\sigma_{xx}}{\partial x} + \frac{\partial\sigma_{xy}}{\partial y}\right)\frac{\partial f}{\partial x}\\ &\quad + \left(\frac{\partial\sigma_{xy}}{\partial x} + \frac{\partial\sigma_{yy}}{\partial y}\right)\frac{\partial f}{\partial y}\end{aligned} \tag{91}$$

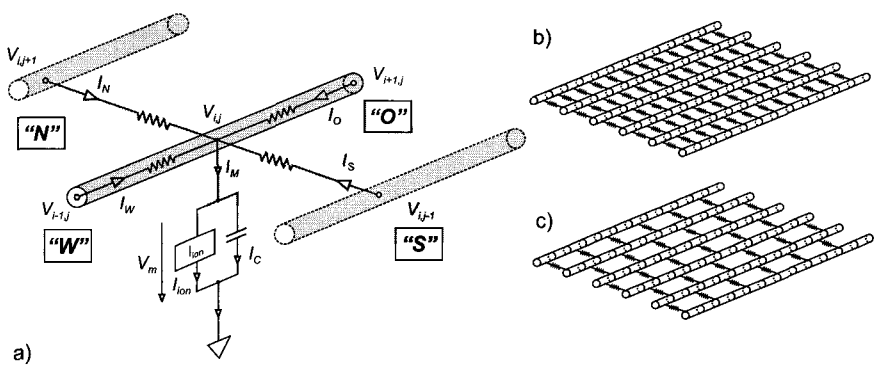


Fig. 4. (a) Cardiac tissue as a set of parallel fibers: Using a coordinate system aligned with the principal axes of the fibers allows a representation as a regular lattice of resistor S. Application of Kirchhoff's current law to the central node (i, j) by simply summing up the currents $I_M = I_N + I_E + I_S + I_W$ leads to the same equation as the spatial discretization of the diffusion part of monodomain equation. (b) Two-dimensional monodomain representation of a uniform continuous tissue model and (c) of a coupled cable tissue model.

4.2.3. *Discontinuous monodomain models*

From a macroscopic point of view, cardiac tissue can be considered as a continuum based on the assumption of syncytial behavior. At a microscopic size scale, however, this is definitely not the case. The intracellular region of a cardiac cell is connected with the intracellular space of the neighboring cells by low-ohmic gap junctions, i.e. the intracellular space is continuous in the sense that a moving ion does not have to pass to the extracellular space to get from one cell to another, but the distribution of electrical parameters is discontinuous. A typical value of a junctional resistance is approximately equal to that of an entire cell region in the range of 0.5–5 MΩ [14, 71].

One-dimensional monodomain models were developed to account for these discontinuities. The influence of the junctions was taken into account by subdividing a cell into several segments and adding the junctional resistances at the terminal segments of the cell where the gap junctions are assumed to be located [23, 42, 69, 103, 128]. Based on one-dimensional discontinuous cables two- and three-dimensional discontinuous monodomain models were built as well. Leon *et al.* developed a multidimensional discontinuous model by connecting one-dimensional cables laterally [59, 60]. The cables are allowed to be discontinuous and the lateral interconnections are placed in a staggered way so that no cable is connected with more than one neighboring cable at a certain junction site (see Fig. 4(c)). The arrangement of the lateral connections is a limitation of the computational technique which prevents the simulation of completely continuous tissue. This method was extended to three dimensions as well [121, 122].

Other studies started from the regular resistive network approach (Fig. 4(b)) and varied longitudinal and transverse resistances within certain bounds. Probably the most detailed approach of this kind was proposed by Spach *et al.* [113–115]. They constructed a 2D model of multiple cells based on an approximation of the naturally occurring variations in size and shape of isolated cardiac cells and the distribution of the cell-to-cell connections [126, 127] and adjusted the resistive grid to reflect both cell geometry and junctional coupling.

4.3. *Numerical solution of monodomain equations*

When modelling propagation of the cardiac excitation spread, simplifying assumptions of structural and functional details of the tissue. A computer model which takes into account all known details of cardiac tissue would exceed by far available computational resources. Therefore computer

models are tailored for specific problems. Macroscopic models are capable of simulating large pieces of tissue, but the spatial resolution of these models is coarse and the functional aspects of the membrane do not regard too much physiological details. On the other hand, microscopic models are more detailed concerning the ionic current and have a fine spatial resolution, but they are restricted to the simulation of small pieces of tissue. Today there is no model available equally well suited for all questions and the amount of detail necessary for a specific problem, should be carefully selected, since the computational complexity increases with the amount of details included in the model. This is particularly true for ionic current based models, whose complexity increased considerably within the last years. Such membrane models [19, 62, 65, 74, 91] permit the simulation of complicated electrophysiological phenomena [117] like early and delayed afterdepolarizations [55, 66, 106], effects of antiarrhythmic drugs, the dynamical tracking of ionic movements [75], ischemia [11, 34, 110] and other pathological conditions. Unfortunately, the price to pay for such a variety of physiological details is high. If we consider a monodomain model using a modern description of the membrane behavior like e.g. [65] up to 85% of the overall computational workload is spent on the integration of the ordinary differential equations of the kinetic model [90]. Particularly if multidimensional tissue with ionic current based models is considered, simulations are usually limited to phenomena occurring within the time frame of several milliseconds or seconds. The observation of various phenomena like fibrillation or ischemia require the simulation of much longer periods.

As a consequence of the computational costs of these models most three-dimensional models reported in the literature use membrane models with a reduced number of ionic currents [33] or non-physiological models like the FitzHugh-Nagumo-model [1, 36] with a small number of state variables and a slow membrane kinetic.

Efficient integration is rather hard to achieve owing to the very fast time scale and the short length scale of propagating action potentials. Physiological membrane kinetic models lead to an exceedingly stiff system of equations, a phenomenon which often occurs when some components in the solution change at a very fast time scale in comparison with other components and the overall time scale of the solution [52].

In the following a survey of the numerical methods reported in the research literature for action potential propagation in monodomain models will be presented. Although nowadays bidomain models are considered as the state of the art, understanding of numerical methods used for the integration of monodomain models is still of great importance for several

reasons: the monodomain model represents a special case of a bidomain model, only when intracellular and interstitial anisotropy ratios are different the bidomain model will yield qualitatively different results; methods for monodomain integration were developed first and bidomain methods evolved from those methods which proved to be efficient there; besides monodomain models are still frequently used, partly because a given problem may not require a bidomain formulation and partly because the computational costs of the monodomain are significantly less compared with a bidomain.

For the purpose of illustration action potential propagation in a two-dimensional sheet of tissue in the domain $\Omega = [0, a] \times [0, b]$ will be considered. The monodomain model should be represented by the following set of equations

$$\nabla \cdot (\mathbf{D_i} \nabla V(t, \mathbf{x})) = \beta \left[\tilde{c}_m \frac{\partial V(t, \mathbf{x})}{\partial t} + \tilde{i}_{ion}(V, m_n) \right] - I_s(t, \mathbf{x}) \tag{92}$$

$$\frac{dm_n}{dt} = f(V, m_n), \quad n = 1 \ldots N \tag{93}$$

where $V(t, \mathbf{x})$ is the membrane potential at time t and location $\mathbf{x} = (x, y)$, $\tilde{c}_m$ is the membrane capacitance per unit area, β is the surface-to-volume ratio, $\mathbf{D_i}$ the conductivity tensor, $\tilde{i}_{ion}$ the total ionic current per unit area and $I_s(t, \mathbf{x})$ a stimulus current per unit volume to initiate propagation. The variable m_n represents the gating variables whereby N depends on the used kinetic model. The initial values of $V(t_0)$ and $m_n(t_0)$ are typically computed by assuming steady-state conditions $dV/dt = dm_n/dt = 0$. Neumann boundary conditions are imposed at all boundaries by

$$\mathbf{n} \cdot (\mathbf{D_i} \nabla V) = 0 \tag{94}$$

4.3.1. *Spatial discretization of equations*

The majority of monodomain models has been implemented using finite difference schemes for the spatial discretization. The major drawback of this approach are difficulties regarding the modelling of complex geometries and realistic descriptions of the fiber direction. Since mostly regular domains were considered, the ease of implementation of finite differences was the crucial argument to prefer this method. There are just a handful studies reporting the use of other techniques. Finite elements have been used in [4, 98, 100, 101, 108, 109] and a finite volume method in [40, 41] to model a two-dimensional domain with curved boundaries.

The most common approach to discretize the problem has been the use of a static uniform grid. Since the appropriate degree of spatial resolution varies with time when solving for a propagating action potential, the simplest means of obtaining sufficient spatial resolution is the use of a uniform fine grid for the entire domain.

The spatial operators are mostly expressed in cartesian or cylindrical coordinates. The finite difference approximations to the Laplacian operator in cartesian coordinates typically use the centered, second order accurate approximation

$$\left.\frac{\partial^2 f}{\partial x^2}\right|_{x_i} \approx \frac{f_{i-1} - 2f_i + f_{i+1}}{\Delta x^2} \tag{95}$$

in the interior of the domain where f_{i-1}, f_i and f_{i+1} are grid points and Δx is the spatial discretization step. The boundaries of the domain are typically assumed as sealed ends. At the right boundary $x = a$ of the domain Ω an expression for the boundary condition $\partial f/\partial x = 0$ is obtained by using a ghost point x_{i+1} (which lies actually outside the domain, but substitution of the boundary condition into (95) at the boundary nodes will permit to let the ghost points disappear) with the first order accurate approximation

$$\left.\frac{\partial f}{\partial x}\right|_{x=a} \approx \frac{f_{i+1} - f_i}{\Delta x} \tag{96}$$

or the second order accurate approximation

$$\left.\frac{\partial f}{\partial x}\right|_{x=a} \approx \frac{f_{i+1} - f_{i-1}}{2\Delta x} \tag{97}$$

Approximation (96), although less accurate, is useful because it corresponds to the discrete electrical network model [56, 61, 97] which has been used in several simulation studies.

For the discretization of the diffusion term at the left hand side of (92) we assume that $\bar{\Omega}$ is defined as a set of nodes (x_i, y_j) with $\{x_i \,|\, x_i = i\Delta x, i = 0 \ldots N_x\}$ and $\{y_j \,|\, y_j = j\Delta y, j = 0 \ldots N_y\}$ representing a mesh on Ω with N_x cells along and N_y cells across the fibers, each cell of size $\Delta x \times \Delta y$ with $\Delta x = \frac{a}{N_x}$ and $\Delta y = \frac{b}{N_y}$. We will seek approximations of the solution V defined everywhere in Ω at the mesh points (x_i, y_i) of $\bar{\Omega}$ and will denote these values with $V_{i,j}$. If we rewrite the diffusion term, given in the invariant form in (92), in cartesian coordinates, assuming straight fibers and alignment of the fiber axis with x, we obtain

$$\nabla \cdot (\mathbf{D_i} \nabla V) = \sigma_{i_L} \frac{\partial^2 V}{\partial x^2} + \sigma_{i_T} \frac{\partial^2 V}{\partial y^2} \tag{98}$$

and for the boundary condition in (94)

$$\sigma_{i_L} \frac{\partial V}{\partial x}\bigg|_{x=0,a} = 0, \quad \sigma_{i_T} \frac{\partial V}{\partial y}\bigg|_{y=0,b} = 0 \tag{99}$$

where $\mathbf{D_i}$ is a diagonal matrix $\mathrm{diag}(\sigma_{i_L}, \sigma_{i_T})$ with σ_{i_L} designating the conductivity along the fibers and σ_{i_T} across. The differential operators can be approximated by finite differences given in (95) and (97). Imposing of (97) on (95) at the boundaries permits to get rid of the ghost points. This yields the following approximations

$$\sigma_{i_L} \frac{\partial^2 V}{\partial x^2}\bigg|_{i,j} \approx \sigma_{i_L} \frac{V_{i-1,j} - 2V_{i,j} + V_{i+1,j}}{\Delta x^2} \tag{100}$$

$$\sigma_{i_T} \frac{\partial^2 V}{\partial y^2}\bigg|_{i,j} \approx \sigma_{i_T} \frac{V_{i,j-1} - 2V_{i,j} + V_{i,j+1}}{\Delta y^2} \tag{101}$$

in the interior of Ω and

$$\sigma_{i_L} \frac{\partial^2 V}{\partial x^2}\bigg|_{x=0,y} \approx \sigma_{i_L} \frac{2(V_{i+1,j} - V_{i,j})}{\Delta x^2} \quad i = 0,\ j = 0 \ldots N_y \tag{102}$$

$$\sigma_{i_L} \frac{\partial^2 V}{\partial x^2}\bigg|_{x=a,y} \approx \sigma_{i_L} \frac{2(V_{i-1,j} - V_{i,j})}{\Delta x^2} \quad i = N_x,\ j = 0 \ldots N_y \tag{103}$$

$$\sigma_{i_T} \frac{\partial^2 V}{\partial y^2}\bigg|_{x,y=0} \approx \sigma_{i_T} \frac{2(V_{i,j+1} - V_{i,j})}{\Delta y^2} \quad i = 0 \ldots N_x,\ j = 0 \tag{104}$$

$$\sigma_{i_T} \frac{\partial^2 V}{\partial y^2}\bigg|_{x,y=b} \approx \sigma_{i_T} \frac{2(V_{i,j-1} - V_{i,j})}{\Delta y^2} \quad i = 0 \ldots N_x,\ j = N_y \tag{105}$$

at the boundaries of Ω.

Projection of the remaining variables $m_n, \tilde{i}_{ion}$ onto $\bar{\Omega}$ and arrangement in vector form permits to rewrite Eqs. (92)–(93) in a spatially discretized form

$$\frac{d\mathbf{v}}{dt} = \frac{1}{\beta \tilde{c}_m}\left[-\mathbf{G_i}\mathbf{v} - \beta \tilde{\mathbf{i}}_{\mathbf{ion}} + \mathbf{I_s}\right] \tag{106}$$

$$\frac{d\mathbf{m_n}}{dt} = f(\mathbf{m_n}, \mathbf{v}) \tag{107}$$

where the spatial derivatives are represented by means of the conductance matrix $\mathbf{G_i}$ as $\mathbf{G_i} \cdot \mathbf{v} \approx -\nabla \cdot (\mathbf{D_i} \nabla V)$. The vectors $\mathbf{v}, \mathbf{m_n}, \tilde{\mathbf{i}}_{\mathbf{ion}}$ and $\mathbf{I_s}$ are of length $L = N_x \cdot N_y$ (the number of mesh points), the conductance matrix $\mathbf{G_i}$ is $L \times L$. $\mathbf{G_i}$ is sparse and pentadiagonal in the two-dimensional case.

4.3.2. *Monodomain integration methods*

Due to the computational expense of integrating ionic models the use of efficient and inexpensive methods becomes important. One of the simplest methods is the explicit forward Euler method. Equations (106)–(107) are integrated with the following scheme:

$$\mathbf{v}^{k+1} = \mathbf{v}^k - \frac{\Delta t}{\beta \tilde{c}_m}\left[\mathbf{G_i}\,\mathbf{v}^k + \tilde{\mathbf{i}}_{\mathbf{ion}}\left(\mathbf{m}_\mathbf{n}^k, \mathbf{v}^k\right) - \mathbf{I}_\mathbf{s}^k\right] \tag{108}$$

$$\mathbf{m}_\mathbf{n}^{k+1} = \mathbf{m}_\mathbf{n}^k + \Delta t\, f(\mathbf{m}_\mathbf{n}^k, \mathbf{v}^k) \tag{109}$$

where k denotes the time instant $t = k\Delta t$. Although only first order accurate, this method has often been used [8, 9, 29, 30, 86, 89] owing to its simplicity and the ease with which vectorizable code may be written [86]. A severe restriction of this method is its limited stability. Explicit methods for parabolic partial differential equations impose a constraint on the size of the time step for numerical stability. A stability constraint using the mesh ratio, appropriate for the diffusion equation, given by

$$\frac{\sigma \Delta t}{\beta \tilde{c}_m \Delta x^2} \leq \frac{1}{2d} \tag{110}$$

where d is the number of space dimensions [16, 73], has been found helpful for the monodomain equations as well [85, 112]. An explanation why the stability constraints of diffusion and monodomain equation are linked is found in [52]. According to (110) the time step Δt has to be kept very small to prevent the solution from becoming unstable regardless of whether there are transients in the solution whose accurate solution would require such a small time step.

An alternative to low-cost integration schemes like forward Euler is the use of implicit or semi-implicit methods which reduce the dependence of the time step on the spatial resolution substantially. Most studies report the use of a semi-implicit method referred to in the electrophysiological literature as Crank–Nicholson method. The linear term $\mathbf{G_i} \cdot \mathbf{v}$ is treated implicitly as the weighted average (with weighting factor $1/2$) of the spatial derivatives at the instants k and $k+1$ like in the Crank–Nicholson method [20], the nonlinear term $\tilde{\mathbf{i}}_{\mathbf{ion}}(\mathbf{m_n}, \mathbf{v})$, however, is treated explicitly. Hence Eq. (106) becomes

$$\left[\frac{1}{2}\mathbf{G_i} + \frac{\Delta t}{\beta \tilde{c}_m}\mathbf{I}\right]\mathbf{v}^{k+1} = -\left[\frac{1}{2}\mathbf{G_i} - \frac{\Delta t}{\beta \tilde{c}_m}\mathbf{I}\right]\mathbf{v}^k - \tilde{\mathbf{i}}_{\mathbf{ion}}\left(\mathbf{m}_\mathbf{n}^k, \mathbf{m}_\mathbf{n}^k\right) - \mathbf{I}_\mathbf{s}^k. \tag{111}$$

The stability of this semi-implicit method does not depend on the spatial resolution any longer and numerical stability is much better compared

to forward Euler. The cost of this increased stability is that an algebraic system has to be solved at each integration time step. Direct methods [27, 39] factoring the coefficient matrix at the left-hand side of (111) once and using the factors at each time step to generate the result, have been used. For one-dimensional domains the coefficient matrix is tridiagonal and banded methods are optimal. For higher dimensions the equations should first be reordered [105] to reduce the cost of both the factor and solve steps. Beyond a certain number of nodes (i.e. large two-dimensional or three dimensional models) the storage costs of a direct method are prohibitive and iterative methods have to be applied. Classical iterative methods like Gauss–Seidel, Jacobi or SOR have been applied to that problem as well as semi-iterative Krylov subspace methods. These methods usually require a preconditioner to be effective such as diagonal preconditioning or incomplete Cholesky factorization [105]. For multidimensional domains, iterative methods [67, 68, 97, 113, 114] or implicit treatment in one dimension by the ADI method [7, 31, 32, 35, 56, 61, 72, 78, 92–94, 121, 123] have generally been preferred.

The integration step of the state variables $\mathbf{m_n}$ is performed separately using an explicit method. Several researchers used a method developed by Rush and Larsen [104]. Their technique is based on the observation that the equations for the gating variables can be viewed as linear equations, assuming that the voltage-dependent parameters used to update them change slowly. Other approaches to improve the efficiency of the integration include the variation of the time step depending on the time scale of the variable being integrated [121] and the use of lookup tables to store voltage-dependent variables. It was demonstrated that with a sufficiently fine resolution of the lookup tables the resulting deviations are negligible and the computation time is just a third compared to the direct calculations of the coefficients α and β which involve the costly evaluation of several exponential terms [113].

In contrast to the common usage of semi-implicit methods, fully implicit methods (i.e. those that are implicit in both $\mathbf{v}$ and $\mathbf{m_n}$) have rarely been used. Cooley and Dodge [18] used the Trapezoidal Rule method in one of the earliest computer simulations of a propagating action potential and solved the resulting equation system using the Gauss–Seidel method. Mascagni [70] applied the backward Euler method to the one-dimensional problem with the Hodgkin–Huxley membrane. Hooke developed a fully implicit algorithm based on a Trapezoidal and Backward Euler integration with a nonlinear Newton solver [52, 88]. This method, second order accurate in time, can be augmented for variable time stepping based on a rigorous error estimate.

4.3.3. *Advanced techniques*

As stated above high temporal and spatial resolutions are required to resolve membrane depolarization with accurate upstroke and wave front conduction velocities [57, 112]. Solution times for sufficient cycles of electrical activity and the corresponding memory requirements are the main constraints to problem size. Solution time is proportional to the number of mesh points $L = N_x \cdot N_y \cdot N_z$ and the number of time steps $T/\Delta t$. Memory requirements are proportional to L. Since desktop processor speeds have increased dramatically over the last decade such simulations might be carried out on desktop computers up to a certain level of computational workload, beyond this, particularly if three-dimensional tissue should be simulated, high performance computing environments are required.

Sophisticated numerical methods that address these constraints include adaptivity [3, 15, 95, 121], iterative linear system solution [53] and table lookup to accelerate gating variable evaluation [25, 113, 120, 121]. Parallel computing techniques become more and more important and have been applied with success to large scale monodomain problems on shared memory supercomputers [17, 35, 86]. Recently the use of a cluster of workstations as a practical alternative to supercomputers was suggested to distribute calculation and data storage among several machines connected by a local area network [90].

Most approaches try to reduce the computational costs using adaptivity in space, in time or in both based on the following considerations: the fast components within the activated regions of the tissue determine the smallest possible time step, although the major part of the tissue is not in active state and would allow integration with much bigger time steps. Standard algorithms applied to this problem advance all the nodes of the domain with the same small Δt, whose size is limited by the time scale of activation. Hence, adjusting Δt locally is a key factor in improving computational efficiency, particularly since most of the computing time is spent calculating the reaction term which is a purely local function without spatial dependencies. A further reduction of computational costs can be achieved by dynamically adapting the spatial resolution of the mesh with a very coarse grid in regions of quiescence and a fine grid around the activated regions.

Barr and Plonsey [83] describe a physiologically based technique known as dynamically tracking of the active region. Again, a uniform grid is assumed throughout the domain, but at each integration step calculations are performed only on a subset of the grid points which includes only cells

located near the active wavefront. The choice of these cells is based on heuristic rules linked to the presumed shape of the depolarization waveform. The solution is found with substantially less computational efforts compared to conventional techniques and propagation velocity and the shape of the transthreshold waveform is preserved, but distortions of the initial electrotonic part of the waveform occurs. This method has proven to be efficient in large-scale simulation [86–88], however, it is only useful when gross features of propagation are of interest.

A new method to control the time step locally was presented recently by Quan *et al.* [95]. The integration procedure consists of two stages: integration over the whole domain and integration over subdomains. In the first stage between two consecutive integrations over the whole domain the implicit integration method of Cooley–Dodge [18] with modified alternating-direction-implicit (ADI) is used. In the second stage the model is spatially decomposed into many subdomains and an explicit Euler integration method is applied. Since Δt is defined locally, a priority queue is used to store and order next update time for each subdomain, the subdomain with the earliest update time is given the highest priority and advanced first. Domain decomposition and priority queue integration allow a large integration time step for nonactive subdomains. A performance improvement between 3 and 17 has been reported for the integration of a two-dimensional domain with the Luo–Rudy-II kinetic model.

The coupled cable model developed by Leon *et al.* [60] uses a clever numerical algorithm that is amenable to a parallel implementation. Particularly, if the number of lateral connections between neighboring cables is low, this method is highly efficient and has been used successfully to simulate three-dimensional tissue with fiber rotation. It is not clear if the algorithm is directly extensible to a true continuous structure (with lateral connections at all cells) or twisting fibers without performance degradation.

Another adaptive approach based on this technique was presented by Vigmond *et al.* [121]. The identification of subdomains where update of variables with larger time steps is possible was done by using both temporal and spatial methods. Integration of the gating variables was optimized by exploiting the fact that different gates respond on different time scales. Fast responding gates were updated more frequently than slow responding gates. A performance improvement of 2 was reported in this study.

Until recently adaptive methods were implemented by varying either the spatial or temporal resolution, but not both, locally and dynamically. A new approach involving a dynamical adaption of temporal as well as spatial resolution was proposed by Cherry *et al.* [6, 15]. The basic idea of this

adaptive mesh refinement approach is to focus the computational efforts on areas with large spatial and temporal gradients. It has been demonstrated that this method is able to reduce memory and computation time requirements of a simulation of complex cardiac dynamics compared to a uniform space-time mesh by a factor of 5 to 15.

Otani and Allexandre proposed a different approach based on a modified backward Euler method that allows unsynchronized time steps across the domain. Neighboring values at new time steps are extrapolated linearly in time from earlier values when they are not known. An increase of the maximum stable time step by around three orders of magnitude was reported [76, 77].

Qu *et al.* [93] suggested an advanced method for solving reaction-diffusion-type equations for cardiac conduction using operator splitting [116] and adaptive time step methods. Operator splitting allows to separate diffusion and reaction term. The advantage is that in one step the simplest possible diffusion equation has to be solved and in a further step the integration of the ordinary differential equations can be done adaptively.

Veronese and Othmer have formulated an efficient algorithm for monodomain and bidomain problems that uses an alternating direction implicit (ADI) step with a Multigrid step [119]. This method has been used on extremely large scale, three-dimensional problems, but may be limited to parallel fibers.

5. Recovery of Extracellular Potentials and Fields

Monodomain simulations do not provide extracellular potentials immediately since $\Phi_e \approx 0$ is assumed. Nevertheless it is common practice to use the transmembrane currents obtained from monodomain simulations to compute Φ_e. This procedure of first ignoring the effect of extracellular potentials to compute the excitation spread in the tissue, and then recover non-zero potentials Φ_e based on these data seems to be paradox, but it is justified by the fact that ignoring Φ_e introduces negligible errors in shape and velocity of action potential propagation [43, 46].

Although the approximation $\Phi_e \approx 0$ is often satisfactory it is not applicable under all circumstances. Then it is more adequate to use the somewhat more general bidomain model which accounts explicitly for current flow in both extracellular and intracellular spaces. A typical problem falling in this category is, for instance, the stimulation of cardiac tissue with external currents (defibrillation). Unfortunately the step from a monodomain to a putatively similar and closely related bidomain model is accompanied by

a significant increase in computational costs. Thus whenever there is no need to account explicitly for the extracellular flow, except in its contribution to the overall conduction velocity, a treatment of the tissue as a computationally more tractable monodomain [47] is preferred.

The following considerations should elucidate the procedure of extracellular field recovery from data obtained with monodomain simulations. A two-dimensional sheet of cardiac cells in contact with an extensive homogeneous isotropic bathing fluid (volume conductor) will be assumed. Under these circumstances the extracellular resistance is small and the assumption of a grounded extracellular region is quite well satisfied.

5.1. *Source-field concept*

For the computation of extracellular potentials and fields from monodomain simulations it is necessary to establish a relation between the electrical activation of a fiber (the source) and the concomitant volume conductor field [84]. For the basic understanding of this concept it is helpful to regard first the source-field relation of simple point sources. This will facilitate the interpretation of the more sophisticated relation of fields evoked by the sources of a cylindrical fiber.

It has been shown that the current flow field in an electrophysiological volume conductor is quasi-static [12, 81]. This permits to derive the electric field $\mathbf{E}$ as the gradient of a scalar potential Φ. Consequently, the electric field can be expressed as

$$\mathbf{E} = -\nabla\Phi \tag{112}$$

and the current density, which is related to $\mathbf{E}$ by Ohm's law, as

$$\mathbf{J} = \sigma_e \mathbf{E} \tag{113}$$

Assuming a point source of strength I_0, lying in an uniform conducting medium of infinite extent of conductivity σ_e, the radial current density results in

$$\mathbf{J} = \frac{I_0}{4\pi\sigma_e r}\,\mathbf{e_r} \tag{114}$$

as a consequence of symmetry, where $\mathbf{e_r}$ is a radial unit vector. Using (112) we find the associated potential field of a single point or monopole source at site $\mathbf{x_f}$ as

$$\Phi_m(\mathbf{x_f}) = \frac{1}{4\pi\sigma_e}\, I_0\, \frac{1}{r} \tag{115}$$

where r is the distance from source point $\mathbf{x}_s$ to field point $\mathbf{x}_f$. A further source configuration of interest is a point source and a point sink of equal strength I_0 very close together, separated by the displacement $\mathbf{d}$, where $I_0 \to \infty$ and $d \to 0$ such that their product $p = I_0\, d$ remains finite. Such a source is known as a dipole source $\mathbf{p} = p\,\mathbf{e_d}$ where $\mathbf{e_d}$ is the unit vector of the dipole axis and $p = I_0\, d$ the magnitude of the dipole source strength. A mathematical expression can be found by superposition of two monopole sources located at $\mathbf{x}_s$ and $\mathbf{x}_s + \mathbf{d}$ using (115). This is most conveniently evaluated by taking the directional derivative of (115), since $\Phi(\mathbf{x}_f + \mathbf{d}) - \Phi(\mathbf{x}_f) = \Phi(\mathbf{x}_f) + \nabla\Phi \cdot \mathbf{d} - \Phi(\mathbf{x}_f)$ with $\nabla\Phi \cdot \mathbf{d} = d\,(\partial\Phi/\partial d)$. Consequently, the potential field of a dipole source may be written as

$$\Phi_d(\mathbf{x}_f) = \mathbf{d} \cdot \nabla\Phi_m = \frac{1}{4\pi\sigma_e}\,(I_0 d\,\mathbf{e_d}) \cdot \nabla\left(\frac{1}{r}\right) = \frac{1}{4\pi\sigma_e}\,\mathbf{p} \cdot \mathbf{e_r}\left(\frac{1}{r^2}\right) \tag{116}$$

5.2. *Volume conductor fields of cylindrical fibers*

If we consider again a single cell of cylindrical geometry, infinitely long and axially symmetric, which is immersed in a uniform medium of conductivity σ_e, the extracellular potential is given by

$$\Phi(\mathbf{x}_f) = \frac{1}{4\pi\sigma_e} \int_\Gamma \left[\sigma_e \Phi_e(x) - \sigma_i \Phi_i(x)\right] \nabla\left(\frac{1}{r}\right) d\,\mathbf{\Gamma} \tag{117}$$

where r is the distance between the field point $\mathbf{x}_f$ and a source element located at the surface Γ of the cylinder and Φ_i and Φ_e are the intracellular and extracellular surface potentials [82, 84]. Assuming that the potentials $\Phi_i(x)$ and $\Phi_e(x)$, defined at the membrane, take on the same value throughout the entire cross-section, permits to apply the divergence theorem to (117) and integrate over the fiber volume V. This results in

$$\Phi(\mathbf{x}_f) = \frac{1}{4\pi\sigma_e} \int_V \nabla \cdot \left[\left[\sigma_e \Phi_e(x) - \sigma_i \Phi_i(x)\right] \nabla\left(\frac{1}{r}\right)\right] dV \tag{118}$$

which is equivalent to

$$\Phi(\mathbf{x}_f) = \frac{1}{4\pi\sigma_e} \int_V \nabla\left[\sigma_e \Phi_e(x) - \sigma_i \Phi_i(x)\right] \cdot \nabla\left(\frac{1}{r}\right) dV \tag{119}$$

since $\mathbf{x}_f$ lies outside the cell, hence r cannot become zero and $\nabla^2(1/r) = 0$ holds. Taking into account that the dot product in (119) will be zero for all components apart from x we are allowed to further simplify the integrand to

$$\Phi(\mathbf{x}_f) = \frac{1}{4\pi\sigma_e} \int_A \int_x \frac{\partial\left[\sigma_e \Phi_e(x) - \sigma_i \Phi_i(x)\right]}{\partial x} \frac{\partial 1/r}{\partial x}\, dx\, dA \tag{120}$$

where dA is an element of the cross-sectional area and $dV = dA\,dx$ represents a volume element. Integration by parts of (120) yields the expression

$$\Phi(\mathbf{x}_f) = -\frac{1}{4\pi\sigma_e}\int_A\int_x \frac{\partial^2\left(\sigma_e\Phi_e(x) - \sigma_i\Phi_i(x)\right)}{\partial x^2}\frac{1}{r}\,dx\,dA \tag{121}$$

for the potential field, since the integrated parts, evaluated at $x = \pm\infty$, are zero (the membrane is assumed to be in resting state there).

Comparison with expressions (115)–(116) obtained for monopole and dipole point sources permits the following interpretation of equations (120) and (121): in (120) the field arises from a source that consists of stacks of double layer disks of thickness dx oriented in the positive x-direction. The dipole strength of one such disk is $-a^2\pi\,dx\partial(\sigma_e\Phi_e - \sigma_i\Phi_i)/\partial x$, where the term $-\partial(\sigma_e\Phi_e - \sigma_i\Phi_i)/\partial x$ is identified as a volume density function varying with x (see Fig. 5I(c)). Comparison of (115) and (121) reveals that (121) is a single layer representation of the same source, where the monopole source strength of a disk is $-a^2\pi\,dx\,\partial^2[\sigma_e\Phi_e - \sigma_i\Phi_i]/\partial x^2$, and the volume density function is given by $\partial^2[\sigma_e\Phi_e - \sigma_i\Phi_i]/\partial x^2$ (see Fig. 5I(d)). These disk sources are not real sources, but they are equivalent in the sense that their evaluation yields the same field outside the fiber like expression (117), where the sources are assumed to be located at the membrane only. Both representation (120) and (121) are fully equivalent [63, 82].

If the field point $\mathbf{x}_f$ is located sufficiently far from the fiber in relation to the fiber radius a, the function $1/r$ in (120) and (121) can be considered essentially as constant over the cross-sectional area. Using the monodomain approximation $\Phi_e \approx 0$ (hence $V_m = \Phi_i - \Phi_e \cong \Phi_i$), Eqs. (120) and (121) simplify to

$$\Phi(\mathbf{x}_f) = -\frac{1}{4\pi\sigma_e}\int_x a^2\pi\sigma_i\frac{\partial V_m}{\partial x}\frac{\partial(1/r)}{\partial x}\,dx \tag{122}$$

$$\Phi(\mathbf{x}_f) = \frac{1}{4\pi\sigma_e}\int_x a^2\pi\sigma_i\frac{\partial^2\Phi_i}{\partial x^2}\frac{1}{r}\,dx \tag{123}$$

Substitution of expression (42) for the transmembrane current per unit length, $i_m = a^2\pi\sigma_i\,\partial^2\Phi_i/\partial x^2$, permits to rewrite (123) as

$$\Phi(\mathbf{x}_f) = \frac{1}{4\pi\sigma_e}\int_x \frac{i_m}{r}\,dx \tag{124}$$

The assumption of a constant $1/r$ over the cross-section is equivalent to assume that all the sources are concentrated at the axis of the cylindrical fiber. Therefore the sources in Eqs. (122)–(124) are line densities and referred to as a "line-source" models.

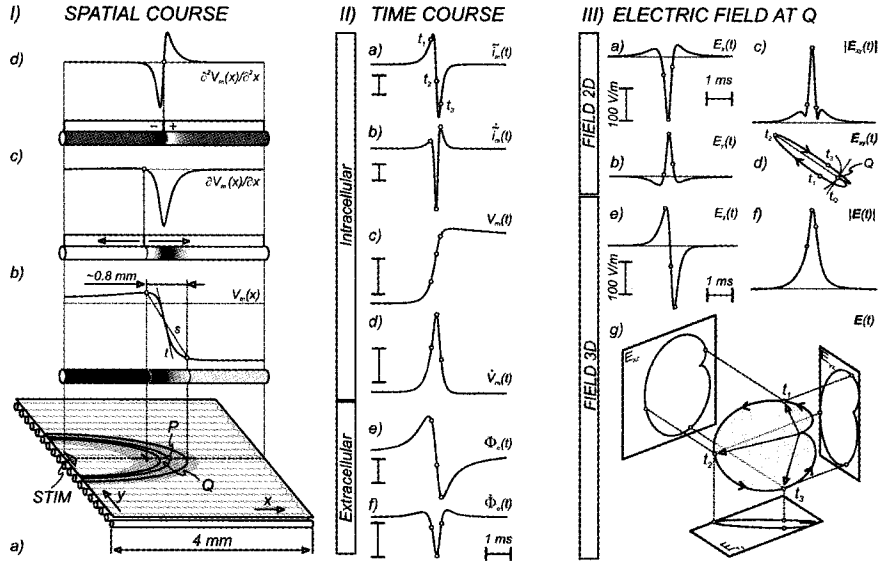


Fig. 5. (I) Simulation of action potential propagation with a monodomain model: (a) Monodomain model with observation sites P and Q. Propagation is initiated at site *STIM*. The central fiber is marked with a dashed line. (b) Spatial course of V_m along the central fiber. (c) Distribution of dipole source density (arrows indicate dipole orientation) and (d) monopole source density (+/− indicates source polarity). (II) Time course of intra- and extracellular signals at Q: Instants t_1, t_2 and t_3 mark the maxima and minima of $\ddot{\tilde{i}}_m$. (a) transmembrane current $\tilde{i}_m$ and (b) derivative $\dot{\tilde{i}}_m$, (c) transmembrane potential V_m and (d) derivative $\dot{V}_m$, (e) extracellular potential Φ_e and (f) derivative $\dot{\Phi}_e$. Scales are (a) $100\,\mu\mathrm{A/cm^2}$, (b) $1000\,\mu\mathrm{A/cm^2/ms}$, (c) 50 mV, (d) $100\,V/s$, (e) 10 mV, and (f) $100\,V/s$. (III) The electric field at Q: (a)–(d) The electric field $\mathbf{E}_{xy}$ in a plane parallel to the tissue surface. (a) time course of the components E_x and (b) E_y, and (c) the magnitude of $\mathbf{E}_{xy}$. (d) The maximum vector $\mathbf{E}_{xy}$ occurs at the instant t_2 and points opposite the direction of propagation (the vector is orthogonal to the tangent t_Q of the local isochrone (curved line) at site Q). (e)–(g) The three-dimensional electric field at Q: (e) Component E_z of the field orthogonal to the tissue surface. (f) Magnitude of the field $\mathbf{E}$. (g) Temporal evolution of a three-dimensional vector loop of $\mathbf{E}$.

If we assume a set of cylindrical fibers arranged in parallel, each fiber of radius a with a current density per unit area of $\tilde{i}_m = i_m/2\pi a$ at the site $\mathbf{x}_\mathrm{s}$, the extracellular potential at the field point $\mathbf{x}_\mathrm{f}$ is found with

$$\Phi(\mathbf{x}_\mathrm{f}) = \frac{1}{4\pi\sigma_e}\int_\Gamma \frac{\tilde{i}_m(\mathbf{x}_\mathrm{s})}{|\mathbf{r}_\mathrm{sf}|}\,d\Gamma \tag{125}$$

where $\mathbf{x}_\mathrm{s}$ designate the source point, $\mathbf{r}_\mathrm{sf} = \mathbf{x}_\mathrm{f} - \mathbf{x}_\mathrm{s}$ the vector from source to field point and $d\Gamma = 2a\pi dx$ is the surface area of one fiber element of

length dx. According to (112) the electric field $\mathbf{E}$ at $\mathbf{x}_f$ can be expressed by taking the negative gradient of Φ_e. Applying the gradient to (125) we obtain

$$\mathbf{E}(\mathbf{x}_f) = -\nabla\Phi_e(\mathbf{x}_f) = \frac{1}{4\pi\sigma_e}\int_\Gamma \frac{\tilde{i}_m(\mathbf{x}_s)\mathbf{r}_{sf}}{|\mathbf{r}_{sf}|^3}\, d\Gamma \tag{126}$$

where ∇ operates at the field coordinates.

6. Volume Conductor Potentials and Fields during Depolarization

This final section serves to demonstrate methods presented in this paper by means of a simulation example. Action potential propagation in a small piece of tissue will be simulated using a monodomain computer model. Extracellular potentials and fields will be recovered from transmembrane current data obtained from the monodomain model. Basic properties of intra- and extracellular signals will be discussed in terms of their relevance for electrophysiological measurements.

6.1. *Two-dimensional tissue model*

For the analysis of extracellular potentials and fields during depolarization in the immediate vicinity of a cardiac tissue surface action potential propagation was simulated using a two-dimensional monodomain computer model representing a thin sheet of cardiac tissue of size $4 \times 4\,\mathrm{mm}^2$ (see Fig. 5I(a)). The sheet was assembled with a set of cable-like cylindrical elements (radius $a = 6\,\mu\mathrm{m}$) arranged in parallel with a center-to-center distance of $15\,\mu\mathrm{m}$. Cables were transversely connected by a network of resistances. The dynamic membrane behavior was described by a capacitance in parallel with the ionic currents corresponding to the Luo-Rudy-I model [64]. The model parameters according to Eq. (90) were chosen as follows: anisotropy in conduction was represented by different intracellular conductances along x (longitudinal to fibers) with $\sigma_{iL} = 4\,\mathrm{mS/cm}$, and along y (transversal to fibers) with $\sigma_{iT} = 0.44\,\mathrm{mS/cm}$, the surface-to-volume ratio was set to $\beta = 2/a$ and the specific membrane capacitance was chosen with $\tilde{c}_m = 1\,\mu\mathrm{F/cm}^2$.

The spatial discretization steps were set to $\Delta x = \Delta y = 15\,\mu\mathrm{m}$, the temporal integration step to $\Delta t = 2\,\mu\mathrm{s}$. Boundaries of the sheet were considered to be sealed ends. Equation (90) was discretized using a semi-implicit scheme. The ordinary differential equations of the ionic currents

were solved with an Euler Predictor-Corrector method. The equation system was reordered using the reverse Cuthill–McKee method, preconditioned with incomplete LU-factorization and solved with a conjugate gradient method. The model was implemented with Matlab, all simulations were carried out on a Linux-PC (Dual Pentium II, $2 \times 850\,\mathrm{MHz}$, 2 GB RAM).

Action potential propagation was initiated by pacing five central elements at the left edge of the sheet. Propagation velocity of the elliptic wavefront at the center of the sheet was 0.68 m/s with a longitudinal-to-transverse velocity ratio of $\theta_L/\theta_T \approx 3/1$.

The extracellular potential and field was computed at site Q (see Fig. 5I(a)) at a distance of $30\,\mu\mathrm{m}$ from the tissue surface by evaluating Eqs. (125) and (126).

6.2. *Spatial source distribution at the central fiber*

The following matter relates to the initial phase (depolarization) of the action potential and its propagation. The spatial course of the transmembrane voltage V_m and the volume source density functions $\partial V_m/\partial x$ and $\partial^2 V_m/\partial x^2$ were examined along the central fiber at the instant of local activation at site P in the center of the tissue (see Fig. 5I(a)).

The action potential propagates in the positive x direction, therefore we have activated tissue with $V_m \approx +20\,\mathrm{mV}$ at the left side of the upstroke and resting condition with $V_m \approx -85\,\mathrm{mV}$ at the right side. The spatial course of the depolarization of $V_m(x)$ is monophasic. The distance between the isopotential contours of $V_m = +20\,\mathrm{mV}$ and $V_m = -80\,\mathrm{mV}$ is about 0.8 mm on the central fiber (see Fig. 5I(b)). For uniform propagation the space-time behavior of $V_m(x,t)$ satisfies the wave equation $V_m(x,t) = V_m(x-\theta t)$ where θ is the propagation velocity. Consequently, the spatial course $V_m(x)$ is a scaled and right-left reversed image of $V_m(t)$ (compare Fig. 5I(b) and 5II(c)).

As discussed earlier the equivalent double layer source density is proportional to $-\partial V_m/\partial x$ whereas the single-layer source density is proportional to $\partial^2 V_m/\partial x^2$. Both functions are shown in Fig. 5I(c) and (d).

Since the central fiber (marked in Fig. 5I(a) with a dashed line) represents a symmetry axis in the given arrangement, the derivative $\partial^2 V_m/\partial y^2$ vanishes there and the transmembrane current $\tilde{i}_m$ is proportional to $\partial^2 V_m/\partial x^2$ (see Eq. (90)). Taking into account that the time course of $\tilde{i}_m(t)$ is a right-left reversed image of $\tilde{i}_m(x)$ allows a qualitative verification of this proportionality by comparing the waveforms of the signal $\tilde{i}_m(t)$ at site Q (Fig. 5II(a)) and $\tilde{i}_m(x) \sim \partial^2 V_m/\partial x^2$ along the central fiber (Fig. 5I(d)).

6.3. *Time course of intra- and extracellular signals*

For the examination of the cardiac excitation spread it is required to determine the activation pattern with electrophysiological measurements, that is to find out when activation takes place at a certain site and how fast and in which direction the wavefront propagates. A standard procedure is to determine local activation times (LAT) simultaneously at a given set of recording sites and construct isochrone maps from these time markers. Each isochrone corresponds to a local activation time, the direction of propagation is orthogonal to the isochrone and the distance between two neighboring isochrones is inversely related to the conduction velocity (dense isochrones correspond to zones of slow propagation and vice versa).

The most accurate determination of the instant of activation is achieved with intracellular measurements. Quantities measurable during action potential propagation are the intracellular potential Φ_i (cells are impaled with microelectrodes) or the transmembrane potential (measurements with optical methods which give a relative measure of V_m). Extracellular measurements reflect the activation sequence as well, but deviations from the actual intracellular activation pattern will occur for several reasons: if conduction is discontinuous due to inhomogeneities of the tissue like for instance ischemic zones or conduction obstacles like connective tissue, propagation delays will occur which are not reflected in the same way in the extracellular space. Furthermore, due to the integration effect of the volume conductor (all the sources of the tissue contribute to the extracellular signal, but to those sources which are closer to the recording site is given more weight by virtue of the factor $1/r$) signals are smoothed out somewhat (compare Fig. 5II(a) and 5II(e)).

Since in clinical routine intracardially only extracellular measurements are possible (by means of catheters), it is important to have a reliable marker for the local activation time. The most common practice to determine LAT is by means of unipolar measurements of $\dot{\Phi}_e$ (Fig. 5II(f)) or bipolar measurements of voltage differences of two closely spaced electrodes yielding signals similar to those in Fig. 5III(a)–(b). In the uniform case the temporal coincidence of the instants of maximum respectively minimum of the derivatives is very close, a property which is routinely exploited to determine LAT. This is shown at site Q in the lateral part of the elliptic wavefront (see Fig. 5I(a)). The time course of the intracellular signals V_m and $\tilde{i}_m$, of the extracellular signal Φ_e and of the corresponding derivatives $\dot{V}_m$, $\dot{\tilde{i}}_m$ and $\dot{\Phi}_e$ were computed. A very close coincidence of the maximum

derivative of V_m with the minima of the derivatives of $\tilde{i}_m$ and Φ_e is observed (see Fig. 5II(b), (d), (e)) [112].

6.4. *The electric field evoked by an elliptic wavefront*

The electric field near the surface of cardiac tissue is a three-dimensional phenomenon which may be represented as a vector. The field vector $\mathbf{E}$ varies magnitude and direction during depolarization. The trajectory described by the tip of the vector is referred to as vector loop. At a given instant t the electric field vector $\mathbf{E}$ is accurately determined by its components, that is $\mathbf{E}(t) = [E_x(t), E_y(t), E_z(t)]^T$, where E_x, E_y and E_z are the projections of $\mathbf{E}$ onto the cartesian coordinate axes defined in Fig. 2(a).

In contrast to measurements of uni- and bipolar extracellular signals, measurements of the electric field are rarely used. This may be partly due to the somewhat more difficult interpretation of vector representations and partly due to the considerable technical efforts required for such measurements. The two-dimensional electric field in a plane parallel to the surface of the tissue can be measured with a square arrangement of four tightly spaced electrodes, very small inter-electrode distances of less than 100 μm and high sampling rates of at least 50 kHz are required [80]. This is explained as follows: like stated in (112), the electric field $\mathbf{E}$ corresponds to the gradient of the potential. In the one-dimensional case the gradient is given by $\partial\Phi/\partial x$, graphically represented by the slope of the tangent t (Fig. 5I(b)). Measurements are based on the approximation of the tangent t by a secant s like $(\Phi(x) - \Phi(x + dx))/dx$. From Fig. 5I(b) it is evident that for the accurate determination of the tangent in the steep part of $V_m(x)$ a spatial sampling interval much smaller than the distance of 0.8 mm between the isopotential contours of -80 and $+20$ mV is required. The high required sampling rate is explained in a similar manner. The elapsed time from t_1 to t_3 is just $\approx 400\,\mu$s and the movement of the tip of the vector is extremely fast (Fig. 5III(d)). Thus a too low sampling frequency would cut off the tip of the vector loop.

In Fig. 5III(a)–(b) the time course of the two-dimensional field components E_x and E_y are shown, in 5III(c)–(d) the magnitude of $|\mathbf{E}_{xy}|$ and the vector loop $\mathbf{E}_{xy}$. The instant of the maximum field of $\mathbf{E}_{xy}$ coincides closely with the instant of local activation t_2 in the tissue. A further interesting property is that the maximum field vector $\mathbf{E}_{xy}$ points opposite the local direction of propagation. This is illustrated in Fig. 5III(d) for the vector loop at site Q. The vector $\mathbf{E}_{xy}(t_2)$ is orthogonal to the tangent t_Q of the local isochrone (curved line) at site Q.

The three-dimensional field **E** is shown in 5III(e)–(g). During the initial lobe of the loop the field points upwards since the potential gradient during this phase is mainly sustained by the capacitive outward current $\tilde{c}_m \, \partial V_m / \partial t$ (compare location of instant t_2 in Fig. 5II(a) and 5III(e), (g)). The terminal lobe of **E** is mainly driven by the sodium inward current and the field vector points downwards (compare location of instant t_3 in Fig. 5II(a) and 5III(e),(g)). The time course of the magnitude of **E** is monophasic. The field strength **E** increases when the wavefront approaches, shows a maximum, when activation takes place at the recording site, and finally decreases when the wavefront departs.

Acknowledgments

The author would like to acknowledge Ernst Hofer for his excellent support during the last years and my colleagues Marta Monserrat, Blanca Rodríguez and Javier Saiz for their helpful comments and assistance in preparing this manuscript.

References

[1] R. R. Aliev and A. V. Panfilov, Modeling of heart excitation patterns caused by a local inhomogeneity, *J. Theor. Biol.* **181**(1) (1996) 33–40.

[2] R. E. McAllister, D. Noble and R. W. Tsien, Reconstruction of the electrical activity of cardiac Purkinje fibres, *J. Physiol.* **251** (1975) 1–59.

[3] R. C. Barr and R. Plonsey, Propagation of excitation in idealized anisotropic two-dimensional tissue, *Biophys. J.* **45**(6) (1984) 1191–1202.

[4] J. Beaumont, N. Davidenko, J. M. Davidenko and J. Jalife, Spiral waves in two-dimensional models of ventricular muscle: Formation of a stationary core, *Biophys. J.* **75**(1) (1998) 1–14.

[5] G. W. Beeler and H. Reuter, Reconstruction of the action potential of ventricular myocardial fibres, *J. Physiol.* **268** (1977) 177–210.

[6] M. J. Berger and R. J. Leveque, Adaptive mesh refinement using wave-propagation algorithms for hyperbolic systems, *SIAM J. Numer. Anal.* **35** (1998) 2298.

[7] M. J. Burgess, B. M. Steinhaus, K. W. Spitzer and P. R. Ershler, Nonuniform epicardial activation and repolarization, properties of in vivo canine pulmonary conus, *Circ. Res.* **62**(2) (1988) 233–246.

[8] C. Cabo, A. M. Pertsov, W. T. Baxter, J. M. Davidenko, R. A. Gray and J. Jalife, Wave-front curvature as a cause of slow conduction and block in isolated cardiac muscle, *Circ. Res.* **75** (1994) 1014–1028.

[9] C. Cabo, A. M. Pertsov, J. M. Davidenko, W. T. Baxter, R. A. Gray and J. Jalife, Vortex shedding as a precursor of turbulent electrical activity in cardiac muscle, *Biophys. J.* **70**(3) (1996) 1105–1111.

[10] F. J. L. van Capelle and D. Durrer, Computer simulation of arrhythmias in a network of coupled excitable elements, *Circ. Res.* **47**(3) (1980) 464–466.
[11] E. Carmeliet, Cardiac ionic currents and acute ischemia: From channels to arrhythmias, *Physiol. Rev.* **79**(3) (1999) 917–1017.
[12] J. Clark and R. Plonsey, A mathematical evaluation of the core conductor model, *Biophys. J.* **6**(1) (1966) 95–112.
[13] L. Clerk, Directional differences of impulse spread in trabecular muscle from mammalian heart, *J. Physiol.* **255** (1976) 335–346.
[14] R. A. Chapman and C. H. Fry, An analysis of the cable properties of frog ventricular myocardium, *J. Physiol. (Lond.)* **283** (1978) 263–282.
[15] E. M. Cherry, H. S. Greenside and C. S. Henriquez, A space-time adaptive method for simulating complex cardiac dynamics, *Phys. Rev. Lett.* **84**(6) (2000) 1343–1346.
[16] E. M. Cherry, A space-time adaptive mesh refinement method for simulating complex cardiac electrical dynamics, Ph.D. dissertation, Duke University (2000).
[17] E. Chudin, A. Garfinkel, J. Weiss, W. Karplus and B. Kogan, Wave propagation in cardiac tissue and effects of intracellular calcium dynamics, *Prog. Biophys. Mol. Biol.* **69**(2–3) (1998) 225–236.
[18] J. W. Cooley and F. A. Dodge, Digital computer solutions for excitation and propagation of the nerve impulse, *Biophys. J.* **6** (1966) 583–599.
[19] M. Courtemanche, R. J. Ramirez and S. Nattel, Ionic mechanisms underlying human atrial action potential properties: Insights from a mathematical model, *Am. J. Physiol.* **275** (1998) H301–H321.
[20] J. Crank and P. Nicholson, A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type, *Proc. Camb. Phil. Soc.* **43** (1947) 50–67.
[21] J. Deleze, The recovery of resting potential and input resistance in sheep heart injured by knife or laser, *J. Physiol (Lond.)* **208**(3) (1970) 547–562.
[22] S. S. Demir, J. W. Clark, C. R. Murphey and W. R. Giles, A mathematical model of a rabbit sinoatrial node cell, *Am. J. Physiol.* **266** (1994) C832–C852.
[23] P. J. Diaz, Y. Rudy and R. Plonsey, The intercalated discs as a cause for discontinuous propagation in cardiac musle: A theoretical simulation, *Ann. Biomed. Eng.* **121** (1983) 177–190.
[24] D. DiFrancesco and D. Noble, A model of cardiac electrical activity incorporating ionic pumps and current changes, *Philos. Trans. Royal Soc. London* **307** (1985) 353–398.
[25] J. P. Drouhard and F. A. Roberge, A simulation study of the ventricular myocardial action potential, *IEEE Trans. Biomed. Eng.* **29**(7) (1982) 494–502.
[26] J. P. Drouhard and F. A. Roberge, Revised formulation of the Hodgkin–Huxley representation of the sodium current in cardiac cells, *Comput. Biomed. Res.* **20**(4) (1987) 333–350.
[27] I. S. Duff, A. M. Erisman and J. K. Reid, *Direct Methods for Sparse Matrices* (Clarendon Press, 1986).

[28] L. Ebihara and E. A. Johnson, Fast sodium current in cardiac muscle — a quantitative description, *Biophys. J.* **32** (1980) 779–790.

[29] W. S. Ellis, D. M. Auslander and M. D. Lesh, Effects of coupling heterogeneity on fractionated electrograms in a model of nonuniformly anisotropic ventricular myocardium, *J. Electrocardiol.* **27**(Suppl.) (1994) 171–178.

[30] W. S. Ellis, D. M. Auslander and M. D. Lesh, Fractionated electrograms from a computer model of heterogeneously uncoupled anisotropic ventricular myocardium, *Circulation* **92**(6) (1995) 1619–1626.

[31] V. Fast and A. G. Kléber, Microscopic conduction in cultured strands of neonatal rat heart cells measured with voltage-sensitive dyes, *Circ. Res.* **73** (1993) 914–925.

[32] V. Fast and A. G. Kléber, Block of impulse propagation at an abrupt tissue expansion: Evaluation of the critical strand diameter in 2- and 3-dimensional computer models, *Cardiovasc. Res.* **20** (1995) 449–459.

[33] F. Fenton and A. Karma, Vortex dynamics in three-dimensional continuous myocardium with fiber rotation: Filament instability and fibrillation, *Chaos* **8** (1998) 20–47.

[34] J. M. Ferrero, Jr., J. Saiz, J. M. Ferrero and N. V. Thakor, Simulation of action potentials from metabolically impaired cardiac myocytes. Role of ATP-sensitive K^+ current, *Circ. Res.* **79**(2) (1996) 208–221.

[35] M. G. Fishler and N. V. Thakor, A massively parallel computer model of propagation through a two-dimensional cardiac syncytium, *Pacing Clin. Electrophysiol.* **14**(11)2 (1991) 1694–1699.

[36] R. FitzHugh, Impulses and physiological states in theoretical models of nerve membrane, *Biophys. J.* **1** (1961) 445–465.

[37] H. A. Fozzard, Conduction of the action potential, *Handbook of Physiology — The Cardiovascular System I*, pp. 335–355.

[38] D. B. Geselowitz and W. T. Miller, A bidomain model for anisotropic cardiac muscle, *Ann. Biomed. Eng.* **11** (1984) 191–206.

[39] G. H. Golub and C. F. V. Loan, *Matrix Computations* (Johns Hopkins University Press, 1989).

[40] D. M. Harrild and C. S. Henriquez, A finite volume model of cardiac propagation, *Ann. Biomed. Eng.* **25**(2) (1997) 315–334.

[41] D. M. Harrild, R. C. Penland and C. S. Henriquez, A flexible method for simulating cardiac conduction in three-dimensional complex geometries, *J. Electrocardiol.* **33**(3) (2000) 241–251.

[42] C. S. Henriquez and R. Plonsey, Effect of resistive discontinuities on waveshape and velocity in a single cardiac fibre, *Med. Biol. Eng. Comput.* **25** (1987) 428–438.

[43] C. S. Henriquez and R. Plonsey, The effect of the extracellular potential on propagation in excitable tissue, *Comments Theoret. Biol.* **1** (1988) 47–64.

[44] C. S. Henriquez, N. Trayanova and R. Plonsey, A planar slab bidomain model for cardiac tissue, *Ann. Biomed. Eng.* **18** (1990) 367–376.

[45] C. S. Henriquez, Simulating the electrical behavior of cardiac tissue using the bidomain model, *Crit. Rev. Biomed. Eng.* **21**(1) (1993) 1–77.

[46] C. S. Henriquez, A. L. Muzikant and C. K. Smoak, Anisotropy, fiber curvature, and bath loading effects on activation in thin and thick cardiac

tissue preparations: Simulations in a three-dimensional bidomain model, *J. Cardiovasc. Electrophysiol.* **7**(5) (1996) 424–444.

[47] C. S. Henriquez and A. A. Papazoglou, Using computer to understand the roles of tissue structure and membrane dynamics in arrhythmogenesis, *Proceedings of the IEEE* **84**(3) (1996) 334–354.

[48] A. Hodgkin and A. Rushton, The electrical constants of a crustacean nerve fibre, *Proc. Royal. Soc. B* (1946) 133.

[49] A. Hodgkin and A. Huxley, A quantitative description or membrane current and its application to conduction and excitation in nerve, *J. Physiol.* **117** (1952) 500–544.

[50] A. Hodgkin, A note on conduction velocity, *J. Physiol.* **125** (1954) 221–224.

[51] A. Hodgkin, *Conduction of the Nervous Impulse* (Liverpool University Press, 1964).

[52] N. F. Hooke, Efficient simulation of action potential propagation in a bidomain. Ph.D. dissertation, Duke University (1992).

[53] N. Hooke, C. S. Henriquez, P. Lanzkron and D. Rose, Linear algebraic transformations of the bidomain equations: Implications for numerical methods, *Math. Biosci.* **120**(2) (1994) 127–145.

[54] J. J. B. Jack, D. Noble and R. W. Tsien, *Electric Current Flow in Excitable Cells* (Clarendon, Oxford, 1975).

[55] C. T. January, V. Chau and J. C. Makielski, Triggered activity in the heart: Cellular mechanisms of early after- depolarizations, *Eur. Heart J.* **12**(Suppl. F) (1991) 4–9.

[56] R. W. Joyner, F. Ramon and J. W. Moore, Simulation of action potential propagation in an inhomogeneous sheet of coupled excitable cells, *Circ. Res.* **36** (1975) 654–661.

[57] R. W. Joyner, Effects of the discrete pattern of electrical coupling on propagation through an electrical syncytium, *Circ. Res.* **50**(2) (1982) 192–200.

[58] L. J. Leon and F. A. Roberge, A new cable model formulation based on Green's theorem, *Ann. Biomed. Eng.* **18** (1990) 1–17.

[59] L. J. Leon and F. A. Roberge, Directional characteristics of action potential propagation in cardiac muscle, *Circ. Res.* **69** (1991) 378–395.

[60] L. J. Leon and F. A. Roberge, Structural complexity effects on transverse propagation in a two-dimensional model of myocardium, *IEEE Trans. Biomed. Eng.* **38**(10) (1991) 997–1009.

[61] M. D. Lesh, M. Pring and J. F. Spear, Cellular uncoupling can unmask dispersion of action potential duration in ventricular myocardium. A computer modeling study, *Circ. Res.* **65** (1989) 1426–1440.

[62] D. S. Lindblad, C. R. Murphey, J. W. Clark and W. R. Giles, A model of the action potential and underlying membrane currents in a rabbit atrial cell, *Am. J. Physiol.* **271**(4) (1996) H1666–H1696.

[63] R. Lorente de Nó, Analysis of the distribution of action currents of nerve in volume conductors, *Studies from the Rockefeller Institute for Medical Research* **132** (1947) 384–447.

[64] C. H. Luo and Y. Rudy, A model of the ventricular cardiac action potential. Depolarization, repolarization, and their interaction, *Circ. Res.* **68** (1991) 1501–1526.

[65] C. H. Luo and Y. Rudy, A dynamic model of the cardiac ventricular action potential. I. Simulations of ionic currents and concentration changes, *Circ. Res.* **74**(6) (1994) 1071–1096.

[66] C. H. Luo and Y. Rudy, A dynamic model of the cardiac ventricular action potential. II. Afterdepolarizations, triggered activity, and potentiation, *Circ. Res.* **74**(6) (1994) 1097–1113.

[67] N. Maglaveras, J. M. de Bakker, F. J. Van Capelle, C. Pappas and M. J. Janse, Activation delay in healed myocardial infarction: A comparison between model and experiment, *Am. J. Physiol.* **269**(4)2 (1995) H1441–H1449.

[68] N. Maglaveras, F. Offner, F. J. Van Capelle, M. A. Allessie and A. V. Sahakian, Effects of barriers on propagation of action potentials in two-dimensional cardiac tissue. A computer simulation study, *J. Electrocardiol.* **28**(1) (1995) 17–31.

[69] N. Maglaveras, F. J. Van Capelle and J. M. de Bakker, Wave propagation simulation in normal and infarcted myocardium: Computational and modelling issues, *Med. Inform. (Lond.)* **23**(2) (1998) 105–118.

[70] M. Mascagni, The backward Euler method for numerical solution of the Hodgkin–Huxley equations of nerve conduction, *SIAM J. Numer. Anal.* **27** (1991) 941–962.

[71] P. Metzger and R. Weingart, Electric current flow in cell pairs isolated from adult rat hearts, *J. Physiol. (Lond.)* **366** (1985) 177–195.

[72] M. Monserrat, J. Saiz, J. M. Ferrero and N. V. Thakor, Ectopic activity in ventricular cells induced by early afterdepolarizations developed in Purkinje cells, *Ann. Biomed. Eng.* **28**(11) (2000) 1343–1351.

[73] K. W. Morton and D. F. Mayers, *Numerical Solution of Partial Differential Equations* (Cambridge University Press, 1994).

[74] A. Nygren, C. Fiset, L. Firek, J. W. Clark, D. S. Lindblad, R. B. Clark and W. R. Giles, Mathematical model of an adult human atrial cell, *Circ. Res.* **82** (1998) 63–81.

[75] A. Nygren and J. A. Halter, A general approach to modeling conduction and concentration dynamics in excitable cells of concentric cylindrical geometry, *J. Theor. Biol.* **199** (1999) 329–358.

[76] N. F. Otani, Computer modeling in cardiac electrophysiology, *J. Comput. Phys.* **161** (2000) 21–34.

[77] N. F. Otani and D. Allexandre, A new variable timestep numerical method for excitable systems, submitted to *Phys. Rev. Lett.*, http://reentry.cwru.edu/otani/oolrsim/prl01.html.

[78] D. W. Peaceman and H. H. Rachford Jr., The numerical solution of parabolic and elliptic differential equations, *J. Soc. Industrial Appl. Math.* **3** (1955) 28–41.

[79] A. M. Pertsov, J. M. Davidenko, R. Salomonsz, W. T. Baxter and J. Jalife, Spiral waves of excitation underlie reentrant activity in isolated cardiac muscle, *Circ. Res.* **72**(3) (1993) 631–650.

[80] G. Plank and E. Hofer, Model study of vector-loop morphology during electrical mapping of microscopic conduction in cardiac tissue, *Ann. Biomed. Eng.* **28**(10) (2000) 1244–1252.

[81] R. Plonsey and D. Heppner, Considerations of quasi-stationarity in electrophysiological systems, *Bull. Math. Biophys.* **29** (1967) 657–664.

[82] R. Plonsey, The active fiber in a volume conductor, *IEEE Trans. Biomed. Eng.* **21**(5) (1974) 371–381.

[83] R. C. Barr and R. Plonsey, Propagation of excitation in idealized anisotropic two-dimensional tissue, *Biophys. J.* **45** (1984) 1191–1202.

[84] R. Plonsey, Bioelectric sources arising in excitable fibers, *Ann. Biomed. Eng.* **16** (1988) 519–546.

[85] R. Plonsey and R. C. Barr, *Bioelectricity — A Quantitative Approach* (Plenum Press, New York, 1988).

[86] A. E. Pollard and R. C. Barr, Computer simulations of activation in an anatomically based model of the human ventricular conduction system, *IEEE Trans. Biomed. Eng.* **38** (1991) 982–996.

[87] A. E. Pollard, J. M. Burgess and K. W. Spitzer, Computer simulations of three-dimensional propagation in ventricular myocardium. Effects of intramural fiber rotation and inhomogeneous conductivity on epicardial activation, *Circ. Res.* **72**(4) (1991) 744–756.

[88] A. E. Pollard, N. Hooke and C. S. Henriquez, Cardiac propagation simulation, *Crit. Rev. Biomed. Eng.* **20**(3–4) (1992) 171–210.

[89] A. E. Pollard, M. J. Burgess and K. W. Spitzer, Computer simulations of three-dimensional propagation in ventricular myocardium. Effects of intramural fiber rotation and inhomogeneous conductivity on epicardial activation, *Circ. Res.* **72**(4) (1993) 744–756.

[90] D. Porras, J. M. Rogers, W. M. Smith and A. E. Pollard, Distributed computing for membrane-based modeling of action potential propagation, *IEEE Trans. Biomed. Eng.* **47**(8) (2000) 1051–1057.

[91] L. Priebe and D. J. Beuckelmann, Simulation study of cellular electric properties in heart failure, *Circ. Res.* **82**(11) (1998) 1206–1223.

[92] Z. Qu, J. N. Weiss and A. Garfinkel, Cardiac electrical restitution properties and stability of reentrant spiral waves: A simulation study, *Am. J. Physiol.* **276**(1) (1999) H269–H283.

[93] Z. Qu and A. Garfinkel, An advanced algorithm for solving partial differential equation in cardiac conduction, *IEEE Trans. Biomed. Eng.* **46**(9) (1999) 1166–1168.

[94] Z. Qu, J. Kil, F. Xie, A. Garfinkel and J. N. Weiss, Scroll wave dynamics in a three-dimensional cardiac tissue model: Roles of restitution, thickness and fiber rotation, *Biophys. J.* **78**(6) (2000) 2761–2775.

[95] W. Quan, J. F. Spear, S. J. Stevens and H. M. Hastings, Efficient integration of a realistic two-dimensional cardiac tissue model by domain decomposition, *IEEE Trans. Biomed. Eng.* **45** (1998) 372–385.

[96] R. L. Rasmusson, J. W. Clark, W. R. Giles, K. Robinson, R. B. Clark, E. F. Shibata and D. L. Campbell, A mathematical model of electrophysiological activity in a bullfrog atrial cell, *Am. J. Physiol.* **259** (1990) H370–H389.

[97] F. A. Roberge, A. Vinet and B. Victorri, Reconstruction of propagated electrical activity with a two-dimensional model of anisotropic heart muscle, *Circ. Res.* **58**(4) (1986) 461–475.

[98] F. A. Roberge, L. Boucher and A. Vinet, Model study of the spread of electrotonic potential in cardiac tissue, *Med. Biol. Eng. Comput.* **27**(4) (1989) 405–415.

[99] F. A. Roberge, S. Wang, H. Hogues and L. J. Leon, Propagation on a central fiber surrounded by inactive fibers in a multifibered bundle, *Ann. Biomed. Eng.* **24** (1996) 647–661.

[100] J. M. Rogers and A. D. McCulloch, A collocation-Galerkin finite element model of cardiac action potential propagation, *IEEE Trans. Biomed. Eng.* **41**(8) (1994) 743–757.

[101] J. M. Rogers and A. D. McCulloch, Nonuniform muscle fiber orientation causes spiral wave drift in a finite element model of cardiac action potential propagation, *J. Cardiovasc. Electrophysiol.* **5**(6) (1994) 496–506.

[102] P. Rosenfalck, Intra- and extracellular potential fields of active nerve and muscle fibres. A physico-mathematical analysis of different models, *Acta Physiol. Scand.* **321**(Suppl.) (1969) 1–168.

[103] Y. Rudy and W. L. Quan, A model study of the effects of discrete cellular structure on electrical propagation in cardiac tissue, *Circ. Res.* **61** (1987) 815–823.

[104] S. Rush and H. Larsen, A practical algorithm for solving dynamic membrane equations, *IEEE Trans. Biomed. Eng.* **25** (1978) 389–392.

[105] Y. Saad, *Iterative Methods for Sparse Linear Systems* (PWS Publishing Company, Boston, 1995).

[106] J. Saiz, J. M. Ferrero Jr., M. Monserrat, J. M. Ferrero and N. V. Thakor, Influence of electrical coupling on early afterdepolarizations in ventricular myocytes, *IEEE Trans. Biomed. Eng.* **46**(2) (1999) 138–147.

[107] J. E. Saffitz, H. L. Kanter, K. G. Green, T. K. Tolley and E. C. Beyer, Tissue-specific determinants of anisotropic conduction velocity in canine atrial and ventricular myocardium, *Circ. Res.* **74**(6) (1994) 1065–1070.

[108] N. G. Sepulveda and J. P. Wikswo Jr., Electric and magnetic fields from two-dimensional anisotropic bisyncytia, *Biophys. J.* **51** (1987) 557–568.

[109] N. G. Sepulveda, B. J. Roth and J. P. Wikswo Jr., Current injection into a two-dimensional anisotropic bidomain, *Biophys. J.* **55**(5) (1989) 987–999.

[110] R. M. Shaw and Y. Rudy, Electrophysiologic effects of acute myocardial ischemia: A theoretical study of altered cell excitability and action potential duration, *Cardiovasc. Res.* **35**(2) (1997) 256–272.

[111] J. R. Sommer and B. Scherer, Geometry of cell and bundle appositions in cardiac muscle: Light microscopy, *Am. J. Physiol.* **248**(6) (1985) H792–H803.

[112] M. S. Spach and J. M. Kootsey, Relating the sodium current and conductance to the shape of transmembrane and extracellular potentials by simulation: Effects of propagation boundaries, *IEEE Trans. Biomed. Eng.* **10** (1985) 743–755.

[113] M. S. Spach and J. F. Heidlage, A multidimensional model of cellular effects on the spread of electrotonic currents and on propagating action potentials, *Crit. Rev. Biomed. Eng.* **20**(3, 4) (1992) 141–169.

[114] M. S. Spach and J. F. Heidlage, The stochastic nature of cardiac propagation at a microscopic level — electrical description of myocardial architecture and its application to conduction, *Circ. Res.* **76** (1995) 366–380.

[115] M. S. Spach, Discontinuous cardiac conduction — its origin in cellular connectivity with long-term adaptive changes that cause arrhythmias, in *Discontinuous Conduction in the Heart*, eds. P. M. Spooner, R. W. Joyner and J. Jalife (Futura Publishing Company, Inc., Armonk, NY, 1997).

[116] G. Strang, On the construction and comparison of difference schenes, *SIAM J. Numer. Anal.* **5** (1968) 506–517.

[117] N. V. Thakor, J. M. Ferrero, Jr., J. Saiz, B. I. Gramatikov and J. M. Ferrero, Electrophysiologic models of heart cells and cell networks, *IEEE Eng. Med. Biol. Mag.* **17**(5) (1998) 73–83.

[118] N. Trayanova, C. S. Henriquez and R. Plonsey, Extracellular potentials and currents of a single active fiber in a restricted volume conductor, *Ann. Biomed. Eng.* **18** (1990) 219–238.

[119] S. Veronese and H. G. Othmer, A computational study of wave propagation in a model for anisotropic cardiac ventricular tissue, Lecture Notes in Computer Science (Springer-Verlag, Berlin), Vol. 919, pp. 248–253.

[120] B. Victorri, A. Vinet, F. A. Roberge and J. P. Drouhard, Numerical integration in the reconstruction of cardiac action potentials using Hodgkin–Huxley-type models, *Comput. Biomed. Res.* **18**(1) (1985) 10–23.

[121] E. J. Vigmond and L. J. Leon, Computationally efficient model for simulation electrical activity in cardiac tissue with fiber rotation, *Ann. Biomed. Eng.* **27** (1999) 160–170.

[122] E. J. Vigmond and L. J. Leon, Effect of fibre rotation on the initiation of re-entry in cardiac tissue, *Med. Biol. Eng. Comput.* **39** (2001) 455–464.

[123] N. Virag, J. M. Vesin and L. Kappenberger, A computer model of cardiac electrical activity for the simulation of arrhythmias, *Pacing Clin. Electrophysiol.* **21**(11–2) (1998) 2366–2371.

[124] S. Wang, J. L. Leon and F. A. Roberge, Interactions between adjacent fibers in a cardiac muscle bundle, *Ann. Biomed. Eng.* **24** (1996) 662–674.

[125] S. Weidmann, The electrical constants of purkinje fibres, *J. Physiol.* **118** (1952) 348–360.

[126] R. Weingart, Electrical properties of the nexal membrane studied in rat ventricular cell pairs, *J. Physiol. (London)* **370** (1986) 267–284.

[127] R. Weingart and P. Maurer, Action potential transfer in cell pairs isolated from adult rat and guinea pig ventricles, *Circ. Res.* **63** (1988) 72–80.

[128] J. Wu and D. P. Zipes, Effects of spatial segmentation in the continuous model of excitation propagation in cardiac muscle, *J. Cardiovasc. Electrophysiol.* **10**(7) (1999) 965–972.

CHAPTER 10

FLOW IN TUBES WITH COMPLICATED GEOMETRIES WITH SPECIAL APPLICATION TO BLOOD FLOW IN LARGE ARTERIES

GIRIJA JAYARAMAN
Centre for Atmospheric Sciences, Indian Institute of Technology
New Delhi-110016, India
jgirija@cas.iitd.ernet.in

0. Introduction

The study of flow in tubes of complex geometry has invited a lot of attention due to its application to wide range of problems. Tubes of complex geometry are of common occurrence in almost all piping systems, several engineering devices such as heat and mass exchangers, chemical reactors, chromatography columns, processing equipments and human cardiovascular system. Fluid dynamic principles have been successfully applied to understand and solve many engineering problems and it is our belief that those ideas can be applied with confidence to understand physiological flows in general and blood flows in particular [7]. However, it is important to remember that fluid flows within the human body raise problems very different from those raised by engineering flows [35]. Modelling blood flows in the circulatory system would require incorporating various essential features like the properties of blood, the pulsatility of flow, multiplicity of vessel branching and variation in pressure/velocity of the flow.

The flow of a fluid in tubes of complex geometry can be used to understand the flow characteristics of blood flow in the cardiovascular system which provides the means for circulation of materials throughout the body. The anatomy of the canine aorta and its main branches is described in Fig. 1. The blood vessels are of different dimensions, are curved, elastic and branched. Hence, the flow, which is highly influenced by the geometry of the vessel, is never a Poiseuille flow. The repeated branching keeps it as

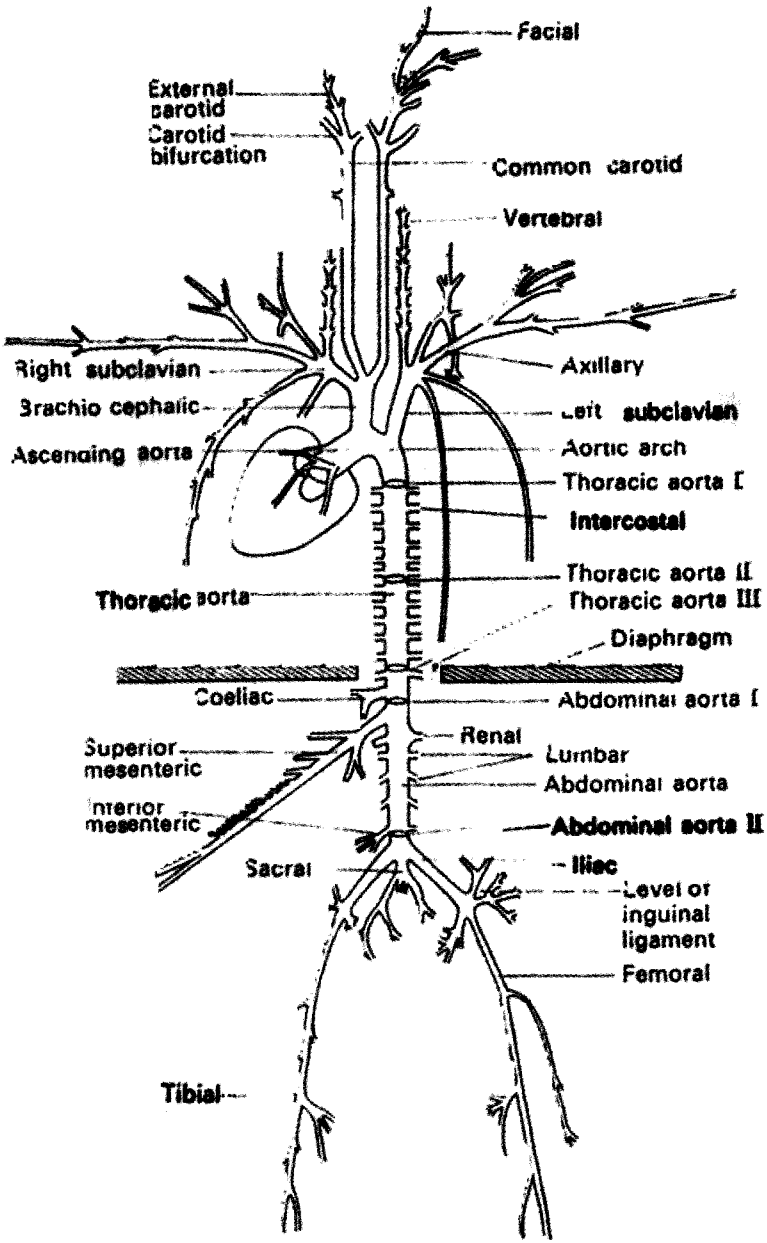


Fig. 1. A diagrammatic representation of the major branches of the canine arterial tree (After McDonald (1974)).

an entry flow and due to the curvature, secondary motions are developed in addition to the primary flow. Localized hydrodynamic effects such as pressure distribution along the blood vessel, wall shear stress distribution, velocity distribution, secondary vortices, separation phenomena and onset of turbulence are crucial in understanding the nature of the flow in the cardiovascular system.

0.1. *Wall shear stress and atherosclerosis*

The possibility that arterial fluid mechanics may explain why certain vascular sites are more prone than others to the development of atherosclerosis — a disease due to the thickening of the artery wall — is responsible for the interest in modelling the cardiovascular system through fluid dynamics. The belief that apart from physiological factors, additional factors like the magnitude of wall shear stress and physics of the blood flow in vessels of complex geometry may contribute to the initiation of the disease brought together a lot of fluid dynamicists to work in this area [22]. The flow details are found to be important in determining the distribution of wall shear stress in arteries, a major factor in atherogenesis. It is established that low and/oscillatory wall shear stress are the associated hemodynamic conditions at these sites where the disease is found to develop [6]. Departures from unidirectional, laminar and symmetrical flow patterns in the blood circulation are found to encourage plaque formation. The coronary arteries are exposed to greater fluctuations in flow direction and amplitude during systole than are other systemic arteries and increases in heart rate result in decreased diastolic time while systolic time remains nearly constant [24]. The hemodynamic investigations in the living systems being inadequate at the moment, the emphasis should be on models, both theoretical as well as experimental, in order to correlate the distribution of lesions in given anatomical regions of human subjects with quantitative fluid dynamic measurements.

0.2. *Arterial stenosis*

A consequence of the thickening of the arterial wall due to the atherosclerotic plaques is the occlusion of the artery — it is then called a stenosed artery. As a result of the occlusion, the blood supply to the corresponding organs is impaired. Arterial constriction/stenosis is associated with significant changes in the blood flow, pressure distribution and resistance to the flow. In regions of narrowing arterial constriction the flow accelerates and consequently the velocity gradient near the wall region is steeper due to

the increased core velocity resulting in relatively large shear stress on the wall even for a mild constriction. The large pressure loss across the stenosis is essentially dependent on the flow rate and the geometry (size as well as shape) of the stenosis. Thus, the important characteristics to be studied in problems of blood flow through a stenosed artery are (i) reduction in blood supply/enhanced impedance to flow (ii) changed flow pattern causing separation of flow or turbulence and (iii) changed properties of the artery walls like post stenotic dilatation. The mathematical study of stenosed blood flow was initiated by Young [55, 57]. Using a simple order of magnitude analysis, the impedance factor was obtained. Subsequently, there have been numerous studies, both theoretical and experimental [34, 38, 41, 42, 44, 56].

0.3. *Entry flows*

When a fluid enters a tube with a flat velocity profile, the portion which comes into contact with the wall is forced to be motionless in view of the "no slip" condition. Immediately, a velocity gradient is established between the motionless fluid at the wall and the adjacent fluid in the core. As the flow proceeds along the tube, viscosity progressively modifies the blunt profile. The original velocity gradient at the wall becomes reduced, and more of the core is sheared. The core fluid is accelerated to maintain the constant flux across a cross section. There has been a keen interest in the study of entry flows and several publications [21, 36, 50–52] came up due to its application to blood flows, since, as in the words of Lighthill [35], blood flow in "large arteries is almost all entry region". This is due to the repeated branchings of the blood vessels which do not let the flow to become a developed one. The problem gets more acute when a stenosis is at the very entrance of a blood vessel [28, 29].

0.4. *Influence of curvature*

The interest in understanding the fluid dynamics of blood circulation has also initiated a lot of study on problems related to flow in curved tubes. The aorta, which takes origin from the left ventricle, curves in a complicated three dimensional way, through about 180°, giving branches to the heart, head and upper limbs (Fig. 1). The exact distribution of velocity and pressure at these vessel entrances is dependent on the ventricular contractions and the heart valves.

The flow in a curved tube is very different from that in a straight tube. In addition to the primary flow along the axis of the tube, there exist

secondary components of velocity due to the lateral forces. This secondary flow is in the form of the fluid in the core being swept to the outside of the bend and that near the wall returning towards the inside. This induces a pressure gradient, called the centrifugal pressure gradient, directed towards the inner bend of the curved tube. Computed axial velocity contours and secondary flow streamlines are shown in Figs. 2 and 3. The first theoretical study on steady fully developed flow of an incompressible Newtonian fluid in a loosely coiled curved pipe was made by Dean [15, 16] for values of Dean number D (similarity parameter) up to 96.

Subsequently, there have been numerous theoretical and experimental investigations on steady, unsteady, developing, and fully-developed flows in curved tubes of circular and non-circular cross-sections which are extensively reviewed by Pedley [46], Berger *et al.* [3], Ito [26].

The numerical solution of the Dean's problem for intermediate and higher values of Dean number D were obtained by McConalogue and Srivastava [37] for $96 \leq D \leq 605$, Truesdell and Alder [54] for $96 \leq D \leq 3578$, Greenspan [25], Collins and Dennis [11], and Dennis [18] for $96 \leq D \leq 5000$ by using finite difference schemes of different accuracy. The significance of the numerical solution by Collins and Dennis [11] was that it was of second order accuracy with respect to grid sizes and it established

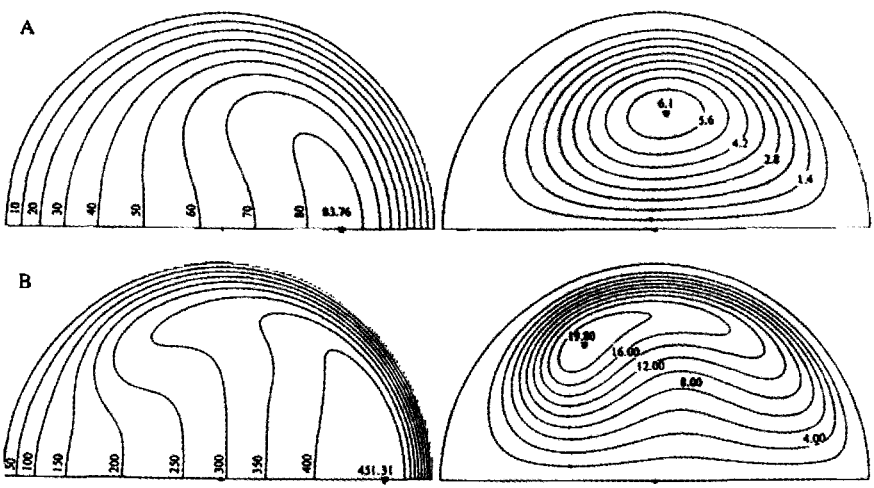


Fig. 2. Computed axial velocity contours (left) and two-vortex secondary flow streamlines (right) for steady flow in a curved tube of small curvature at two values of the Dean number. (A) $D = 500$; (B) $D = 5000$ (Daskopoulos and Lenhoff [14]).

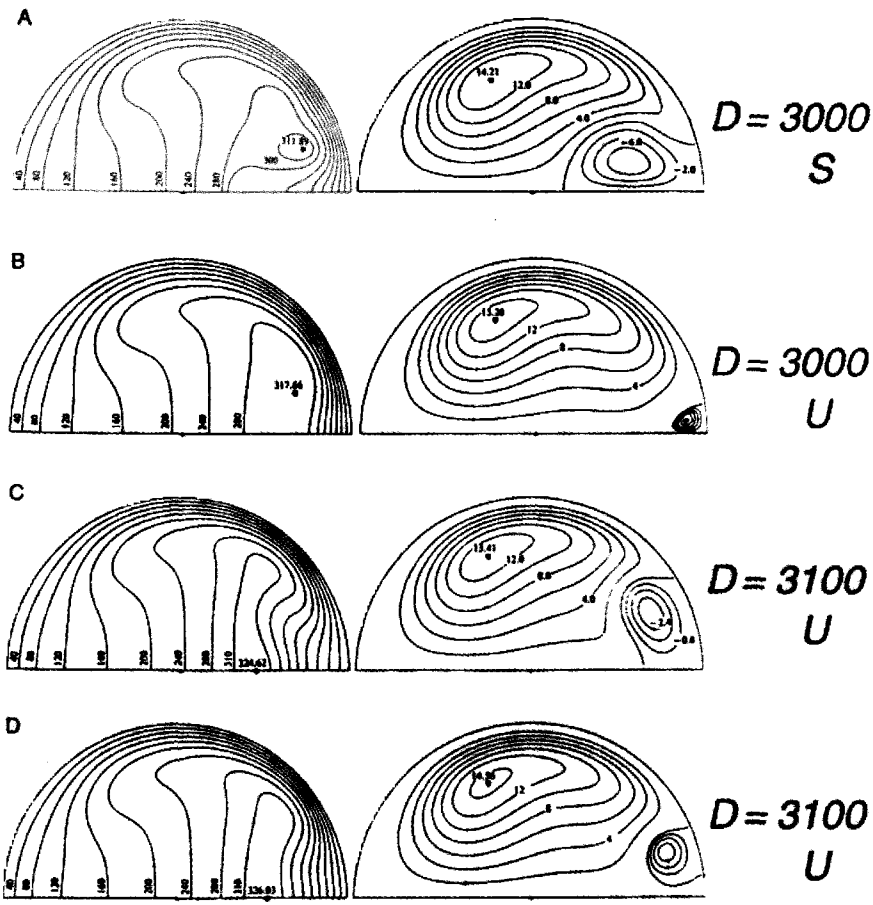


Fig. 3. Computed axial Velocity contours (left) and a secondary flow streamlines (right) for steady flow is a curved tube of small curvature at comparable values of the Dean number D, showing the non-uniqueness of the flow. The flows marked S are stable, those marked U are unstable (Daskopoulos and Lenhoff [14]).

the asymptotic structure of the solution for $D \to \infty$. The multiple solution for flow in curved pipes/ducts were obtained by Dennis and Ng [19], Daskopoulos and Lenhoff [14], Kao [32], and Mees *et al.* [40].

With an objective to understand the role of fluid mechanics on cardiovascular diseases like atherosclerosis (the disease occurs at certain preferential sites like curved portions and flow dividers of the arterial tree), several studies related to flow in curved pipes were taken to estimate the wall shear stress. Padmanabhan and Jayaraman [44] and Jain and Jayaraman

[27] studied the flow characteristics in curved constricted tubes of circular and elliptic cross-sections, respectively, based on a double series perturbation analysis for small curvature and mild constriction. The numerical simulation of pulsatile or oscillatory flow in a curved tube was made by Chang and Tarbell [8, 9] and Schilt *et al.* [49].

0.5. *Artefacts of catheters*

Catheters attached with various functional tools have extensive use in contemporary medical sciences. The measurements of various physiological flow characteristics (such as arterial blood pressure or pressure gradient and flow velocity or flow rate) as well as the diagnosis and treatment of various arterial diseases (such as X-ray angiography, intravascular ultrasound, and coronary balloon angioplasty) are done through an appropriate catheter-tool device by inserting the device into a peripheral artery and positioning it in the desired part of arterial network [47, 48]. In addition, when a catheter is inserted into a stenosed artery, it will further increase the impedance or frictional resistance to flow and will alter the pressure distribution. Thus, the pressure/pressure gradient or flow velocity/flow rate recorded by the device will certainly differ from that of uncatheterized artery. Recent interests in flow in curved annulus [13, 20, 30, 33] are due to their applicability to understand the changed flow pattern in a catheterized artery and to introduce corrections to the measured pressure or pressure gradient using catheters.

With the evolution of coronary balloon angioplasty, there has been a considerable increase in the use of catheters of various sizes. It has been shown that, by reducing the obstruction through balloon angioplasty, the mean translesional pressure drop $\Delta\tilde{p}$, i.e. the difference of mean pressure between coronary ostium as measured through the guiding catheter (2.6 mm diameter) and just distal to the stenosis as measured through the angioplasty catheter (1.4 mm diameter), is reduced and the coronary blood flow as well as the coronary flow reserve is increased. The magnitude of mean translesional pressure drop $|\Delta\tilde{p}|$ is often used by clinicians to gauge the severity of the lesion and the reduction in $|\Delta\tilde{p}|$ due to angioplasty is used to judge the effectiveness of the interventions [23]. It is important to mention here that relatively large mean translesional pressure drop of about 51 mm Hg (nearly 50% of $\approx$100 mm Hg, mean overall pressure drop across the coronary artery) has been observed at basal flow before angioplasty.

It is well-known that the standard angioplasty catheters cause coronary flow obstruction, and therefore, will certainly magnify the true pressure

drop. In a series of papers, Bjorno and Pettersson [4, 5] studied extensively the hydro- and hemodynamic effects of catheterization of vessels with and without stenosis through various experimental models. Back [1] and Back *et al.* [2] studied the important hemodynamical characteristics like the wall shear stress, pressure drop and frictional resistance in catheterized coronary arteries under normal as well as the pathological situation of a stenosis present. The effect of catheterization on various flow characteristics in a curved artery was studied by Jayaraman and Tiwari [30]; it was shown that catheterization led to an increase in the axial wall shear stress and formation of increased number of secondary vortices. The experimental study on flow characteristics in a curved vessel with an aneurysm was made by Niimi *et al.* [43]; it was shown that the vortices induced in the aneurysm influenced and modified the axial velocity and secondary flow due to the vessel curvature.

It is fairly obvious from the foregoing that in the past four decades, a lot of emphasis has been laid on internal flows, especially, in tubes of complex geometries with an objective of understanding the flow in the human blood circulation. We shall discuss in the following an example which takes into account most of the complexities discussed so far and explores the possibilities of mathematical modelling becoming a part of the procedures in clinical medicine. The details of mathematical formulation including the simplification of the governing equations of motion are given in Sec. 2. The methods of approach (i) a perturbed solution and (ii) a numerical scheme for the simplified equations are discussed in Sec. 3. The effects of Dean number D and radii ratio k on various flow characteristics — i.e. flow rate, pressure gradient, pressure drop, frictional resistance, friction factor, wall shear stress, and the primary and secondary flow patterns — are discussed in Sec. 4. Finally, we have discussed the important application of this study to the clinical problem — flow in a stenosed artery with an inserted catheter — as required to model in balloon angioplasty and during blood pressure measurement using catheters. The effect of catheterization on various physiologically important flow characteristics — i.e. pressure drop, impedance, wall shear stress and the change in flow. The results are used to obtain the estimates of increased pressure drop (and hence, impedance) and wall shear stress across a coronary artery stenosis during catheterization. In addition, many interesting fluid mechanical phenomena, i.e. the modification of secondary streamlines due to the combined action of stenosis and curvature, and formation of increased number of secondary vortices due to catheterization, are brought out.

1. Mathematical Formulation

1.1. *Flow geometry*

The mathematical formulation models the curved artery as a rigid circular tube of radius "a", coiled in the form of a circle of radius "b", and the catheter as a coaxial rigid tube with radius "ka" with $k < 1$. It is assumed that the stenosis has developed in an axi-symmetric manner due to some abnormal growth over a length "L" of the artery given by

$$\frac{\bar{\eta}(\bar{z})}{a} = 1 - \frac{h}{a} \sin \pi \left(\frac{\bar{z} - d}{L} \right), \quad d \leq \bar{z} \leq d + L, \tag{1}$$

where $\bar{\eta}(\bar{z})$ is the radius of the stenosis, $\bar{z}$ is along the axis of the artery and "h" is the maximum projection of the stenosis into the lumen. Since we are interested in the instantaneous condition of the stenosis during catheterization, the growth of stenosis with time, which is very slow, can be neglected. The *schematic* diagram of the flow geometry corresponding to a catheterized curved artery with stenosis is shown in Fig. 4. It is further assumed that the flow geometry lies in a plane so that the effect of torsion can be neglected.

1.2. *Co-ordinate system*

Figure 5 shows the system of toroidal co-ordinates $(\bar{r}, \theta, \phi)$ used to analyze the flow field in the geometry mentioned above.

"C" is the center of the cross-section of the tube which makes an angle ϕ with the fixed axial plane and "P" is an arbitrary point in the cross-section

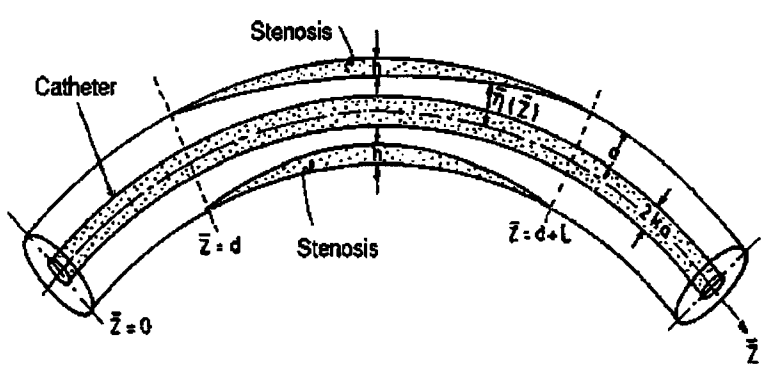


Fig. 4. The schematic diagram of a catheterized curved artery with stenosis.

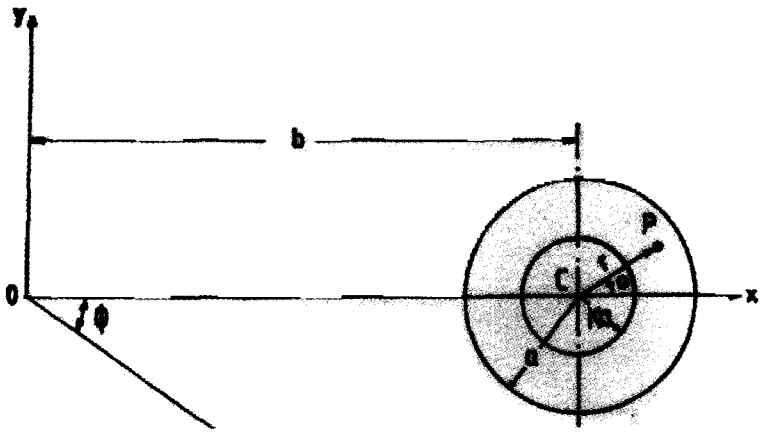


Fig. 5. The Toroidal co-ordinate of the flow geometry.

whose polar co-ordinate is $(\bar{r}, \theta)$. OC is the length "b" which is the radius of curvature of the curved tube. The axial co-ordinate is defined by $\bar{z} = b\phi$.

1.3. *Governing equations of motion*

Blood is modelled as an incompressible Newtonian fluid and the flow is assumed to be steady and laminar. The equations of motion governing the flow, in the co-ordinates system described above, are given in non-dimensional form as [31, 46].

$$\Delta_1 u - \frac{v^2}{r} - \frac{\varepsilon w^2 \cos\theta}{H} = -\frac{\partial p}{\partial r} + \frac{1}{\mathrm{Re}}\left[\nabla^2 u - \frac{u}{r^2} - \frac{2}{r^2}\frac{\partial v}{\partial \theta} + \frac{\varepsilon v \sin\theta}{rH} - \frac{\varepsilon \sin\theta}{rH}\frac{\partial u}{\partial \theta} + \frac{\varepsilon \cos\theta}{H}\frac{\partial u}{\partial r} - \frac{2\varepsilon\delta\cos\theta}{H^2}\frac{\partial w}{\partial z} - \frac{\varepsilon^2 u \cos^2\theta}{H^2} + \frac{\varepsilon^2 v \sin\theta\cos\theta}{H^2}\right], \tag{2a}$$

$$\Delta_1 v + \frac{uv}{r} + \frac{\varepsilon w^2 \sin\theta}{H} = -\frac{1}{r}\frac{\partial p}{\partial \theta} + \frac{1}{\mathrm{Re}}\left[\nabla^2 v - \frac{v}{r^2} + \frac{2}{r^2}\frac{\partial u}{\partial \theta} - \frac{\varepsilon u \sin\theta}{rH} - \frac{\varepsilon \sin\theta}{rH}\frac{\partial v}{\partial \theta} + \frac{\varepsilon \cos\theta}{H}\frac{\partial v}{\partial r} + \frac{2\varepsilon\delta\cos\theta}{H^2}\frac{\partial w}{\partial z} - \frac{\varepsilon^2 v \sin^2\theta}{H^2} + \frac{\varepsilon^2 u \sin\theta\cos\theta}{H^2}\right], \tag{2b}$$

$$\Delta_1 w + \frac{\varepsilon w}{H}(u\cos\theta - v\sin\theta) = -\frac{\delta}{H}\frac{\partial p}{\partial z} + \frac{1}{\mathrm{Re}}\left[\nabla^2 w - \frac{\varepsilon^2 w}{H^2} - \frac{\varepsilon\sin\theta}{rH}\frac{\partial w}{\partial\theta} + \frac{\varepsilon\cos\theta}{H}\frac{\partial w}{\partial r} - \frac{2\varepsilon\delta\sin\theta}{H^2}\frac{\partial v}{\partial z} + \frac{2\varepsilon\delta\cos\theta}{H^2}\frac{\partial u}{\partial z}\right], \tag{2c}$$

$$\frac{\partial u}{\partial r} + \frac{u}{r} + \frac{1}{r}\frac{\partial v}{\partial\theta} + \frac{\varepsilon u\cos\theta}{H} - \frac{\varepsilon v\sin\theta}{H} + \frac{\delta}{H}\frac{\partial w}{\partial z} = 0, \tag{2d}$$

where

$$\nabla^2 = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2} + \frac{\delta^2}{H^2}\frac{\partial^2}{\partial z^2}, \tag{3a}$$

$$\Delta_1 = u\frac{\partial}{\partial r} + \frac{v}{r}\frac{\partial}{\partial\theta} + \frac{\delta w}{H}\frac{\partial}{\partial z}, \tag{3b}$$

$\vec{q} = (u, v, w)$ is the velocity field in (r, θ, z) co-ordinates, p is the pressure field and $H = 1+\varepsilon r\cos\theta$. The parameters occurring in the problem through these equations are the Reynolds number Re, curvature parameter ε, and the geometric parameter δ, defined respectively, as

$$\mathrm{Re} = \frac{aU_0}{\nu}, \quad \varepsilon = \frac{a}{b}, \quad \delta = \frac{a}{L}, \tag{4}$$

where U_0 is the characteristic velocity (centerline velocity in a straight tube), $\nu = \mu/\rho$ is the kinematic viscosity, μ is the dynamic viscosity and ρ is the density of the fluid. The non-dimensionalization of the various variables has been performed as follows;

$$(u, v, w) = \left(\frac{\bar{u}}{U_0}, \frac{\bar{v}}{U_0}, \frac{\bar{w}}{U_0}\right), \quad z = \frac{\bar{z} - d}{L}, \quad r = \frac{\bar{r}}{a}, \quad p = \frac{\bar{p}}{\rho {U_0}^2}. \tag{5}$$

1.4. *Boundary conditions*

The appropriate boundary conditions for the problem under study are the no-slip conditions at the arterial wall and the catheter wall, i.e.

$$u = v = w = 0 \quad \text{at } r = \eta(z) \quad \text{and} \quad r = k, \tag{6}$$

where

$$\eta(z) = 1 - \delta_1 \sin\pi z, \quad 0 \le z \le 1, \tag{7}$$

$\eta(z)$ is the non-dimensional radius of the stenosed artery and $\delta_1 = h/a$ is the non-dimensional maximum height of the stenosis (stenotic parameter). The pressure field is obtained by using the unity flux condition over the length

of the stenosis. It is to be noted that two more parameters, i.e. the non-dimensional catheter radius k and the stenotic parameter δ_1, enter into the problem through the boundary conditions. The other boundary condition includes the symmetric condition for the flow field about the central plane (the plane passing through OX and perpendicular to OY in Fig. 2), i.e.

$$\frac{\partial u}{\partial \theta} = v = \frac{\partial w}{\partial \theta} = 0 \quad \text{at } \theta = 0 \quad \text{and} \quad \theta = \pi. \tag{8}$$

2. Methods of Solution

2.1. *Perturbation analysis*

Equations (2a)–(2d) are non-linear in nature, and hence, it is difficult to obtain a closed-form solution for all values of ε, δ and Re. Nevertheless, it is possible to find an approximate solution for small values of ε and δ through a double series perturbation analysis; small ε (i.e. $\varepsilon = a/b \ll 1$) refers to small curvature ratio and small δ (i.e. $\delta = a/L \ll 1$) corresponds to slowly varying cross-section and enables the use of lubrication theory. Thus, "z" will appear as a parameter in the problem and the solution in the stenotic region will correspond to $0 \leq z \leq 1$. We perturb all the physical variables in powers of ε and δ and seek the solutions in the form of series expansions, as in Padmanabhan and Jayaraman [44] given by

$$\begin{aligned}\vec{q} = {} & (\vec{q}_{00} + \delta\vec{q}_{01} + \delta^2\vec{q}_{02} + \cdots) + \varepsilon(\vec{q}_{10} + \delta\vec{q}_{11} + \delta^2\vec{q}_{12} + \cdots) \\ & + \varepsilon^2(\vec{q}_{20} + \delta\vec{q}_{21} + \delta^2\vec{q}_{22} + \cdots) + \cdots,\end{aligned} \tag{9a}$$

$$\begin{aligned}p = {} & \left(\frac{1}{\delta}p_{0-1} + p_{00} + \delta p_{01} + \cdots\right) + \varepsilon\left(\frac{1}{\delta}p_{1-1} + p_{10} + \delta p_{11} + \cdots\right) \\ & + \varepsilon^2\left(\frac{1}{\delta}p_{2-1} + p_{20} + \delta p_{21} + \cdots\right) + \cdots,\end{aligned} \tag{9b}$$

with $v_{00} = v_{01} = v_{02} = 0$ (since the flow is two-dimensional in the absence of curvature, i.e. for $\varepsilon = 0$) and $u_{00} = 0$. Thus, $\partial u_{0j}/\partial\theta = \partial w_{0j}/\partial\theta = \partial p_{0j}/\partial\theta = 0$, for all j, and also, $\partial p_{1-1}/\partial r = \partial p_{1-1}/\partial\theta = 0 = \partial p_{2-1}/\partial r = \partial p_{2-1}/\partial\theta$. The equations of O(1) with $\delta_1 = 0$ corresponds to the Poiseuille flow. The differential equations of various orders of ε and δ can be obtained by substituting expansions (9a) and (9b) into Eqs. (2a)–(2d) and equating the coefficients of various orders on both sides. Details of these calculations are in Dash *et al.* [13].

2.2. *Numerical approach*

The three-dimensional non-linear elliptic partial differential equations (2a)–(2d) are not amenable directly to numerical solution. So, we simplify these equations through the following steps:

(I) We define the characteristic velocity as $U_0 = \mu/\rho a$ so that the Reynolds number Re defined by Eq. (4) becomes unity.

(II) Since the centrifugal force terms drive the secondary motion, we need to rescale the velocities to make the centrifugal force terms to be of the same order of magnitude as the viscous and inertial terms. This is accomplished through the transformation $(u, v, w) \rightarrow (u, v, (2\varepsilon)^{-1/2}w)$.

(III) We assume that the radius of curvature of the outer pipe is large compared to its mean radius (i.e. $\varepsilon = a/b \ll 1$) so that the terms of $O(\varepsilon)$ and higher order terms in ε can be neglected. The effect of curvature is taken into account through the terms of $O(\varepsilon^{1/2})$. This is the loosely coiled approximations in curved pipe flows [3, 15, 16, 46].

(IV) Again, we assume that the length of the constriction is very large as compared to the mean radius of the outer pipe (i.e. $\delta = a/L \ll 1$) so that the terms of $O(\delta)$ and higher order terms in δ can also be neglected compared to the terms of $O(1)$. Nevertheless, the effect of constriction is taken into consideration through the no-slip boundary condition (6) at the outer wall defined by Eq. (7) in the constricted region $0 \leq z \leq 1$. This is, in fact, the order of magnitude approach of Young [57] modified for the flow characteristics in a curved constricted pipe/annulus. This assumption makes the governing equations locally two-dimensional and axial co-ordinate z appears as a parameter in the problem.

Under the above assumptions, the pressure field can be approximated by

$$p \approx \frac{1}{\delta} G f(z) + p_1(r, \theta, z),$$

where G is a constant and $f(z)$ is an unknown function to be determined by using the constant flux condition; $f(z) = z$ when $\delta_1 = 0$. Then the governing Eqs. (2a)–(2d) of motion are reduced to

$$\nabla_0 u - \frac{v^2}{r} - \frac{w^2}{2}\cos\theta = -\frac{\partial p_1}{\partial r} + \Delta_0 u - \frac{u}{r^2} - \frac{2}{r^2}\frac{\partial v}{\partial \theta}, \tag{10a}$$

$$\nabla_0 v + \frac{uv}{r} + \frac{w^2}{2}\sin\theta = -\frac{1}{r}\frac{\partial p_1}{\partial \theta} + \Delta_0 u - \frac{v}{r^2} + \frac{2}{r^2}\frac{\partial u}{\partial \theta}, \tag{10b}$$

$$\nabla_0 w = D\frac{df}{dz} + \Delta_0 w, \tag{10c}$$

$$\frac{\partial u}{\partial r} + \frac{u}{r} + \frac{1}{r}\frac{\partial v}{\partial \theta} = 0, \tag{10d}$$

where $\nabla_0 = \nabla_\delta$ at $\delta = 0$, $\Delta_0 = \Delta_\delta$ at $\delta = 0$, and $D = (2\varepsilon)^{1/2}G = 4(2\varepsilon)^{1/2}\,\mathrm{Re}^s$ is the Dean number. Here, $\mathrm{Re}^s = G/4$ is the Reynolds number defined with respect to the centerline velocity in a Poiseuille flow. D is regarded as the dynamical similarity parameter for curved pipe flows and is a measure of secondary flow. Now if we introduce the secondary stream function ψ and the vorticity function Ω defined through

$$u = \frac{1}{r}\frac{\partial \psi}{\partial \theta}, \quad v = -\frac{\partial \psi}{\partial r}, \quad \text{and} \quad \Omega = \frac{\partial v}{\partial r} + \frac{v}{r} - \frac{1}{r}\frac{\partial u}{\partial \theta}, \tag{11}$$

then the equation of continuity (10d) is identically satisfied and the momentum equations (10a)–(10c) are reduced to

$$\frac{\partial^2 \psi}{\partial r^2} + \frac{1}{r}\frac{\partial \psi}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \psi}{\partial \theta^2} = -\Omega, \tag{12a}$$

$$\frac{\partial^2 \Omega}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 \Omega}{\partial \theta^2} + 2\mu(r,\theta)\frac{\partial \Omega}{\partial \theta} + 2\lambda(r,\theta)\frac{\partial \Omega}{\partial r} = w\left[\frac{\partial w}{\partial r}\sin\theta + \frac{1}{r}\frac{\partial w}{\partial \theta}\cos\theta\right], \tag{12b}$$

$$\frac{\partial^2 w}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 w}{\partial \theta^2} + 2\mu(r,\theta)\frac{\partial w}{\partial \theta} + 2\lambda(r,\theta)\frac{\partial w}{\partial r} = -D\frac{df}{dz}, \tag{12c}$$

where

$$\lambda(r,\theta) = \frac{1}{2r}\left(1 - \frac{\partial \psi}{\partial \theta}\right), \quad \text{and} \quad \mu(r,\theta) = \frac{1}{2r}\frac{\partial \psi}{\partial r}. \tag{12d}$$

The boundary conditions (6) and (8) are reduced to

$$w = \psi = \frac{\partial \psi}{\partial r} = 0, \quad \text{at } r = \eta(z) \quad \text{and} \quad r = k, \tag{13a}$$

$$\frac{\partial w}{\partial \theta} = 0, \quad \psi = 0 = \Omega, \quad \text{at } \theta = 0 \quad \text{and} \quad \theta = \pi, \tag{13b}$$

$$w(r,-\theta) = w(r,\theta), \quad \psi(r,-\theta) = -\psi(r,\theta), \quad \Omega(r,-\theta) = -\Omega(r,\theta). \tag{13c}$$

As mentioned before, df/dz is an unknown function of z which will be determined by using the constant flux condition given by

$$\frac{(2\varepsilon)^{-1/2}}{\pi}\int_k^{\eta(z)}\int_0^{2\pi} w(r,\theta,z)r\,dr\,d\theta = Q(k,D), \tag{14}$$

where $Q(k,D)$ is the flow rate in a curved annulus without constriction. In view of the symmetric condition (13b)–(13c), it is necessary to determine the flow field only within the semi-annular region $k \le r \le \eta(z)$, $0 \le \theta \le \pi$. The simplified equation of motion (12a)–(12d), and the

boundary conditions (13a)–(13d) are considered in Collins and Dennis [11]. But, those were for flow in a non-constricted curved pipe in which $f(z) = z$. In contrary, the present analysis deals with the flow in a constricted curved annulus and $f(z)$ is required to be determined using the constant flux condition (14). So, the flow variables will depend on the axial distance z and radii ratio k through the boundary condition (13a) in addition to their dependence on Dean number D. Details of the computational procedure are given in Jayaraman and Dash [31].

3. Discussions

Since the heart surface is usually curved, the coronary arteries also tend to be curved as they follow the surface contour of the heart. The radius of curvature "b" of the coronary arteries is about 10 times their radius "a", and therefore, the value of the curvature parameter, $\varepsilon = a/b$, is about 0.1. The average radius of the coronary arteries is about 1.5 mm. Although the mean Reynolds number in the coronary arteries is about 150 under resting condition under the pathological situation of a stenosis present, the mean Reynolds number can be even as small as 50. The values of density ρ and kinematic viscosity ν are assumed to be $1.05\,\mathrm{gm/cm^3}$ and $0.035\,\mathrm{cm^2/sec}$ respectively.

For our mathematical analysis, based on perturbation method, which is valid for small values of the geometric parameter δ, we fix it at 0.1. The stenotic parameter δ_1 is also fixed at 0.1. These values correspond to a typical mild stenosis in which the stenosis has a maximum height of 10% of the radius with a length of 100 times that of the height. This case, though it corresponds to a slowly varying mild stenosis, is expected to give an insight into the actual situation in which the values of δ and δ_1 are larger.

The angiographic data on coronary artery shows that the proximal vessel diameter ranged about from 2 to 4.7 mm and therefore, for an angioplasty catheter (1.4 mm diameter), a rough range of values for k of interest is about from 0.3 to 0.7 Back [1]. For smaller infusion catheters 0.66 mm diameters as used by Ganz *et al.* [23], this range is even smaller and is about from 0.14 to 0.33. Therefore, in our calculations based on numerical scheme, the results are obtained for different values of radii ratio $0.1 \leq k \leq 0.7$ and Dean number $50 \leq D \leq 2000$ based on 11 grid points (i.e. $N_1 = 10$) in radial direction and 19 grid points (i.e. $N_2 = 18$) in azimuthal direction. Since our primary goal is towards the application of the model to blood flow in a catheterized stenosed artery, we have not done grid independent

test of the results. As discussed in Collins and Dennis [11], 11×19 number of grid points should be enough to obtain the desired accurate results.

3.1. *Pressure drop and impedance*

We define the pressure drop Δp_z over a stenotic length z, averaged over a cross-section, by

$$\Delta pz = \frac{\int_0^z \int_0^{2\pi} \int_k^\eta \left(-\frac{\partial p}{\partial z}\right) r\, dr\, d\theta\, dz}{\int_0^{2\pi} \int_k^\eta r\, dr\, d\theta}. \tag{15}$$

Table 1 shows the comparison of the maximum pressure drop $\Delta p_{\max}$ across a stenosis with the pressure drop Δp over the whole stenotic length (i.e. Δp_z at $z = 1$) for $\varepsilon = 0.1$, $\delta = 0.1$ and $\delta_1 = 0.1$, and different values of k and Re. From the table, the increase in the actual pressure drop due to the curvature and stenosis as well as the catheterization can be estimated. The estimated increased pressure drop due to catheterization can be used to find the error involved in the measured pressure gradient using catheters.

Table 2 shows the ratio of the pressure drop in a catheterized artery to that in an uncatheterized artery as a function of catheter radius k and Reynolds number Re for $\delta = 0.1$, $\varepsilon = 0$ and $\varepsilon = 0.1$, $\delta_1 = 0$ (without stenosis) and $\delta_1 = 0.1$ (with stenosis). In the absence of curvature and stenosis, the pressure drop is inversely proportional to the Reynolds number Re. The pressure drop ratio can then be obtained as

$$\frac{\Delta p_c}{\Delta p_u} = \frac{\ln k}{(1-k^4)\ln k + (1-k^2)^2} \tag{16}$$

Table 1. Comparison of the maximum pressure drop (Δp max) across a stenosis with the pressure drop (Δp) over the whole stenotic length for $\varepsilon = 0.1$, $\delta = 0.1$ and $\delta_1 = 0.1$, and different values of k and Re.

K	Re = 25		Re = 50		Re = 75		Re = 100	
	$\Delta p_{\max}$	Δp	$\Delta p_{\max}$	Δp	$\Delta p_{\max}$	Δp	$\Delta p_{\max}$	Δp
0.0	2.25	2.23	1.13	1.09	0.74	0.69	0.57	0.50
0.1	3.69	3.64	1.89	1.74	1.54	1.09	1.31	0.78
0.2	5.14	5.08	2.94	2.44	2.34	1.52	2.05	1.08
0.3	7.72	7.42	5.04	3.57	4.23	2.23	3.83	1.59
0.4	13.77	11.64	9.97	5.63	8.80	3.52	8.18	2.52
0.5	30.38	20.44	24.29	9.92	22.23	6.24	21.26	4.48

Table 2. The ratio of pressure drop in a catheterized artery to that in an uncatheterized artery ($\Delta p_c/\Delta p_u$) as a function of Re and k for $\delta = 0.1$, $\varepsilon = 0$ and $\varepsilon = 0.1$, $\delta_1 = 0$ and $\delta_1 = 0.1$.

K	$\varepsilon = 0.1, \delta_1 = 0.1$	$\varepsilon = 0.1,\ \delta_1 = 0.1$			
	For all Re	Re = 25	Re = 50	Re = 75	Re = 100
0.0	1.00	1.00	1.00	1.00	1.00
0.1	1.74	1.63	1.60	1.58	1.56
0.2	2.35	2.28	2.24	2.20	2.17
0.3	3.29	3.33	3.28	3.23	3.18
0.4	4.89	5.22	5.16	5.10	5.03
0.5	7.94	9.17	9.10	9.04	8.96

where the subscripts "c" and "u" refers to a "catheterized" and an "uncatheterized" artery respectively. Thus, the pressure drop ratio is independent of the Reynolds number Re and depends only on the catheter radius k, as seen in Table 2 for $\varepsilon = 0$ and $\delta_1 = 0$, which agrees with the estimation of increased mean pressure drop obtained by Back [1]. In the presence of curvature and stenosis with $\varepsilon = 0.1$ and $\delta_1 = 0.1$, the pressure drop over the whole stenotic length is seen to be higher than the corresponding drop in pressure over an unit length in the absence of curvature and stenosis. Again, the insertion of a catheter into the artery leads to a considerable increase in their magnitudes. The increase in the pressure drop due to catheterization depends on the catheter radius k as well as the Reynolds number Re.

The impedance or frictional resistance FR_z over a stenotic length z is defined by

$$FR_z = \frac{\Delta p_z}{zQ} \tag{17}$$

where Δp_z is the pressure drop over the stenotic length z and Q is the total flow rate over a cross-section which is taken as unity. This, in fact, gives a measure of reduction to blood flow rate for a given pressure gradient, and hence, may be interpreted as the resistance to blood flow (frictional resistance per unit length; Back *et al.* [2]) offered by the stenosis.

The axial variation of frictional resistance FR_z for $D = 1000$ and different values of k is shown in Fig. 6(A), while the variation of frictional resistance FR_z with the radii ratio k at the entrance ($z = 0$) and exit ($z = 1$) of stenosis for different values of the Dean number D ($D = 100$ and $D = 2000$) is shown in Fig. 6(B). The frictional resistance FR in a curved annulus without constriction (i.e. for $\delta_1 = 0$ or $\eta = 1$) corresponds to the value of FR_z at $z = 0$. It is seen from Fig. 6(A) that the frictional resistance

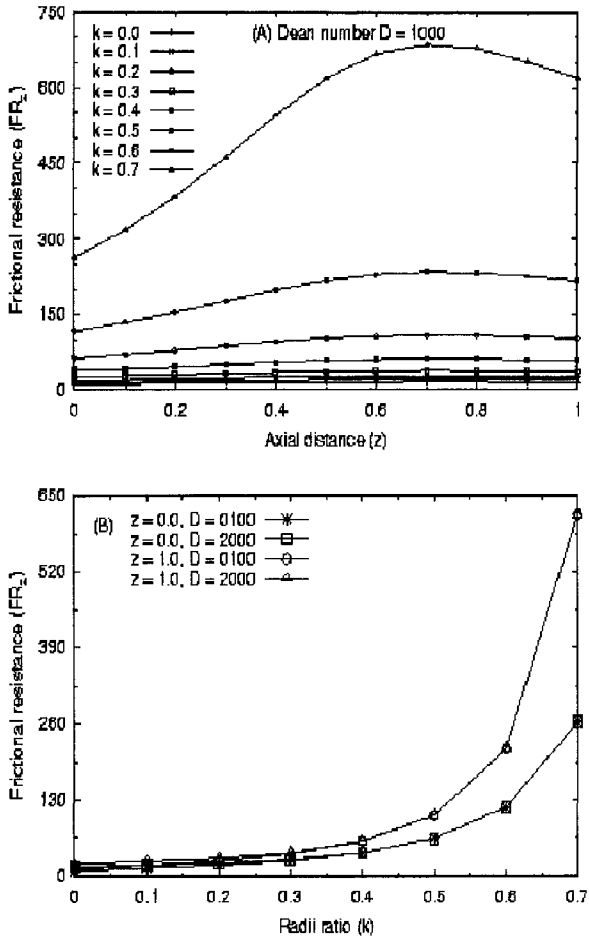


Fig. 6. (A) Axial variation of frictional resistance FR_2 in the presence of constriction with constricted parameter $\delta_1 = 0.1$, curvature parameter $\varepsilon = 0.1$, Dean number $D = 1000$, and different values of radii ratio k; (B) Variation of FR_2 with k at the entrance $z = 0$ and exit $z = 1$ of stenosis for $\delta_1 = 0.1, \varepsilon = 0.1$, and different values of D.

FR_z in a constricted curved tube (i.e. for $k = 0$) does not vary much over the length of the constriction. But, in a constricted curved annulus with relatively higher value of radii ratio k, it varies significantly over the length of the constriction. It is further depicted that, the frictional resistance FR_z in the downstream of the constriction is higher than the corresponding value in the upstream. It is observed from Fig. 6(B) that, the frictional resistance FR_z does not vary much with the Dean number D, and it becomes almost

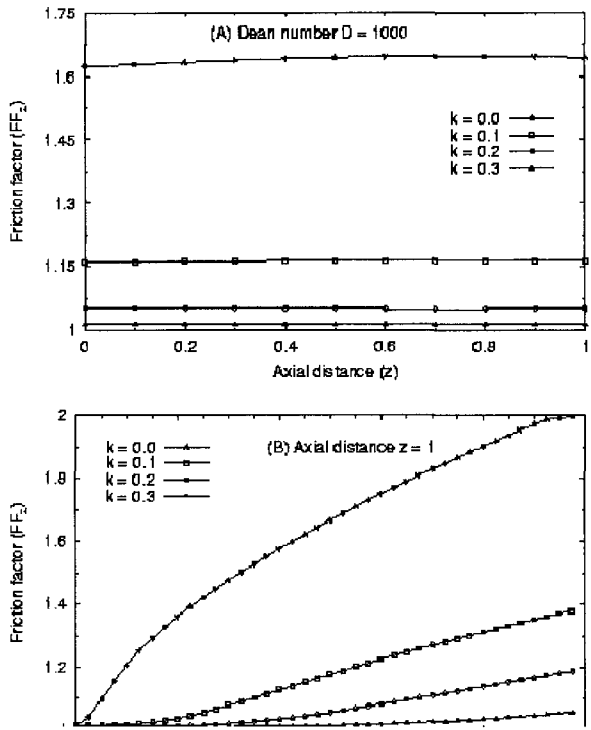


Fig. 7. (A) Axial variation of frictional factor FF_2 in the presence of constriction with constricted parameter $\delta_1 = 0.1$, curvature parameter $\varepsilon = 0.1$, Dean number $D = 1000$, and different values of radii ratio k; (B) Variation of FF_2 with D at the exit $z = 1$ of stenosis for $\delta_1 = 0.1, \varepsilon = 0.1$, and different values of D.

independent of D (i.e. the effect of curvature is nullified) for higher values of radii ratio k (e.g. for $k \geq 0.4$). It is further depicted that, for higher values of k, the frictional resistance in the presence of constriction (i.e. FR_z at $z = 1$) is considerably higher than the corresponding value in the absence of constriction (i.e. FR_z at $z = 0$). However, the frictional resistance FR_z increases with the increase in the value of radii ratio k.

It is inferred from our present results that the insertion of a catheter into an artery leads to an increase in the frictional resistance. The factor by which frictional resistance increases due to catheterization can be estimated by obtaining the ratio of frictional resistance in a catheterized artery to that in an uncatheterized artery.

The comparison of frictional resistance ratio FRR in a curved catheterized artery ($\varepsilon = 0.1$) with that in a straight catheterized artery

Table 3. Frictional resistance ratio FRR in a straight ($\varepsilon = 0$) and curved ($\varepsilon = 0.1$) catheterized artery without stenosis ($\delta_1 = 0$) (i.e. FRR_z at $z = 0$) for different values of catheter radius k and Dean number D.

$K \backslash D$	$\varepsilon = 0.1,\ \delta_1 = 0.1$	$\varepsilon = 0.1,\ \delta_1 = 0.1$				
	For all Re	$D = 100$	$D = 500$	$D = 1000$	$D = 1500$	$D = 2000$
0.1	1.741	1.705	1.323	1.246	1.220	1.202
0.2	2.349	2.289	1.732	1.523	1.445	1.406
0.3	3.289	3.201	2.414	2.059	1.865	1.756
0.4	4.894	4.760	3.590	3.049	2.723	2.499
0.5	7.938	7.719	5.821	4.943	4.408	4.033
0.6	14.586	14.182	10.695	9.082	8.099	7.409
0.7	32.611	31.709	23.910	20.304	18.105	16.563

Table 4. Frictional resistance ratio FRR in a straight ($\varepsilon = 0$) and curved ($\varepsilon = 0.1$) catheterized artery with stenosis ($\delta_1 = 0.1$) (i.e. FRR_z at $z = 1$) for different values of catheter radius k and Dean number D.

k/D	$\varepsilon = 0.1,\ \delta_1 = 0.1$	$\varepsilon = 0.1,\ \delta_1 = 0.1$				
	For all Re	$D = 100$	$D = 500$	$D = 1000$	$D = 1500$	$D = 2000$
0.1	1.779	1.734	1.339	1.261	1.233	1.230
0.2	2.461	2.389	1.797	1.574	1.489	1.466
0.3	3.571	3.463	2.596	2.206	1.990	1.893
0.4	5.596	5.424	4.066	3.442	3.066	2.845
0.5	9.815	9.511	7.189	6.034	5.368	4.970
0.6	20.491	19.854	14.882	12.596	11.206	10.373
0.7	58.266	56.452	42.315	35.813	31.862	29.495

($\varepsilon = 0$) corresponding to $\delta_1 = 0$ (without stenosis) and $\delta_1 = 0.1$ (with stenosis) and different values of k and D are shown in Tables 3 and 4, respectively. It is observed that the frictional resistance ratio FRR increases with the increase in value of radii ratio k. Again, in a curved catheterized artery, FRR is smaller than the corresponding value in a straight catheterized artery, and it decreases further with the increase in Dean number D. Thus, these results indicate that the increase in the frictional resistance (or equivalently, the increase in the pressure gradient at a constant flow rate) due to catheterization at a higher value of D is less than that at a lower value of D. For $k = 0.5$, the frictional resistance in the catheterized artery without stenosis (i.e. for $\delta_1 = 0$ or $\eta = 1$) is about 5.8 times of the value in the uncatheterized artery at $D = 500$ and 4 times of the value in the uncatheterized artery at $D = 2000$. In the presence of stenosis with $\delta_1 = 0.1$, this increase factor

(7.2 at $D = 500$ and 4.97 at $D = 2000$) is even higher than the corresponding value in the absence of stenosis for which $\delta_1 = 0$ or $\eta = 1$.

The friction factor FF in a curved tube (i.e. for $k = 0$), is considerably higher than the corresponding value in a curved annulus. For $D = 2000$, the friction factor FF in a curved tube (i.e. for $k = 0$) is about 2 (implying a 50% reduction in flow rate due to curvature effect). But, for the same value of D, the friction factor FF in a curved annulus with the value $k = 0.1$ is about 1.35 (implying a 26% reduction in flow rate due to curvature effect).

For higher values of k, the friction factor decreases further, and for $k \geq 0.4$, it becomes almost independent of Dean number D, implying that the curvature effect is almost nullified for $k \geq 0.4$. It is further depicted that, for all values of $k \geq 0.1$, the variation of friction factor FF with the Dean number D is insignificant compared to that in a curved tube (i.e. for $k = 0$) whenever $D \leq 500$.

3.2. *Wall shear stress*

The non-dimensional wall shear stress (shear stress non-dimensionalized with respect to $\rho(\mu/(\rho a)^2)$ at any axial point z is approximated by

$$\tau_z \approx -(2\varepsilon)^{-1/2} \left.\frac{\partial w}{\partial r}\right|_{r=1}. \tag{18}$$

It is evaluated numerically using four-point backward difference formula. The axial variation of wall shear stress τ_z for $D = 1000$, $D = 2000$ and different values of radii ratio k is shown in Fig. 8(A, B), where as the variation of wall shear stress τ_z with the Dean number D at the entrance $z = 0$ and peak $z = 0.5$ of stenosis for different values of radii ratio k is shown in Fig. 8(C, D). The wall shear stress τ in a curved annulus without constriction (i.e. for $\delta_1 = 0$ or $\eta = 1$) corresponds to the value of τ_z at $z = 0$ or $z = 1$. It is seen from Fig. 8(A, B) that, for smaller values of radii ratio k, the wall shear stress τ_z varies markedly along the length of the stenosis. It is further depicted that the wall shear stress τ_z increases with the increase in Dean number D but decreases with the increase in radii ratio k. For a fixed pressure gradient G, or equivalently for a fixed value of D, increase in the value of k results in the decrease in flow rate Q), which in fact results in the decrease in wall shear stress τ_z. But, if the flow rate Q is maintained to be constant (independent of k), then the increase in the value of k would result in an increase in the pressure gradient G which would, in fact, lead to an increase in the wall shear stress τ_z. It is seen from Fig. 8(A, B) that the wall shear stress τ_z remains positive over the entire stenotic length. Thus,

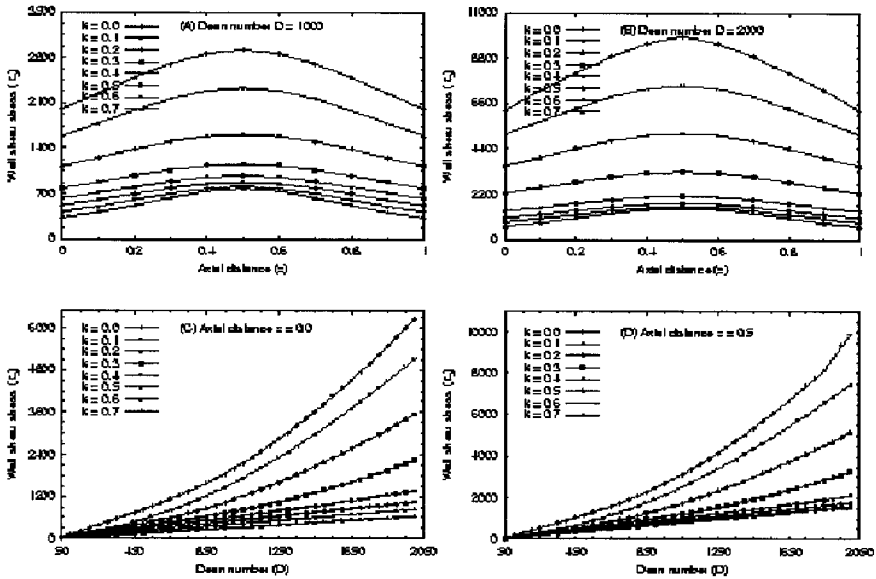


Fig. 8. (A, B) Axial variation of wall shear stress τ_2 in the presence of constriction with constricted parameter $\delta_1 = 0.1$, curvature parameter $\varepsilon = 0.1$, Dean number $D = 1000$, and $D = 2000$, and different values of radii ratio k; (C, D) Variation of τ_2 with D at the entrance $z = 0$ and peark $z = 0.5$ of stenosis for $\delta_1 = 0.1$ $\varepsilon = 0.1$, and different values of k.

this analysis could not detect the point of separation in the downstream of the flow field. To capture the separation points, the governing equations have to include all the neglected terms, and a better numerical procedure has to be adopted.

3.3. *Flow behavior*

Figure 9 shows the secondary streamlines (ψ = constant) in $r - \theta$ plane of a curved annulus without constriction (i.e. for $\delta_1 = 0$ or $\eta = 1$) for (a) $k = 0.1$, $D = 1000$, (b) $k = 0.1$, $D = 2000$, (c) $k = 0.3$, $D = 1000$, (d) $k = 0.3$, $D = 2000$, (e) $k = 0.5$, $D = 1000$, and (f) $k = 0.5$, $D = 2000$. It is observed that the streamline pattern divides each half of the cross-sectional plane into two parts forming two loops which is in contrast to the streamlines pattern in a curved tube where only one loop formation occurs unless a dual solution exists [3, 11, 25, 37, 46]. The loop near the inner wall is smaller for lower values of radii ratio k. But, as the value of radii ratio k

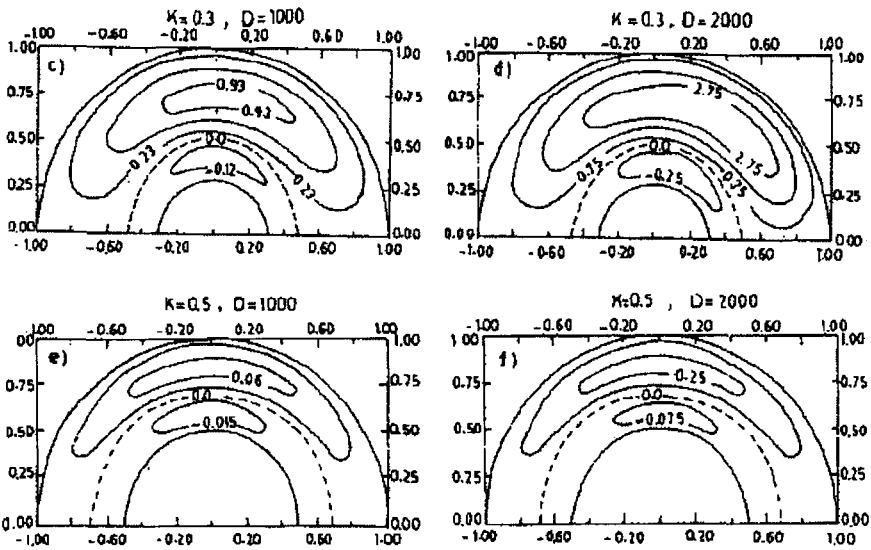


Fig. 9. Secondary streamlines.

increases, the loop near the inner wall becomes larger and the loop near the outer wall becomes smaller.

4. Concluding Remarks

The example of flow through a curved annulus with a local constriction at the outer wall brings out many interesting fluid mechanical phenomena due to the effect of flow geometry (inner wall radius, outer wall variation and curvature) as well as the dynamics of flow governed by the Dean number (dynamical similarity parameter) D. These results can be used to estimate the increase in frictional resistance or pressure drop in an artery when a catheter is inserted into it. It is found that, because of the curvature, the increase in frictional resistance due to catheterization depends on the catheter size (radii ratio k) as well as the Dean number D. In the absence of constriction and depending on the value of k ranging from 0.1 to 0.7, the frictional resistance increases by a factor ranging from 1.32 to 23.91 for $D = 500$ and 1.20 to 16.56 for $D = 2000$. But, in the presence of constriction and with the same range for k, the increase in frictional resistance is by a factor ranging from 1.34 to 42.32 for $D = 500$ and 1.18 to 29.5 for $D = 2000$. These estimate for the increased frictional resistance can be used to correct the error involved in the measured pressure gradients using catheters.

The study gives a lot of hope that fluid dynamic principles can be used effectively to model physiological phenomena as well as the procedures in clinical medicine. The ultimate success in making these sort of studies as a part of the clinical medicine or procedures will require a wholehearted interdisciplinary research involving engineers, physiologists, applied mathematicians and physicists.

References

[1] L. H. Back, Estimated mean flow resistance increase during coronary artery catheterization, *J. Biomech.* **27** (1994) 169–175.

[2] L. H. Back, E. Y. Kwack and M. R. Back, Flow rate-pressure drop relation in coronary angioplasty: Catheter obstruction effect, *J. Biomech. Eng. Trans. ASME.* **118** (1996) 83–89.

[3] S. A. Berger, L. Talbot and L. S. Yao, Flow in curved pipes, *Ann. Rev. Fluid Mech.* **15** (1983) 461–512.

[4] L. Bjorno and H. Pattersson, Hydro- and hemodynamic effects of catheterization of vessels — I. An experimental model, *Acta Radiol. Diagnosis* **17** (1976a) 511.

[5] L. Bjorno and H. Pattersson, Hydro- and hemodynamic effects of catheterization of vessels — II. Model experiments comparing circular and annular human area reduction, *Acta Radiol. Diagnosis* **17** (1976b) 749–762.

[6] C. G. Caro, J. M. Fitz-Gerland and R. C. Schroter, Atheroma and arterial wall shear: Observation correlation and proposal of a shear dependent mass transfer mechanism for atherogenesis, *Proc. R. Soc. London B* **177** (1971) 109–159.

[7] C. G. Caro, T. J. Pedley, R. C. Schroter and W. A. Seed, *The Mechanics of Circulation* (Oxford University Press, 1978).

[8] L. J. Chang and J. M. Tarbell, Numerical simulation of fully-developed sinusoidal and pulsatile (physiological) flow in curved tubes, *J. Fluid Mech.* **161** (1985) 175–198.

[9] L. J. Chang and J. M. Tarbell, A numerical study of flow in curved tubes simulating coronary arteries, *J. Biomech.* **21** (1988) 927–937.

[10] C. Y. Chow, *An Introduction to Computational Fluid Dynamics*, Chap. 4 (John Wiley and Sons, New York, 1979).

[11] W. M. Collins and S. C. R. Dennis, The steady motion of a viscous fluid in a curved tube, *Quart. J. Mech. Appl. Math.* **28** (1975) 133–156.

[12] R. K. Dash, G. Jayaraman and K. N. Mehta, Estimation of increased flow resistance in a narrow catheterized artery, *J. Biomech.* **29** (1996) 917–930.

[13] R. K. Dash, G. Jayaraman and K. N. Mehta, Flow in a catheterised curved artery with stenosis, *J. Biomech.* **32** (1999) 49–61.

[14] P. Daskopoulos and A. M. Lenhoff, Flow in curved ducts: Bifurcation structure for stationary ducts, *J. Fluid Mech.* **203** (1989) 125–148.

[15] W. R. Dean, Note on the motion of fluid in a curved pipe, *Phil. Mag. J. Sci.* **4** (1927) 208–223.

[16] W. R. Dean, The streamline motion of fluid in a curved pipe, *Phil. Mag. J. Sci.* **5** (1928) 673–693.

[17] S. C. R. Dennis and G. Z. Chang, Numerical integration of the Navier-Stokes equations for steady two-dimensional flow, *Phys. Fluids* Supplement II, **12** (1969) 88–93.

[18] S. C. R. Dennis, Calculation of the steady flow through a curved tube using a new finite-difference method, *J. Fluid Mech.* **99** (1980) 449–467.

[19] S. C. R. Dennis and M. Ng, Dual solutions for steady laminar flow through a curved tube, *Quart. J. Mech. Appl. Math.* **35** (1982) 305–324.

[20] M. A. Ebadian, Rate of flow in a concentric pipe of circular cross-section, *J. Appl. Mech. Trans. ASME.* **57** (1990) 1073–1076.

[21] D. Fargie and B. W. Martin, Developing laminar flow in a pipe of circular cross-section, *Proc. Roy. Soc. London A* **321** (1971) 461–476.

[22] D. L. Fry, Mass transport, atherogenesis and risk arteriosclerosis, *Arteriosclerosis* **7** (1987) 88–100.

[23] P. Ganz, R. Abben, P. L. Friedman, J. D. Garnic, W. H. Barry and D. C. Levin, Usefulness of transstenotic coronary pressure gradient measurements during diagnostic catheterization, *Am. J. Cardiol.* **55** (1985) 910–914.

[24] D. P. Giddens, C. K. Zarins and S. Glagov, The role of fluid mechanics in the localization and detection of atherosclerosis, *J. Biomech. Eng.* **115** (1993) 588–594.

[25] A. D. Greenspan, Secondary flow in a curved tube, *J. Fluid Mech.* **57** (1973) 167–176.

[26] H. Ito, Flow in curved pipes, *JSME Int. J. II* **30** (1987) 543–552.

[27] R. Jain and G. Jayaraman, On the steady laminar flow in a curved pipe of varying elliptic cross-section, *Fluid Dynamics Research* **5** (1990) 351–362.

[28] G. Jayaraman, M. P. Singh, N. Padmanabhan and A. Kumar, Reversing flow in the aorta — a theoretical model, *J. Biomech.* **17** (1984) 479–490.

[29] G. Jayaraman, R. Mehrotra and N. Padmanabhan, Entry flow into a circular cylinder of varying cross-section, *Fluid Dynamic Research* **1** (1986) 131–144.

[30] G. Jayaraman and K. Tiwari, Flow in a catheterised curved artery, *Med. Biol. Eng. Comput.* **33** (1995) 720–724.

[31] G. Jayaraman and R. K. Dash, Numerical study of flow in a constricted curved annulus: An application to flow in a catheterized artery, *J. Eng. Maths* **40(4)** (2001) 355–375.

[32] H. C. Kao, Some aspects of bifurcation structure of laminar flow in curved ducts, *J. Fluid Mech.* **243** (1992) 519–539.

[33] G. T. Karahalios, Some possible effects of a catheter on the arterial wall, *Med. Phys.* **17** (1990) 922–925.

[34] J. S. Lee and Y. C. Fung, Flow in locally constricted tube at low Reynolds numbers, *J. App. Mechs. Trans. ASME* **37** (1970) 9–16.

[35] M. J. Lighthill, Physiological fluid dynamics: A survey, *J. Fluid Mech.* **52** (1972) 475–497.

[36] T. S. Lundegren, E. M. Sparrow and J. B. Starr, Pressure drop due to the entry region in ducts of arbitrary cross-section, *J. Basic Eng. Trans. ASME* **86** (1964) 620–629.

[37] D. J. McConalogue and R. S. Srivastava, Motion of fluid in a curved tube, *Proc. Roy. Soc. Lond. A* **307** (1968) 37–53.
[38] D. A. Macdonald, On steady flow through modelled vascular stenosis, *J. Biomech.* **12** (1979) 13–20.
[39] D. A. Macdonald, *Blood Flow in Arteries* (Arnold, London, 1974).
[40] P. A. J. Mees, K. Nandakumar and J. H. Masliyah, Instability and transitions of flow in a curved squire duct: The development of two pairs of Dean vortices, *J. Fluid Mech.* **314** (1996) 227–246.
[41] R. Mehrotra, G. Jayaraman and N. Padmanabhan, Pulsatile blood flow in a stenosed artery — a theoretical model, *Med. Biol. Eng. Comput.* **23** (1985) 55–62.
[42] B. E. Morgan and D. F. Young, An integral method for the analysis of flow in arterial stenosis, *Bull. Math. Bio.* **36** (1974) 39–53.
[43] H. Niimi, Y. Kawano and I. Sugiyama, Structure of blood flow through a curved vessel with an aneurysm, *Biorheology* **21** (1984) 603–615.
[44] N. Padmanabhan and G. Jayaraman, Flow in a curved tube with constriction — an application to the arterial system, *Med. Biol. Eng. Comput.* **22** (1984) 216–224.
[45] T. J. Pedley, High Reynolds number flow in tubes of complex geometry with application to wall shear stress in arteries, *Biological Fluid Dynamics* (the Society for Experimental Biology, 1995), pp. 219–239.
[46] T. J. Pedley, *The Fluid Mechanics of Large Blood Vessels* (Cambridge University Press, London, 1980).
[47] A. Sarkar and G. Jayaraman, Correction to flow rate — pressure drop relation in coronary angioplasty: Steady streaming effect, *J. Biomech.* **31** (1998) 781–791.
[48] A. Sarkar and G. Jayaraman, Nonlinear analysis of oscillatory flow in the annulus of an elastic tube — application to catheterized artery, *Phys. Fluids* **13** (2001) 2901–2911.
[49] S. Schilt, J. E. Moore, Jr., A. Delfino and J. J. Meister, The effects of time-varying curvature on velocity profile in a model of the coronary arteries, *J. Biomech.* **29** (1996) 469–474.
[50] M. P. Singh, Entry flow in a curved pipe, *J. Fluid Mech.* **65** (1974) 517.
[51] M. P. Singh, P. C. Sinha and M. Agarwal, Flow in the entrance of the aorta, *J. Fluid Mech.* **87** (1978) 197.
[52] F. T. Smith, On entry flow effects in bifurcating, blocked or constricted tubes, *J. Fluid Mech.* **78** (1976b) 709–736.
[53] F. T. Smith, Pipe flows distorted by non-symmetric indentation or branching, *Mathematika* **23** (1976c) 62–83.
[54] L. C. Truesdell and R. J. Adler, Numerical treatment of fully developed laminar flow in helically coiled tubes, *AICHE J.* **16** (1970) 1010–1015.
[55] D. F. Young, Fluid mechanics of arterial stenoses, *J. Biomech. Eng.* **101** (1979) 157–175.
[56] D. F. Young and F. Y. Tsai, Flow characteristics in models of arterial stenosis: I steady flow, *J. Biomech.* **6** (1973) 395–254.
[57] D. F. Young, Effect of a time dependent stenosis on flow through a tube, *J. Eng. Ind. Trans. ASME* **90** (1968) 248–260.

CHAPTER 11

MATHEMATICAL MODELING IN REPRODUCTIVE BIOMEDICINE

SHIVANI SHARMA
Centre for Biomedical Engineering, Indian Institute of Technology New Delhi 110016

SUJOY K. GUHA
School of Medical Science and Technology, Indian Institute of Technology Kharagpur 721302
guha_sk@yahoo.com

Biomechanistic and theoretical models can be applied to investigate complex reproductive processes. The application of Laplace equation to follicle rupture helps in obtaining insight into the simultaneous effect of a number of factors pertaining to bursting of the follicle. Gamete transport dynamics via peristaltic analysis for cilia beat has been outlined for the oviduct. The metachronal wave generated in the wall of the vas deferens contradicts peristalsis to be the dominating factor in spermatozoa transport. Biomechanical characterization of the forces involved in the mechanics of sperm-egg interactions are outlined. Generalized Hooks law to obtain displacements for specific load conditions during fetal head moulding and formulation of a fertilization index based on Von Foerster's equation accounting for both epididymal spermatozoa reserves and spermatozoa numbers for a specific ejaculation frequency are described.

Keywords: Reproductive processes; theoretical models; mathematical analysis biomechanics; spermatozoa.

1. Introduction

Bioengineering is the application of engineering and technology to the problems of biology. Engineering methodology application to reproductive biology attempts to describe quantitatively the different reproductive phenomenon and to assess the effects of different mechanisms. Biomechanics of reproductive biology describes the mathematical modeling of processes and events in reproductive biology such as descriptions of processes of ovulation, models of cell cleavage, models of effects of contractions and cilia on gamete transport, models of contractile pattern in the uterus, relation

of viscosity of sperm transport in the cervix, computer models and simulation of hormone interactions, models of fluid flow in the tract and blood flow in the tract, analysis of sperm motility and relation of sperm motility to energetics and morphology. The mechanics of some reproductive processes, feedback control of spermatozoa maturation and fertilization index formulation are being presented.

2. Mechanics of Ovulation

The biochemical aspects of ovulation have been most extensively studied. However, in order to elucidate fully the mechanism of this complex process such as enlargement and rupture of follicle, it is important to study biomechanical aspects of the process to get an insight into the mechanistic dynamics of ovulation physiology [10]. This is facilitated by the use of mathematical and physical model of mammalian follicle.

Experimental observations of follicular maturation suggest a hypothesis that distensibility of the follicular wall is compatible with constancy in pressure and increase in intrafollicular fluid volume together. Mathematically, bursting of the follicle without a pressure increase suggests the presence of a trigger at the time of rupture. Lardner *et al.* [7] have put forward a simple but elegant model correlating all the above mentioned findings and thoughts. They base the model on a thin shell approximation representation. With this approximation the wall stress l is given by Laplace equation

$$\sigma = \frac{pR}{2t}$$

where p is the internal pressure, R the inner radius of the shell and t is the wall thickness. The hypothesis proposed suggests that the distensibility of the follicular wall material begins to increase near the period of ovulation. An increase in distensibility implies a reduction in the modulus of elasticity. The decrease can be estimated with the help of the equations formed. At the time of ovulation $R/Ro = 1.5$ where Ro is the unstressed radius. Current experimental techniques do not allow a reliable quantitation of the modulus of elasticity of the follicular wall. But some verification of the model can emerge from an estimation of the follicular wall thickness. Indeed by ultrasonography, it is noted that during follicular maturation the follicular wall thickness reduces and there is a rapid decline in the thickness, close to the time of ovulation. Therefore, in spite of the gross approximations involved, the biomechanical model does

help in obtaining an insight into some of the factors pertaining to follicle rupture.

3. Transport of Gametes

The dynamics of transport of gametes is an important aspect of the reproductive process and controlled fertility. Hence, there is a need to understand the intrinsic physiology of the transport process. Major investigative methods of experimental study cannot provide all details of the transport process without significantly altering the physiological process itself. Experimental studies on gamete transport are accomplished through common imaging systems such as ultrasonics, X-rays and magnetic resonance. However, gamete imaging across the walls of reproductive tract cannot be performed. An alternative is to examine the problem from an analytical viewpoint with parameters and boundary conditions being obtained from experimental observations whenever possible. By coupling experimental observation with theoretical modeling, various limitations can be surpassed. Experiments provide information regarding the movement patterns of the reproductive tract. Theoretical model correlates this data with gamete transport mechanism.

3.1. *Ovum transport*

Ovum descends the fallopian tube at speed characteristic to the animal species and it is apparent that ovum does not possess a steady one directional progression. The back and forth motions, periods of hold up at some sites and phases of rapid and slow movement seem to follow a stochastic pattern [13]. However, since the random motion is transient, the motion of the ovum undergoes a deterministic pattern. No intrinsic motility is evident from the structural details of the ovum. The movement of ovum is passive under the influence of external forces. Considering the structure and function of the fallopian tube, the factors which may contribute to ovum propulsion include: drag by secretory fluid flow, cilia action or the wall contractions [1].

The secretory fluid volume being small, theoretically an object suspended in this fluid can be dragged in the direction of flow. Gupta and Sheshadri [5] have analysed peristalsis for a sinusoidal waveform. It is assumed that the tips of cilia in the testis form a sinusoidal envelope. If $p(x,t)$ is the pressure at time t at any point (x,o) along the axis of the tube, and $u(x,r,t)$ is the velocity of fluid within the tube at a point x, at

a distance r from the central axis at time t, the final flow rate equation is given as

$$U(x,r,t) = \frac{2a_0^2 c}{h^4}(h^2 - r^2) - x - \left[\frac{\Phi^2}{2} + \frac{2kx}{a_0}\Phi \sin\frac{2\pi}{\lambda}(x - ct) + 2\Phi \sin\frac{2\pi}{\lambda}(x - ct) + \Phi^2 \sin^2\frac{2\pi}{\lambda}(x - ct)\right]$$

where c is wave velocity of cilia metachronal wave; λ the wavelength; k the taper coefficient; a_0 is the resting radius. The average flow velocity (U) over the ovum is given by

$$U = \frac{1}{2RT}\int_0^T \cdot \int_0^R u(x,r,t)dr\,dt$$

where T is the total time period and R is ovum radius. The instantaneous ovum velocity is approximately equal to U because of the viscous drag. The expressions have been evaluated taking appropriate values: $c = 200\,\mu s^{-1}$; $b = 5\,\mu m$; $\lambda = 500\,\mu m$; $k = -0.04$; $a_0 = 2500\,\mu m$; $T = 2.5\,s$; and $R = 75\,\mu m$. The flow velocity has been calculated to be $0.1\,\mu s^{-1}$. Hence, fluid flow generated by ciliary action cannot be a significant contributor to ovum propulsion.

3.2. *Transport of spermatozoa*

The dynamics of transport mechanism of spermatozoa is an important aspect of the reproductive process and controlled fertility. Spermatozoa as they emerge from the epididymis have already acquired a definite flagellar swimming character but the linear progression on account of this movement is too slow to contribute significantly to the rapid transport required at the time of ejaculation. Guha *et al.* [3] examined the mechanisms involved in the transport of spermatozoa in quiet and ejaculatory state. Based on biomechanical analysis of the morphological structures involved in transport mechanism, the following factors can be identified as possibilities for sources of pressure gradient: pressure exerted by the epididymis; negative pressure or a sucking action produced by the flow of seminal vesicle fluid; relaxation of elastic recoil of the stretched walls of the vas deferens and active contraction of the wall of the vas deferens constricting the lumen. The above factors may operate simultaneously, but the various parameters were initially analysed independently. The fluid transport parameters indicate that peristalsis plays an important role for slow filling of the ampulla in the vas deferens. A strong contraction of the ampullar end was observed

and it was estimated that the volume within the ampulla reduces by 90%. Thus 0.018 ml of spermatic fluid will be expelled for ampullar volume of 0.02 ml, which may either move distally or proximally. At the proximal end is present the long lumen diameter vas deferens main segment, preventing any possibility of spermatic fluid flow due to high flow resistance. However, the wide ejaculatory duct at the distal region allows practically the entire volume to flow into the ejaculatory duct.

For peristaltic analysis of spermatozoa the metachronal wave may be considered to be generated in the wall of the vas deferens [4]. The time period T of the cycle of the peristaltic waves is 7.5 sec, wave velocity $c = 8\,\mathrm{mm\,s^{-1}}$, peristaltic ratio $\phi = 0.2$ and lumen radius $a = 0.16\,\mathrm{mm}$. Using low Reynolds-number infinite-wavelength analysis with appropriate parameter values the maximum flow for no pressure build-up is

$$Q_{\max} = \frac{a^2 c}{2}\left\{\frac{8\phi^2(1-\phi^2/16}{2+3\phi^2}\right\} = 1.56 \times 10^{-5}\,\mathrm{ml/sec}$$

And the maximum pressure per unit wavelength for no flow is

$$\begin{aligned}\Delta P_{\max} &= 0.098\left(\frac{\lambda\eta c}{2\pi a^2}\right)\left[\frac{64\pi\phi^2(1-\phi^2/16}{(1-\phi^2)^{7/2}}\right] \\ &= 0.108\,\mathrm{KN/m^2}.\end{aligned}$$

Thus the reduction in the peristaltic action cannot be considered as a potent factor for the failure of pregnancy following vasectomy. Also, in designing reversible occlusion valves for the vas deferens, the interruption of peristaltic wave at the site of implantation of the device need not be taken as a contraindication for sectioning the vas deferens.

4. Mechanics of Sperm-Egg Interaction

The fusion between sperm and egg plasma membrane, during fertilization has been well studied. However, the molecular mechanism of the fusion process needs further investigation. The mammalian sperm traverses various barriers before fertilization can occur [11]. The mammalian egg is encapsulated by the outermost cumulus and inner zona. Before reaching the egg plasma membrane, a sperm must first bind to the zona, reorient towards the egg, and penetrate the zona. The head of a motile mammalian sperm tends to move in three dimensions. Hence when bound to the zona, it not only pushes on the zona, but also pulls away from it. For the sperm to remain bound, the tensile strength (Fs) of the sperm zona bonds must be

great enough to withstand the maximum pulling force (Fp) exerted by the sperm.

To mimic the mechanics of a motile sperm stuck to the surface of the zona, Baltz and David [6] employed a suction micropipet to exert a suction force (Fs) analogous to the adhesive force (Fp). Since the magnitude of Fs could be manipulated, it was possible to measure the minimum net strength required to tether the sperm to the zona.

The force exerted by a suction pipet on a surface occluding the opening at its tip is

$$Fs = (P_{\text{out}} - P_{\text{in}})A$$

where P_{in} is the pressure inside the pipet, P_{out} is the pressure outside the pipet (i.e. 1 atmosphere) and A is the area of the opening.

They proposed that if a motile sperm is being held on the pipet by suction alone, the suction force exerted by the pipet on the sperm must be greater than the force with which the sperm pulls against the pipet opening. If the suction force is then lowered slowly, the sperm first begins to swim freely when the suction force drops below the maximum pulling force extended by that sperm. Thus using the above equation together with the measured area of the opening, the pressure at which each sperm is first able to swim free yields the maximum pulling force exerted by that sperm. In a separate method the force exerted by a sperm was determined on the basis of flagellar beat parameters (length of the flagellum, beat frequency and beat shape that were measured experimentally).

For motile sperm (at 35° to 37°C), the maximum pulling forces were found to range between 11 and 28 μdyn, with a mean of 20 ± 1.5 μdyn (mean ± SEM, $n = 15$ sperm) using five different pipets whose openings had diameters from 1.0 to 1.5 μm.

5. Fetal Head Molding

There have been various speculations regarding the forces producing molding of the fetal head during labour. Limitations of force measurement data and its reliability and reproduction are a major hindrance in biomechanical analysis of fetal head molding. The phenomena of head molding has then to be considered on the basis of direct experimental data on forces as well as inferences derived from other related observations. The parameters involved include: structure of the system, movement of the structures, measured forces and physical properties of the materials forming the fetal

head and the maternal pelvis. For biomechanical study, Hooks law, that is stress is proportional to strain, is the basis of all analysis.

$$\text{Modulus of Elasticity } E = \frac{Stress}{Strain}.$$

Taking notations in the Cartesian coordinates, in the general case in an elemental volume of a stressed body, there are six components of stress expressed as a vector

$$\{\sigma\}^T = [\sigma_x \sigma_y \sigma_{z\tau} \tau_{xy} \tau_{yz} \tau_{zx}]$$

where σ_x, σ_y and σ_z are the normal components of stress and τ_{xy}, τ_{yz} and τ_{zx} are the components of shear stress. At a point there are six normal components of strain given by the strain vector

$$\{\varepsilon\}^\Upsilon = [\varepsilon_x \varepsilon_y \varepsilon_z \gamma_{xy} \gamma_{yz} \gamma_{zx}]$$

where ε_x, ε_y and ε_z are the normal strain and γ_{xy}, γ_{yz} and γ_{zx} are the shear strains.

Each of the six stress components may be expressed as a linear function of the six components of strain, to obtain a generalized Hooke's law.

$$\begin{bmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ \tau_{xy} \\ \tau_{yz} \\ \tau_{zx} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16} \\ C_{21} & C_{22} & C_{23} & C_{24} & C_{25} & C_{26} \\ C_{31} & C_{32} & C_{33} & C_{34} & C_{35} & C_{36} \\ C_{41} & C_{42} & C_{43} & C_{44} & C_{45} & C_{46} \\ C_{51} & C_{52} & C_{53} & C_{54} & C_{55} & C_{56} \\ C_{61} & C_{62} & C_{63} & C_{64} & C_{65} & C_{66} \end{bmatrix} \begin{bmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \\ \gamma_{xy} \\ \gamma_{yz} \\ \gamma_{zx} \end{bmatrix}$$

The terms Cs incorporate the relationship between the modulus of elasticity and the Poisson's ratio which is the ratio between the lateral strain and the longitudinal strain. In general case where the material is anisotropic 21 elastic constants come into the picture. The present analysis is simplified considering the orthotropic nature of the material and that the thickness of the skull bone is small in relation to the overall dimension so that plane stress conditions will be applicable. Under these approximations all terms relating to the z direction disappear and stresses and strains are taken in the radial direction (r) and the tangential direction (t). These steps give the simplified expressions of McPherson and Kriewall [9].

$$\begin{bmatrix} \sigma_r \\ \sigma_t \\ \sigma_{rt} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & 0 \\ C_{21} & C_{22} & 0 \\ 0 & CO_0 & C_{33} \end{bmatrix} \begin{bmatrix} \varepsilon_r \\ \varepsilon_t \\ \gamma_{rt} \end{bmatrix}$$

$$C_{11} = \frac{E_r}{1 - \nu_{rt}\nu_{tr}}$$
$$C_{22} = \frac{Et}{1 - \nu_{rt}\nu tr}$$
$$C_{12} = \frac{E_t\nu_{tr}}{1 - \nu_{rt}\nu_{tr}}$$
$$C_{21} = \frac{E_t\nu_{tr}}{1 - \nu_{rt}\nu tr}$$
$$C_{33} = G_{rt}$$

where ν_{rt} is the Poisson's ratio for load in the r direction with lateral contraction in the t direction; and νtr is the Poisson's ratio in the t direction with lateral contraction in the r direction. G_{rt} is the shear modulus, which may be approximated by

$$G_{rt} = \frac{E_r}{2(1 + \nu_{rt})}$$

By symmetry $C_{12} = C_{21}$ and hence

$$\frac{E_r}{\nu_{rt}} = \frac{E_t}{\nu_{tr}}.$$

The equation is solved by matrix inversion to obtain the strain and thereby the displacements for specific load conditions as derived from experimental study during labour.

6. Fertility Index of Spermatozoa

In analysing the contraceptive efficiency of a drug, an index for the fertilizing capability of the aggregate of spermatozoa in the ejaculate is required [3]. Von Foerster's equation can be applied to the mode of sperm maturation in epididymis and sperm number in the ejaculate [12]. Here 'a' is the age of spermatozoa, λ_1 is the loss function, K_1 = proportionality factor, K_2 = factor representing biochemical rate processes and K_3 = factor associated with destruction process of the spermatozoa. If $\eta(t, a)$ is defined as the number of spermatozoa present of age between a and $a + da$ at time t, then the relationship of this function with the loss function, neglecting ejaculatory loss, can be represented as

$$\frac{\partial \eta}{\partial t} + \frac{\partial \eta(t, a)}{\partial a} = -\lambda_1(a) \cdot \eta(t, a).$$

The input into the system is at a constant rate. Therefore the input function $\alpha(t)$ is given by

$$\alpha(t) = \eta(t, o) = K_4$$

where K_4 is the input rate constant and having a typical value of 50 million per day. Thus

$$\eta(t,a) = \alpha(t-a)\exp\left[-\int_{x=0}^{a}\chi(x)dx\right].$$

Since $\alpha(t)$ is time independent function $\eta(t,a)$ also becomes independent of time and may be written as $\eta(a)$. Therefore

$$\eta(a) = K_4\exp\left[-\int_{x=0}^{a}\frac{K_3}{K_2}\left\{1-\frac{1}{1+Kx^2}\right\}dx\right].$$

If no loss of ejaculation occurs, the total number of spermatozoa present in the epididymis at any time t is

$$N(t) = \int_0^t n(a)\cdot da.$$

The quantity is also termed as the epididymal reserve. The integration limit of t is taken because the upper bound of spermatozoa age is the time elapsed from the reference, that is, time equal to zero. At this instant the epididymis is considered as being empty.

The average age of the spermatozoa in the epididymal reservoir at any instant of time t is given by

$$a^{-}(t) = \frac{\int_0^t a\eta(a)da}{\int_0^t \eta(a)da}.$$

As non-integral forms are involved, the expressions were numerically evaluated for discrete time intervals equal to a day. Different values are ascribed to the parameter K, to obtain the results. A value of $K = 0.0004$ meets the physiological requirements that the average age stabilizes at around 10 days. That is, if an ejaculation occurred following stabilization, the average epididymal transit time of the ejaculated spermatozoa would be 10 days. If the epididymis is emptied by two ejaculations per day continued through a week and then the subject is given sexual rest, the epididymis reserve builds up and stabilizes in about three weeks time [8]. For $K = 0.0004$, such characteristic is obtained. Based on these correlation's, it may be concluded that the above given expression with $K = 0.0004$ is appropriate.

7. Conclusion

The use of mathematical modeling is an effective method to link quantitative engineering techniques with qualitative physiological descriptions. The integration of engineering with biology and physiology has invariably

brought new understanding to complex biological processes as is evident from the recent resurgence of utilization of physiological system modeling in most areas of biomedicine.

References

[1] S. Anand and S. K. Guha, Mechanics of transport of the ovum in the oviduct, *Med. Biol. Eng. Comput.* **16** (1978) 256.

[2] S. K. Guha, Bioengg in Reproductive medicine (CRC Press, Boca Raton US, 1990) pp. 4–15; 30; 189–196.

[3] S. K. Guha, H. Kaur and M. A. Ahmed, Formulation of fertilization index for male fertility studies, *Medical Life Science Engg.* **1**(1) (1975) 17–27.

[4] S. K. Guha, H. Kaur and M. A. Ahmed, Mechanics of spermatic fluid transport in the vas deferens, *Medical Biological Engineering* **7** (1975) 518–522.

[5] B. B. Gupta and V. Sheshadri, Peristaltic pumping in nonuniform tubes, *J. Biomechanics* **9** (1976) 105.

[6] J. M. Baltz, D. F. Katz and R. A. Cone, Mechanics of sperm-egg interaction at the zona pellucida, *Biophys. J.* **54** (1988) 643–654.

[7] T. J. Lardner W. J. Shack and P. Y. Tam, A note on the mechanics of ovulation, *J. Theor. Biol.* **48** (1974) 481.

[8] R. M. Levin, J. Latimore, A. J. Wein and K. N. Van Arsdalen, Correlation of sperm count with frequency of ejaculation, *Fertil. Steril.* **45**(5) (1986) 732.

[9] G. K. McPherson and T. J. Kriewall, Fetal head molding: an investigation utilizing a finite element model of the parietal bone, *J. Biomechanics* **13** (1980) 17.

[10] P. Rondell, Biophysical aspects of ovulation, *Biol Reprod. Suppl.* **2** (1970) 64.

[11] M. J. Rowley, F. Teshima and C. G. Heller, Duration of transit of spermatozoa through the male ductular system, *Fertil. Steril.* **21**(5) (1970) 390.

[12] E. Trucco, Mathematical models for cellular systems, The Von Foersters' equation. I, *Bull. Math. Biophys.* **27** (1965) 285.

[13] P. Verdugo, W. I. Lee, S. A. Halbert, R. J. Blandan and P. Y. Tam, A stochastic model for oviductal egg transport, *Biophysics J.* **29** (1980) 257.

CHAPTER 12

IMAGE THEORY AND APPLICATIONS IN BIOELECTROMAGNETICS

P. D. EINZIGER

Department of Electrical Engineering
Technion Israel Institute of Technology Haifa
32000, Israel

L. M. LIVSHITZ and J. MIZRAHI*

Department Biomedical Engineering
Technion Israel Institute of Technology Haifa
32000, Israel
**jm@biomed.technion.ac.il*

Investigation of bioelectricity phenomena has gained recently a steadily increasing interest in medical and engineering applications. This chapter deals with the computational aspects of bioelectromagnetic interactions and their related bioelectric processes, aiming to provide a better physical understanding of the effect of functional electrical stimulation (FES) on biological tissue and to set-up models that can provide quantitative insights into this bioelectromagnetic phenomenon. These goals are achieved here by an explicit image series construction of the macroscopic electromagnetic field within the multilayer tissue. The novel image series expansion scheme, outlined here for quasistatic Green's function in multilayer media, utilizes a unique and explicit recursive representation for Green's function. Our recursive construction convergences under rather general constraints on the media parameters. The usefulness and effectiveness of the proposed analysis is demonstrated through an hybrid scheme, combining image series and moment method procedures, that are capable of handling effectively layered medium problems excited by an electrode array. The inclusion of a collective image term, expressed in a closed form asymptotic evaluation of the series remainder integral, significantly accelerated the image series convergence and the overall algorithm speed and accuracy. This proposed computational procedure can be used as a simple tool for producing analytical data for testing numerical subroutines applied to simulate direct (FES) and inverse (bioelectromagnetic imaging) problems in biomedical applications.

1. Introduction

1.1. *Bioelectromagnetic interaction between electric field and biological tissue: computational aspects*

Knowledge of the potential distribution caused by an electrode array during functional electrical stimulation (FES) (forward problem) or detection of the potential caused by activation of excitable cells (inverse problem) is a very important topic in biomedical engineering [34, 35, 41, 42, 50, 54, 66, 73]. For instance, when surface electrodes are used, for either stimulation or detection, the current has to pass through non-excitable regions such as skin, fat and connective tissues as well as through the actual excitable tissue.

The major difficulties in the application of field theory on biological tissues include: (a) non-homogeneity, i.e. different tissues have different electrical properties; (b) anisotropy, i.e. properties depend on direction of measurement; (c) dispersivity, i.e. properties are frequency dependent [24]; (d) complex electrode/tissue interface [74].

Analytical electromagnetic field theory is well developed for handling problems involving isotropic, homogeneous infinite media. Field theory is also reasonably well developed for finite media of specific and limited types of geometry [18, 43, 88, 92]. Planarly layered media remain the most studied non-homogeneous media due to simplicity of modeling. Meaningful results can be obtained from modeling without intensive computational calculations. A layered medium may serve as a simplified first-order prototype model for a variety of realistic biomedical problems, where the dependence on the number of electrodes, tissue layers and their electrical properties has to be accounted for [36, 87]. Furthermore, closed-form solutions in terms of spectral integrals (Sommerfeld integrals) allow for asymptotic approximations, providing a more physical insight into the problems.

Models of field distribution in a region of planar stratification have been applied in several disciplines such as geophysical prospecting [33, 45, 75, 98], remote sensing [6, 23], microstrip circuits and antennas [2, 3, 13, 14, 47, 59, 95, 96], acoustical engineering [103], wave propagation theory [4, 18, 94, 101, 102, 105], and electrode grounding in power systems [19, 20, 49, 104].

Most of the investigations dealing with electric field propagation theory in plane stratified media, fall into two main categories: the "harmonic school" and the "image school".

1.2. *Harmonic school*

Olendorff [76] gave the solution for the electric field in plane stratified media in the form of an integral involving Bessel functions. Later, Stefanesco and

Schlumberger [86] gave for the case of three layer media a particular derivation of the Hummel formulae [40] by direct solution of Laplace's equation in terms of Bessel's functions. They also indicated the nature of the solution for the general case of n layers. Direct calculation of the determinants of order $2n$, which occur in this solution, is computationally inefficient. Thus, Sunde [89] suggested instead a recursive relation based on the transmission line theory.

The major problem with the harmonic approach has been the numerical computation of the potential expression, written as an infinite integral containing a combination of Bessel and kernel functions. This form does not lend itself to analytical integration. The kernel (Green's function) exhibits the well-known singularity $1/r$. Recently, one of the following three techniques has been used to compute the space domain of Green's function:

(1) Discrete complex image [19, 39, 48]. This method approximates the smooth part of the spectral domain Green's function by a sum of complex exponentials (Prony's method [37]). This technique heavily depends upon the accurate approximation of the kernel function. In the general case, this leads to a system of nonlinear equations whose solution is rather difficult. In addition, serious accuracy and stability problems make it impractical for application in more than two layers.
(2) FFT scheme. These methods enforce Dirichlet or Neumann boundary conditions on a bounding box, and use 3D FFT to perform the transformation. They require, however, extremely fine sampling to sufficiently cover the spectral contents in order to achieve a reasonable accuracy [11].
(3) Numerical integration. Green's function may be represented via a one-dimensional Hankel transform [29, 105]. Although robust, the numerical evaluation of this transform, is relatively time consuming. However, the use of the Fast Hankel Transform reduces this cost significantly. Nevertheless, numerical evaluation of the potential integral is not effective, in particular, when this integral has to be evaluated repeatedly in a numerical algorithm. Such is the situation in the solution of a 3-D potential problem for a large number of mesh points to calculate the potential distribution in multi-layered media due to a finite electrode array.

1.3. *Image school*

In the static case, Laplace equation for electrostatic (or magnetostatic) potential can be solved in a space region using a technique, in which the

required boundary conditions for the potential are simulated by using one or more *image* charges placed in a different region.

The physical interpretation via the wave/transmission line analog can be done as follows: At the beginning, the current source generates two starting voltage waves one up and one down at the excitation point. The two waves can independently propagate. If the waves meet an impedance discontinuity, they split into reflected and transmission waves according to the boundary conditions. The reflected and transmission waves then independently propagate, each being attenuated by the splitting process and the distance it has traveled. Waves passing the observation point are called *images.*

In the quasi-optic limit, the physical interpretation via ray optics analogy can be used. There is some analogy between the way in which a current travels through a medium and the way light rays spread through space. For instance, both the current density and light ray intensities obey the law of the inverse square of the distance from the point source. This analogy, however, does not imply, that the principles of geometric optics can be used to solve each and every problem in electrical current flow. In fact, the use of images is valid only in solving a limited number of potential problems [44]. In setting up the optical analogy, current sources are replaced by light sources and the planes with different conductivities are replaced by semi-transparent mirrors having reflection and transmission coefficients correlated to the resistivities of the layers. Accordingly, the light intensity at a point in a given medium is due partly to the point source and partly to its images from the other layers.

For a horizontally stratified medium consisting of two parallel layers, Maxwell [64] first expressed the electric field due to a current from a point source in terms of an infinite series of images by adopting the method of images first proposed by Lord Kelvin [90]. Several authors extended the method to three layers. For instance, Hummel [40] published, without an explicit derivation, formulae in terms of images. Other authors [7, 49, 79, 82, 105] extended the solution for the multilayer case by using different modifications of the image method. An extensive list of contributions in this field can be found [45, 98]. Evidently, the image method has also been asymptotically applied in optics [8] and wave propagation [31, 55].

The main advantage of the image technique is that a clear physical interpretation can be attributed to the mathematical terms. This may help to easily simplify the resulting algebraic expressions, without significant loss in the calculation precision. A major difficulty, though, is the complexity in taking into account all the reflection combinations from the primary and secondary images. In addition, the image series schemes existing in the

literature neither led to explicit closed form expansions for $n > 2$ (where n is the number of interfaces), nor contained information regarding its convergence properties. Furthermore, while the existence of a direct link between quasioptic and quasistatic image series expansions is accepted [55], this link has not been outlined systematically and explicitly.

The Institute of Scientific Information (ISI) database mentions more than 250 references dealing with this topic over the past decade alone, indicating that work on multiple-layer stratification is continuing and is far from being accomplished.

1.4. *Brief summary*

The first part of the chapter outlines the image series expansion of Green's function for a medium with planar stratification. The second part presents a hybrid model, dealing with the current density distribution in biological media. This latter model combines a novel image series expansion algorithm with the moment method [38] and illustrates an application of the image theory in bioelectromagnetics [62].

Upon developing the image series method, the following drawbacks were addressed: (1) For more than two layers, the infinite series expression is relatively hard to be derived and to implement; (2) Up to date no robust convergence testing procedure exists for the image series method. We thus focused on the following three goals: (i) rigorous closed form image series expansions for n (number of interfaces) ≥ 3, (ii) series convergence properties via a truncation-error estimation, (iii) bridging analytically between the ray-optics and quasistatic regimes. These targets were achieved by introducing first an integral representation for the frequency-dependent Hertz potential via a recursively constructed characteristic Green's function, in terms of reflection and transmission coefficients. The benefits of this construction were evidently overlooked in previous investigations, in which transmission-line impedances were used instead [65]. Next, the Hertz potential is expanded in finite image integral sums, often labeled as exact images [55] or ray integrals [31], and collective image integral (remainder) terms. The expansions utilize the unique recursive construction for Green's function which is a generic characteristic of the stratification and are explicitly constructed for $n \leq 3$. While results for $0 \leq n \leq 2$ are given for reference only, the expansion scheme for a double slab configuration, $n = 3$, is quite general and outlines the procedure for $n > 3$, without any increase in the

complexity. The n-layer finite expansion scheme is outlined by recursively extending the $n = 3$ procedure.

The finite expansions lead to rigorous image series expansions in the quasistatic limit where the remainder terms can be made negligibly small for sufficiently large summation indices. Evidently, our expansion scheme, relying on the recursive construction of the associated Green's function, bridges smoothly between the low and the high-frequency regimes. The collective image, representing an error-estimation term, is a closed form expression obtained via an asymptotic evaluation of the series remainder integral and is valid for sufficiency large summation indices. The image series convergence is accelerated by including a collective image term.

The fast convergence proved important when we further used this expansion, in conjunction with the moment method [38], to calculate the potential distribution due to a finite electrode array in multiple-layered media, as shown in Sec. 3 of the chapter.

Electrode excitation of a biological tissue is a well known and fundamental phenomenon, related to almost every FES application. Nevertheless, the literature reports only a few elementary models dealing with either a single finite electrode in infinite space [81, 99] (both reference make use of Jackson's derivation for circular electrode [43]) [12, 77, 78, 83], or arrays of point electrodes [84, 93]. More general electrode models (analytical or numerical) in biomedical applications are not reported in the existing literature.

Thus, we present in this part of the chapter the electromagnetic field interaction with multilayered biological medium as investigated for an electrode array excitation. A layered medium may serve as a simplified first-order prototype model for many realistic biomedical problems where the dependence on the number of electrodes, tissue layers and their electrical properties has to be accounted for [36]. Mathematically, the addressed problem may be reduced into a system of integral equations of the electrodes' current distribution [61]. The solution of the integral equation is accomplished by using the hybrid method. This enables the construction of the electrodes voltage/current relations via the impedance (admittance) matrix of the electrode array and consequently, also to evaluate the electrode power. The inversion of the integral operator is carried out in a two-step procedure: first its kernel is succinctly expanded in an image series expansion with a remainder term (collective image). Next, the moment matrix elements are calculated through an explicit analytic integration of the image terms. The hybrid scheme is further applied for numerical calculations.

The simulations are selected to address simple, yet fundamental, concepts associated with electromagnetic field interaction with biological tissues such as the potential distribution and electrode array impedance calculations.

2. Rigorous Image Series Expansion of Green's Function for Plane-Stratified Media

2.1. *Finite image integral expansion*

The physical configuration of our problem, depicted in Fig. 1, consists of a time-harmonic point-source S located at $\mathbf{r}' = (0, 0, z')$, an observation point P located at $\mathbf{r} = (x, y, z)$, and $n + 1$ isotropic homogeneous layers. The electromagnetic vector fields $\mathbf{E}(\mathbf{r}, \mathbf{r}')$ and $\mathbf{H}(\mathbf{r}, \mathbf{r}')$ are assumed to be excited by z-directed electric and magnetic point current elements, of length ℓ,

$$\mathbf{J}_e(\mathbf{r}) = I_e \ell \delta(\mathbf{r} - \mathbf{r}')\hat{\mathbf{z}}, \tag{1}$$

and

$$\mathbf{J}_m(\mathbf{r}) = I_m \ell \delta(\mathbf{r} - \mathbf{r}')\hat{\mathbf{z}}, \tag{2}$$

respectively. In the above equations $\hat{\mathbf{z}}$ is a z-directed unit vector, $\mathbf{J}_e(\mathbf{r})$ and $\mathbf{J}_m(\mathbf{r})$ are the electric and magnetic source current densities, I_e and

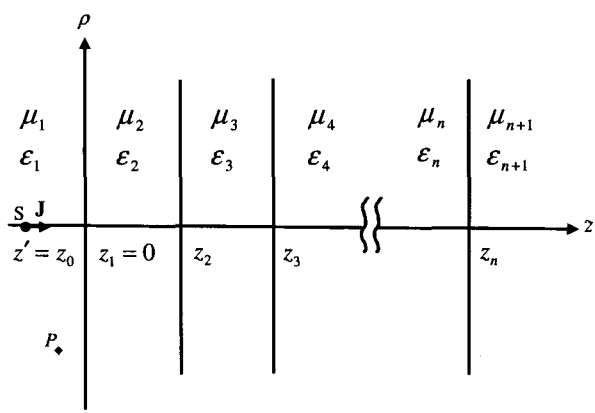


Fig. 1. Physical configuration for plane-stratified media, consisting of $n + 1$ layers, with n planar boundaries between the layers. The observation point P, the source point S, and the transverse coordinate ρ are defined via; $\mathbf{r} = \boldsymbol{\rho} + z\hat{\mathbf{z}} = (x, y, z)$, $\mathbf{r}' = \boldsymbol{\rho}' + z\hat{\mathbf{z}}' = (0, 0, z')$ and $\rho = \sqrt{x^2 + y^2}$, respectively. The parameters ε and μ denote the medium permittivity and permeability, respectively.

I_m electric and magnetic source currents, and $\delta(\mathbf{r}-\mathbf{r}')$ is the Dirac delta function. The fields may be expressed at $\mathbf{r} \neq \mathbf{r}'$, assuming time-dependence $e^{\mathcal{J}\omega t}$ [9, 30, 97], as:

$$\mathbf{E}(\mathbf{r},\mathbf{r}') = \nabla \times \nabla \times \hat{\mathbf{z}}\Pi_e(\mathbf{r},\mathbf{r}') - \mathcal{J}\omega\alpha_m(z)\nabla \times \hat{\mathbf{z}}\Pi_m(\mathbf{r},\mathbf{r}'), \tag{3}$$

$$\mathbf{H}(\mathbf{r},\mathbf{r}') = \mathcal{J}\omega\alpha_e(z)\nabla \times \hat{\mathbf{z}}\Pi_e(\mathbf{r},\mathbf{r}') + \nabla \times \nabla \times \hat{\mathbf{z}}\Pi_m(\mathbf{r},\mathbf{r}'), \tag{4}$$

where $\Pi_e(\mathbf{r},\mathbf{r}')$ and $\Pi_m(\mathbf{r},\mathbf{r}')$ represents the E-mode and H-mode Hertz potentials [88], respectively, $\mathcal{J}$ is the imaginary unit and ω is the angular frequency. The piecewise constant, generally complex, permittivity $\epsilon(z)$ and permeability $\mu(z)$ of the medium are denoted by the parameters $a_e(z)$ and $a_m(z)$, i.e.

$$\alpha_e(z) = \varepsilon(z) = \varepsilon_i, \quad \alpha_m(z) = \mu(z) = \mu_i, \tag{5}$$

in the ith layer,

$$z_{i-1} < z < z_i, \quad i = 0, 1, \ldots, n+1, \quad z_{-1} = -\infty, \quad z_0 = z', \quad z_{n+1} = \infty. \tag{6}$$

The Hertz potential in (3) and (4) can be expressed most effectively via Green's function,

$$\Pi(\mathbf{r},\mathbf{r}') = \frac{Q\ell}{\alpha(z)} G(\mathbf{r},\mathbf{r}'), \tag{7}$$

where

$$Q = I/\mathcal{J}\omega. \tag{8}$$

The distinguishing subscripts e and m in (7) and (8) have been omitted since the equations apply to both modes. This rule is adopted throughout the entire paper for all the equations that apply to both modes.

Equation (7) illuminates the role played in selecting longitudinal electric and magnetic current elements in (1) and (2), respectively. In particular, a longitudinal electric current element ($I_e \neq 0$, $I_m = 0$) generates E modes only and a longitudinal magnetic current elements ($I_e = 0$, $I_m \neq 0$) excites only H modes, whereas both modes types are generated by a transversely directed source of either electric or magnetic current [9, 30, 97]. Thus, only a single scalar potential function is involved in either the E- or H-mode quasistatic limit discussed later in Sec. 2.2.

2.1.1. *Integral representation*

The point-source coordinates selection, $\mathbf{r}' = (0, 0, z')$ (Fig. 1), leads to a circularly symmetric Green's function (Eq. (7)) which can be expressed

via a single spectral integral, known as the Fourier–Bessel representation [9, 30, 97]

$$G(\mathbf{r}, \mathbf{r}') = \frac{1}{2\pi} \int_0^\infty \xi g(z, z') J_0(\xi\rho) d\xi, \tag{9}$$

where J_0 is the Bessel function of the first kind and order zero, ρ is the radial coordinate (Fig. 1) and $g(z, z')$, the characteristic Green's function, is given via

$$g_i(z, z') = \begin{cases} \frac{\alpha_i e^{\mathcal{J}\beta_1 z'}}{\mathcal{J}2\alpha_1\beta_i} \left[\prod_{p=1}^{i} T_p(\xi)\right] [e^{-\mathcal{J}\beta_i z} - R_i(\xi) e^{\mathcal{J}\beta_i z}], & i > 0, \\ \frac{e^{-\mathcal{J}\beta_1 z'} - R_1(\xi) e^{\mathcal{J}\beta_1 z'}}{\mathcal{J}2\beta_1} e^{\mathcal{J}\beta_1 z}, & i = 0. \end{cases} \tag{10}$$

The subscript i (or p) denotes specific expressions or values that are valid in the ith (pth) layer, defined in (6). Both $G(\mathbf{r}, \mathbf{r}')$ and $g(z, z')$, in (9) satisfy, 3-D and 1-D wave equations, respectively, and appropriate constraints, as summarized in Table 1.

The reflection and transmission coefficients $R(\xi)$ and $T(\xi)$ in (10), respectively, are obtained by imposing the constraints listed in the rightmost column of Table 1:

$$R_i(\xi) = \left\{ K_i(\xi) + \frac{[1 - K_i^2(\xi)] R_{i+1}(\xi) e^{\mathcal{J}2\beta_{i+1} z_i}}{1 + K_i(\xi) R_{i+1}(\xi) e^{\mathcal{J}2\beta_{i+1} z_i}} \right\} e^{-\mathcal{J}2\beta_i z_i}, \quad R_{n+1}(\xi) = 0, \tag{11}$$

$$T_i(\xi) = \frac{1 - R_{i-1}(\xi) e^{\mathcal{J}2\beta_{i-1} z_{i-1}}}{1 - R_i(\xi) e^{\mathcal{J}2\beta_i z_{i-1}}} e^{\mathcal{J}(\beta_i - \beta_{i-1}) z_{i-1}}$$

$$= \frac{[1 + K_{i-1}(\xi)] e^{\mathcal{J}(\beta_i - \beta_{i-1}) z_{i-1}}}{1 + K_{i-1}(\xi) R_i(\xi) e^{\mathcal{J}2\beta_i z_{i-1}}}, \quad T_1(\xi) = 1, \tag{12}$$

where $K_i(\xi)$ denotes the local reflection coefficient of the ith interface,

$$K_i(\xi) = \frac{\alpha_i \beta_{i+1} - \alpha_{i+1} \beta_i}{\alpha_i \beta_{i+1} + \alpha_{i+1} \beta_i}, \quad K_0(\xi) = 0, \ \alpha_0 = \alpha_1, \ \beta_0 = \beta_1. \tag{13}$$

The representation of the z-directed transmission-line characteristic Green's function $g(z, z')$ in terms of reflection and transmission coefficients $R(\xi)$ and $T(\xi)$ constitutes a unique recursive construction, via the intrinsic reflection coefficient $K_i(\xi)$ in (13). Finally, substituting (9) in (7) results in an explicit integral representation for the Hertz potential,

$$\Pi(\mathbf{r}, \mathbf{r}') = \frac{Q\ell}{4\pi\alpha_1} \int_0^\infty \frac{\xi}{\mathcal{J}\beta(z)} \pi(z, z') J_0(\xi\rho) d\xi, \tag{14}$$

where the characteristic Hertz potential $\pi(z, z')$ is obtained via (10)

$$\pi_i(z, z') = \frac{\mathcal{J}2\alpha_1\beta_i}{\alpha_i} g_i(z, z'). \tag{15}$$

Table 1. Wave equations and boundary/continuity conditions for $G(\mathbf{r},\mathbf{r}')$ and $g(z,z')$.

	$G(\mathbf{r},\mathbf{r}')$	$g(z,z')$
Differential equation	$[\nabla^2 + k^2(z)]G(\mathbf{r},\mathbf{r}') = -\delta(\mathbf{r},\mathbf{r}')$	$\left[\dfrac{d^2}{dz^2} - \beta^2(z)\right] g(z,z') = -\delta(z-z')$
Wave number $k(z)$	$k(z) = \omega\sqrt{\mu(z)\epsilon(z)},\ \Im m[k(z)] \le 0$	
Propag. constant $\beta(z)$		$\beta(z) = \sqrt{k^2(z) - \xi^2},\quad \Im m[\beta(z)] \le 0$
[value in ith layer]	$[k_i = \omega\sqrt{\mu_i\epsilon_i}]$	$\left[\beta_i = \sqrt{k_i^2 - \xi^2}\right]$
Source condition	$\displaystyle\int_{V\to 0} \nabla^2 G(\mathbf{r},\mathbf{r}')dV = \oint_{A\to 0} \nabla G(\mathbf{r},\mathbf{r}')\cdot d\mathbf{A} = -1$	$\dfrac{dg(z'+h,z')}{dz} - \dfrac{dg(z'-h,z')}{dz}\Big\|_{h\to 0} = -1$
Continuity condition	$G_i(\mathbf{r},\mathbf{r}') = G_{i+1}(\mathbf{r},\mathbf{r}')$	$g_i(z,z') = g_{i+1}(z,z')$
at $z = z_i, i > 0$	$\dfrac{1}{\alpha_i}\dfrac{\partial G_i(\mathbf{r},\mathbf{r}')}{\partial z} = \dfrac{1}{\alpha_{i+1}}\dfrac{\partial G_{i+1}(\mathbf{r},\mathbf{r}')}{\partial z}$	$\dfrac{1}{\alpha_i}\dfrac{dg_i(z,z')}{dz} = \dfrac{1}{\alpha_{i+1}}\dfrac{dg_{i+1}(z,z')}{dz}$
Radiation condition	$r\left[\dfrac{\partial}{\partial r} + \mathcal{J}k(z)\right] G(\mathbf{r},\mathbf{r}')\Big\|_{r\to\infty} = 0$	$\left[\dfrac{d}{dz} + \mathcal{J}\beta(z)\right] g(z,z')\Big\|_{\|z\|\to\infty} = 0$

2.1.2. *Image integral expansions*

The recursive characteristics of $R_i(\xi)$ and $T_i(\xi)$ are now implemented to obtain finite expansions of $\pi(z, z')$ and $\Pi(\mathbf{r}, \mathbf{r}')$ in (14), for an arbitrary number of boundaries (Fig. 1). While results for $0 \leq n \leq 2$ are well known (see [9, 30, 55, 97] for $n \leq 1$), the expansion scheme developed for a double slab configuration, $n = 3$, is quite general and outlines the procedure for $n > 3$, without any increase in the complexity. The n-layer finite expansion schemes are obtained here by recursively extending the $n = 3$ procedure, as outlined at the end of this section [26].

The representation of the finite expansion scheme is given, without any loss of generality, for the observation points (P in Fig. 1) that are embedded in a single layer selected here as $-\infty < z < z'$ (Eq. (6)) to assist the applications carried out in Sec. 3.4.

The finite expansion is represented in terms of finite image integral expansions, often labeled as exact images [55] or ray integrals [31], and remainder terms. The remainder terms represents $n - 1$ collective summations of the individual images related to each of the $n - 1$ slabs, that were excluded from the image series expansions due to truncation. These remainders can be made negligibly small for sufficiently large summation indices: (1) in the quasistatic limit (see later 2.4), as $k(z) \to 0$, leading to rigorous image series expansions [26] and (2) asymptotically in the quasioptic (or the so-called geometric optical) limit, as $k(z) \to \infty$, leading to asymptotic (ray-optical) image series expansions [27, 28, 31]. Evidently, our expansion scheme, relying on the recursive construction of $\pi(z, z')$, has a generic characteristic that bridges smoothly between the low-frequency and high-frequency image series expansions, via an appropriate frequency adjustment. This generic characteristic is obscured in alternative procedures, which first apply either one of the frequency limits and only then the image series expansion scheme [8, 10].

The series expansion of (15) via (10) and consequently the image integral expansion of (14) are carried out by utilizing the binomial series expansion

$$\frac{1}{(1-x)^{N+1}} = \sum_{m=0}^{M} \binom{m+N}{m} x^m + \sum_{p=0}^{N} \binom{M+N+1}{M+p+1} \frac{x^{M+p+1}}{(1-x)^{p+1}}, \quad (16)$$

containing M terms and remainder. The series converges as $M \to \infty$, for all $|x| < 1$.

This is illustrated in the following examples.

2.1.3. *Unbounded medium, $n = 0$*

In an unbounded medium $R_1(\xi) = 0$ (Eq. (11)). Hence, the characteristic Hertz potential $\pi_1(z, z')$ consists of a single outgoing wave in free space,

$$\pi_1(z, z') = e^{-\mathcal{J}\beta_1|z-z'|}. \tag{17}$$

A closed form expression for the Hertz potential can be obtained by substituting (17) in (14) and applying the Sommerfeld integral identity,

$$\Pi_1(\mathbf{r}, \mathbf{r}') = \frac{Q\ell}{4\pi\alpha_1}\int_0^\infty \frac{\xi}{\mathcal{J}\beta_1} e^{-\mathcal{J}\beta_1|z-z'|} J_0(\xi\rho)d\xi = \frac{Q\ell}{4\pi\alpha_1}\frac{e^{-\mathcal{J}k_1|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}. \tag{18}$$

2.1.4. *Semi-infinite medium, $n = 1$*

In a semi-infinite medium, $R_1(\xi) = K_1(\xi)$ (Eqs. (11) and (13)). Thus, the potential in (15) contains two terms representing outgoing and reflected waves,

$$\pi_1(z, z') = e^{-\mathcal{J}\beta_1|z-z'|} - K_1(\xi)e^{\mathcal{J}\beta_1(z+z')}. \tag{19}$$

The potential $\Pi_1(\mathbf{r}, \mathbf{r}')$ in (14) can be expressed as,

$$\Pi_1(\mathbf{r}, \mathbf{r}') = \frac{Q\ell}{4\pi\alpha_1}\left[\frac{e^{-\mathcal{J}k_1|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|} - \int_0^\infty \frac{\xi}{\mathcal{J}\beta_1} K_1(\xi)e^{\mathcal{J}\beta_1(z+z')} J_0(\xi\rho)d\xi\right]. \tag{20}$$

2.1.5. *Single slab configuration, $n = 2$*

In a three layers medium, $R_1(\xi)$ can be expanded into a finite geometric series in $K_1(\xi)R_2(\xi)$, where $R_2(\xi) = K_2(\xi)e^{-\mathcal{J}2\beta_2 z_2}$, and a remainder term. The 1-D potential function $\pi_1(z, z')$ can be expressed, similarly, as

$$\begin{aligned}
\pi_1(z, z') &= e^{-\mathcal{J}\beta_1|z-z'|} - K_1(\xi)e^{\mathcal{J}\beta_1(z+z')} - \sum_{m_1=0}^{M_1}[1 - K_1^2(\xi)] \\
&\quad \times [-K_1(\xi)]_1^m K_2^{m_1+1}(\xi)e^{\mathcal{J}[\beta_1(z+z')-2(m_1+1)\beta_2 z_2]} + \gamma_{M_1}(z, z') \\
&= e^{-\mathcal{J}\beta_1|z-z'|} - e^{\mathcal{J}\beta_1(z+z')}\sum_{l_1=0}^{1}\sum_{m_1=0}^{M_1}\binom{1}{l_1}\binom{m_1-l_1}{m_1} \\
&\quad \times K_1^l(\xi)[1 - K_1^2(\xi)]^{-l_1+1}[-K_1(\xi)]^{m_1} K_2^{m_1-l_1+1}(\xi) \\
&\quad \times e^{-\mathcal{J}2(m_1-l_1+1)\beta_2 z_2} + \gamma_{M_1}(z, z'),
\end{aligned} \tag{21}$$

where the remainder term $\gamma_{M_1}(z,z')$ is given by means of (16),

$$\gamma_{M_1}(z,z') = -e^{\mathcal{J}\beta_1(z+z')}[1-K_1^2(\xi)]\frac{[-K_1(\xi)]^{M_1+1}K_2^{M_1+2}(\xi)e^{-\mathcal{J}2(M_1+2)\beta_2 z_2}}{1+K_1(\xi)K_2(\xi)e^{-\mathcal{J}2\beta_2 z_2}}$$

$$= e^{\mathcal{J}\beta_1(z-z')}\sum_{m_0=0}^{M_0}\sum_{l_1=0}^{1}\sum_{p_1=0}^{-l_1}\binom{0}{m_0}\binom{1}{l_1}K_1^{l_1}(\xi)$$

$$\times[1-K_1^2(\xi)]^{-l_1+1}[-K_0(\xi)]^{m_0}e^{-\mathcal{J}2\beta_1(z_1-z')}\binom{M_1-l_1+1}{M_1+p_1+1}$$

$$\times\frac{[-K_1(\xi)]^{M_1+p_1+1}[K_2(\xi)e^{-\mathcal{J}2\beta_2(z_2-z_1)}]^{M_1+p_1-l_1+2}}{[1+K_1(\xi)K_2(\xi)e^{-\mathcal{J}2\beta_2(z_2-z_1)}]^{p_1+1}}. \qquad (22)$$

The last expressions in (21) and (22), though identical to the preceding terms, render the generalization to n layered media straightforward.

Substituting (21) and (22) in (14) and changing the order of the integration and summation, results in a finite expansion for $\Pi_1(\mathbf{r},\mathbf{r}')$,

$$\Pi_1(\mathbf{r},\mathbf{r}') = \frac{Q\ell}{4\pi\alpha_1}\left\{\frac{e^{-\mathcal{J}k_1|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|} - \int_0^\infty \frac{\xi}{\mathcal{J}\beta_1}K_1(\xi)e^{\mathcal{J}\beta_1(z+z')}J_0(\xi\rho)d\xi\right.$$

$$-\sum_{m_1=0}^{M_1}\int_0^\infty\frac{\xi}{\mathcal{J}\beta_1}[1-K_1^2(\xi)][-K_1(\xi)]^{m_1}K_2^{m_1+1}(\xi)$$

$$\left.\times e^{\mathcal{J}[\beta_1(z+z')-2(m_1+1)\beta_2 z_2]}J_0(\xi\rho)d\xi + \Gamma_{M_1}(\mathbf{r},\mathbf{r}')\right\}$$

$$= \frac{Q\ell}{4\pi\alpha_1}\left\{\frac{e^{-\mathcal{J}k_1|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|} - \sum_{l_1=0}^{1}\sum_{m_1=0}^{M_1}\binom{1}{l_1}\binom{m_1-l_1}{m_1}\right.$$

$$\times\int_0^\infty\frac{\xi}{\mathcal{J}\beta_1}K_1^{l_1}(\xi)[1-K_1^2(\xi)]^{-l_1+1}[-K_1(\xi)]^{m_1}K_2^{m_1+1}(\xi)$$

$$\left.\times\ e^{\mathcal{J}[\beta_1(z-z')-2(m_1-l_1+1)\beta_2 z_2}J_0(\xi\rho)d\xi + \Gamma_{M_1}(\mathbf{r},\mathbf{r}')\right\}, \qquad (23)$$

where the remainder integral $\Gamma_{M_1}(\mathbf{r},\mathbf{r}')$ is,

$$\Gamma_{M_1}(\mathbf{r},\mathbf{r}') = \int_0^\infty\frac{\xi}{\mathcal{J}\beta_1}\gamma_{M_1}(z,z')J_0(\xi\rho)d\xi. \qquad (24)$$

Note that for $l_1 = 1$ the series expansions in (21) and (23) are reduced into single terms, and the corresponding contributions to the remainder terms in (22) and (24) are zero.

2.1.6. *Double slab configuration, $n = 3$*

In a four-layer medium $R_1(\xi)$ can be expanded into a finite geometric series in $K_1(\xi)R_2(\xi)$ followed by two finite binomial expansions in $K_2(\xi)R_3(\xi)$,

where $R_3(\xi) = K_3(\xi)e^{-\mathcal{J}2\beta_3 z_3}$, and two remainder terms. The characteristic Hertz potential $\pi_1(z, z')$ can be expressed, similarly, as

$$\begin{aligned}
\pi_1(z,z') &= e^{-\mathcal{J}\beta_1|z-z'|} - K_1(\xi)e^{\mathcal{J}\beta_1(z+z')} \\
&\quad - \sum_{m_1=0}^{M_1}\sum_{l_2=0}^{m_1+1}\sum_{m_2=0}^{M_2}[1-K_1^2(\xi)][-K_1(\xi)]^{m_1}\binom{m_1+1}{l_2} \\
&\quad \times K_2^{l_2}(\xi)[1-K_2^2(\xi)]^{m_1-l_2+1}\binom{m_1+m_2-l_2}{m_2} \\
&\quad \times [-K_2(\xi)]^{m_2}[K_3(\xi)]^{m_1+m_2-l_2+1} \\
&\quad \times e^{\mathcal{J}[\beta_1(z+z')-2(m_1+1)\beta_2 z_2-2(m_1+m_2-l_2+1)\beta_3(z_3-z_2)]} \\
&\quad + \gamma_{M_1}(z,z') + \gamma_{M_1,M_2}(z,z') \\
&= e^{-\mathcal{J}\beta_1|z-z'|} - e^{\mathcal{J}\beta_1(z+z')}\sum_{l_1=0}^{1}\sum_{m_1=0}^{M_1}\sum_{l_2=0}^{m_1-l_1+1}\sum_{m_2=0}^{M_2}\binom{1}{l_1} \\
&\quad \times \binom{m_1-l_1}{m_1}\binom{m_1-l_1+1}{l_2}\binom{m_1+m_2-l_1-l_2}{m_2} \\
&\quad \times K_1^{l_1}(\xi)K_2^{l_2}(\xi)[1-K_1(\xi)]^{-l_1+1}[1-K_2(\xi)]^{m_1-l_1-l_2+1} \\
&\quad \times [-K_1(\xi)]^{m_1}[-K_2(\xi)]^{m_2}[K_3(\xi)]^{m_1+m_2-l_1-l_2+1} \\
&\quad \times e^{-\mathcal{J}2[(m_1-l_1+1)\beta_2 z_2+(m_1+m_2-l_1-l_2+1)\beta_3(z_3-z_2)]} \\
&\quad + \gamma_{M_1}(z,z') + \gamma_{M_1,M_2}(z,z'), \qquad (25)
\end{aligned}$$

where the remainder terms $\gamma_{M_1}(z, z')$ and $\gamma_{M_1,M_2}(z, z')$ are given by means of (16),

$$\begin{aligned}
\gamma_{M_1}(z,z') &= -e^{\mathcal{J}\beta_1(z+z')}[1-K_1^2(\xi)] \\
&\quad \times \frac{[-K_1(\xi)]^{M_1+1}[R_2(\xi)e^{\mathcal{J}2\beta_2 z_2}]^{M_1+2}e^{-\mathcal{J}2(M_1+2)\beta_2 z_2}}{1+K_1(\xi)R_2(\xi)}, \\
&= e^{\mathcal{J}\beta_1(z-z')}\sum_{m_0=0}^{M_0}\sum_{l_1=0}^{1}\sum_{p_1=0}^{-l_1}\binom{0}{m_0}\binom{1}{l_1}K_1^{l_1}(\xi) \\
&\quad \times [1-K_1^2(\xi)]^{-l_1+1}[-K_0(\xi)]^{m_0}e^{-\mathcal{J}2\beta_1(z_1-z')}\binom{M_1-l_1+1}{M_1+p_1+1} \\
&\quad \times \frac{[-K_1(\xi)]^{M_1+p_1+1}[R_2(\xi)e^{\mathcal{J}2\beta_2 z_1}]^{M_1+p_1-l_1+2}}{[1+K_1(\xi)R_2(\xi)e^{\mathcal{J}2\beta_2 z_1}]^{p_1+1}}, \qquad (26)
\end{aligned}$$

and

$$
\begin{aligned}
&\gamma_{M_1,M_2}(z,z') \\
&= -e^{\mathcal{J}\beta_1(z+z')} \sum_{m_1=0}^{M_1} \sum_{l_2=0}^{m_1} \sum_{p_2=0}^{m_1-l_2} \binom{m_1+1}{l_2} K_2^{l_2}(\xi)[1-K_1^2(\xi)] \\
&\quad \times [1-K_2^2(\xi)]^{m_1-l_2+1}[-K_1(\xi)]^{m_1} e^{-\mathcal{J}2(m_1+1)\beta_2 z_2} \binom{M_2+m_1-l_2+1}{M_2+p_2+1} \\
&\quad \times \frac{[-K_2(\xi)]^{M_2+p_2+1}[K_3(\xi)]^{M_2+p_2+m_1-l_2+2} e^{-\mathcal{J}2(M_2+p_2+m_1-l_2+2)\beta_3(z_3-z_2)}}{[1+K_2(\xi)K_3(\xi)e^{-\mathcal{J}2\beta_3(z_3-z_2)}]^{p_2+1}} \\
&= e^{\mathcal{J}\beta_1(z-z')} \sum_{m_0=0}^{M_0} \sum_{l_1=0}^{1} \sum_{m_1=0}^{M_1} \sum_{l_2=0}^{m_1-l_1+1} \sum_{p_2=0}^{m_1-l_1-l_2} \\
&\quad \times \binom{0}{m_0} \binom{m_1-l_1}{m_1} \binom{1}{l_1} \binom{m_1-l_1+1}{l_2} K_1^{l_1}(\xi) K_2^{l_2}(\xi) \\
&\quad \times [1-K_1^2(\xi)]^{-l_1+1}[1-K_2^2(\xi)]^{+m_1-l_1-l_2+1}[-K_0(\xi)]^{m_0}[-K_1(\xi)]^{m_1} \\
&\quad \times e^{-\mathcal{J}2[\beta_1(z_1-z')+(m_1-l_0-l_1+1)\beta_2(z_2-z_1)]} \binom{M_2+m_1-l_1-l_2+1}{M_2+p_2+1} \\
&\quad \times \frac{[-K_2(\xi)]^{M_2+p_2+1}[R_3(\xi)e^{\mathcal{J}2\beta_3 z_2}]^{M_2+p_2+m_1-l_1-l_2+2}}{[1+K_2(\xi)R_3(\xi)e^{\mathcal{J}2\beta_3 z_2}]^{p_2+1}},
\end{aligned}
\tag{27}
$$

respectively. As indicated previously ($n=2$ configuration), the last expressions in (25), (26) and (27), though identical to the preceding terms, are given merely to render the generalization to $n+1$ layered problem straightforward. Substitution of (25)–(27) in (14) and changing the order the of integration and summation, results in a finite expansion for $\Pi_1(\mathbf{r},\mathbf{r}')$,

$$
\begin{aligned}
\Pi_1(\mathbf{r},\mathbf{r}') = \frac{Q\ell}{4\pi\alpha_1} \Bigg[& \frac{e^{-\mathcal{J}k_1|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|} - \int_0^\infty \frac{\xi}{\mathcal{J}\beta_1} K_1(\xi) e^{\mathcal{J}\beta_1(z+z')} J_0(\xi\rho) d\xi \\
& - \sum_{m_1=0}^{M_1} \sum_{l_2=0}^{m_1+1} \sum_{m_2=0}^{M_2} \binom{m_1+1}{l_2} \binom{m_1+m_2-l_2}{m_2} \\
& \times \int_0^\infty \frac{\xi}{\mathcal{J}\beta_1} [1-K_1^2(\xi)][1-K_2^2(\xi)]^{m_1-l_2+1}[-K_1(\xi)]^{m_1} \\
& \times [-K_2(\xi)]^{m_2} K_2^{l_2}(\xi)[K_3(\xi)]^{m_1+m_2-l_2+1} \\
& \times e^{\mathcal{J}[\beta_1(z+z')-2\beta_2 z_2(m_1+1)-2\beta_3(z_3-z_2)(m_1+m_2-l_2+1)]} \\
& \times J_0(\xi\rho) d\xi + \Gamma_{M_1}(\mathbf{r},\mathbf{r}') + \Gamma_{M_1,M_2}(\mathbf{r},\mathbf{r}') \Bigg],
\end{aligned}
$$

$$= \frac{Q\ell}{4\pi\alpha_1}\left[\frac{e^{-\mathcal{J}k_1|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|} - \sum_{l_1=0}^{1}\sum_{m_1=0}^{M_1}\sum_{l_2=0}^{m_1-l_1+1}\sum_{m_2=0}^{M_2}\right.$$
$$\times \binom{1}{l_1}\binom{m_1-l_1}{m_1}\binom{m_1-l_1+1}{l_2}\binom{m_1+m_2-l_1-l_2}{m_2}$$
$$\times \int_0^\infty \frac{\xi}{\mathcal{J}\beta_1}[1-K_1^2(\xi)]K_1^{l_1}(\xi)K_2^{l_2}(\xi)[1-K_1(\xi)]^{-l_1+1}$$
$$\times [1-K_2(\xi)]^{m_1-l_1-l_2+1}[-K_1(\xi)]^{m_1}[-K_2(\xi)]^{m_2}$$
$$\times [K_3(\xi)]^{m_1+m_2-l_1-l_2+1}$$
$$\times e^{\mathcal{J}[\beta_1(z+z')-2(m_1-l_1+1)\beta_2 z_2+(m_1+m_2-l_1-l_2+1)\beta_3(z_3-z_2)]}$$
$$\left.\times J_0(\xi\rho)d\xi + \Gamma_{M_1}(\mathbf{r},\mathbf{r}') + \Gamma_{M_1,M_2}(\mathbf{r},\mathbf{r}')\right], \tag{28}$$

where the remainder integrals $\Gamma_{M_1}(\mathbf{r},\mathbf{r}')$ and $\Gamma_{M_1,M_2}(\mathbf{r},\mathbf{r}')$ are given via,

$$\Gamma_{M_1}(\mathbf{r},\mathbf{r}') = \int_0^\infty \frac{\xi}{\mathcal{J}\beta_1}\gamma_{M_1}(z,z')J_0(\xi\rho)d\xi, \tag{29}$$

and

$$\Gamma_{M_1,M_2}(\mathbf{r},\mathbf{r}') = \int_0^\infty \frac{\xi}{\mathcal{J}\beta_1}\gamma_{M_1,M_2}(z,z')J_0(\xi\rho)d\xi, \tag{30}$$

respectively. It should be noted that for either $l_1 = 1$ or $l_2 = m_1 + 1$ the series expansion in (25) and (28) are reduced into single terms or single slab series, respectively, and both corresponding contributions to the remainder terms in (27), (29) and (30) are zero.

2.1.7. *Generalized image integral expansion* ($n \geq 1$)

The finite expansion scheme for a triple slab configuration ($n = 4$) can readily be confirmed to repeat the double slab procedure ($n = 3$) and then followed by two additional finite binomial expansions in $K_3(\xi)R_4(\xi)$ and additional remainder term. The extended procedure is due to the recursive expression (11) in which $K_3(\xi)$ is replaced by $R_3(\xi)$ and $R_4(\xi) = K_4(\xi)e^{-\mathcal{J}2\beta_4 z_4}$. Extension of the finite expansion scheme for n-interface configurations is straightforward, in view of the last expressions contained

in (21), (25), (22), (26) and (27), and can be carried out by induction:

$$\pi_1(z,z') = e^{-\mathcal{J}\beta_1|z-z'|} - e^{\mathcal{J}\beta_1(z+z')} \sum_{l_1=0}^{1} \sum_{m_1=0}^{M_1} \cdots \sum_{l_{n-1}=0}^{s_{n-2}+1} \sum_{m_{n-1}=0}^{M_{n-1}}$$
$$\times \left\{ \prod_{k=1}^{n-1} \binom{s_{k-1}+1}{l_k} \binom{s_k}{m_k} \right.$$
$$\times K_k^{l_k}(\xi)[1-K_k^2(\xi)]^{s_{k-1}-l_k+1}$$
$$\left. \times [-K_k(\xi)]^{m_k}[K_n(\xi)]^{m_k-l_k} e^{-\mathcal{J}2(s_k+1)\beta_{k+1}(z_{k+1}-z_k)} \right\} K_n(\xi)$$
$$+ \sum_{k=1}^{n-1} \gamma_{M_1,\ldots,M_k}(z,z'), \tag{31}$$

where

$$s_n = \sum_{k=1}^{n}(m_k - l_k), \quad s_0 = 0, \tag{32}$$

and

$$\gamma_{M_1,\ldots,M_k}(z,z') = -e^{\mathcal{J}\beta_1(z-z')} \sum_{m_0=0}^{M_0} \sum_{l_1=0}^{1} \cdots \sum_{m_{k-1}=0}^{M_{k-1}} \sum_{l_k=0}^{s_{k-1}+1} \sum_{p_k=0}^{s_{k-1}-l_k}$$
$$\times \left\{ \prod_{j=1}^{k} \binom{s_{j-1}+1}{l_j} \binom{s_{j-1}}{m_{j-1}} \right.$$
$$\times K_j^{l_j}(\xi)[1-K_j^2(\xi)]^{s_{j-1}-l_j+1}$$
$$\left. \times [-K_{j-1}(\xi)]^{m_{j-1}} e^{-\mathcal{J}2(s_{j-1}+1)\beta_j(z_j-z_{j-1})} \right\}$$
$$\times \binom{M_k + s_{k-1} - l_k + 1}{M_k + p_k + 1}$$
$$\times \frac{[-K_k(\xi)]^{M_k+p_k+1}[R_{k+1}(\xi)e^{\mathcal{J}2\beta_{k+1}z_k}]^{M_k+p_k+s_{k-1}-l_k+2}}{[1+K_k(\xi)R_{k+1}(\xi)e^{\mathcal{J}2\beta_{k+1}z_k}]^{p_k+1}}. \tag{33}$$

Finally,

$$\Pi_1(\mathbf{r},\mathbf{r}') = \frac{Q\ell}{4\pi\alpha_1}\left[\frac{e^{-\mathcal{J}\beta_1|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|} - \sum_{l_1=0}^{1}\sum_{m_1=0}^{M_1}\cdots\sum_{l_{n-1}=0}^{s_{n-2}+1}\sum_{m_{n-1}=0}^{M_{n-1}}\right.$$
$$\times\left\{\prod_{k=1}^{n-1}\binom{s_{k-1}+1}{l_k}\binom{s_k}{m_k}\int_0^\infty K_k^{l_k}(\xi)[1-K_k^2(\xi)]^{s_{k-1}-l_k+1}\right.$$
$$\left.\times[-K_k(\xi)]^{m_k}[K_n(\xi)]^{m_k-l_k}e^{-\mathcal{J}2(s_k+1)\beta_{k+1}(z_{k+1}-z_k)}\right\}$$
$$\left.\times K_n(\xi)\frac{\xi}{\mathcal{J}\beta_1}e^{\mathcal{J}\beta_1(z+z')}J_0(\xi\rho)d\xi + \sum_{j=1}^{n-1}\Gamma_{M_1,\ldots,M_k}(\mathbf{r},\mathbf{r}')\right], \quad (34)$$

where

$$\Gamma_{M_1,\ldots,M_k}(\mathbf{r},\mathbf{r}') = \int_0^\infty \frac{\xi}{\mathcal{J}\beta_1}\gamma_{M_1,\ldots,M_k}(z,z')J_0(\xi\rho)d\xi. \quad (35)$$

It should be noted that, for $l_k = s_{k-1}+1$ the series expansions in (31) and (34), carried out for $n-1$ slabs, are reduced into expansions for $k-1$ slabs, and the corresponding contributions to the remainder terms (33) and (35) are zero. In general, the total number of summations N in these equations is twice the number of slabs $(n-1)$. However, since the observation point lies within the seminfinite layer $i=1$, the number of summation is reduced by one, i.e.

$$N = 2(n-1)-1 = 2n-3. \quad (36)$$

The number N is an intrinsic characteristic of the stratification, and thus, invariant to $\mathbf{r}$ which can be arbitrarily embedded in any layer. The summation indices l_k and m_k denote the number of bounces on the kth layer from the left and the right sides, respectively (Fig. 1). It can readily be verified that for a single interface problem $(n=1)$ the image series expansion in (34) contributes a single term and zero remainder, since $\prod_{k=1}^{0} \equiv 1$ and $\sum_{k=1}^{0} \equiv 0$.

2.2. *Image series expansion*

The quasistatic limit $(k(z) = \omega\sqrt{\mu(z)\epsilon(z)} \to 0)$ of the electromagnetic vector fields $\mathbf{E}(\mathbf{r},\mathbf{r}')$, $\mathbf{H}(\mathbf{r},\mathbf{r}')$ and the Hertz potential $\Pi(\mathbf{r},\mathbf{r}')$ is completely specified by the leading terms of their low-frequency power series approximations, in $k(z)$. Since the propagation constant $\beta(z)$ is an even function of the wave number $k(z)$ (Table 1), it can be readily shown by means of (7)

that the quasistatic limit of $\mathbf{E}(\mathbf{r},\mathbf{r}')$ and $\mathbf{H}(\mathbf{r},\mathbf{r}')$ in (3) and (4), respectively, is given by the following relations,

$$\begin{aligned}\nabla\times\nabla\times\hat{\mathbf{z}}\Pi(\mathbf{r},\mathbf{r}') &= \nabla\times\nabla\times\hat{\mathbf{z}}\mathcal{P}(\mathbf{r},\mathbf{r}')+O[k^2(z)]\\ &= \nabla\frac{\partial\mathcal{P}(\mathbf{r},\mathbf{r}')}{\partial z}+O[k^2(z)],\quad k(z)\to 0, \qquad (37)\end{aligned}$$

$$\mathcal{J}\omega\alpha(z)\nabla\times\hat{\mathbf{z}}\mathcal{P}(\mathbf{r},\mathbf{r}') = \mathcal{J}\omega\alpha(z)\{\nabla\times\hat{\mathbf{z}}\mathcal{P}(\mathbf{r},\mathbf{r}')+O[k^2(z)]\},\quad k(z)\to 0. \qquad (38)$$

Substituting $\beta=-\mathcal{J}\xi$ $(k(z)\to 0)$ into the Hertz potential $\Pi(z,z')$ (Eq. (14)), the characteristic Hertz potential $\pi(z,z')$ (Eq. (15)), the reflection coefficient $R(\xi)$ (Eq. (11)), the transmission coefficient $T(\xi)$ (Eq. (12)) and the intrinsic reflection coefficient $K(\xi)$ (Eq. (13)), result in their quasistatic limit in the following expressions:

$$\mathcal{P}(\mathbf{r},\mathbf{r}') = \frac{Q\ell}{4\pi a_1}\int_0^\infty \mathcal{P}(z,z')J_0(\xi\rho)d\xi, \qquad (39)$$

$$\mathcal{P}_i(z,z') = \begin{cases}\left[\prod_{p=1}^{i}\mathcal{T}_p(\xi)\right][e^{-\xi z}-\mathcal{R}_i(\xi)e^{\xi z}]e^{\xi z'}, & i>0,\\ [e^{-\xi z'}-\mathcal{R}_1(\xi)e^{\xi z'}]e^{\xi z}, & i=0,\end{cases} \qquad (40)$$

$$\mathcal{R}_i(\xi) = \left[\mathcal{K}_i+\frac{(1-\mathcal{K}_i^2)\mathcal{R}_{i+1}(\xi)e^{2\xi z_i}}{1+\mathcal{K}_i\mathcal{R}_{i+1}(\xi)e^{2\xi z_i}}\right]e^{-2\xi z_i},\quad \mathcal{R}_{n+1}(\xi)=0, \qquad (41)$$

$$\mathcal{T}_i(\xi) = \frac{1-\mathcal{R}_{i-1}(\xi)e^{2\xi z_{i-1}}}{1-\mathcal{R}_i(\xi)e^{2\xi z_{i-1}}} = \frac{1+\mathcal{K}_{i-1}}{1+\mathcal{K}_{i-1}\mathcal{R}_i(\xi)e^{2\xi z_{i-1}}},\quad \mathcal{T}_1(\xi)=1, \qquad (42)$$

and

$$\mathcal{K}_i = \frac{a_i-a_{i+1}}{a_i+a_{i+1}},\quad \mathcal{K}_0=0,\quad a_0=a_1, \qquad (43)$$

respectively. If the parameter $a(z)$ is not singular, then the electroquasistatic or magnetoquasistatic fields are obtained by setting $Q_e\neq 0$, $Q_m=0$ or $Q_e=0$, $Q_m\neq 0$, respectively. If, however, $a(z)$ is singular as $k(z)\to 0$ then $\mathcal{J}\omega a(z)=\sigma(z)$, the latter denoting the conductivity of the medium [88, 98] in the stationary-current regime. In this case Eq. (39) reduces, utilizing Eq. (8), to $\mathcal{J}\omega Q/\mathcal{J}\omega a(z)=I/\sigma(z)$ and Eq. (43) reduces to $\mathcal{K}_i=\mathcal{J}\omega(a_i-a_{i+1})/\mathcal{J}\omega(a_i+a_{i+1})=(\sigma_i+\sigma_{i+1})/(\sigma_i+\sigma_{i+1})$.

2.2.1. *Properties of* $\mathcal{R}(\xi)$

The reflection coefficients $\mathcal{R}_n(\xi)$ and $\mathcal{R}_{n-1}(\xi)$ in (41) can be evaluated, at $\xi=0$, as,

$$\mathcal{R}_n(0)=\mathcal{K}_n=\frac{a_n-a_{n+1}}{a_n+a_{n+1}}, \qquad (44)$$

and

$$\mathcal{R}_{n-1}(0) = \frac{a_{n-1} - a_{n+1}}{a_{n-1} + a_{n+1}}, \tag{45}$$

respectively. Consequently, $\mathcal{R}_m(0)$ is obtained by induction leading to,

$$\mathcal{R}_m(0) = \frac{a_m - a_{n+1}}{a_m + a_{n+1}}. \tag{46}$$

Utilizing Eq. (46) and an alternative representation for the reflection coefficient $\mathcal{R}_m(\xi)$ in (41),

$$\mathcal{R}_i(\xi) = \frac{1 - \frac{1-\mathcal{K}_i}{1+\mathcal{K}_i}\frac{1-\mathcal{R}_{i+1}(\xi)e^{2\xi z_i}}{1+\mathcal{R}_{i+1}(\xi)e^{2\xi z_i}}}{1 + \frac{1-\mathcal{K}_i}{1+\mathcal{K}_i}\frac{1-\mathcal{R}_{i+1}(\xi)e^{2\xi z_i}}{1+\mathcal{R}_{i+1}(\xi)e^{2\xi z_i}}} e^{-2\xi z_i} \tag{47}$$

results in,

$$|\mathcal{R}_{n-1}(\xi)e^{2\xi z_{n-1}}| < 1, \quad \text{if both } |\mathcal{K}_{n-1}| < 1 \text{ and } |\mathcal{K}_n| < 1. \tag{48}$$

The general rule for $1 \leq m < n-1$ is obtained by means of induction in conjunction with the continuity conditions (Table 2),

$$\frac{1 - \mathcal{R}_m(\xi)e^{2\xi z_m}}{1 + \mathcal{R}_m(\xi)e^{2\xi z_m}} = \frac{1 - \mathcal{K}_m}{1 + \mathcal{K}_m}\frac{1 - \mathcal{R}_{m+1}(\xi)e^{2\xi z_m}}{1 + \mathcal{R}_{m+1}(\xi)e^{2\xi z_m}} \tag{49}$$

Table 2. Laplace equations and boundary/continuity conditions for $\mathcal{G}(\mathbf{r}, \mathbf{r}')$ and $\mathcal{G}(z, z')$.

	$\mathcal{G}(\mathbf{r}, \mathbf{r}')$	$\mathcal{G}(z, z')$
Differential equation	$\nabla^2\mathcal{G}(\mathbf{r}, \mathbf{r}') = -\delta(\mathbf{r}, \mathbf{r}')$	$\left(\frac{d^2}{dz^2} - \xi^2\right)\mathcal{G}(z, z') = -\delta(z - z')$
Source condition at $z = z'$	$\int_{V\to 0} \nabla^2\mathcal{G}(\mathbf{r}, \mathbf{r}')dV$ $= \oint_{A\to 0} \nabla\mathcal{G}(\mathbf{r}, \mathbf{r}') \cdot d\mathbf{A} = -1$	$\frac{d\mathcal{G}(z'+h, z')}{dz} - \frac{d\mathcal{G}(z'-h, z')}{dz}\Big\vert_{h\to 0}$ $= -1$
Continuity condition at $z = z_i$ $i > 0$	$\mathcal{G}_i(\mathbf{r}, \mathbf{r}') = \mathcal{G}_{i+1}(\mathbf{r}, \mathbf{r}')$ $a_i\frac{\partial\mathcal{G}_i(\mathbf{r}, \mathbf{r}')}{\partial z} = a_{i+1}\frac{\partial\mathcal{G}_{i+1}(\mathbf{r}, \mathbf{r}')}{\partial z}$	$\mathcal{G}_i(z, z') = \mathcal{G}_{i+1}(z, z')$ $a_i\frac{d\mathcal{G}_i(z, z')}{dz} = a_{i+1}\frac{d\mathcal{G}_{i+1}(z, z')}{dz}$
Decay at infinity	$r\mathcal{G}(\mathbf{r}, \mathbf{r}')_{r\to\infty} < \infty$	$\left(\frac{d}{dz} + \xi\right)\mathcal{G}(z, z')_{\vert z\vert\to\infty} = 0$

leading to

$$|\mathcal{R}_m(\xi)e^{2\xi z_m}| < 1, \quad \text{if both } |\mathcal{K}_m| < 1 \text{ and } |\mathcal{R}_{m+1}(\xi)e^{2\xi z_m}| < 1. \tag{50}$$

Note that $|\mathcal{K}_i| < 1$, $1 \leq i \leq n$, if $\Re e[a_i] > 0$ whereas $\Im m[a_i] \leq 0$, since $\Im m[k(z)] \leq 0$ (Table 1). Similarly, in a conductive medium where $\sigma(z) = \mathcal{J}\omega a \neq 0$ (following Eq. (43)), $|\mathcal{K}_i| < 1$, if $\Re e[\sigma_i] = \Re e[\mathcal{J}\omega a_i] > 0$ whereas $\Im m[a_i] \leq 0$, since $\Im m[\sigma_i] = \Im m[\mathcal{J}\omega a_i] \geq 0$. A very important property of the global reflection and intrinsic coefficient is thus proved which makes converging of the image series representation possible, as discussed in Sec. 2.4.

2.2.2. *Quasistatic point-charge potential*

The longitudinal derivative of the quasistatic Hertz potential $\partial\mathcal{P}(\mathbf{r},\mathbf{r}')/\partial z$, in (37), identified as a dipole potential, can be expressed alternatively as the difference between two point-charge responses [88, 98],

$$\frac{\partial\mathcal{P}(\mathbf{r},\mathbf{r}')}{\partial z} = \lim_{\ell\to 0}[\Phi(\mathbf{r},\mathbf{r}' - \hat{\mathbf{z}}\ell/2) - \Phi(\mathbf{r},\mathbf{r}' + \hat{\mathbf{z}}\ell/2)] = -\frac{\partial\Phi(\mathbf{r},\mathbf{r}')}{\partial z'}\ell. \tag{51}$$

The point-charge potential $\Phi(\mathbf{r},\mathbf{r}')$ can be expressed via the quasistatic Green's function as

$$\Phi(\mathbf{r},\mathbf{r}') = \frac{Q}{a_1}\mathcal{G}(\mathbf{r},\mathbf{r}'), \tag{52}$$

where

$$\mathcal{G}(\mathbf{r},\mathbf{r}') = \frac{1}{2\pi}\int_0^\infty \xi\mathcal{G}(z,z')J_0(\xi\rho)d\xi. \tag{53}$$

Finally the quasistatic characteristic Green's function $\mathcal{G}(z,z')$ is given by

$$\mathcal{G}_i(z,z') = \frac{1}{2\xi}\begin{cases}[e^{\xi z'}\prod_{p=1}^{i}\mathcal{T}_p(\xi)][e^{-\xi z} + \mathcal{R}_i(\xi)e^{\xi z}], & i > 0,\\ [e^{-\xi z'} + \mathcal{R}_1(\xi)e^{\xi z'}]e^{\xi z}, & i = 0.\end{cases} \tag{54}$$

The potential function $\Phi(\mathbf{r},\mathbf{r}')$, representing a normalized point-source response ($Q/a_1 = 1$), is more suitable for implementation in quasistatic problems than the potential difference, i.e. the dipole response [60, 98] represented by $\partial\mathcal{P}(\mathbf{r},\mathbf{r}')/\partial z$ in (37). Furthermore, since the expressions for $\Phi(z,z')$ and $\mathcal{P}(z,z')$ in (52) and (40), respectively, are similar up to the sign of the reflection coefficient $\mathcal{R}(\xi)$, the image series expansions of $\Phi(\mathbf{r},\mathbf{r}')$ and $\mathcal{P}(\mathbf{r},\mathbf{r}')$ in (54) and (39), respectively, are expected to be identical up to the sign of the corresponding terms (and the multiplication constant ℓ). Both $\mathcal{G}(\mathbf{r},\mathbf{r}')$ and $\mathcal{G}(z,z')$, in (53), satisfy 3-D and 1-D Laplace equations, respectively, and appropriate constraints, as summarized in Table 2.

2.3. *Infinite image series expansions*

In the quasistatic limit, $\mathcal{K}_i$, the intrinsic reflection coefficient in (43) is independent of the integration variable ξ, thus, the individual integrals contained in the finite expansions (Sec. 2.1) can be evaluated in closed form explicit expressions, interpreted as properly weighted and shifted point-source (image-source) responses. Since, the remainder integrals (i.e. (24), (29), (30) and (35)) can be made negligibly small by increasing the number of the expansion terms, under rather general set of constraints, to be discussed in the following section, the image series expansions may be regarded as converging representations. The complete image expansion derived here for $-\infty < z < z'$ is quite general and outlines the procedure for $i > 1$ without any increase in the complexity.

The following cases illustrate the procedure.

2.3.1. *Unbounded medium, $n = 0$*

The reduction of the radiation integral in (18) into a close form expression,

$$\mathcal{P}_1(\mathbf{r}, \mathbf{r}') = \frac{Q\ell}{4\pi a_1} \int_0^\infty e^{-\xi|z-z'|} J_0(\xi\rho) d\xi = \frac{Q\ell}{4\pi a_1} \frac{1}{|\mathbf{r} - \mathbf{r}'|} \tag{55}$$

is carried out via the Weber-Lipschitz integral identity [98] which is the quasistatic limit of the Sommerfeld identity ($\beta \to -\mathcal{J}\xi$ as $k(z) \to 0$). The integral identity is a vital tool in converting the finite integral summations into infinite image series representations.

2.3.2. *Semi-infinite medium, $n = 1$*

The quasistatic potential for $n = 1$, obtained from Eq. (20), is given by,

$$\mathcal{P}_1(\mathbf{r}, \mathbf{r}') = \frac{Q\ell}{4\pi a_1} \left[\frac{1}{|\mathbf{r} - \mathbf{r}'|} - \frac{\mathcal{K}_1}{|\mathbf{r} - \tilde{\mathbf{r}}|} \right], \quad \tilde{\mathbf{r}} = -\mathbf{r}' = (0, 0, -z'). \tag{56}$$

As depicted in Fig. 2, both contributions from the point-source at $\mathbf{r}' = (0, 0, z')$ and the image-source at $\tilde{\mathbf{r}}' = -\mathbf{r}' = (0, 0, \tilde{z}') = (0, 0, -z')$ reach the observation point P. The image-source contribution can be interpreted as the point-source contribution undergoing a single reflection ($\mathcal{K}_1$) at $z = z_1 = 0$.

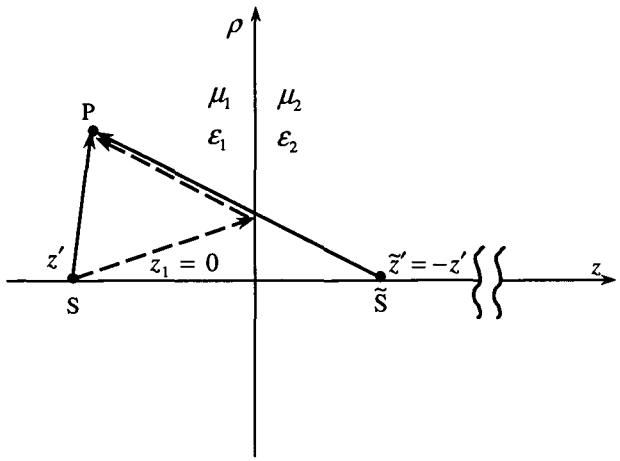


Fig. 2. Physical configuration for a semiinfinite medium, $n = 1$. Both contributions from the point-source S at $\mathbf{r}' = (0, 0, z')$ and the image-source $\tilde{S}$ at $\tilde{\mathbf{r}}' = -\mathbf{r}' = (0, 0, \tilde{z}') = (0, 0, -z')$ reach the observation point P. The image-source ($\tilde{S}$) contribution (solid line) can be interpreted as the point-source (S) contribution undergoing a single reflection ($\mathcal{K}_1$) at $z = z_1 = 0$ (dashdot line).

2.3.3. *Single slab configuration, $n = 2$*

The quasistatic Hertz potential for a three layers medium, obtained from Eq. (23), is given [10] by,

$$\mathcal{P}_1(\mathbf{r}, \mathbf{r}') = \frac{Q\ell}{4\pi a_1}\left[\frac{1}{|\mathbf{r}-\mathbf{r}'|} - \sum_{l_1=0}^{1}\sum_{m_1=0}^{\infty}\binom{1}{l_1}\binom{m_1-l_1}{m_1} \times \frac{K_1^{l_1}[1-\mathcal{K}_1^2]^{-l_1+1}[-\mathcal{K}_1(\xi)]^{m_1}\mathcal{K}_2^{m_1+1}}{|\mathbf{r}-\tilde{\mathbf{r}}_{l_1,m_1}|}\right],$$

$$\tilde{\mathbf{r}}_{l_1,m_1} = [0, 0, 2z_2(m_1 - l_1 + 1) - z']. \tag{57}$$

As depicted in Fig. 3, both contributions from the point-sources at $\mathbf{r}' = (0, 0, z')$ and the image-source set located at $\tilde{\mathbf{r}}'_{m_1,l_1}, (l_1, m_1) = (1, 0), (0, 0), (1, 0)$, reach the observation point P. The contribution of the image-source set can be interpreted as a summation over all the point-source responses undergoing, at $z = z_1 = 0$, either a single reflection ($\mathcal{K}_1$, $l_1 = 1$, $m_1 = 0$) or a single transmission ($l_1 = 0$) in $(1 + \mathcal{K}_1)$ and out $(1 - \mathcal{K}_1)$ accompanied by single reflection ($\mathcal{K}_2$) at $z = z_2$ and m_1 ($m_1 \geq 0$) bounces both at z_1 ($(-\mathcal{K}_1)^{m_1}$) and z_2 ($\mathcal{K}_2^{m_1+1}$). Note that for $l_1 = 1$ the series expansion in (57) is reduced into a single term.

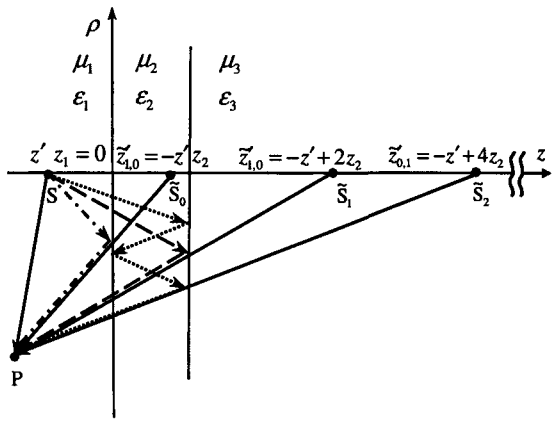


Fig. 3. Physical configuration for a single slab, $n = 2$. Both contributions from the point-sources S at $\mathbf{r}' = (0, 0, z')$ and the image-source set $\tilde{S}_{l_1,m_1}$ at $\tilde{\mathbf{r}}'_{m_1,l_1}, (l_1, m_1) = (1, 0), (0, 0), (1, 0)$, reach the observation point P. The image-source set $(\tilde{S}_{l_1,m_1})$ contribution (solid line) can be interpreted as a point-source (S) contribution undergoing, at $z = z_1 = 0$, either a single reflection ($\mathcal{K}_1$, dashdot line, $l_1 = 1$, $m_1 = 0$) or a single transmission in $(1 + \mathcal{K}_1)$ and out $(1 - \mathcal{K}_1)$ accompanied by single reflection $(\mathcal{K}_2)$ at $z = z_2$ ($l_1 = 0$)and m_1 ($m_1 \geq 0$) bounces both at z_1 ($(-\mathcal{K}_1)^{m_1}$) and z_2 ($(-\mathcal{K}_2)^{m_1}$), dashed and dotted lines for $m_1 = 0$, and $m_1 = 1$, respectively.

2.3.4. *Double slab configuration, $n = 3$*

The quasistatic Hertz potential for a four layers medium, obtained from (28), is expressed as,

$$\mathcal{P}_1(\mathbf{r}, \mathbf{r}') = \frac{Q\ell}{4\pi a_1} \left[\frac{1}{|\mathbf{r} - \mathbf{r}'|} - \sum_{l_1=0}^{1} \sum_{m_1=0}^{\infty} \sum_{l_2=0}^{m_1-l_1+1} \sum_{m_2=0}^{\infty} \times \binom{1}{l_1} \binom{m_1 - l_1}{m_1} \binom{m_1 - l_1 + 1}{l_2} \binom{m_1 + m_2 - l_1 - l_2}{m_2} \times \frac{\mathcal{K}_1^{l_1} \mathcal{K}_2^{l_2} [1 - \mathcal{K}_1]^{-l_1+1} [1 - \mathcal{K}_2]^{m_1-l_1-l_2+1} [-\mathcal{K}_1]^{m_1}}{|\mathbf{r} - \tilde{\mathbf{r}}'_{l_1,m_1,l_2,m_2}|} \right],$$

$$\tilde{\mathbf{r}}'_{l_1,m_1,l_2,m_2} = [0, 0, 2(m_1 - l_1 + 1)z_2 + (m_1 + m_2 - l_1 - l_2 + 1)(z_3 - z_2) - z']. \quad (58)$$

As depicted in Fig. 4, both contributions from the point-sources S, at $\mathbf{r}' = (0, 0, z')$ and the image-source set at $\tilde{\mathbf{r}}'_{l_1,m_1,l_2,m_2}$, reach the observation point P. The contribution of the image-source set can be interpreted as a summation, over all the point-source responses undergoing, at $z = z_1 = 0$, either a single reflection ($\mathcal{K}_1$, $l_1 = 1$) or a single transmission ($l_1 = 0$)

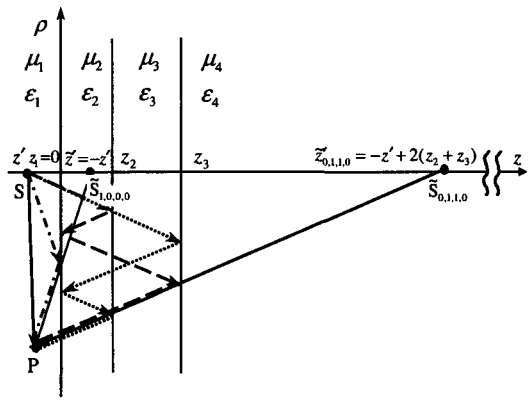


Fig. 4. Physical configuration for a double slab geometry, $n = 3$. Both contributions from the point-sources S at $\mathbf{r}' = (0, 0, z')$ and the image-source set $\tilde{S}_{l_1,m_1,l_2,m_2}$ at $\tilde{\mathbf{r}}'_{l_1,m_1,l_2,m_2}$, where $(l_1, m_1, l_2, m_2) = (1, 0, 0, 0), (0, 1, 1, 0)$ reach the observation point P. The image-source set $(\tilde{S}_{l_1,m_1,l_2,m_2})$ contribution (solid line) can be interpreted as a point-source (S) contribution undergoing, at $z = z_1 = 0$, either a single reflection ($\mathcal{K}_1$, dashdot line, $l_1 = 1$) or a single transmission ($l_1 = 0$) in $(1 + \mathcal{K}_1)$ and out $(1 - \mathcal{K}_1)$, accompanied by all possible combinations of bounces and transmissions $z = z_1 = 0$, $z = z_2$ and $z = z_3$, $\binom{m_1+1}{l_2}\binom{m_1+m_2-l_2}{m_2}$, where m_1 $(m_1 \geq 0)$ and m_2 denote the number of internal reflections at $z = z_1 = 0$ $((-\mathcal{K}_1)^{m_1})$ and $z = z_2$ $(-\mathcal{K}_2)^{m_2})$, respectively, and $m_1 - l_2 + 1$ is the number of transmission in $(1 + \mathcal{K}_2)$ and out $(1 - \mathcal{K}_2)$ at $z = z_2$ $((l_1, m_1, l_2, m_2) = (0, 1, 1, 0)$ associated with the two only combinations depicted by dashed and dotted lines).

in $(1 + \mathcal{K}_1)$ and out $(1 - \mathcal{K}_1)$, accompanied by all possible combinations of bounces and transmissions at $z = z_1 = 0$, $z = z_2$ and $z = z_3$, $\binom{m_1+1}{l_2}\binom{m_1+m_2-l_2}{m_2}$, where m_1 $(m_1 \geq 0)$ and m_2 denote the number of internal reflections at z_1 $((-\mathcal{K}_1)^{m_1})$ and z_2 $((-\mathcal{K}_2)^{m_2})$, respectively, and $m_1 - l_2 + 1$ is the number of transmission in $(1 + \mathcal{K}_2)$ and out $(1 - \mathcal{K}_2)$ at $z = z_2$. It should be noted that for either $m_1 - l_2 + 1 = 0$ or $\mathcal{K}_3 = 0$ the series expansion in (58) is reduced into a single geometric series.

2.3.5. $n + 1$ *layered media*

Upon utilizing (34) one obtain

$$\mathcal{P}_1(\mathbf{r}, \mathbf{r}')$$
$$= \frac{Q\ell}{4\pi a_1}\left[\frac{1}{|\mathbf{r} - \mathbf{r}'|} - \sum_{l_1=0}^{1}\sum_{m_1=0}^{\infty}\cdots\sum_{l_{n-1}=0}^{s_{n-2}}\sum_{m_{n-1}=0}^{\infty}\left\{\prod_{k=1}^{n-1}\binom{s_{k-1}+1}{l_k}\binom{s_k}{m_k}\right.\right.$$
$$\left.\left.\times\,\mathcal{K}_k^{l_k}[1 - \mathcal{K}_k^2]^{s_{k-1}-l_k+1}[-\mathcal{K}_k]^{m_k}\mathcal{K}_n^{m_k-l_k}\right\}\frac{\mathcal{K}_n}{[x^2 + y^2 + (z - z'_k)^2]^{1/2}}\right], \tag{59}$$

where

$$z'_k = 2\sum_{k=1}^{n-1}(s_k+1)(z_{k+1}-z_k) - z'. \tag{60}$$

Expressions in the quasistatic limit for any layer and any number of layers within the medium were recently reported [61].

2.4. *Convergence and truncation-error estimation: the collective image approach*

The quasistatic reduction process which was successfully utilized to generate closed form image series expansions, in the previous section, is applied here for reduction of the remainder integral terms, such as, $\Gamma_{M_1}(\mathbf{r},\mathbf{r}')$, and $\Gamma_{M_1,M_2}(\mathbf{r},\mathbf{r}')$, in (24) (29), and (30), respectively. Two goals are readily accomplished: (i) the remainders are shown to be negligibly small for sufficiently large summation indices warranting the convergence of associated image series under rather general physically interpretable, set of constraints; (ii) closed form asymptotic expressions (end-point contributions), obtained via integration by parts, enable accurate truncation-error estimations, for sufficiently large number of summation indices. The asymptotic remainders are regarded here as collective image contributions. Image series expansions including a finitely small number of ordinary image terms together with (asymptotic) collective image contributions, are shown to converge faster than expansions containing ordinary image terms only. The following examples illustrate the procedure.

2.4.1. *Single slab configuration, $n = 2$*

Equations (22) and (24) are reduced, in the quasi-static limit, into,

$$\lambda_{M_1}(z,z') = -\frac{(1-\mathcal{K}_1^2)(-\mathcal{K}_1)^{M_1+1}\mathcal{K}_2^{M_1+2}e^{\xi[z+z'-2z_2(M_1+2)]}}{1+\mathcal{K}_1\mathcal{K}_2e^{-\xi z_2}} \tag{61}$$

and

$$\Lambda_{M_1}(\mathbf{r},\mathbf{r}') = \int_0^\infty \lambda_{M_1}(z,z')J_0(\xi\rho)\mathrm{d}\xi, \tag{62}$$

respectively. Equation (62) can readily be reduced into the following inequality,

$$|\Lambda_{M_1}(\mathbf{r},\mathbf{r}')| < \frac{(1-\mathcal{K}_1^2)|\mathcal{K}_1|^{M_1+1}|\mathcal{K}_2|^{M_1+2}}{(1-|\mathcal{K}_1\mathcal{K}_2|)[2z_2(M_1+2)-z-z']}, \tag{63}$$

which guarantees the convergence of the image series expansion in (57), if at least one of $|\mathcal{K}_1|$ or $|\mathcal{K}_2|$ is less than unity, i.e. $\Lambda_{M_1}(\mathbf{r},\mathbf{r}') \to 0$ as $M_1 \to \infty$, if either $|\mathcal{K}_1| < 1$ or $|\mathcal{K}_2| < 1$. The collective image contribution can be obtained asymptotically (for large M_1) via integration by the parts of (62), yielding,

$$\Lambda_{M_1}(\mathbf{r},\mathbf{r}') \sim \bar{\Lambda}_{M_1}(z,z') = -\frac{[1-\mathcal{K}_1^2](-\mathcal{K}_1)^{M_1+1}\mathcal{K}_2^{M_1+2}}{(1+\mathcal{K}_1\mathcal{K}_2)[2z_2(M_1+2)-z-z']}, \tag{64}$$

where $\bar{\Lambda}_{M_1}(z,z')$ is the asymptotic collective image contribution.

2.4.2. *Double slab configuration, $n = 3$*

Equations (26), (27), (29) and (30) are reduced, in the quasistatic limit, into,

$$\lambda_{M_1}(z,z') = -\frac{[1-\mathcal{K}_1^2](-\mathcal{K}_1)^{M_1+1}[\mathcal{R}_2(\xi)]^{M_1+2}e^{\xi(z+z')}}{1+\mathcal{K}_1\mathcal{R}_2(\xi)}, \tag{65}$$

$$\begin{aligned}
&\lambda_{M_1,M_2}(z,z') \\
&= -\sum_{m_1=0}^{M_1}\sum_{l_2=0}^{m_1+1}\sum_{p=0}^{m_1-l_2}\binom{m_1+1}{l_2}\binom{M_2+m_1-l_2+1}{M_2+p_2+1} \\
&\quad\times(1-\mathcal{K}_1^2)(-\mathcal{K}_1)^{m_1}(\mathcal{K}_2)^{l_2}\{[1-\mathcal{K}_2^2]\mathcal{K}_3\}^{m_1-l_2+1} \\
&\quad\times\frac{(-\mathcal{K}_2\mathcal{K}_3)^{M_2+p_2+1}e^{-\xi[2(z_3-z_2)(M_2+p_2+m_1-l_2+2)+2z_2(m_1+1)-z-z']}}{[1-\mathcal{K}_2\mathcal{K}_3e^{-2\xi(z_3-z_2)}]^{p_2+1}},
\end{aligned} \tag{66}$$

$$\Lambda_{M_1}(\mathbf{r},\mathbf{r}') = \int_0^\infty \lambda_{M_1}(z,z')J_0(\xi\rho)d\xi \tag{67}$$

and

$$\Lambda_{M_1,M_2}(\mathbf{r},\mathbf{r}') = \int_0^\infty \lambda_{M_1,M_2}(z,z')J_0(\xi\rho)d\xi, \tag{68}$$

respectively. Equations (67) and (68) can be manipulated into the following inequalities:

$$|\Lambda_{M_1}(\mathbf{r},\mathbf{r}')| < \frac{[1-\mathcal{K}_1^2]|\mathcal{K}_1|^{M_1+1}|\mathcal{R}_2(\xi_0)e^{2\xi_0 z_2}|^{M_1+2}}{[1-|\mathcal{K}_1\mathcal{R}_2(\xi_1)|][2z_2(M_1+2)-z-z']}, \tag{69}$$

where ξ_0 and ξ_1 are defined via $|1+\mathcal{K}_1\mathcal{R}_2(\xi)| \geq 1-|\mathcal{K}_1\mathcal{R}_2(\xi)| \geq 1-|\mathcal{K}_1\mathcal{R}_2(\xi_1)|$ and $\mathcal{R}_2(\xi)e^{2\xi z_2} \leq |\mathcal{R}_2(\xi_0)e^{2\xi_0 z_2}| \leq 1$, $0 \leq \xi \leq \infty$ respectively,

and

$$|\Lambda_{M_1,M_2}(\mathbf{r},\mathbf{r}')| < [1-\mathcal{K}_1^2]\sum_{m_1=0}^{M_1}\sum_{l_2=0}^{m_1+1}\sum_{p_2=0}^{m_1-l_2}\binom{m_1+1}{l_2}\binom{M_2+m_1-l_2+1}{M_2+p_2+1}$$
$$\times\frac{|-\mathcal{K}_1|^{m_1}|\mathcal{K}_2|^{l_2}\{[1-\mathcal{K}_2^2]|\mathcal{K}_3|\}^{m_1-l_2+1}|\mathcal{K}_2\mathcal{K}_3|^{M_2+p_2+1}}{(1-|\mathcal{K}_2\mathcal{K}_3|)^{p_2+1}[2(z_3-z_2)(M_2+p_2+m_1-l_2+2)+2z_2(m_1+1)-z-z']}. \tag{70}$$

The image series expansion in (58) converges if at least two of $|\mathcal{K}_1|$, $|\mathcal{K}_2|$ $|\mathcal{K}_3|$ are less than unity, i.e. $\Lambda_{M_1}(\mathbf{r},\mathbf{r}')\to 0$ as $M_1\to\infty$, if either $|\mathcal{K}_1|<1$ or $|\mathcal{R}_2(\xi_0)e^{2\xi_0 z_2}|<1$. Note that $|\mathcal{R}_2(\xi_0)e^{2\xi_0 z_2}|<1$ if both $|\mathcal{K}_2|<1$ and $|\mathcal{K}_3|<1$ (see Sec. 2.2). Similarly, $\Lambda_{M_1,M_2}(\mathbf{r},\mathbf{r}')\to 0$ as $M_2\to\infty$, if either $|\mathcal{K}_2|<1$ or $|\mathcal{K}_3|<1$.

The collective image contributions can be obtained asymptotically (for large M_1 and M_2) via integration by the parts of (67) and (68), yielding,

$$\Lambda_{M_1}(\mathbf{r},\mathbf{r}')\sim\bar{\Lambda}_{M_1}(z,z')=-\frac{[1-\mathcal{K}_1^2](-\mathcal{K}_1)^{M_1+1}[\mathcal{R}_2(0)]^{M_1+2}}{[1+\mathcal{K}_1\mathcal{R}_2(0)][2z_2(M_1+2)-z-z']},$$
$$\mathcal{R}_2(0)=\frac{a_2-a_4}{a_2+a_4}, \tag{71}$$

where $\mathcal{R}_2(0)$ is derived in 2.2, and

$$\Lambda_{M_1,M_2}(\mathbf{r},\mathbf{r}')\sim\bar{\Lambda}_{M_1,M2}(z,z')$$
$$=-\sum_{m_1=0}^{M_1}\sum_{l_2=0}^{m_1+1}\sum_{p_2=0}^{m_1-l_2}\binom{m_1+1}{l_2}\binom{M_2+m_1-l_2+1}{M_2+p_2+1}$$
$$\times\frac{(1-\mathcal{K}_1^2)(-\mathcal{K}_1)^{m_1}\mathcal{K}_2^{l_2}[(1-\mathcal{K}_2^2)\mathcal{K}_3]^{m_1-l_2+1}(-\mathcal{K}_2\mathcal{K}_3)^{M_2+p_2+1}}{(1+\mathcal{K}_2\mathcal{K}_3)^{p_2+1}[2(z_3-z_2)(M_2+p_2+m_1-l_2+2)+2z_2(m_1+1)-z-z']}. \tag{72}$$

The asymptotic collective image contribution in (71) and (72) is denoted by $\bar{\Lambda}_{M_1}(z,z')$ and $\bar{\Lambda}_{M_1,M2}(z,z')$, respectively. Note that for either $m_1-l_2+1=0$ $\mathcal{K}_3=0$, $\bar{\Lambda}_{M_1,M2}(z,z')$ in (72) is equal to zero.

2.4.3. $n+1$ *layered media*

The remainder term obtained via end-point integration is given as

$$\bar{\Lambda}_{M_1,\ldots,M_k}(z,z')=-\sum_{m_0=0}^{M_0}\sum_{l_1=0}^{1}\cdots\sum_{m_{k-1}=0}^{M_{k-1}}\sum_{l_k=0}^{s_{k-1}+1}\sum_{p_k=0}^{s_{k-1}-l_k}$$
$$\times\left\{\prod_{j=1}^{k}\binom{s_{j-1}+1}{l_j}\binom{s_{j-1}}{m_{j-1}}\right.$$

$$\left.\times \mathcal{K}_j^{l_j}[1-\mathcal{K}_j^2]^{s_{j-1}-l_j+1}[-\mathcal{K}_{j-1}]^{m_{j-1}}\right\}$$

$$\times \begin{pmatrix} M_k+s_{k-1}-l_k+1 \\ M_k+p_k+1 \end{pmatrix}$$

$$\times \frac{[-\mathcal{K}_k^{M_k+p_k+1}[\mathcal{R}_{k+1}(0)]^{M_k+p_k+s_{k-1}-l_k+2}]}{[1+K_k\mathcal{R}_{k+1}(0)]^{p_k+1}(z-\tilde{Z}_k')}, \quad (73)$$

where

$$\tilde{Z}_k' = 2\sum_{j=2}^{k}[(z_j-z_{j-1})(s_{j-1}+1)]$$
$$+(z_{k+1}-z_k)(s_{k-1}-l_k+M_k+p_k+2)-z'. \quad (74)$$

In n-layered media, the finite expansion converges in the quasistatic limit, if at least $n-1$ of the intrinsic reflection coefficients $|\mathcal{K}_i|$, $1 \le i \le n$, are less then unity, i.e. $\Lambda_{M_1}(\mathbf{r},\mathbf{r}') \to 0$ as $M_1 \to \infty$, if either $|\mathcal{K}_1| < 1$ or $|\mathcal{R}_2(\xi_0)e^{2\xi_0 z_2}| < 1$. Note that $|\mathcal{R}_2(\xi_0)e^{2\xi_0 z_2}| < 1$ if both $|\mathcal{K}_2| < 1$ and $|\mathcal{R}_3(\xi)e^{2\xi z_3}| < 1$, etc. (see Sec. 2.2). Finally $\Lambda_{M_1,M_2,\ldots,M_{n-1}}(\mathbf{r},\mathbf{r}') \to 0$ as $M_{n-1} \to \infty$, if either $|\mathcal{K}_{n-1}| < 1$ or $|\mathcal{K}_n| < 1$.

The effectiveness of the collective image approach is demonstrated in Figs. 5 and 6, for $n = 2$ and $n = 3$, respectively. Both figures show the normalized truncation error dependence of the image series expansion on

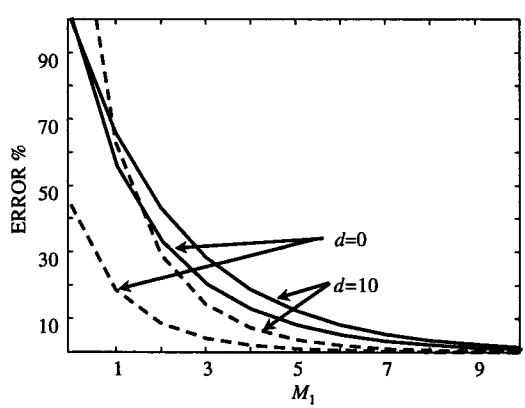


Fig. 5. Normalized truncation error of finite image series expansion truncated at M_1, for a single slab configuration, $n = 2$. Contributions incorporating either ordinary image terms only, $100|\Lambda_{M_1}(\mathbf{r},\mathbf{r}')|/|\Lambda_0(\mathbf{r},\mathbf{r}')|$ (Eq. (62)), or both the ordinary and collective image terms, $100|\Lambda_{M_1}(\mathbf{r},\mathbf{r}')-\bar{\Lambda}_{M_1}(z,z')|/|\Lambda_0(\mathbf{r},\mathbf{r}')|$ (Eq. (64)), are denoted by solid and dashed lines, respectively. The simulation parameters: $\mathcal{K}_1 = -0.89, \mathcal{K}_2 = 0.75$, $z_1 = 0.0$, $z_2 = 0.03\,m$, $z' = -0.01\,m$, $\mathbf{r}' = (0,0,z')$, $\mathbf{r} = (0,0,dz')$.

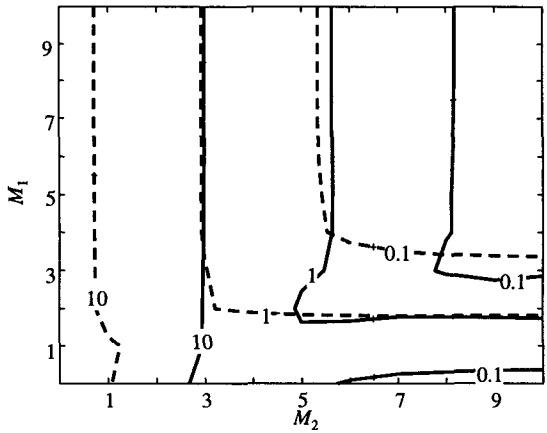


Fig. 6. Truncation error contours of finite image series expansion truncated at M_1, M_2, for a double slab configuration, $n = 3$. Contributions incorporating either ordinary image terms only, $100|\Lambda_{M_1}(\mathbf{r},\mathbf{r}') + \Lambda_{M_1,M_2}(\mathbf{r},\mathbf{r}')|/|\Lambda_0(\mathbf{r},\mathbf{r}') + \Lambda_{0,0}(\mathbf{r},\mathbf{r}')|$ (Eqs. (67), (68)), or both ordinary and collective image terms, $100|\Lambda_{M_1}(\mathbf{r},\mathbf{r}') + \Lambda_{M_1,M_2}(\mathbf{r},\mathbf{r}') - \bar{\Lambda}_{M_1}(z,z') - \bar{\Lambda}_{M_1,M_2}(z,z')|/|\Lambda_0(\mathbf{r},\mathbf{r}') + \Lambda_{0,0}(\mathbf{r},\mathbf{r}')|$ (Eqs. (71), (72)), are denoted by solid and dashed lines, respectively. The simulation parameters: $\mathcal{K}_1 = 0.65$, $\mathcal{K}_2 = -0.75$, $\mathcal{K}_3 = 0.85$, $z_1 = 0$, $z_2 = 0.01m$, $z_3 = 0.02m$, $z' = -0.01m$, $\mathbf{r}' = (0,0,z')$, $\mathbf{r} = (0,0,10z')$.

the summation indices (M_1 for $n = 2$ and M_1, M_2 for $n = 3$) using biological medium parameters [60, 62].

The highest convergence rate is always achieved via contributions incorporating both ordinary and collective image terms, thereby, establishing the superiority of the collective image approach over ordinary image summation. This proved important when we further used this expansion, through the moment method, to calculate the potential distribution due to finite electrode array in multilayered media (Sec. 3).

3. Electrode Array in Layered Media

3.1. *Integral equation formulation*

The physical configuration of our problem, depicted in Fig. 7 consists of a stratified biological medium with n boundaries separating between the $n+1$ homogeneous and isotropic layers. Each layer is characterized by its thickness, conductivity $\sigma(z)$ (generally complex), where $\sigma(z) = \sigma_i$ as defined in (6). An array of P rectangular electrodes is placed in the first layer ($i = 1$). The evaluation of the electrodes' current distributions and potentials is carried out within the quasistatic (low-frequency) regime. Assuming that all

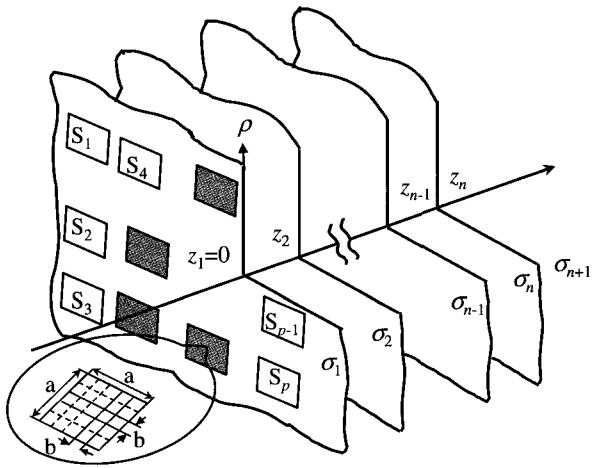


Fig. 7. Physical configuration for a layered biological tissue excited by an array of finite electrodes.

the P electrode plates (Fig. 7) are perfect conductors, i.e. constant potential patches, the potential of each electrode V is specified. Hence, the problem constitutes a system of P Fredholm integral equations of the first kind [61] for the electrodes' current distribution $i_p(\mathbf{r}_p)$,

$$V_q = \Phi(\mathbf{r}_q) = \frac{1}{\sigma_1} \sum_{p=1}^{P} \oint_{S_p} i_p(\mathbf{r}_p) \mathcal{G}(\mathbf{r}_q, \mathbf{r}_p) ds_p, \quad q = 1, 2, \ldots, P, \tag{75}$$

where $i_p(\mathbf{r})$, the pth electrode current distribution, $\Phi(\mathbf{r})$ can be expressed as a superposition over all the electrode potentials $\Phi_p(\mathbf{r})$,

$$\Phi(\mathbf{r}) = \sum_{p=1}^{P} \Phi_p(\mathbf{r}), \tag{76}$$

defined via the convolution integral

$$\Phi_p(\mathbf{r}) = \frac{1}{\sigma_1} \oint_{S_p} i_p(\mathbf{r}_p) \mathcal{G}(\mathbf{r}, \mathbf{r}_p) ds_p. \tag{77}$$

The point-source response $\mathcal{G}(\mathbf{r}, \mathbf{r}_p)$ (Green's function, Table 2) can be represented most effectively as the image series expansion i.e. a collection of properly weighted and shifted point-source responses and a remainder term (collective image) as presented in Sec. 2.2 [60, 61]. We note that $\sigma(z)$ should be replaced by the complex conductivity $\varsigma(z) = \sigma(z) + \jmath\omega\epsilon(z)$, in every layer for which the inequality $\sigma(z) \gg \jmath\omega\epsilon(z)$ is not satisfied. The parameter

ω denotes the angular frequency corresponding to the electrode excitation. To simplify the notation we allow σ to be complex in the remainder of the paper.

3.2. *Electrode array*

The total current of each electrode I_p is obtained by integration of the electrode current density distribution $i_p(\mathbf{r}_p)$ over the electrode surface,

$$I_p = \oint_{S_p} i_p(\mathbf{r}_p) ds_p. \tag{78}$$

The uniqueness of the solution of system (75) in conjunction with the superposition principle leads to the following linear relation between the electrode currents and electrode voltages

$$\mathbf{V} = \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_P \end{pmatrix} = \begin{pmatrix} \mathrm{R}_{11} & \mathrm{R}_{12} & \mathrm{R}_{13} & \dots & \mathrm{R}_{1P} \\ \mathrm{R}_{21} & \mathrm{R}_{22} & \mathrm{R}_{23} & \dots & \mathrm{R}_{2P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{R}_{P1} & \mathrm{R}_{P2} & \mathrm{R}_{P3} & \dots & \mathrm{R}_{PP} \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \\ \vdots \\ I_P \end{pmatrix} = \mathbf{RI}, \tag{79}$$

or, alternatively

$$\mathbf{I} = \begin{pmatrix} \mathrm{G}_{11} & \mathrm{G}_{12} & \mathrm{G}_{13} & \dots & \mathrm{G}_{1P} \\ \mathrm{G}_{21} & \mathrm{G}_{22} & \mathrm{G}_{23} & \dots & \mathrm{G}_{2P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{G}_{P1} & \mathrm{G}_{P2} & \mathrm{G}_{P3} & \dots & \mathrm{G}_{PP} \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_P \end{pmatrix} = \mathbf{GV}, \tag{80}$$

where $\mathbf{G}$ and $\mathbf{R} = \mathbf{G}^{-1}$ denote the input conductance (admittance) and the input resistance (impedance) $P \times P$ matrices of the electrode array feeding network, respectively. It can be readily shown that the matrices $\mathbf{R}$ and $\mathbf{G}$ are symmetric due to the reciprocity property of $\mathcal{G}(\mathbf{r}, \mathbf{r}_p)$, i.e. $\mathcal{G}(\mathbf{r}_q, \mathbf{r}_p) = \mathcal{G}(\mathbf{r}_p, \mathbf{r}_q)$. Furthermore, all the diagonal elements of both matrices are positive, whereas, the off-diagonal elements are positive for $\mathbf{R}$ and negative for $\mathbf{G}$.

In view of Kirchhoff's current law, the sum of all the electrode currents must be zero,

$$(1, 1, \dots, 1)\mathbf{I} = (1, 1, \dots, 1)\mathbf{GV} = 0, \tag{81}$$

i.e. (LHS),

$$\sum_{p=1}^{P} I_p = 0, \tag{82}$$

or, equivalently (RHS),

$$\sum_{p=1}^{P} \alpha_p V_p = 0, \quad \alpha_p = \sum_{q=1}^{P} \mathrm{G}_{qp}. \tag{83}$$

This restriction leads to the conclusion that only $P-1$ of the elements of either the vector $\mathbf{V}$ or the vector $\mathbf{I}$ can be arbitrarily selected. Thus, the remaining $P+1$ elements of $\mathbf{V}$ and $\mathbf{I}$ are explicitly specified via either (82) or (83), and (80).

The total complex power S, delivered by a P-electrode array, can be expressed in terms of the vector $\mathbf{V}$ and complex conjugate of the vector $\mathbf{I}$,

$$S = \frac{1}{2}\mathbf{V}^T\mathbf{I}^*. \tag{84}$$

Note that for $\omega = 0$, the real input power is $S = \mathbf{V}^T\mathbf{I}$.

3.3. *Moment method*

The integral equation system in (75) can be inverted using the moment method with pulse base for the electrode current distribution and point match for the potential [38]. The discretized electrode potential $\mathbf{\Phi}$ is a linear transformation of the discretized current density distribution $\boldsymbol{i}$ via $\mathbf{L}$,

$$\mathbf{\Phi} = \mathbf{L}\boldsymbol{i}. \tag{85}$$

The moment matrix $\mathbf{L}$ is a square matrix specified by its elements ℓ_{mn} (representing the potential at the center of the subsection m due to unit current density distribution on the subsection n), given as,

$$\ell_{mn} = \frac{1}{\sigma_1} \int_{x_n-b/2}^{x_n+b/2} \int_{y_n-b/2}^{y_n+b/2} \mathcal{G}(\mathbf{r}_m, \mathbf{r}'_n) \mathrm{d}x'_n \mathrm{d}y'_n, \tag{86}$$

where $\mathbf{r}_m$ and $\mathbf{r}_n$ represent the location of the observation and source points, respectively. It can be readily verified that the discretization quantum is a square element of size $b \times b$, thus a square electrode of size $a \times a$ contains $N = (a/b)^2$ subdivisions (Fig. 7). Hence, a problem involving P identical square electrodes associates with vectors $\boldsymbol{i}$ and $\mathbf{\Phi}$ of size PN and a moment matrix $\mathbf{L}$ of size $(PN) \times (PN)$. An explicit closed-form expression for the moment matrix element can be obtained by substituting (59) and (73) from Sec. 2 in (86) and utilizing the identity,

$$\begin{aligned} f(x,y,z) &= \int^{x} \int^{y} \frac{\mathrm{d}x'\mathrm{d}y'}{r'} \\ &= x \ln(y+r) + y \ln(x+r) - z \arctan(xy/zr), \end{aligned} \tag{87}$$

where $r' = (x'^2 + y'^2 + z^2)^{1/2}$. In this derivation we make use of reported procedures [25, 46], or more efficiently, of symbolic software *Mathematica* 3 (Wolfram Research Corp.). The expression can be reduced into Bancroft's result [5], upon setting $z = 0$. The resultant element ℓ_{mn} is given by,

$$\begin{aligned}\ell_{mn} = \frac{1}{4\pi\sigma_1}[&h(x_m - x_n + b/2, y_m - y_n + b/2, z_m)\\ &- h(x_m - x_n + b/2, y_m - y_n - b/2, z_m)\\ &- h(x_m - x_n - b/2, y_m - y_n + b/2, z_m)\\ &+ h(x_m - x_n - b/2, y_m - y_n - b/2, z_m)],\end{aligned} \tag{88}$$

where $h(x, y, z_m)$ is expressed via the (59)

$$\begin{aligned}h(x,y,z) = {} & f(x,y,z-z')\\ &+ \sum_{l_1=0}^{1}\sum_{m_1=0}^{M_1}\cdots\sum_{l_{n-1}=0}^{s_{n-2}}\sum_{m_{n-1}=0}^{M_{n-1}}\left\{\prod_{k=1}^{n-1}\binom{s_{k-1}+1}{l_k}\binom{s_k}{m_k}\right.\\ &\left.\times\, \mathcal{K}_k^{l_k}[1-\mathcal{K}_k^2]^{s_{k-1}-l_k+1}[-\mathcal{K}_k]^{m_k}\mathcal{K}_n^{m_k-l_k} f(x,y,z+z_k')\right\}\mathcal{K}_n\\ &+ \sum_{k=1}^{n-1}\bar{\Lambda}_{M_1,\ldots,M_k}(z,z')b^2.\end{aligned} \tag{89}$$

The last term in the RHS of (89) represents an asymptotic error estimation (Eq. (73)) of ℓ_{mn} due to the truncated image series expansion in (59). This collective image term significantly accelerates the image series convergence and the overall algorithm speed.

Note that, the electrode voltages and currents in (80) are related to the discretized electrode potential and current density distribution in (85) via,

$$\mathbf{\Phi} = \mathbf{UV} = \mathbf{UG}^{-1}\mathbf{I}, \tag{90}$$

and

$$\mathbf{I} = b^2\mathbf{U}^T\boldsymbol{i}, \tag{91}$$

where $\mathbf{U}$ is a $(PN) \times P$ rectangular matrix,

$$U_{ij} = \begin{cases}1 & \text{if}, 1 + N(j-1) \le i \le Nj\\ 0 & \text{otherwise},\ i = 1, 2, \ldots, PN,\ j = 1, 2, \ldots, P.\end{cases} \tag{92}$$

Thus, using (90), Eq. (85) can be uniquely inverted once either the $P-1$ electrode voltages or electrode currents are specified (Eqs. (81)–(83)). Furthermore, upon utilizing (91) as well, the conductance matrix $\mathbf{G}$ is completely determined via $\mathbf{L}^{-1}$

$$\mathbf{G} = b^2\mathbf{U}^T\mathbf{L}^{-1}\mathbf{U}. \tag{93}$$

3.4. *Electrode array excitation of layered biological tissue: numerical simulations*

The hybrid image series and moment method scheme that has been outlined in the previous sections is applied herein for numerical calculations. The simulations are selected to address simple, yet fundamental, concepts associated with low-frequency interaction between electromagnetic field and biological tissues. Thereby, they demonstrate the potential promise of the hybrid scheme that is capable of efficiently handling 3-D problems in layered media excited by an array of finite electrodes of arbitrary (generally non-planar) shape.

Since, all of the calculations are carried out for the physical configuration depicted in Fig. 7 where $z_p = z_1 = 0$, $p = 1, 2, \ldots, P$ (z_p is a component of $\mathbf{r}_p = (x_p, y_p, z_p)$), $\omega = 0$, and $\sigma_1 = 0$ (air layer), the expression for $\mathcal{G}_1(\mathbf{r}, \mathbf{r}')$ in (59) has to be modified in accordance with the identity,

$$\frac{1+\mathcal{K}_1}{\sigma_1} = \frac{1-\mathcal{K}_1}{\sigma_2}, \quad \lim_{\sigma_1 \to 0} \frac{1+\mathcal{K}_1}{\sigma_1} = \frac{2}{\sigma_2}. \tag{94}$$

Furthermore, the simulations are calculated assuming perfect conducting electrode plates discretized as $b = 0.05a$ ($N = (a/b)^2 = 400$) and the following typical FES parameters [32]: $n = 4$, $\sigma_1 = 0$ (air), $\sigma_2 = 0.4$ S/m $(wet\ skin)$, $\sigma_3 = 0.04$ S/m (fat), $\sigma_4 = 0.7$ S/m $(muscle)$, $\sigma_5 = 0.07$ S/m $(bone/fascia)$, $z_1 = 0$, $z_2 = 0.005$ m, $z_3 = 0.01$ m, $z_4 = 0.04$ m.

3.4.1. *Potential map*

The potential in the mth layer ($m = 1, 2, \ldots, n+1$) is obtained via (76) through discretization of (77),

$$\mathbf{\Phi}_m(\mathbf{r}) = \frac{b^2}{\sigma_1} \sum_{k=1}^{PN} i_k \mathcal{G}_m(\mathbf{r}, \mathbf{r}_k), \quad \mathbf{r}_k \in S_p, \tag{95}$$

where i_k is a component of the PN-dimensional vector $\boldsymbol{i}$ in (85) and $\mathcal{G}_m(\mathbf{r}, \mathbf{r}_k)$ is the corresponding mth layer Green's function. The current density $\mathbf{J}(\mathbf{r})$ and electric field $\mathbf{E}(\mathbf{r})$ are related via $\mathbf{J}(\mathbf{r}) = \sigma(z)\mathbf{E}(\mathbf{r})$ and $\mathbf{E}(\mathbf{r}) = -\nabla\Phi(\mathbf{r})$, respectively, and obtained through explicit (analytic) closed-form differentiation of $\Phi(\mathbf{r})$ and $\mathcal{G}(\mathbf{r}, \mathbf{r}_p)$ (i.e. term by term differentiation of the image series expansion). This is more accurate and stable than the numerical differentiation generally used in other solution schemes. A sketch of the computational algorithm is depicted in Fig. 8.

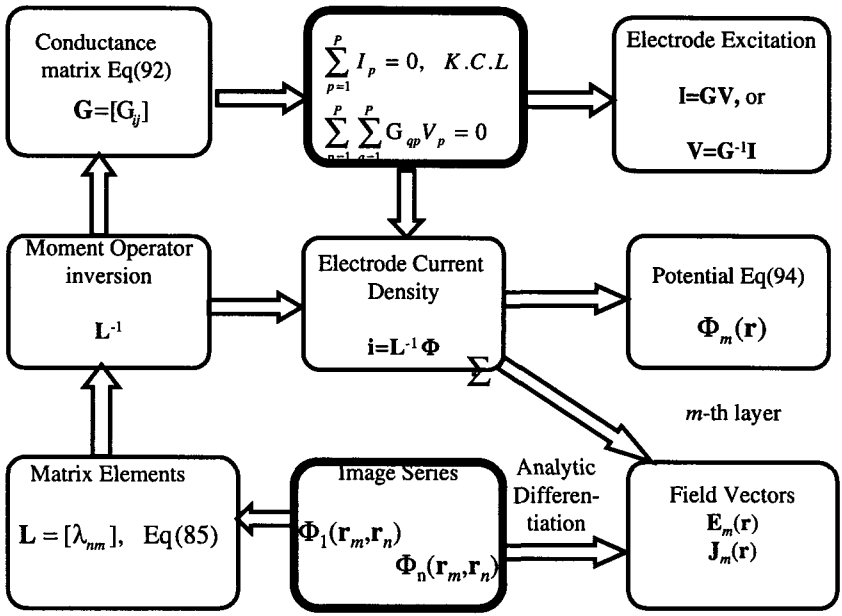


Fig. 8. Block scheme of the algorithm: (1) The kernel (Green's function) expanded in image series; (2) Moment matrix elements calculated through analytical integration of the image terms; (3) Moment matrix inversion; (4) Impedance matrix calculation; (5) Electrode current density distribution calculation; (6) Potential or field distribution at any point in any layer calculation. Note that we use image expansion on two different occasions: (a) To obtain the electrode current density distribution (using the image series expansion corresponding to the layer where the electrode array is placed), and (b) to obtain, after moment matrix inversion, the potential at any layer, using the image series expansion corresponding to that layer. Image series can be analytically differentiated to obtain the electric field and current density vectors. Since the impedance matrix depends only on the problem geometry, we need to perform matrix inversion only once, and then study the electrode array current–voltage relation for any given input voltage or current.

The potential distribution and vector plot of the $x-y$ components of the electric field, depicted in Fig. 9, is calculated for a four electrode array $P = 4$, $PN = 4 \times 400 = 1600$, $m = 1$ and electrode size $a = 0.04m$. The map illustrates efficiently the complete excitation status of the biological tissue at the electrode plane $z_p = z_1 = 0$, $p = 1, 2, 3, 4$. The electrode $x-y$ spacings and the electrode potentials $\mathbf{V}$ are specified in Fig. 9. The plot of the $x-y$ projection of the field vector $\mathbf{E}(\mathbf{r})$ is obtained through explicit (analytic) closed-form differentiation of $\Phi(\mathbf{r})$ and $\mathcal{G}_m(\mathbf{r}, \mathbf{r}_k)$ in (95), i.e. term by term differentiation of the image series expansion. The method is illustrated for the case of a two electrodes.

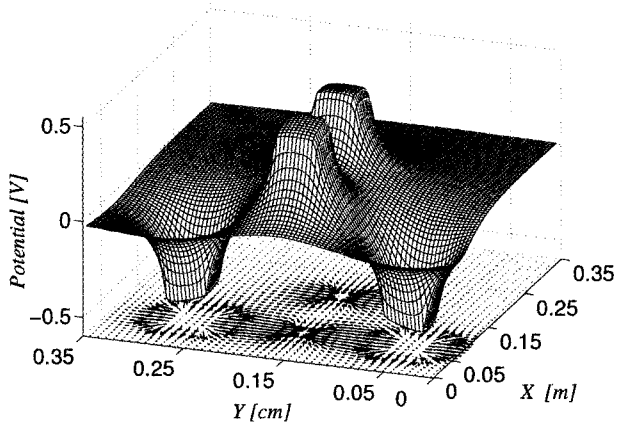


Fig. 9. Potential map and vector plot of the $x-y$ components of the electric field for a four-electrode array at $z = z_p = z_1 = 0$, $p = 1, 2, 3, 4$. The electrodes' size are $a_p = 0.04\,m$, their $x-y$ centers are $x = (x_1, x_2, x_3, x_4) = (0.1, 0.1, 0.1, 0.2)$ $[m]$, $y = (y_1, y_2, y_3, y_4) = (0.05, 0.15, 0.25, 0.15)[m]$, and their potentials are $\mathbf{V} = (-0.5, 0.5, -0.5, 0.5)[V]$, respectively.

3.4.2. *Two-electrode configuration*

We focus here on the dependence of electrode array and biological tissue interaction on the following three parameters: 1. electrode size, 2. electrode separation, 3. number of layers and their conductivities. The array configuration, therefore, is reduced to the simplest possible, i.e. a two-electrode system.

We focus herein on the evaluation of the conductance matrix elements G_{11} and G_{12} in (93) as well as the electrode input admittance G_{in}, given via,

$$G_{in} = \frac{I_2}{V_2 - V_1} = \frac{1}{2}(\mathrm{G}_{11} - \mathrm{G}_{12}), \tag{96}$$

for the symmetrical two-electrode problem ($V_1 = -V_2 \Leftrightarrow \mathrm{G}_{11} = \mathrm{G}_{22}$, Eq. (83)). The dependence of $\mathrm{G}_{11}/\mathrm{G}_{11\max}$, $\mathrm{G}_{12}/\mathrm{G}_{11\max}$ and $\mathrm{G}_{in}/\mathrm{G}_{11\max}$ on the electrodes' normalized center spacing $d/a \geq 1$, is given in Fig. 10. Note that for $d/a > 3$, there is practically no interaction between the electrodes, i.e. $\mathrm{G}_{12} = 0$, and G_{11} reaches the single electrode limit. The normalized conductivity dependence on either σ_2 or σ_3 is depicted in Figs. 11 and 12, respectively. While the conductivities strongly depend on the skin layer conductivity (σ_2), that is in contact with the electrode array, the somewhat more moderate dependence on the fat layer conductivity (σ_3) cannot be ignored.

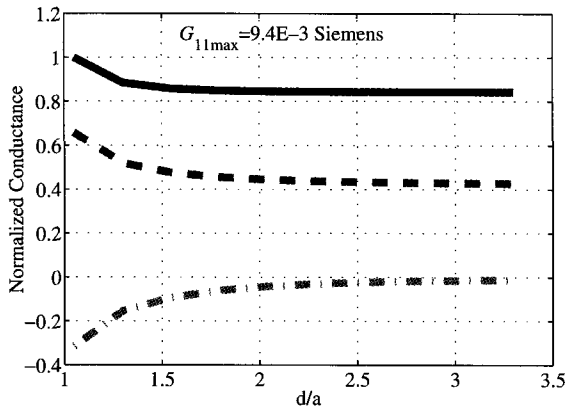


Fig. 10. Dependence of the normalized electrodes input conductance (G_{in} in Eq. (96), dashed line) and conductance matrix elements (G_{11} solid line and G_{12} dotted-dashed line, Eq. (93)) on the normalized distances between the electrode centers.

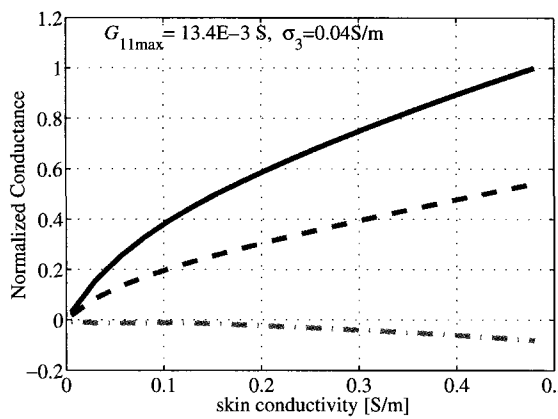


Fig. 11. Dependence of the normalized electrodes input conductance (dashed line, Eq. (96)) and conductance matrix elements G_{11} (solid line) and G_{12} (dotted-dashed line, Eq. (93)) on the second layer conductivity, σ_2, (skin).

4. Conclusion

A major outcome from this work was a novel image series expansion scheme for quasistatic Green's function in media with arbitrary number of layers. The expansions utilized a unique recursive representation for Green's function that is a generic characteristic of the stratification and were explicitly

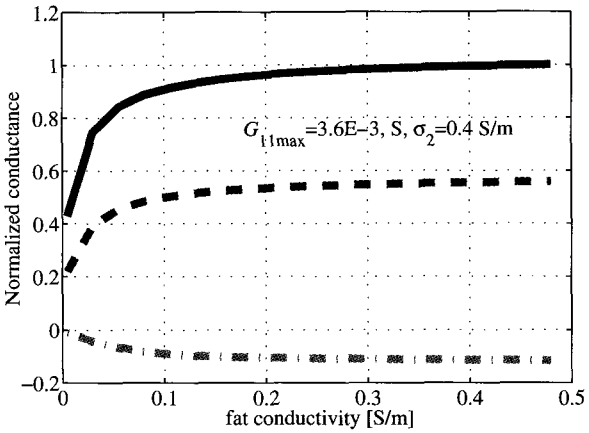


Fig. 12. Dependence of the normalized electrodes input conductance (dashed line, Eq. (96)) and conductance matrix elements (G_{11} solid line) and G_{11} dotted-dashed line, Eq. (93) on the third layer conductivity, σ_3 (fat).

constructed for multilayer media. Our recursive construction allowed us to prove analytically, to our knowledge for the first time, the convergence of n-layer image series under general conditions.

The numerical simulations demonstrated the importance of appropriate modeling of the tissue layers for FES application. The proposed hybrid model was shown capable to handle effectively layered medium problems with any number of layers. Thus a decision whether a particular layer should be included in the model could be accurately made.

The efficient handling of 3-D problems was achieved with the finite electrode arrays of arbitrary (generally non-planar) geometry due to the following: (a) The moment matrix elements were expressed explicitly via the analytical integration of image series terms combined with an asymptotic truncation error estimation; (b) The field was obtained through an analytic closed-form differentiation of the potential (i.e. term by term differentiation of the image series expansion); (c) Utilization of complex conductivity enabled generation of low-frequency field data for layered media rather than for the DC component only.

The inclusion of a collective image term, representing a closed form asymptotic expression of the series remainder integral, significantly accelerated the image series convergence and the overall algorithm speed. The numerical simulations signify the importance of the appropriate modeling of the tissue layers. Oversimplified models in FES problems, utilizing

a reduced number of layers, may result in inaccurate simulations, which greatly deviate from the real problem. Since our hybrid model can effectively handle layered medium problems with any number of layers, a decision whether a particular layer should be included in the model can be accurately made. The simulation results can be readily implemented for the classification, calibration, verification and interpretation of reported numerical and experimental data. The proposed computational procedure can thus be used as a simple tool for producing analytical data for testing numerical subroutines applied to simulate direct (FES) and inverse (bio electromagnetic imaging) problems in biomedical application.

Acknowledgments

This study was supported in part by the Segal and Isler foundations.

References

[1] M. Abramovitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover Publication Inc, New York, 1972).
[2] A. K. Abdelmageed and A. A. K. Mohsen, *Microwave Optical Tech. Letters* **29** (2001) 130.
[3] M. I. Aksun, *IEEE Trans. Microwave Theory Tech.* **44** (1996) 651.
[4] A. Badawi and A. Sebak, *J. Electromag. Waves Appl.* **14** (2000) 285.
[5] R. Bancroft, *IEEE Trans. Antenn. Propagat.* **45** (1997) 1704.
[6] V. G. Baranovskii, *Russian J. Nondestructive Testing* **29** (1993) 283.
[7] C. J. Bergeron *et al.*, *Geophysics* **66** (2001) 125.
[8] M. Born and E. Wolf, *Principles of Optics* (Pergamon Press, Oxford, 1998).
[9] L. M. Brekhovskikh, *Waves in Layered Media* (Academic Press, New York, 1980).
[10] B. M. Budak *et al.*, *A Collection of Problems on Mathematical Physics* (Pergamon Press, Oxford, 1964).
[11] E. Charbon *et al.*, *IEEE Trans. CAD Integrated Circuits Systems* **18** (1999) 172.
[12] P. M. Caruso *et al.*, *J. Biomech. Eng.* **104** (1982) 324.
[13] C. H. Chan and R. A. Kipp, *Int. J. Microwave Millimeter-Wave CAD Engin.* **7** (1997) 368.
[14] C. N. Chang and J. F. Cheng, *IEE Proc.-H Microwaves Antennas Propagat* **140** (1993) 79.
[15] J. Y. Chen *et al.*, *Electromagnetics* **20** (2000) 1.
[16] H. Y. Chen *et al.*, *Radio Science* **34** (1999) 1013.
[17] K-A. Cheng *et al.*, *IEEE Trans. Biomed. Eng.* **36** (1989) 918.
[18] W. C. Chew, *Waves and Fields in Inhomogeneous Media* (Van Nostrand Reinhold, New York, 1990).
[19] Y. L. Chow and W. C. Tang, *IEEE Trans. Microwave Theory Tech.* **49** (2001) 1483.

[20] Y. L. Chow *et al.*, *IEEE Trans. Power Delivery* **10** (1995) 707.
[21] Y. L. Chow *et al.*, *IEE Proc. Microwaves Antennas Propagat.* **145** (1998) 85.
[22] M. L. Chuang and F. L. Chu, *Int. J. Electronics* **84** (1998) 403.
[23] T. J. Cui *et al.*, *IEEE Trans. Geoscience Remote Sensing* **36** (1998) 526.
[24] F. A. Duck, *Physical Properties Tissue* (Academic Press, New York, 1990).
[25] H. B. Dwight, *Tables of Integrals and Other Mathematical Data* (Macmillan Co., New York, 1947).
[26] P. D. Einziger *et al.*, *IEEE Trans. Antennas Propagat* **50** (2002) 1813.
[27] P. D. Einziger and L. B. Felsen, *IEEE Trans. Antennas Propagat.* **31** (1983) 863.
[28] P. D. Einziger and L. B. Felsen, *IEEE Trans. Antennas Propagat.* **31** (1983) 870.
[29] T. F. Eibert and V. Hansen, *IEEE Trans. Microwave Theory Techniques* **45** (1997) 1105.
[30] L. B. Felsen and N. Marcuvitz, *Radiation and Scattering of Waves* (Englewood Cliffs, New Jersey, 1973).
[31] L. B. Felsen, ed. *Hybrid Formulation of Wave Propagation and Scattering* (Martinus Nijhoff Publishers, Dordrecht, 1984).
[32] K. R. Foster and H. P. Schwan, in C. Polk and E. Postow, *CRC Handbook of Biological Effects of Electromagnetic Field* (CRC Press, Boca Raton, FL, 1986).
[33] L. Fernandez Alvarez *et al.*, *Surface Science* **369** (1996) 367.
[34] Y. Giat *et al.*, *IEEE Trans. Biomed. Eng.* **40** (1993) 664.
[35] Y. Giat *et al.*, *ASME J. Biomechanical Engineering* **118** (1996) 357.
[36] W. M. Grill, *IEEE Trans. Biomed. Eng.* **46** (1999) 918.
[37] R. W. Hamming, *Numerical Methods for Scientists and Engineers*, 2nd edn. (McGraw-Hill, New York, 1973).
[38] R. F. Harrington, *Field Computation by Moment Methods* (Macmillan, New York, 1968).
[39] N. Hojjat *et al.*, *IEE Proceedings-Microwaves Antennas and Propagation* **145** (1998) 449.
[40] J. N. Hummel, *Trans. Am. Inst. Min. Met. Eng.* **97** (1932) 392 (first published in German in 1928).
[41] E. Isakov *et al.*, *J. Rehab. Research and Development* **23** (1986) 9.
[42] E. Isakov and J. Mizrahi, *Clinical Rehabilitation* **7** (1993) 39.
[43] J. D. Jackson, *Classical Electodynamics*, 3rd edn. (Wiley, New York, 1999).
[44] J. B. Keller, *Commun. Pure Appl. Math.* **6** (1953) 505.
[45] G. V. Keller, and F. C. Frischknecht, *Electrical Methods in Geophysical Prospecting* (Pergamon Press, Oxford, 1970).
[46] O. D. Kellogg, *Foundation of Potential Theory* (Dover Publication Inc., New York, 1953).
[47] N. D. Kinayman *et al.*, *IEEE Trans. Microwave Theory Tech.* **46** (1998) 430.
[48] R. A. Kipp and C. H. Chan, *IEEE Trans. Microwave Theory Tech.* **42** (1994) 860.

[49] P. J. Lagace *et al.*, *IEEE Trans. Power Delivery* **11** (1996) 1349.
[50] O. Levin *et al.*, *Hong Kong J. Physio.* **18** (2000) 3.
[51] O. Levin and J. Mizrahi, *IEEE Trans. Rehab. Eng.* **7** (1999) 301.
[52] O. Levin *et al.*, **10** (2000) 47.
[53] M. Levy *et al.*, *Magnetic Resonance Medicine* **29** (1993) 53.
[54] M. Levy *et al.*, *Journal of Biomedical Engineering* **12** (1990) 150.
[55] I. V. Lindell, *Methods for Electromagnetic Field Analysis* (Clarendon Press, Oxford, 1992).
[56] K. R. Li *et al.*, *IEEE Trans. Microwave Theory Techniques* **45** (1997) 2.
[57] I. V. Lindell *et al.*, *IEE Proc.-H Microwaves Antennas Propagation* **139** (1992) 186.
[58] F. J. Ling *et al.*, *IEEE Trans. Microwave Theory Tech.* **47** (1999) 1810.
[59] F. Ling, and J. M. Jin, *IEEE Microwave Guided Wave Letters* **10** (2000) 400.
[60] L. M. Livshitz *et al.*, *Ann. Biomed. Eng.* **28** (2000) 1218.
[61] L. M. Livshitz *et al.*, *ACES Journal, SI: Computational Bio-Electromagnetics* **16** (2001) 145.
[62] L. M. Livshitz *et al.*, *IEEE Trans. Neural System and Rehabilitation Eng.* **9** (2001) 355.
[63] S. K. Lukas *et al.*, *SIAM J. Appl. Math.* **57** (1997) 1615.
[64] J. C. Maxwell, *A Treatise on Electricity and Magnetism*, unabridged 3rd edn. (Dover, New York, 1954).
[65] K. A. Michalski and J. R. Mosig, *IEEE Trans. Antennas and Propagat.* **45** (1997) 508.
[66] J. Minzly *et al.*, *Med. Bio. Engin. Comp.* **31** (1993) 75.
[67] J. Minzly *et al.*, *J. Biomed. Eng.* **15** (1993) 333.
[68] J. Mizrahi, *J. Electro. Kinesiology* **7** (1997) 1.
[69] J. Mizrahi, *Critical Rev. Phys. Rehabilitation Medicine* **9** (1997) 93 .
[70] J. Mizrahi *et al.*, *Basic Appl. Myology* **4** (1994) 147.
[71] J. Mizrahi *et al.*, *J. Electro. Kinesiology* **7** (1997) 51.
[72] J. Mizrahi *et al.*, *IEEE Trans. Rehab. Eng.* **2** (1994) 57.
[73] J. Mizrahi *et al.*, *Artificial Organs* **21** (1997) 236.
[74] M. R. Neuman, in *Medical Instrumentation Application and Design*, ed. J. G. Webster (Wiley, New York, 1998).
[75] S. Niwas and M. Israil, *Geophysics* **51** (1986) 1594.
[76] F. Ollendorff, *Erdströme* (Birkhäuser Verlag, Basel und Stuttgart, 1969, first published in German, 1927).
[77] T. Oostendorp and A. van Oosterom,*IEEE Trans. Biomed. Eng.* **38** (1991) 409.
[78] A. van Oosterom and J. Strackee, *Med. Biol. Eng. Comput.* **21** (1983) 473.
[79] A. B. Oslon and I. N. Stankeeva, *Electric Technology in U.S.S.R.* **4** (1979) 68.
[80] C. Polk, Biological applications of large electric fields: Some history and fundamentals, *IEEE Trans. Plasma Science* **28** (2001) 61.
[81] F. Rattay, *IEEE Trans. Biomed. Eng.* **35** (1988) 199.
[82] I. Roman, *Geophysics* **24** (1959) 485.

[83] J. T. Rubinstein *et al.*, *IEEE Trans. Biomed. Eng.* **34** (1987) 864.
[84] K. L. Rodenhiser and A. Spelman, *IEEE Trans. Biomed. Eng.* **42** (1995) 337.
[85] A. Sommerfeld, *Partial Differential Equations in Physics* (Academic Press, New York, 1949).
[86] S. Stefanesco and C. Schlumberger, *J. Phys. Radium* **1** (1930) 130 (in French).
[87] D. F. Stegeman *et al.*, *Biol. Cybernet* **33** (1979) 97.
[88] J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill, New York, 1941).
[89] E. D. Sunde, *Earth Conduction Effects in Transmission Systems* (Van Nostrand, Toronto, 1948).
[90] W. Thomson (Lord Kelvin), *Proc. R. Soc. Lonson Ser. A*, **7** (1855) 382.
[91] S. Tanaka *et al.*, *JSME Int. J. Series C-Dynamics Control Robotics Design Manufacturing* **39** (1996) 195.
[92] A. N. Tikhonov and A. A. Samarskii, *Equations of Mathematical Physics* (Pergamon Press, Oxford, 1963).
[93] J.-Z. Tsai *et al.*, *IEEE Trans. Biomed. Eng.* **47** (2000) 41.
[94] A. Torabian and Y. L. Chow, *IEEE Trans. Microwave Theory Tech.* **47** (1999) 1777.
[95] M. J. Tsai *et al.*, *IEEE Trans. Microwave Theory Tech.* **45** (1997) 330.
[96] L. Tsang *et al.*, *Microwave Optical Technology Letters* **24** (2000) 247.
[97] J. R. Wait, *Electromagnetic Waves in Stratified Media* (Pergamon Press, Oxford, 1970).
[98] J. R. Wait, *Geo-Electromagnetism* (Academic Press, New York, 1982).
[99] J. D. Wiley and J. G. Webster, *IEEE Trans. Biomed. Eng.* **29** (1982) 381.
[100] J. D. Wiley and J. G. Webster, *IEEE Trans. Biomed. Eng.* **29** (1982) 385.
[101] P. Yla-Oijala *et al.*, *J Electromagnet Wave* **15** (2001) 913.
[102] T. J. Yu and W. Cai, *Radio Sci.* **36** (2001) 559.
[103] B. X. Zhang *et al.*, *JOSA* **100** (1996) 3527.
[104] L. P. Zhang *et al.*, *Comm. Numer. Methods Engin.* **15** (1999) 835.
[105] J. Zhao *et al.*, *Proc. Design Automation Conference* 224, June, San Francisco (1998), pp. 15–19.

This page is intentionally left blank

CHAPTER 13

DYNAMICS OF HUMANOID ROBOTS: GEOMETRICAL AND TOPOLOGICAL DUALITY

VLADIMIR G. IVANCEVIC
Defence Science & Technology Organization
Adelaide, Australia
vladimir.Ivancevic@dsto.defence.gov.au

Humanoid robots are human-like, anthropomorphic mechanisms with biodynamics that resembles human musculo-skeletal dynamics. The present chapter enlightens the underlying unique global mathematical structure beneath the general humanoid dynamics (HD, for short). It presents a parallel development of Hamiltonian and Lagrangian formulations of HD, proves both differential-geometrical and algebraic-topological dualities between these two formulations, and finally establishes a unique functorial relation between HD-geometry and HD-topology.

1. Introduction

Highly complex, many-degree-of-freedom dynamics of humanoid robots resembles human motion dynamics (see [10] for technical details on biomechanically-realistic HD). Since the early papers of Vukobratovic [25–30], the vast body of research has been done in relation to kinematics, dynamics and control of anthropomorphic robots [1, 6, 8, 9, 12, 18, 22–24]. Some of the biped models had the ability of passive dynamic walking [15] and others had powered walking ability [16]. The previous decade was dominated by various solutions to the kinematic problems of redundancy and singularities [31, 21]. The last decade of the twentieth century has been characterized mostly by extensive use of intelligent, adaptive, neuro-fuzzy-genetic control of HD [2, 5, 7, 17, 19, 20].

The present chapter uncovers the underlying unique global geometrico-topological structure beneath the HD. It presents a parallel development of Hamiltonian and Lagrangian formulations of dissipative, muscle-driven HD (see [10]), proves both differential-geometrical and algebraic-topological dualities between these two formulations, and finally establishes a *unique*

functorial relation between HD-geometry and HD-topology (see [13] for the modern unifying mathematical language of categories, functors and natural equivalences).

The finite-dimensional *configuration manifold* Q^N of HD is constructed using direct products of constrained rotational Lie groups. Lagrangian formulation of HD is performed on the *tangent bundle* TQ^N, while Hamiltonian formulation is performed on the *cotangent bundle* T^*Q^N. Both *Riemannian* and *symplectic* geometry are used for these formulations. The geometrical duality (see [11, 3]) of Lie groups and algebras between these two HD-formulations is proved as an existence of natural equivalence between Lie and canonical functors. The topological duality (see [4]) between these two HD-formulations is proved as an existence of natural equivalence between Lagrangian and Hamiltonian functors in both *homology* and *cohomology* categories.

2. Topological Preliminaries

In topology of finite-dimensional smooth (i.e. C^{p+1} with $p \geq 0$) manifolds, a fundamental notion is the duality between p-chains C and p-forms (i.e., p-cochains) ω on the smooth manifold M, or domains of integration and integrands — as an integral on M represents a bilinear functional $\int_C \omega \equiv \langle C, \omega \rangle$ (see [3] and [4]). The duality is based on the classical Stokes formula

$$\int_C d\omega = \int_{\partial C} \omega.$$

This is written in terms of scalar products on M as $\langle C, d\omega \rangle = \langle \partial C, \omega \rangle$, where ∂C is the boundary of the p-chain C oriented coherently with C. While the boundary operator ∂ is a global operator, the coboundary operator, that is, the exterior derivative d, is local, and thus more suitable for applications. The main property of the exterior differential,

$$d^2 = 0 \quad \text{implies } \partial^2 = 0,$$

can be easily proved by the use of Stokes' formula

$$\langle \partial^2 C, \omega \rangle = \langle \partial C, d\omega \rangle = \langle C, d^2\omega \rangle = 0.$$

The analysis of p-chains and p-forms on the finite-dimensional smooth manifold M is usually performed in (co)homology categories (see [4]) related to M.

Let $\mathcal{M}^\bullet$ denote the category of cochains, (i.e., p-forms) on the smooth manifold M. When $\mathcal{C} = \mathcal{M}^\bullet$, we have the category $\mathcal{S}^\bullet(\mathcal{M}^\bullet)$ of generalized

cochain complexes $A^\bullet$ in $\mathcal{M}^\bullet$, and if $A'_n = 0$ for $n < 0$ we have a subcategory $\mathcal{S}^\bullet_{\mathcal{DR}}(\mathcal{M}^\bullet)$ of De Rham differential complexes in $\mathcal{M}^\bullet$

$$A^\bullet_{DR} : 0 \to \Omega^0(M) \xrightarrow{d} \Omega^1(M) \xrightarrow{d} \Omega^2(M)$$
$$\cdots \xrightarrow{d} \Omega^n(M) \xrightarrow{d} \cdots. \tag{1}$$

Here $A'_n = \Omega^n(M)$ is the vector space over $\mathbb{R}$ of all p-forms ω on M (for $p = 0$ the smooth functions on M) and $d_n = d : \Omega^{n-1}(M) \to \Omega^n(M)$ is the exterior differential. A form $\omega \in \Omega^n(M)$ such that $d\omega = 0$ is a closed form or n-cocycle. A form $\omega \in \Omega^n(M)$ such that $\omega = d\theta$, where $\theta \in \Omega^{n-1}(M)$, is an exact form or n-coboundary. Let $Z^n(M) = Ker(d)$ (resp. $B^n(M) = Im(d)$) denote a real vector space of cocycles (resp. coboundaries) of degree n. Since $d_{n+1}d_n = d^2 = 0$, we have $B^n(M) \subset Z^n(M)$. The quotient vector space

$$H^n_{DR}(M) = Ker(d)/Im(d) = Z^n(M)/B^n(M)$$

is the de Rham cohomology group. The elements of $H^n_{DR}(M)$ represent equivalence sets of cocycles. Two cocycles ω_1, ω_2 belong to the same equivalence set, or are cohomologous (written $\omega_1 \sim \omega_2$) if and only if they differ by a coboundary $\omega_1 - \omega_2 = d\theta$. The De Rham cohomology class of any form $\omega \in \Omega^n(M)$ is $[\omega] \in H^n_{DR}(M)$. The De Rham differential complex (1) can be considered as a system of second-order differential equations $d^2\theta = 0$, $\theta \in \Omega^{n-1}(M)$ having a solution represented by $Z^n(M) = Ker(d)$.

Analogously let $\mathcal{M}_\bullet$ denote the category of chains on the smooth manifold M. When $\mathcal{C} = \mathcal{M}_\bullet$, we have the category $\mathcal{S}_\bullet(\mathcal{M}_\bullet)$ of generalized chain complexes $A_\bullet$ in $\mathcal{M}_\bullet$, and if $A_n = 0$ for $n < 0$ we have a subcategory $\mathcal{S}^{\mathcal{C}}_\bullet(\mathcal{M}_\bullet)$ of chain complexes in $\mathcal{M}_\bullet$

$$A_\bullet : 0 \leftarrow C^0(M) \xleftarrow{\partial} C^1(M) \xleftarrow{\partial} C^2(M)$$
$$\cdots \xleftarrow{\partial} C^n(M) \xleftarrow{\partial} \cdots.$$

Here $A_n = C^n(M)$ is the vector space over $\mathbb{R}$ of all finite chains C on the manifold M and $\partial_n = \partial : C^{n+1}(M) \to C^n(M)$. A finite chain C such that $\partial C = 0$ is an n-cycle. A finite chain C such that $C = \partial B$ is an n-boundary. Let $Z_n(M) = Ker(\partial)$ (resp. $B_n(M) = Im(\partial)$) denote a real vector space of cycles (resp. boundaries) of degree n. Since $\partial_{n+1}\partial_n = \partial^2 = 0$, we have $B_n(M) \subset Z_n(M)$. The quotient vector space

$$H^C_n(M) = Ker(\partial)/Im(\partial) = Z_n(M)/B_n(M)$$

is the n-homology group. The elements of $H^C_n(M)$ are equivalence sets of cycles. Two cycles C_1, C_2 belong to the same equivalence set, or are homologous (written $C_1 \sim C_2$), if and only if they differ by a boundary

$C_1 - C_2 = \partial B$). The homology class of a finite chain $C \in C^n(M)$ is $[C] \in H_n^C(M)$.

The dimension of the n-cohomology (resp. n-homology) group equals the nth Betti number b^n (resp. b_n) of the manifold M. *Poincaré lemma* says that on an open set $U \in M$ diffeomorphic to $\mathbb{R}^N$, all closed forms (cycles) of degree $p \geq 1$ are exact (boundaries). That is, the Betti numbers satisfy $b^p = 0$ (resp. $b_p = 0$) for $p = 1, \ldots, n$.

The *De Rham theorem* states the following. The map $\Phi: H_n \times H^n \to \mathbb{R}$ given by $([C], [\omega]) \to \langle C, \omega \rangle$ for $C \in Z_n$, $\omega \in Z^n$ is a bilinear nondegenerate map which establishes the duality of the groups (vector spaces) H_n and H^n and the equality $b_n = b^n$.

3. HD-Configuration Manifold and Its Reduction

3.1. *Configuration manifold*

Kinematics of an n-segment humanoid chain is usually defined as a map between *external* (usually, end-effector) coordinates $x^r (r = 1, \ldots, n)$ and *internal* (joint) coordinates $q^i (i = 1, \ldots, N)$ (see [10]). The *forward kinematics* are defined as a nonlinear map $x^r = x^r(q^i)$ with a corresponding linear vector functions $dx^r = \partial x^r / \partial q^i \, dq^i$ of differentials: and $\dot{x}^r = \partial x^r / \partial q^i \, \dot{q}^i$ of velocities. (Here and subsequently the summation convention over repeated indices is understood.) When the rank of the configuration-dependent Jacobian matrix $J \equiv \partial x^r / \partial q^i$ is less than n the *kinematic singularities* occur; the onset of this condition could be detected by the *manipulability measure*. *Inverse kinematics* are defined conversely by a nonlinear map $q^i = q^i(x^r)$ with a corresponding linear vector functions $dq^i = \partial q^i / \partial x^r \, dx^r$ of differentials and $\dot{q}^i = \partial q^i / \partial x^r \, \dot{x}^r$ of velocities. Again, in the case of *redundancy* ($n < N$), the inverse kinematic problem admits infinite solutions; often the *pseudo-inverse* configuration-control is used instead: $\dot{q}^i = J^* \, \dot{x}^r$, where $J^* = J^T (J J^T)^{-1}$ denotes the Moore–Penrose pseudo-inverse of the Jacobian matrix J.

Humanoid joints, that is, internal coordinates q^i $(i = 1, \ldots, N)$, constitute a smooth configuration manifold Q^N, described as follows. Uniaxial, "hinge" joints represent constrained, rotational Lie groups $SO(2)^i_{cnstr}$, parameterized by constrained angles $q^i_{cnstr} \equiv q^i \in [q^i_{\min}, q^i_{\max}]$. Three-axial, "ball-and-socket" joints represent constrained rotational Lie groups $SO(3)^i_{cnstr}$, parameterized by constrained Euler angles $q^i = q^{\phi_i}_{cnstr}$ (in the following text, the subscript "cnstr" will be omitted, for the sake of simplicity, and always assumed in relation to internal coordinates q^i).

All $SO(n)$-joints are Hausdorff C^∞-manifolds with atlases (U_α, u_α); in other words, they are paracompact and metrizable smooth manifolds, admitting Riemannian metric.

Let A and B be two smooth manifolds described by smooth atlases (U_α, u_α) and (V_β, v_β), respectively. Then the family $(U_\alpha \times V_\beta, u_\alpha \times v_\beta : U_\alpha \times V_\beta \to \mathbb{R}^m \times \mathbb{R}^n)(\alpha, \beta) \in A \times B$ is a smooth atlas for the direct product $A \times B$. Now, if A and B are two Lie groups (say, $SO(n)$), then their *direct product* $G = A \times B$ is at the same time their direct product as smooth manifolds and their direct product as algebraic groups, with the product law

$$(a_1, b_1)(a_2, b_2) = (a_1 a_2, b_1 b_2), \quad a_{1,2} \in A, \quad b_{1,2} \in B.$$

Generalizing the direct product to N rotational joint groups, we can draw an *anthropomorphic product-tree* (see Fig. 1) using a line segment "–" to represent direct products of humanoid's $SO(n)$-joints. This is our basic model of the humanoid configuration manifold Q^N.

Let $T_q Q^N$ be a tangent space to Q^N at the point q. The *tangent bundle* TQ^N represents a union $\bigcup_{q \in Q^N} T_q Q^N$, together with the standard topology on TQ^N and a natural smooth manifold structure, the dimension of which is twice the dimension of Q^N. A vector field X on Q^N represents a section $X : Q^N \to TQ^N$ of the tangent bundle TQ^N.

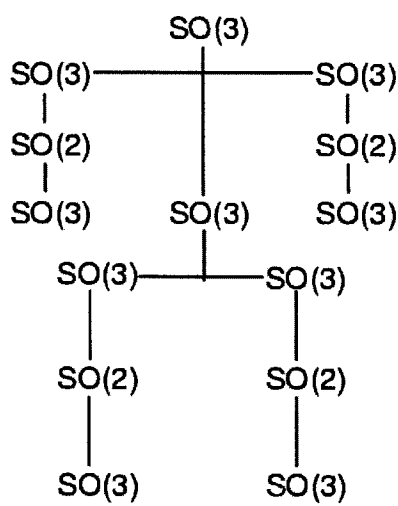


Fig. 1. Configuration HD-manifold Q^N modeled as anthropomorphic product-tree of constrained $SO(n)$ groups.

Analogously let $T_q^*Q^N$ be a cotangent space to Q^N at q, the dual to its tangent space T_qQ^N. The *cotangent bundle* T^*Q^N represents a union $\bigcup_{q\in Q^N} T_q^*Q^N$, together with the standard topology on T^*Q^N and a natural smooth manifold structure, the dimension of which is twice the dimension of Q^N. A one-form θ on Q^N represents a section $\theta : Q^N \to T^*Q^N$ of the cotangent bundle T^*Q^N.

We refer to the tangent bundle TQ^N of HD configuration manifold Q^N as the *velocity phase-space* manifold, and to its cotangent bundle T^*Q^N as the *momentum phase-space* manifold.

3.2. *Reduction of the configuration manifold*

The HD-configuration manifold Q^N (Fig. 1) can be (for the sake of the brain-like motor control [19, 20]) reduced to N-torus T^N, in three steps, as follows.

First, a single three-axial $SO(3)$-joint can be reduced to the direct product of three uniaxial $SO(2)$-joints, in the sense that three hinge joints can produce any orientation in space, just as a ball-joint can. Algebraically, this means reduction (using symbol "$\gtrsim$") of each of the three $SO(3)$ rotation matrices to the corresponding $SO(2)$ rotation matrices

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{pmatrix} \gtrsim \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix},$$

$$\begin{pmatrix} \cos\psi & 0 & \sin\psi \\ 0 & 1 & 0 \\ -\sin\psi & 0 & \cos\psi \end{pmatrix} \gtrsim \begin{pmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix},$$

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \gtrsim \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

In this way we can set the *reduction equivalence relation* $SO(3) \gtrsim SO(2)_\phi \ltimes SO(2)_\psi \ltimes SO(2)_\theta$, where $\ltimes$ denotes the noncommutative semidirect product.

Second, we have a homeomorphism: $SO(2) \sim S^1$, where S^1 denotes the constrained unit circle in the complex plane, which is an Abelian Lie group.

Third, let I^N be the unit cube $[0,1]^N$ in $\mathbb{R}^N$ and "$\sim$" an equivalence relation on $\mathbb{R}^N$ obtained by "gluing" together the opposite sides of I^N, preserving their orientation. The manifold of humanoid configurations depicted

in Fig. 1 can be represented as the quotient space of $\mathbb{R}^N$ by the space of the integral lattice points in $\mathbb{R}^N$, that is a constrained N-dimensional torus T^N:

$$\mathbb{R}^N/Z^N = I^N/\sim \cong \prod_{i=1}^{N} S_i^1$$
$$\equiv \{(q^i, i = 1, \ldots, N) : \text{mod } 2\pi\} = T^N.$$

Since S^1 is an Abelian Lie group, its N-fold tensor product T^N is also an Abelian Lie group, the toral group, of all nondegenerate diagonal $N \times N$ matrices. As a Lie group, the HD-configuration space $Q^N \equiv T^N$ has a natural Banach manifold structure with local internal coordinates $q^i \in U$, U being an open set (chart) in T^N.

Conversely by "ungluing" the configuration space we obtain the primary unit cube. Let "$\sim^*$" denote an equivalent decomposition or "ungluing" relation. By the Tychonoff product-topology theorem, for every such quotient space there exists a "selector" such that their quotient models are homeomorphic, that is, $T^N/\sim^* \approx A^N/\sim^*$. Therefore I^N represents a "selector" for the configuration torus T^N and can be used as an N-directional "command-space" for the topological control of humanoid motion. Any subset of degrees of freedom on the configuration torus T^N representing the joints included in humanoid motion has its simple, rectangular image in the command space — selector I^N. Operationally, this resembles what the brain-motor-controller, the cerebellum, actually performs on the highest level of human motor control (see [20]).

4. Geometrical Duality in Humanoid Dynamics

Theorem 1. *There is a geometrical duality between Lagrangian and Hamiltonian HD-formulations on Q^N. In categorical terms, there is a unique natural geometrical equivalence*

$$\mathbf{Dual}_G : \mathbf{Lie} \cong \mathbf{Can}$$

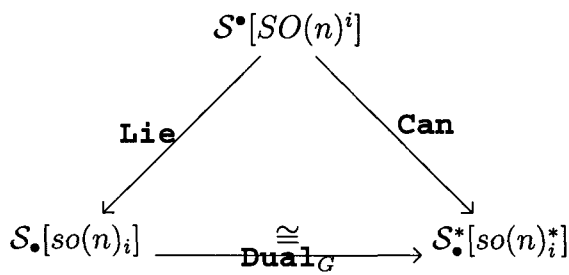
in HD (symbols are described in the next subsection).

Proof. The proof has two parts: Lie-functorial and geometrical. □

4.1. *Lie-functorial proof*

If we apply the functor **Lie** on the category $\mathcal{S}^\bullet[SO(n)^i]$ (for $n = 2, 3$ and $i = 1, \ldots, N$) of rotational Lie groups $SO(n)^i$ (and their homomorphisms) we

obtain the category $\mathcal{S}_\bullet[so(n)_i]$ of corresponding *tangent* Lie algebras $so(n)_i$ (and their homomorphisms). If we further apply the isomorphic functor **Dual** to the category $\mathcal{S}_\bullet[so(n)_i]$ we obtain the dual category $\mathcal{S}^*_\bullet[so(n)^*_i]$ of *cotangent*, or, *canonical* Lie algebras $so(n)^*_i$ (and their homomorphisms). To go directly from $\mathcal{S}^\bullet[SO(n)^i]$ to $\mathcal{S}^*_\bullet[so(n)^*_i]$ we use the canonical functor **Can**. Therefore, we have a commutative triangle:



Applying the functor **Lie** on HD-configuration manifold Q^N (Fig. 1), we get the product-tree of the same anthropomorphic structure, but having tangent Lie algebras $so(n)_i$ as vertices, instead of the groups $SO(n)^i$. Again, applying the functor **Can** on Q^N, we get the product-tree of the same anthropomorphic structure, but this time having cotangent Lie algebras $so(n)^*_i$ as vertices. Both the tangent algebras $so(n)_i$ and the cotangent algebras $so(n)^*_i$ contain infinitesimal group generators: angular velocities $\dot{q}^i = \dot{q}^{\phi_i}$ — in the first case, and canonical angular momenta $p_i = p_{\phi_i}$ — in the second case [10]. As Lie group generators, both the angular velocities and the angular momenta satisfy the commutation relations: $[\dot{q}^{\phi_i}, \dot{q}^{\psi_i}] = \epsilon^{\phi\psi}_\theta \dot{q}^{\theta_i}$ and $[p_{\phi_i}, p_{\psi_i}] = \epsilon^\theta_{\phi\psi} p_{\theta_i}$, respectively, where the structure constants $\epsilon^{\phi\psi}_\theta$ and $\epsilon^\theta_{\phi\psi}$ constitute the totally antisymmetric third-order tensors.

In this way, the functor $\mathbf{Dual}_G : \mathbf{Lie} \cong \mathbf{Can}$ establishes the unique geometrical duality between kinematics of angular velocities $\dot{q}^i$ (involved in *Lagrangian* formalism on the tangent bundle of Q^N) and kinematics of angular momenta p_i (involved in *Hamiltonian* formalism on the cotangent bundle of Q^N), which is analyzed below. In other words, we have two functors, **Lie** and **Can**, from the *category of Lie groups* (of which $\mathcal{S}^\bullet[SO(n)^i]$ is a subcategory) into the *category of* (their) *Lie algebras* (of which $\mathcal{S}_\bullet[so(n)_i]$ and $\mathcal{S}^*_\bullet[so(n)^*_i]$ are subcategories), and a unique natural equivalence between them defined by the functor $\mathbf{Dual}_G$. (As angular momenta p_i are in a bijective correspondence with angular velocities $\dot{q}^i$, every component of the functor $\mathbf{Dual}_G$ is invertible.) □

4.2. *Geometrical proof*

Geometrical proof is given along the lines of Riemannian and symplectic geometry of mechanical systems (see [10] and [14]), as follows. The Riemannian metric $g = \langle , \rangle$ on the configuration manifold Q^N is a positive-definite quadratic form $g : TQ^N \to \mathbb{R}$, given in local coordinates $q^i \in U$ (U open in Q^N) as

$$g_{ij} \mapsto g_{ij}(q, m) dq^i dq^j .$$

Here

$$g_{ij}(q, m) = \sum_{\mu=1}^{n} m_\mu \delta_{rs} \frac{\partial x^r}{\partial q^i} \frac{\partial x^s}{\partial q^j}$$

is the covariant material metric tensor defining a relation between internal and external coordinates and including n segmental masses m_μ. The quantities x^r are external coordinates ($r, s = 1, \ldots, 6n$) and $i, j = 1, \ldots, N \equiv 6n - h$, where h denotes the number of holonomic constraints.

The *Lagrangian* of the system is a quadratic form $L : TQ^N \to \mathbb{R}$ dependent on velocity v and such that $L(v) = \frac{1}{2}\langle v, v \rangle$. It is given by

$$L(v) = \frac{1}{2} g_{ij}(q, m) v^i v^j$$

in local coordinates $q^i, v^i = \dot{q}^i \in U_v$ (U_v open in TQ^N). The *Hamiltonian* of the system is a quadratic form $H : T^*Q^N \to \mathbb{R}$ dependent on momentum p and such that $H(p) = \frac{1}{2}\langle p, p \rangle$. It is given by

$$H(p) = \frac{1}{2} g^{ij}(q, m) p_i p_j$$

in local canonical coordinates $q^i, p_i \in U_p$ (U_p open in T^*Q^N). The inverse (contravariant) metric tensor is defined as

$$g^{ij}(q, m) = \sum_{\mu=1}^{n} m_\mu \delta_{rs} \frac{\partial q^i}{\partial x^r} \frac{\partial q^j}{\partial x^s} .$$

For any smooth function L on TQ^N, the *fiber derivative* or *Legendre transformation* is a diffeomorphism $FL : TQ^N \to T^*Q^N$, $F(w) \cdot v = \langle w, v \rangle$, from the momentum phase-space manifold to the velocity phase-space manifold associated with the metric $g = \langle , \rangle$. In local coordinates $q^i, v^i = \dot{q}^i \in U_v$ (U_v open in TQ^N), FL is given by $(q^i, v^i) \mapsto (q^i, p_i)$.

On the momentum phase-space manifold T^*Q^N exists:

(i) A unique canonical one-form θ_H with the property that, for any one-form β on the configuration manifold Q^N, we have $\beta^*\theta_H = \beta$. In local canonical coordinates $q^i, p_i \in U_p$ (U_p open in T^*Q^N) it is given by $\theta_H = p_i dq^i$.

(ii) A unique nondegenerate Hamiltonian symplectic two-form ω_H, which is closed ($d\omega_H = 0$) and exact ($\omega_H = d\theta_H = dp_i \wedge dq^i$). Each body segment has, in the general $SO(3)$ case, a sub-phase-space manifold $T^*SO(3)$ with

$$\omega_H^{(sub)} = dp_\phi \wedge d\phi + dp_\psi \wedge d\psi + dp_\theta \wedge d\theta.$$

Analogously, on the velocity phase-space manifold TQ^N exists:

(i) A unique one-form θ_L, defined by the *pull-back* $\theta_L = (FL) * \theta_H$ of θ_H by FL. In local coordinates $q^i, v^i = \dot{q}^i \in U_v$ (U_v open in TQ^N) it is given by $\theta_L = L_{v^i} dq^i$, where $L_{v^i} \equiv \partial L / \partial v^i$.

(ii) A unique nondegenerate Lagrangian symplectic two-form ω_L, defined by the pull-back $\omega_L = (FL) * \omega_H$ of ω_H by FL, which is closed ($d\omega_L = 0$) and exact ($\omega_L = d\theta_L = dL_{v^i} \wedge dq^i$).

Both T^*Q^N and TQ^N are orientable manifolds, admitting the standard volumes given respectively by

$$\Omega_{\omega_H} = \frac{(-1)^{\frac{N(N+1)}{2}}}{N!} \omega_H^N, \quad \text{and}$$
$$\Omega_{\omega_L} = \frac{(-1)^{\frac{N(N+1)}{2}}}{N!} \omega_L^N$$

in local coordinates $q^i, p_i \in U_p$ (U_p open in T^*Q^N), resp. $q^i, v^i = \dot{q}^i \in U_v$ (U_v open in TQ^N). They are given by

$$\Omega_H = dq^1 \wedge \cdots \wedge dq^N \wedge dp_1 \wedge \cdots \wedge dp_N, \quad \text{and}$$
$$\Omega_L = dq^1 \wedge \cdots \wedge dq^N \wedge dv^1 \wedge \cdots \wedge dv^N.$$

On the velocity phase-space manifold TQ^N we can also define the *action* $A : TQ^N \to \mathbb{R}$ by $A(v) = FL(v) \cdot v$ and the energy $E = A - L$. In local coordinates $q^i, v^i = \dot{q}^i \in U_v$ (U_v open in TQ^N) we have $A = v^i L_{v^i}$, so $E = v^i L_{v^i} - L$. The Lagrangian vector field X_L on TQ^N is determined

by the condition $i_{X_L}\omega_L = dE$. Classically, it is given by the second-order Lagrange equations

$$\frac{d}{dt}\frac{\partial L}{\partial v^i} = \frac{\partial L}{\partial q^i}. \tag{2}$$

The Hamiltonian vector field X_H is defined on the momentum phase-space manifold T^*Q^N by the condition $i_{X_H}\omega = dH$. The condition may be expressed equivalently as $X_H = J\nabla H$, where

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}.$$

In local canonical coordinates $q^i, p_i \in U_p$ (U_p open in T^*Q^N) the vector field X_H is classically given by the first-order Hamilton's canonical equations

$$\dot{q}^i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q^i}. \tag{3}$$

As a Lie group, the configuration manifold Q^N is Hausdorff. Therefore for $x = (q^i, p_i) \in U_p$ (U_p open in T^*Q^N) there exists a unique one-parameter group of diffeomorphisms $\phi_t : T^*Q^N \to T^*Q^N$ such that $\frac{d}{dt}|_{t=0}\,\phi_t x = J\nabla H(x)$. This is termed *Hamiltonian phase flow* and represents the maximal integral curve $t \mapsto (q^i(t), p_i(t))$ of the Hamiltonian vector field X_H passing through the point x for $t = 0$.

The flow ϕ_t is *symplectic* if ω_H is constant along it (that is, $\phi_t^*\omega_H = \omega_H$) if and only if its Lie derivative vanishes (that is, $L_{X_H}\omega_H = 0$). A symplectic flow consists of canonical transformations on T^*Q^N, that is, local diffeomorphisms that leave ω_H invariant. By Liouville's theorem, a symplectic flow ϕ_t preserves the phase volume on T^*Q^N. Also, the total energy $H = E$ of the system is conserved along ϕ_t, that is, $H \circ \phi_t = \phi_t$.

Lagrangian flow can be defined analogously (see [14]).

For a Lagrangian (resp. a Hamiltonian) vector field X_L (resp. X_H) on Q^N, there is a base integral curve $c_0(t) = (q^i(t), v^i(t))$ (resp. $c_0(t) = (q^i(t), p_i(t))$) if and only if $c_0(t)$ is a geodesic. This is given by the contravariant velocity equation

$$\dot{q}^i = v^i, \quad \dot{v}^i + \Gamma^i_{jk} v^j v^k = 0 \tag{4}$$

in the former case and by the covariant momentum equation

$$\begin{aligned} \dot{q}^k &= g^{ki} p_i, \\ \dot{p}_i + \Gamma^i_{jk} g^{jl} g^{km} p_l p_m &= 0 \end{aligned} \tag{5}$$

in the latter. Here Γ^i_{jk} denote the Christoffel symbols of an affine connection in an open chart U on Q^N, defined on the Riemannian metric $g = \langle , \rangle$ by

$$\Gamma^i_{jk} = g^{il}\Gamma_{jkl},$$
$$\Gamma_{jkl} = \frac{1}{2}\left(\frac{\partial g_{kl}}{\partial q^j} + \frac{\partial g_{jl}}{\partial q^k} - \frac{\partial g_{jk}}{\partial q^l}\right).$$

The left-hand sides $\dot{\dot{v}}^i = \dot{v}^i + \Gamma^i_{jk}v^jv^k$ (resp. $\dot{\dot{p}}_i = \dot{p}_i + \Gamma^i_{jk}g^{jl}g^{km}p_lp_m$) in the second parts of (4) and (5) represent the *Bianchi covariant derivative* of the velocity (resp. momentum) with respect to t. *Parallel transport* on Q^N is defined by $\dot{\dot{v}}^i = 0$, (resp. $\dot{\dot{p}}_i = 0$). When this applies, X_L (resp. X_H) is called the *geodesic spray* and its flow the *geodesic flow.*

For the dynamics in the gravitational potential field $V : Q^N \to \mathbb{R}$, the Lagrangian $L : TQ^N \to \mathbb{R}$ (resp. the Hamiltonian $H : T^*Q^N \to \mathbb{R}$) has an extended form

$$L(v,q) = \frac{1}{2}g_{ij}v^iv^j - V(q),$$
$$(\text{resp. } H(p,q) = \frac{1}{2}g^{ij}p_ip_j + V(q)).$$

A Lagrangian vector field X_L (resp. Hamiltonian vector field X_H) is still defined by the second-order Lagrangian equations (2) and (4) (resp. first-order Hamiltonian equations (3) and (5)).

The fiber derivative $FL : TQ^N \to T^*Q^N$ thus maps Lagrange's equations (2) and (4) into Hamilton's equations (3) and (5). Clearly there exists a diffeomorphism $FH : T^*Q^N \to TQ^N$, such that $FL = (FH)^{-1}$. In local canonical coordinates $q^i, p_i \in U_p$ (U_p, open in T^*Q^N) this is given by $(q^i, p_i) \mapsto (q^i, v^i)$ and thus maps Hamilton's equations (3) and (5) into Lagrange's equations (2) and (4).

A general form of the forced, non-conservative Hamilton's equations (resp. Lagrange's equations) is given as

$$\dot{q}^i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q^i} + F_i(t, q^i, p_i),$$
$$\left(\text{resp. } \frac{d}{dt}\frac{\partial L}{\partial v^i} - \frac{\partial L}{\partial q^i} = F_i(t, q^i, v^i)\right).$$

Here the $F_i(t, q^i, p_i)$ (resp. $F_i(t, q^i, v^i)$) represent any kind of *covariant forces*, including dissipative and elastic joint forces, as well as actuator drives and control forces, as a function of time, coordinates and momenta. In covariant form we have

$$\dot{q}^k = g^{ki}p_i,$$
$$\dot{p}_i + \Gamma^i_{jk}g^{jl}g^{km}p_lp_m = F_i(t, q^i, p_i),$$

$$(\text{resp. } \dot{q}^i = v^i,$$
$$\dot{v}^i + \Gamma^i_{jk} v^j v^k = g^{ij} F_j(t, q^i, v^i)). \qquad \square$$

This proves the existence of the unique natural geometrical equivalence

$$\mathbf{Dual}_G : \mathbf{Lie} \cong \mathbf{Can}$$

in HD.

5. Topological Duality in Humanoid Dynamics

In this section we want to prove that HD can be *equivalently* described in terms of two *topologically dual functors* **Lag** and **Ham**, from **Diff**, the *category of smooth manifolds* (and their smooth maps) of class C^p, into **Bund**, the *category of vector bundles* (and vector-bundle maps) of class C^{p-1}, with $p \geq 1$. **Lag** is physically represented by the second-order Lagrangian formalism on $TQ^N \in \mathbf{Bund}$, while **Ham** is physically represented by the first-order Hamiltonian formalism on $T^*Q^N \in \mathbf{Bund}$.

Theorem 2. *There is a topological duality between Lagrangian and Hamiltonian formalisms on Q^N (Figure 1). In categorical terms, there is a unique natural topological equivalence*

$$\mathbf{Dual}_T : \mathbf{Lag} \cong \mathbf{Ham}$$

in HD.

Proof. The proof has two parts: cohomological and homological.

5.1. *Cohomological proof*

If $\mathcal{C} = \mathcal{H}^\bullet\mathcal{M}$ (resp. $\mathcal{C} = \mathcal{L}^\bullet\mathcal{M}$) represents the Abelian category of cochains on the momentum phase-space manifold T^*Q^N (resp. the velocity phase-space manifold TQ^N), we have the category $\mathcal{S}^\bullet(\mathcal{H}^\bullet\mathcal{M})$ (resp. $\mathcal{S}^\bullet(\mathcal{L}^\bullet\mathcal{M})$) of generalized cochain complexes $A^\bullet$ in $\mathcal{H}^\bullet\mathcal{M}$ (resp. $\mathcal{L}^\bullet\mathcal{M}$) and if $A'_n = 0$ for $n < 0$ we have a subcategory $\mathcal{S}^\bullet_{\mathcal{DR}}(\mathcal{H}^\bullet\mathcal{M})$ (resp. $\mathcal{S}^\bullet_{\mathcal{DR}}(\mathcal{L}^\bullet\mathcal{M})$) of De Rham differential complexes in $\mathcal{S}^\bullet(\mathcal{H}^\bullet\mathcal{M})$ (resp. $\mathcal{S}^\bullet(\mathcal{L}^\bullet\mathcal{M})$)

$$A^\bullet_{DR} : 0 \to \Omega^0(T^*Q^N) \xrightarrow{d} \Omega^1(T^*Q^N)$$
$$\xrightarrow{d} \Omega^2(T^*Q^N) \xrightarrow{d} \cdots \xrightarrow{d} \Omega^N(T^*Q^N) \xrightarrow{d} \cdots$$

(resp.

$$A^\bullet_{DR} : 0 \to \Omega^0(TQ^N) \xrightarrow{d} \Omega^1(TQ^N) \xrightarrow{d} \Omega^2(TQ^N) \xrightarrow{d}$$
$$\cdots \xrightarrow{d} \Omega^N(TQ^N) \xrightarrow{d} \cdots),$$

where $A'_N = \Omega^N(T^*Q^N)$ (resp. $A'_N = \Omega^N(TQ^N)$) is the vector space of all N-forms on T^*Q^N (resp. TQ^N) over $\mathbb{R}$.

Let $Z^N(T^*Q^N) = Ker(d)$ (resp. $Z^N(T) = Ker(d)$) and $B^N(T^*Q^N) = Im(d)$ (resp. $B^N(TQ^N) = Im(d)$) denote respectively the real vector spaces of cocycles and coboundaries of degree N. Since $d_{N+1}d_N = d^2 = 0$, it follows that $B^N(T^*Q^N) \subset Z^N(T^*Q^N)$ (resp. $B^N(TQ^N) \subset Z^N(TQ^N)$). The quotient vector space

$$H^N_{DR}(T^*Q^N) = Ker(d)/Im(d) = Z^N(T^*Q^N)/B^N(T^*Q^N)$$
$$(\text{resp. } H^N_{DR}(TQ^N) = Ker(d)/Im(d) = Z^N(TQ^N)/B^N(TQ^N)),$$

we refer to as the De Rham cohomology group (vector space) of HD on T^*Q^N (resp. TQ^N). The elements of $H^N_{DR}(T^*Q^N)$ (resp. $H^N_{DR}(TQ^N)$) are equivalence sets of cocycles. Two cocycles ω_1 and ω_2 are cohomologous, or belong to the same equivalence set (written $\omega_1 \sim \omega_2$) if and only if they differ by a coboundary $\omega_1 - \omega_2 = d\theta$. Any form $\omega_H \in \Omega^N(T^*Q^N)$ (resp. $\omega_L \in \Omega^N(TQ^N)$ has a De Rham cohomology class $[\omega_H] \in H^N_{DR}(T^*Q^N)$ (resp. $[\omega_L] \in H^N_{DR}(TQ^N)$).

Hamiltonian symplectic form $\omega_H = dp_i \wedge dq_i$ on T^*Q^N (resp. Lagrangian symplectic form $\omega_L = dL_{v^i} \wedge dq^i$ on TQ^N) is by definition both a closed two-form or two-cocycle and an exact two-form or two-coboundary. Therefore the two-dimensional De Rham cohomology group of humanoid motion is defined as a quotient vector space

$$H^2_{DR}(T^*Q^N) = Z^2(T^*Q^N)/B^2(T^*Q^N)$$
$$(\text{resp. } H^2_{DR}(TQ^N) = Z^2(TQ^N)/B^2(TQ^N)).$$

As T^*Q^N (resp. TQ^N) is a compact Hamiltonian symplectic (resp. Lagrangian symplectic) manifold of dimension $2N$, it follows that ω_H^N (resp. ω_L^N) is a volume element on T^*Q^N (resp. TQ^N), and the $2N$-dimensional De Rham cohomology class $[\omega_H^N] \in H^{2N}_{DR}(T^*Q^N)$ (resp. $[\omega_L^N] \in H^{2N}_{DR}(TQ^N)$) is nonzero. Since $[\omega_H^N] = [\omega_H]^N$ (resp. $[\omega_L^N] = [\omega_L]^N$), then $[\omega_H] \in H^2_{DR}(T^*Q^N)$ (resp. $[\omega_L] \in H^2_{DR}(TQ^N)$) and all of its powers up to the Nth must be zero as well. The existence of such an element is a necessary condition for T^*Q^N (resp. TQ^N) to admit a Hamiltonian symplectic structure ω_H (resp. Lagrangian symplectic structure ω_L).

A De Rham complex $A^\bullet_{DR}$ on T^*Q^N (resp. TQ^N) can be considered as a system of second-order differential equations $d^2\theta_H = 0, \theta_H \in \Omega^N(T^*Q^N)$ (resp. $d^2\theta_L = 0, \theta_L \in \Omega^N(TQ^N)$) having a solution represented by $Z^N(T^*Q^N)$ (resp. $Z^N(TQ^N)$). In local coordinates $q^i, p_i \in U_p$ (U_p open in

T^*Q^N) (resp. $q^i, v^i \in U_v$ (U_v open in TQ^N)) we have $d^2\theta_H = d^2(p_i dq^i) = d(dp_i \wedge dq^i) = 0$, (resp. $d^2\theta_L = d^2(L_{v^i} dq^i) = d(dL_{v^i} \wedge dq^i) = 0$). □

5.2. *Homological proof*

If $\mathcal{C} = \mathcal{H}_\bullet\mathcal{M}$, (resp. $\mathcal{C} = \mathcal{L}_\bullet\mathcal{M}$) represents an Abelian category of chains on T^*Q^N (resp. TQ^N), we have a category $\mathcal{S}_\bullet(\mathcal{H}_\bullet\mathcal{M})$ (resp. $\mathcal{S}_\bullet(\mathcal{L}_\bullet\mathcal{M})$) of generalized chain complexes $\mathcal{A}_\bullet$ in $\mathcal{H}_\bullet\mathcal{M}$ (resp. $\mathcal{L}_\bullet\mathcal{M}$), and if $A = 0$ for $n < 0$ we have a sub-category $\mathcal{S}^C_\bullet(H_\bullet M)$ (resp. $\mathcal{S}^C_\bullet(L_\bullet M)$) of chain complexes in $\mathcal{H}_\bullet\mathcal{M}$ (resp. $\mathcal{L}_\bullet\mathcal{M}$)

$$A_\bullet : 0 \leftarrow C^0(T^*Q^N) \xleftarrow{\partial} C^1(T^*Q^N) \xleftarrow{\partial} C^2(T^*Q^N) \xleftarrow{\partial} \cdots \xleftarrow{\partial} C^n(T^*Q^N) \xleftarrow{\partial} \cdots$$

(resp.

$$A_\bullet : 0 \leftarrow C^0(TQ^N) \xleftarrow{\partial} C^1(TQ^N) \xleftarrow{\partial} C^2(TQ^N) \xleftarrow{\partial} \cdots \xleftarrow{\partial} C^n(TQ^N) \xleftarrow{\partial} \cdots).$$

Here $A_N = C^N(T^*Q^N)$ (resp. $A_N = C^N(TQ^N)$) is the vector space of all finite chains C on T^*Q^N (resp. TQ^N) over $\mathbb{R}$, and $\partial_N = \partial : C^{N+1}(T^*Q^N) \to C^N(T^*Q^N)$ (resp. $\partial_N = \partial : C^{N+1}(TQ^N) \to C^N(TQ^N)$). A finite chain C such that $\partial C = 0$ is an N-cycle. A finite chain C such that $C = \partial B$ is an N-boundary. Let $Z_N(T^*Q^N) = Ker(\partial)$ (resp. $Z_N(TQ^N) = Ker(\partial)$) and $B_N(T^*Q^N) = Im(\partial)$ (resp. $B_N(TQ^N) = Im(\partial)$) denote respectively real vector spaces of cycles and boundaries of degree N. Since $\partial_{N-1}\partial_N = \partial^2 = 0$, then $B_N(T^*Q^N) \subset Z_N(T^*Q^N)$ (resp. $B_N(TQ^N) \subset Z_N(TQ^N)$). The quotient vector space

$$H^C_N(T^*Q^N) = Z_N(T^*Q^N)/B_N(T^*Q^N)$$
$$(\text{resp. } H^C_N(TQ^N) = Z_N(TQ^N)/B_N(TQ^N))$$

represents an *N-dimensional homology group (vector space) of humanoid dynamics.* The elements of $H^C_N(T^*Q^N)$ (resp. $H^C_N(TQ^N)$) are equivalence sets of cycles. Two cycles C_1 and C_2 are homologous, or belong to the same equivalence set (written $C_1 \sim C_2$) if and only if they differ by a boundary $C_1 - C_2 = \partial B$. The homology class of a finite chain $C \in C^N(T^*Q^N)$ (resp. $C \in C^N(TQ^N)$) is $[C] \in H^C_N(T^*Q^N)$ (resp. $[C] \in H^C_N(TQ^N)$). □

Particularly, in the case of the N-torus ($Q^N = T^N$), the Betti numbers of HD are given by

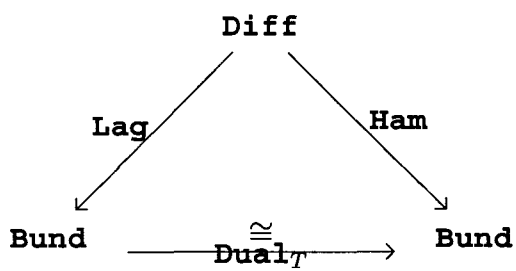
$$\begin{aligned} &b^0 = 1, \\ &b^1 = N, \ldots, b^p = \binom{N}{p}, \ldots, b^{N-1} = N, \\ &b^N = 1 \quad (0 \leq p \leq N). \end{aligned} \tag{6}$$

From the *homotopy axiom* for De Rham cohomologies, it follows that $H^\bullet_{DR}(Q^N) \approx H^\bullet_{DR}(TQ^N) \approx H^\bullet_{DR}(T^*Q^N)$. Also from the *De Rham theorem* it follows that $H^\bullet_{DR}(X) = H_\bullet(X)$ for any smooth manifold X. Therefore, $b^N = b_N$ are given by (6) for all three HD-manifolds $X \equiv T^N, TT^N, T^*T^N$.

Therefore, $b^N = b_N$ are given by (6) for both T^N and T^*T^N, defining also their *Euler-Poincaré characteristic* as [3]

$$\chi(T^N, T^*T^N) = \sum_{p=1}^{N} (-1)^p b_p.$$

In this way, we have proved a commutativity of a triangle:



which implies the existence of the unique natural topological equivalence

$$\mathbf{Dual}_T : \mathbf{Lag} \cong \mathbf{Ham}$$

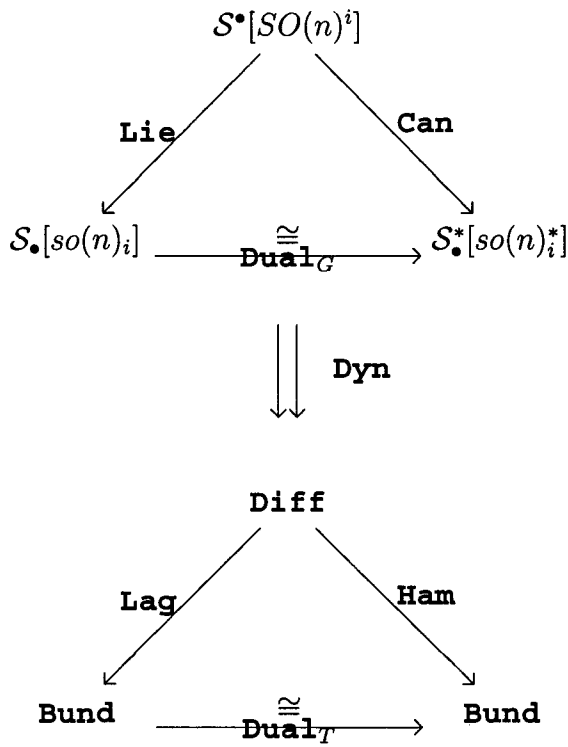
in HD.

6. Global Structure of Humanoid Dynamics

Theorem 3. *Global structure of HD is defined by the unique natural equivalence*

$$\mathbf{Dyn} : \mathbf{Dual}_G \cong \mathbf{Dual}_T.$$

Proof. The unique functorial relation $\mathbf{Dyn}:\mathbf{Dual}_G \cong \mathbf{Dual}_T$, uncovering the natural equivalence between *geometrical* and *topological* structures of HD:



— has been established by parallel development of Lagrangian and Hamiltonian HD-formulations, i.e., functors **Lag(Lie)** and **Ham(Can)**.

References

[1] P. Channon, S. Hopkins and D. Pham, A variational approach to the optimization of gait for a bipedal robot, *J. Mechanical Engin. Sci.* **210** (1996) 177–186.

[2] H. J. Chiel, R. D. Beer, R. D. Quinn and K. S. Espenschied, Robustness of a distributed neural network controller for locomotion in a hexapod robot, *IEEE Trans. Robotics Auto.* **8** (1992) 293–303.

[3] Y. Choquet-Bruhat and C. DeWitt-Morete, *Analysis, Manifolds and Physics*, 2nd edn. (North-Holland, Amsterdam, 1982).

[4] C. T. J. Dodson and P. E. Parker, *A User's Guide to Algebraic Topology* (Kluwer, Dordrecht, Boston, London, 1997).

[5] M. Dorigo and U. Schnepf, Genetics-based machine learning and behavior-based robotics: A new synthesis, *IEEE Trans. Syst. Man Cybernetics* **23b** (1993) 141–154.

[6] S. Hashimoto, Humanoid robots in Waseda University: Hadaly-2 and Wabian, in *IARP First International Workshop on Humanoid and Human Friendly Robotics*, Waseda University (1998), pp. 1–2.

[7] N. G. Hatsopoulos, Coupling the neural and physical dynamics in rhythmic movements, *Neural Comput.* **8** (1996) 567–581.

[8] Y. Hurmuzlu, Dynamics of bipedal gait, *J. Appl. Mechanics* **60** (1993) 331–343.

[9] E. Igarashi and T. Nogai, Study of lower level adaptive walking in the saggital plane by a biped locomotion robot, *Adv. Robotics* **6** (1992) 441–459.

[10] V. Ivancevic and M. Snoswell, Fuzzy-stochastic functor machine for general humanoid-robot dynamics, *IEEE Trans. Syst. Man Cybernetics* **31b**(3) (2001) 319–330.

[11] I. Kolar, P. W. Michor and J. Slovak, *Natural Operations in Differential Geometry* (Springer-Verlag, Berlin, Heidelberg, 1993).

[12] J. Lieh, Computer oriented closed-form algorithm for constrained multibody dynamics for robotics applications, *Mechanism Machine Theory* **29** (1994) 357–371.

[13] S. MacLane, *Categories for the Working Mathematician*, Graduate Texts in Mathematics (Springer-Verlag, New York, 1971).

[14] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Texts in Applied Mathematics, Vol. 17 (Springer-Verlag, New York, 1994).

[15] T. McGeer, Passive dynamic walking, *Int. J. Robotics Res.* **9** (1990) 62–82.

[16] J. Pratt and G. Pratt, Exploiting natural dynamics in the control of a planar bipedal walking robot, in *Proceedings of the 36 Annual Allerton Conference on Communication, Control, and Computing*, Allerton (1998), pp. 739–748.

[17] C. Pribe, S. Grossberg and M. A. Cohen, Neural control of interlimb oscillations. II, Biped and quadruped gaits and bifurcations, *Biological Cybernetics* **77** (1997) 141–152.

[18] P. Sardain, M. Rostami and G. Bessonnet, An anthropomorphic biped robot: Dynamic concepts and technological design, *IEEE Trans. Syst. Man Cybernetics* **28a** (1999) 823–838.

[19] S. Schaal and C. G. Atkeson, Constructive incremental learning from only local information, *Neural Comput.* **10** (1998) 2047–2084.

[20] S. Schaal, Is imitation learning the route to humanoid robots?, *Trends Cognitive Sci.* **3** (1999) 233–242.

[21] H. Seraji, Configuration control of redundant manipulators: Theory and implementation, *IEEE Trans. Robotics Automation* **5** (1989) 437–443.

[22] D. Seward, A. Bradshaw and F. Margrave, The anatomy of a humanoid robot, *Robotica* **14** (1996) 437–443.

[23] C. L. Shih, W. Gruver and T. Lee, Inverse kinematics and inverse dynamics for control of a biped walking machine, *J. Robotic Syst.* **10** (1993) 531–555.

[24] C. L. Shih and C. A. Klein, An adaptive gait for legged walking machines over rough terrain, *IEEE Trans. Syst. Man Cybernetics* **23a** (1993) 1150–1154.

[25] M. Vukobratovic, On the stability of biped locomotion, *IEEE Trans. Biomedical Engin.* **17** (1970) 25–36.

[26] M. Vukobratovic, D. Juricic and A. Frank, On the control and stability of one class of biped locomotion systems, *ASME J. Basic Engin.* **92** (1970) 328–332.

[27] M. Vukobratovic and Y. Stepanenko, On the stability of anthropomorphic systems, *Math. Biosci.* **15** (1972) 1–37.

[28] M. Vukobratovic and Y. Stepanenko, Mathematical models of general anthropomorphic systems, *Math. Biosci.* **17** (1973) 191–242.

[29] M. Vukobratovic, *Legged Locomotion Robots and Anthropomorphic Mechanisms* (Mihailo Pupin, Belgrade, 1975).

[30] M. Vukobratovic, B. Borovac, D. Surla and D. Stokic, *Biped Locomotion: Dynamics, Stability, Control, and Applications* (Springer-Verlag, Berlin, Heidelberg, 1990).

[31] T. Yoshikawa, Analysis and control of robot manipulators with redundancy, in *Robotics Research*, eds. M. Brady and R. Paul (MIT Press, Cambridge, 1984), pp. 735–747.

This page is intentionally left blank

CHAPTER 14

THE EFFECTS OF BODY COMPOSITION ON ENERGY EXPENDITURE AND WEIGHT DYNAMICS DURING HYPOPHAGIA: A SETPOINT ANALYSIS

FRANK P. KOZUSKO
Department of Mathematics, Hampton University, Hampton Virginia, USA 23668
frank.kozusko@hamptonu.edu

1. Introduction

During hypophagia (under eating) loss of body weight is expected. The dynamics of the weight change involve complicated biochemical processes that produce changes in our daily energy needs, the amounts of fat and nonfat tissue stored in the body and the energy efficiency at which we function. If we consume fewer calories than are required for our daily activities, the body is forced to use the energy stored in the fat and nonfat tissues with resultant weight loss. Most of the energy will be supplied by the high energy density fat mass while a smaller quantity will be supplied by consumption of low energy density nonfat body mass. Since it is nonfat that is metabolically active, loss of nonfat reduces the daily required energy, reducing the energy deficit. Chemical changes in the body sense the loss of fat, causing the appetite to increase. The body becomes more efficient at performing its metabolic and physical activities which will further reduce the rate of weight loss. Eventually body weight will decrease to the point where there is no longer a deficit for the dietary calories provided (unless the intake is less than minimum starvation requirements).

Figure 1 is an energy block diagram. The dotted line to the Out arrow indicates the dependence of energy expenditure on the levels of fat and nonfat. What the average person understands about this diagram is that a 3500 Calorie deficit will produce a loss of one pound of fat. Hence, reduce your diet by 500 Calories per day and loose a pound per week. This would

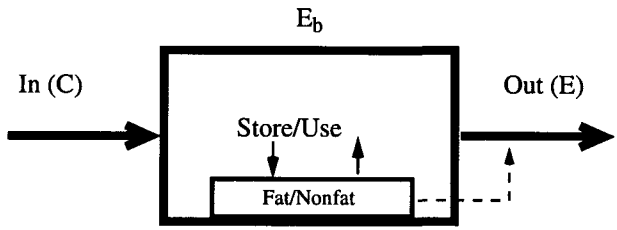


Fig. 1. Energy block diagram.

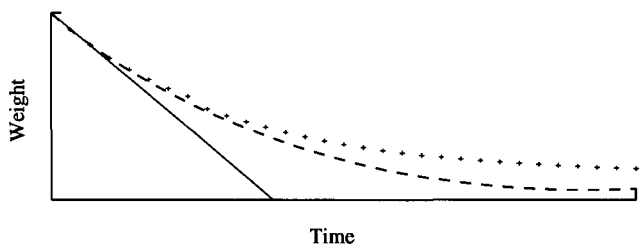


Fig. 2. Representive weight loss model results.

produce a constant rate of weight loss as represented by the solid line in Fig. 2. However, this concept neglects the change in the energy expenditure as well as the fact that not all weight loss is fat. A more refined model allows the energy expenditure to decrease in relationship to the decrease in weight (dashed line in Fig. 2), with the slope determined by the proportion of fat/nonfat in the lost weight. A further refinement would be to model the increased energy efficiency achieved during f weight loss (represented by the +++ curve in Fig. 2).

Figure 1 is a block diagram of the law of conservation of energy represented functionally as [1]:

$$\frac{d}{dt}(\textit{energy stores}) = \frac{d}{dt}(\textit{energy supplied}) - \frac{d}{dt}(\textit{energy consumed}) \qquad (1)$$

and mathematically as

$$\frac{dE_b^*}{dt^*} = C^* - E^*. \qquad (2)$$

At times, it will be convenient to conduct analysis using normalized/nondimensionalized parameters. We use * to indicate dimensional parameters and those without * are nondimensional. Definitions are provided in

Table 1. Definition of Terms (reprinted from [2]).

Term	Definition	Nondimensional
C^*	Daily Calorie intake from diet. (Calories/day)	$C = C^*/C_0^*$
C_0^*	Equilibrium (setpoint) value of C^* ($C_0^* = E_0^*$).	$1 = C_0^*/C_0^*$
E^*	Daily Energy Expenditure. (Calories/day)	$E = E^*/E_0^*$
E_0^*	Equilibrium (setpoint) value of E^* ($E_0^* = C_0^*$).	$1 = E_0^*/E_0^*$
E_b^*	Energy stored in the body (Calories)	
F^*	Body Fat Weight (Lbs)	$F = F^*/W_0^*$
F_0^*	Equilibrium (setpoint) value of F^*	$F_0 = F_0^*/W_0^*$
HB	Harris–Benedict	
k_f^*	Energy Density of Fat (Calories/Lb)	
k_n^*	Energy Density of Nonfat (Calories/Lb)	
k_w^*	Nominal Energy Density of Body Weight (Calories/Lb)	
N^*	Nonfat Body Weight (Lbs)	$N = N^*/W_0^*$
N_0^*	Equilibrium (setpoint) value of N^*	$N_0 = N_0^*/W_0^*$
t^*	Time (Days)	$t = t^*/t_0^*$
t_0^*	Characteristic Time: $t_0^* = k_w^* W_0^*/E_0^*$	
W^*	Total Body Weight (Lbs) $W^* = F^* + N^*$ $W = F + N$ and $1 = F_0 + N_0$	$W = W^*/W_0^*$
W_0^*	Equilibrium (setpoint) value of W^*.	$1 = W_0^*/W_0^*$

Table 1, reprinted from [2]. C^* (Calories/day) is the energy supplied from food consumption, assumed to readily determined. It is the modeling of the other two components of the energy equation which this chapter will explore: (1) How does the body's energy stores (fat/nonfat), E_b^* (Calories), change during a deficit energy balance? and (2) What is the daily energy expenditure, E^* (Calories/day) for a person experiencing an energy deficit induced weight loss?

2. Modeling Human Daily Energy Expenditure

The energy expenditure of the human body consist of energy to digest the food consumed: Thermic Effect of Food (TE), energy to conduct Physical Activity: (PA), and the energy to conduct all other metabolic functions: Resting Metabolic Rate (RMR). Percent estimates for sedentary adults [3] are TE (10%), PA (20–30%) and RMR (60–70%). The sedentary total is called the 24 hour Energy Expenditure (24EE) or the Resting Energy Expenditure (REE). Predictive equations usually involve multilinear regressions of measured REE versus various combinations of body weight, fat free weight, body fat weight, age and gender. Since most of the metabolic activity is carried out by the nonfat mass, models linearly dependent on fat free mass are popular.

2.1. *Equilibrium models*

The classic Harris–Benedict equations [4], first formulated in 1919 and estimating REE based on the subject's age, gender, height and weight, are the most common for research and clinical use [5]. Also popular in the literature is the Brody–Kleiber Law where REE is proportional to body weight to the 3/4 power [6]. Additional energy expenditure for physical activity is directly proportional to body weight [7].

These and other models are based on equilibrium conditions and their use during nonequilibrium, such as weight loss, have been questioned in numerous studies. Leibel *et al.* [8] reported a decrease in both obese and never obese subjects following a 10% weight loss. Foster [9] found a decrease in the relative cost of physical activity (walking) following significant weight loss. Weigle [10] also found a reduction of the energy requirements of walking even after a weighted vest was used to compensate for the lost weight. Stern *et al.* [11], studying rats, stated that the Brody–Kleiber law might not apply in a non-equilibrium state. Other studies of 24EE after weight loss showing reduced levels when compared to that anticipated for changes in body composition include [12] and [13].

2.2. *Setpoint analysis and modeling nonequilibrium energy needs*

This "adaptive thermogenesis" [14] is attributed to the body's defense of a setpoint weight [15, 16]. The setpoint weight may be described as the genotypical "normal" weight of the individual. Hirsch *et al.* [17] define the setpoint weight as that weight for which the energy expenditure is in agreement with the Brody–Kleiber equation. Metabolic adaption to weight loss has been observed in both lean and, to a lesser extent, obese individuals [18]. Although some mathematical models of energy expenditure can be found in the literature, none adequately address this setpoint mechanism.

The simplest model [19] provides energy consumption as a constant proportionality with weight. More robust models described by Alpert [1, 20, 21] track the change in fat and nonfat separately and provide some recognition of setpoint dynamics by using different linear fits for underfed and equilibrium conditions. He also postulates a changing value of the energy supplied per unit of fat weight loss during underfeeding. [22] suggests a model describing physical activity as proportional to weight and a near Kleiber–Brody term for the remaining 24EE. The setpoint model of energy expenditure was introduced in [23].

We ask how does energy expenditure vary from setpoint energy as weight decreases from its setpoint value because of a negative energy balance (undereating). Since weight is the only variable in the HB equations during weight loss, we will model that the energy (E^*) is proportional to weight (W^*). To provide for metabolic adaption, we make the proportionality factor (α^*) a variable depending on weight. Then

$$E^* = \alpha^*(W^*)W^*. \tag{3}$$

It is emphasized that we propose Eq. (3) only for variation around the setpoint and not to calculate setpoint.

E_0^*, W_0^* and α_0^* are the predicting equilibrium values ($E_0^* = \alpha_0^* W_0^*$). Parameters E_s^*, W_s^* and α_s^* are introduced as the equilibrium starvation values. We define the nondimensional parameters

$$E = \frac{E^*}{E_0^*}, \quad W = \frac{W^*}{W_0^*} \quad \text{and} \quad \alpha = \frac{\alpha^*}{\alpha_0^*}. \tag{4}$$

The model assumes a linear fit from (W_0^*, α_0^*) to (W_s^*, α_s^*). The s subscript may also stand for any secondary equilibrium condition established between setpoint and starvation. After nondimensionalization the linear fit yields:

$$\alpha = \frac{(\alpha_s - W_s)}{(1 - W_s)} + \frac{(1 - \alpha_s)}{(1 - W_s)} W. \tag{5}$$

($W = 1 \rightarrow \alpha = 1$ and $W = W_s \rightarrow \alpha = \alpha_s$.) Defining $\beta_1 = \frac{(\alpha_s - W_s)}{(1 - W_s)}$ and $\beta_2 = \frac{(1 - \alpha_s)}{(1 - W_s)}$, the model becomes

$$E = \beta_1 W + \beta_2 W^2. \tag{6}$$

We note that $\beta_1 + \beta_2 = 1$ and devise an analysis to first find β_2. Since $0 < \alpha_s, W_s \leq 1$, β_2 is always positive. We can postulate that $\alpha_s \geq W_s$, that is the body's ability to reduce the per pound energy requirement is limited. Then $0 \leq \beta_1, \beta_2 \leq 1$. Rewriting the setpoint energy expression

$$E = \beta_1 W + \beta_2 W^2 = (1 - \beta_2)W + \beta_2 W^2 = W + \beta_2 W(W - 1). \tag{7}$$

Then

$$\frac{dE}{dW} = 1 + \beta_2(2W - 1). \tag{8}$$

The β_2 term provides for E to decrease faster than the weight and represents the metabolic adaption. The maximum adaption ($\beta_2 = 1$) we assign to the lean individual with a low setpoint body fat ratio ($F_0 = \frac{F_0^*}{W_0^*}$) and the minimum adaption ($\beta_2 = 0$) to the obese with a high body fat ratio.

Because $F_0 = 1$ and $F_0 = 0$ are mathematical but not physiological, we can anticipate an asymptotic approach to $\beta_2 = 0$ and $\beta_2 = 1$ versus F_0. An analytic function that fits this criteria is (Fig. 3):

$$\beta_2 = \frac{\tan h\left(\frac{f(m-F_0)}{F_0(1-F_0)}\right) + 1}{2}. \tag{9}$$

The value of m determines the inflection point and the value of f determines the maximum slope at the inflection point. (See [23] for β_2 define for nonfat ratio.) We predict that the energy expenditure relative to setpoint energy will change with the (setpoint) relative weight change according to Eqs. (8) and (9) depending on the setpoint body fat ratio F_0. Figure 4(a) provides a schema interpreted from [17] showing that weight loss from point 1 does not have a corresponding energy change along the equilibrium relationship to

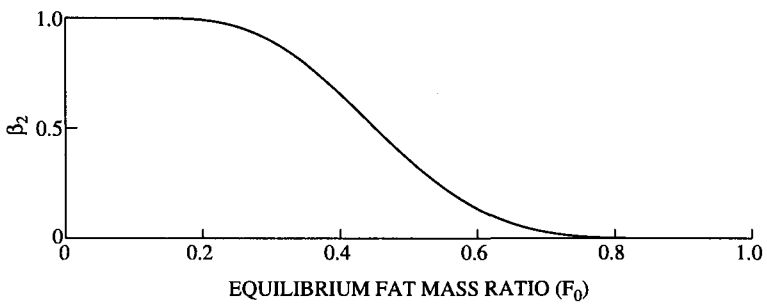


Fig. 3. Metabolic reduction (β_2) factor versus setpoint fat mass ratio ($f = 1.5$ and $m = 0.5$), (reprinted from [2]).

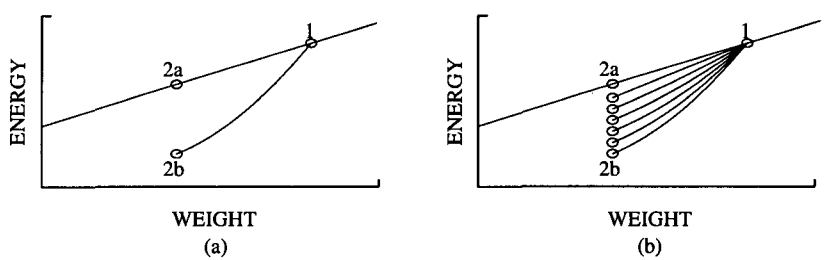


Fig. 4. (a) Hypothetical schema (interpreted from [17]) relating the experimental change of energy for a weight change ($1 \rightarrow 2b$) versus that anticipated from equilibrium relationships ($1 \rightarrow 2a$) and (b) schema depicting many setpoint paths from $1 \rightarrow 2$ depending on the level of metabolic adaption, ($1 \rightarrow 2a$: none and $1 \rightarrow 2b$: maximum).

2a but to a reduce level 2b. Figure 4(b) shows the setpoint model equivalent schema showing many paths from 1 to 2 depending on the level of metabolic adaption: 1 to 2a no adaption (high setpoint F_0) and 1 to 2b maximum adaption (low setpoint F_0).

2.3. *Comparing models*

The Harris–Benedict equation resting energy is of the form

$$REE = \alpha_{r_0} + \alpha_{r_1} W \tag{10}$$

where α_{r0} represents all the linear regression factors for height, weight and gender which remain constant for an individual losing weight. α_{r0} is the linear regression factor for weight dependence. Adding $\alpha_p W$ for additional physical activity [7], the total energy expenditure is

$$E = PA + REE = \alpha_{r_0} + (\alpha_{r_1} + \alpha_p)W = \alpha_{c_0} + \alpha_{c_1} W \tag{11}$$

where

$$\frac{dE}{dW} = \alpha_{c_1} \Rightarrow \textit{constant slope.} \tag{12}$$

We require that $E = 1$ when $W = 1$ for a valid comparison with Eq. (6) and note that $E = W$ will produce a lower limit of any line $E = \alpha_{c_0} + \alpha_{c_1} W$ passing through (1,1) and so provides some reduction of energy from the Harris–Benedict model. The $E = W$ model will be called the constant slope model. The setpoint model is equivalent to the constant slope model when $\beta_2 = 0$. Figure 5 shows E versus W for the setpoint model for $\beta_2 = 1.0, 0.8, 0.6, 0.4$, and 0.2 (bottom to top) and the constant model (dashed). The actual HB energy will be greater than the constant slope value. The figure should be read from right to left. Weight and Energy levels start at nondimensional values 1.0. As the weight decrease (to the left) the energy

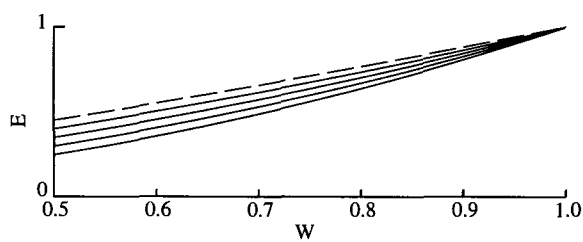


Fig. 5. Energy versus weight for setpoint model with $\beta_2 = 1.0$, 0.8, 0.6, 0.4 and 0.2 (bottom to top) and the constant ($\beta_2 = 0$) model (dashed).

needs decrease, but more slowly for lower values of β_2 and slowest for the constant (dashed) model.

3. Energy from Fat/Nonfat Body Mass

When the body's energy requirements exceed the energy supplied from the diet, the deficit must be made up from the body's stored energy. These stores are in the form of glycogen, protein and fat [24]. Glycogen is stored in the muscles and liver and accounts for only 800–1600 Calories in 200–400 grams, enough energy for less than a day of fasting. Because each gram of glycogen is bound to 2 to 4 grams of water, severe calorie restriction can produce 2 to 3 pounds of weight loss in the first day. We are interested in the long term effects of calorie restriction and will neglect the glycogen energy storage.

The total body weight (W^*) is the sum of the fat (F^*) and the nonfat (N^*)

$$W^* = F^* + N^*. \tag{13}$$

Fat has an energy density (k_f^*) of 9400 Calories/kg (4273 Calories/Lb) while nonfat body mass yields (k_n^*) 1020 Calories/kg (464 Calories/Lb) [25]. Clearly, the way these energy compartments are used to make up the energy deficit will effect the rate of weight loss.

3.1. *The personnel fat ratio*

In the event of energy deficit, the body must respond by combining two disparate means. Consuming high energy density fat mass provides energy to compensate for the intake deficit with a slower loss of body weight. However, reducing the total body fat is not as effective in reducing the energy deficit itself as is reducing the metabolically active nonfat mass. Consuming low energy density nonfat mass produces a more rapid loss of weight. How is this partitioning determined?

Kreitzman [26] defines a Personal Fat Ratio (PFR) as the ratio of fat to fat free mass in the excess weight above the core fat-free body. If we define $W_s^* = N_s^*$ to be a theoretical starvation weight at which body fat is zero then

$$PFR = \frac{F^*}{N^* - N_s^*} \Rightarrow PFR = \frac{\Delta F^*}{\Delta N^*}. \tag{14}$$

Kreitzman reports that this ratio remain stable during weight reduction in both lean and obese individual, even through a reduction in excess of 80 kg.

3.2. *The ratio of nonfat loss to total weight loss*

In a similiar way to Kreitzman, Forbes [27] states that the ratio of nonfat loss to total weight loss is curvilinearly related to the initial percent body fat (F_0). Forbes shows that extremely low calorie diets will have a different curvilinear relationship from that of more reasonable calorie diets. Defining

$$\Phi = \frac{\Delta N^*}{\Delta W^*} \tag{15}$$

and using Eq. (13), shows

$$\frac{1}{1 + PFR} = \Phi. \tag{16}$$

In modeling a relationship between Φ and F_0, we can be comfortable in setting $\Phi = 1$ when $F_0 = 0$. If there is no body fat, then the change in lean mass and the change in total weight must be equal. We are tempted to let $\Phi = 0$ when $F_0 = 1$ (a very theoretical consideration). However, there is a limit to how low Φ can go. The majority of body fat is stored in adipose tissue which is estimated at 80%–85% fat, the rest being water and a small amount of protein. This implies a minimum Φ of 0.2–0.15. Forbes [28] cites a case of a 213 kg weight loss from an initial 304 kg and an estimated 38 kg loss of nonfat weight ($\Phi = 0.18$). Setting $\Phi_{\min} = 0.15$,

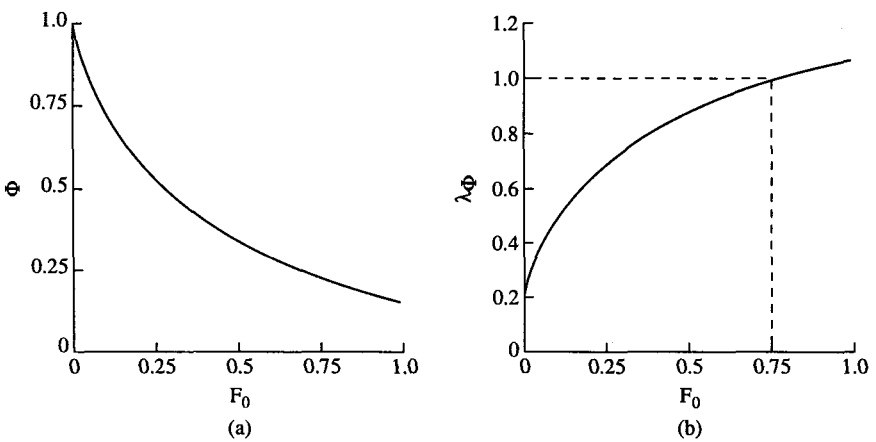


Fig. 6. (a) The ratio of nonfat weight loss/total weight loss (Φ) versus initial body fat ratio and (b) The energy density ratio (λ_Φ) versus initial body fat ratio (reprinted from [2]).

data from [27] suggests a relationship (Fig. 6(a))

$$\Phi = \frac{(1 - F_0)}{(1 + 1.5F_0^{0.75})} + 0.15F_0. \tag{17}$$

3.3. *The energy density and the energy density ratio*

The amount of energy supplied by the consumption of fat and nonfat for a specified change of body weight change is

$$\Delta E_b^* = k_f^* \Delta F^* + k_n^* \Delta N^* \tag{18}$$

using $\Delta W^* = \Delta F^* + \Delta N^*$ and Eq. (15) yields

$$\begin{aligned} \Delta E_b^* = \left(k_f^* - \Phi\left[k_f^* - k_n^*\right]\right)\Delta W^* &= \left(\frac{k_f^*}{k_w^*} - \Phi\frac{[k_f^* - k_n^*]}{k_w^*}\right) k_w^* \Delta W^* \\ &= \lambda_\Phi k_w^* \Delta W^* \end{aligned} \tag{19}$$

$k_f^* - \Phi[k_f^* - k_n^*]$ is the true energy density of the lost body weight. λ_Φ is the energy density ratio and indicates how ΔE_b^* varies from the simple model of $k_w^* \Delta W^*$. Using a common nominal value for k_w^*, 7700 Calories/kg (3500 Calories/Lb), values introduced with Eqs. (13) and (17) yields Fig. 6(b). Figure 6(b) shows that the energy density ratio is less than 1.0 for all but the extremely obese, so most subjects will have an energy density less than the nominal value.

4. The Setpoint/Body Composition Adjusted Energy Rate Equation

Our goal was to develop a model for ΔE_b^* and E^* to use in the conservation of energy equation (2). This has been accomplished and both these quantities modeled as dependent on the initial body fat ratio F_0. We will need to define additional nondimensional variables

$$C = \frac{C^*}{C_0^*} \quad \text{and} \quad t = \frac{t^*}{t_0^*} \tag{20}$$

where t_0^* is defined to simplify the resulting equation $t_0^* = \frac{k_w^*}{\alpha_0^*}$. Making the proper substitutions and completing the nondimensionalization, the setpoint body composition adjusted rate equation is

$$\frac{dW}{dt} = \frac{C - \beta_1 W - \beta_2 W^2}{\lambda_\Phi}; \quad W(t = 0) = 1. \tag{21}$$

Equation (21) shows how the partitioning of fat/nonfat consumption during weight loss effects the rate of weight loss. For most cases $\lambda_\Phi < 1.0$ and the

rate of weight loss will be greater than anticipated using the nominal energy density. Equation (21) has the analytic solution

$$W = \frac{\lambda}{2\beta_2} \coth\left(\frac{\lambda}{2\lambda_\Phi} t + K\right) - \frac{\beta_1}{2\beta_2} \tag{22}$$

with

$$\lambda = \sqrt{\beta_1^2 + 4\beta_2 C}, \quad K = \coth^{-1}\left(\frac{2\beta_2 + \beta_1}{\lambda}\right). \tag{23}$$

5. Analysis

Equation (21) is the generalized nondimensional energy rate equation adjusted for body composition effects on Calories/pound in the lost weight and for the setpoint driven metabolic adaption (reduction) to weight loss. Setting $\beta_2 = 0$ ($\beta_1 = 1$) eliminates the metabolic reduction and provides a constant slope model similiar to HB. When $\lambda_\Phi = 1$, the energy density is adjusted to a constant nominal value (no difference between energy density of fat and nonfat weight loss). We use the robustness of Eq. (21) to analyze and compare weight loss dynamics.

5.1. *The Minnesota experiment*

The seminal Minnesota Semistarvation Experiment was conducted during the last months of World War II using 32 volunteer conscientious objectors. The subjects were monitored during a three month control period establishing an equilibrium body weight and calorie consumption for a specified activity regime. A six months dieting period followed, with the goal of 25% weight loss. The daily average calories were adjusted weekly to provide weight versus time along a parabolic path flattening near the end of the period. The dieting period was followed by a three months rehabilitation period during which calorie consumption was gradually increased. An extensive report of the data taken during these periods is provided in [29]. Figure 7(a) shows the weekly averages of the daily calories consumption relative to setpoint for the 32 individuals of the Minnesota Experiment. The severity of the diet is clear with an average drop in calories of more than 50% during the 24 weeks of weight loss. Calories are increased during the rehabilitation period (weeks 25–36). Figure 7(b) shows the weekly average of the body weights relative to setpoint. The data (+++) shows the parabolic weight loss curve and weight increase during rehabilitation.

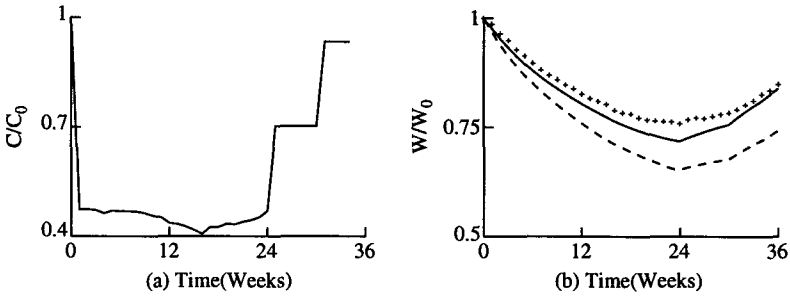


Fig. 7. (a) Average weekly calories relative to setpoint calories for 32 individuals of the Minnesota experiment and (b) Averages of $\left(\frac{W}{W_0}\right)$ for 32 individuals of the Minnesota experiment: data (+++), setpoint model (solid line) and constant model (modified from [30]).

Because C varied with time, Eq. (21) had to be solved numerically. Results as first presented in [30] are shown graphically in Fig. 7(b) displaying the weekly averages of the setpoint normalized weights of the 32 individuals of the Minnesota Experiment and the averages of solutions of Eq. (21) for the setpoint and constant models. The data was sufficient to calculate the characteristic time for each subject. The energy density λ_Φ was set to 1.0. The constant slope model greatly over estimates the amount of weight loss.

5.2. *The characteristic time and rate of weight loss*

Setting $\lambda_\phi = 1$, and $C = 0.75$ to represent a 25% calorie reduction and solving Eq. (21) for various values of β_2 yields results displayed in Fig. 8. This represents the weight loss dynamics if everyone had the same characteristic time, t_0^* and the same energy constant weight loss energy density. Since $t_0^* = k_w^* W_0^*/E_0^*$, a uniform characteristic time requires each individual to have the same setpoint $\frac{W_0^*}{E_0^*}$ value which is not the case.

To get a true picture of the relative rates of weight loss, we need to move out of the nondimensional time units and calculate the individual characteristic time. Forbes [28] provides an estimate for E_0^* based on the initial fat mass (F_0^*) and nonfat mass (N_0^*)

$$E_0^*(\text{Calories/day}) = 35.7N_0^*(\text{kg}) + 15.3F_0^*(\text{kg}) + 198 \tag{24}$$

derived form a study group with a range of body fat from 2 to 74 kg. We convert Eq. (24) to

$$\frac{E_0^*}{W_0^*}(\text{Calories/day/kg}) = 38.7 - 20.4F_0 \tag{25}$$

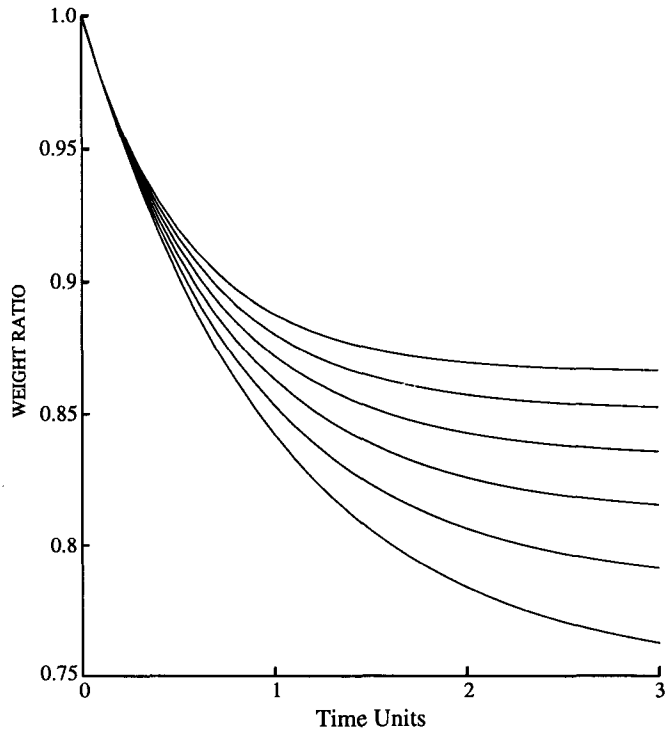


Fig. 8. Weight ratio for $\beta_2 = 1.0, 0.8, 0.6, 0.4, 0.2$ and 0.0 (bottom to top) for $C = 0.75$ versus time units ($\lambda_\Phi = 1$, t_0^* constant) (reprinted from [23]).

then

$$t_0^*(\text{days}) = \frac{k_w^*}{38.7 - 20.4F_0}. \tag{26}$$

We can get a physical interpretation of t_0^* by evaluating Eq. (21) at $t = 0$ (where the rate of weight loss is greatest). In the t^* units (real time)

$$\frac{dW}{dt^*} = \frac{-(1-C)}{t_0^*\lambda_\Phi}. \tag{27}$$

Then the maximum rate of change of weight for a given calorie deficit $(1 - C)$ is inversely proportional to t_0^*. Figure 9(a) show t_0^* versus F_0 and Fig. 9(b) shows the maximum rate of weight loss versus F_0 for diet ratios, $C = 0.50$–0.75.

Considering a 25% diet ($C = 0.75$) and solving Eq. (21) for $F_0 = 10\%$ to $F_0 = 55\%$ in 5% steps ($\beta_2 = 1.0 \rightarrow 0.2$) and using Eq. (26) for characteristic

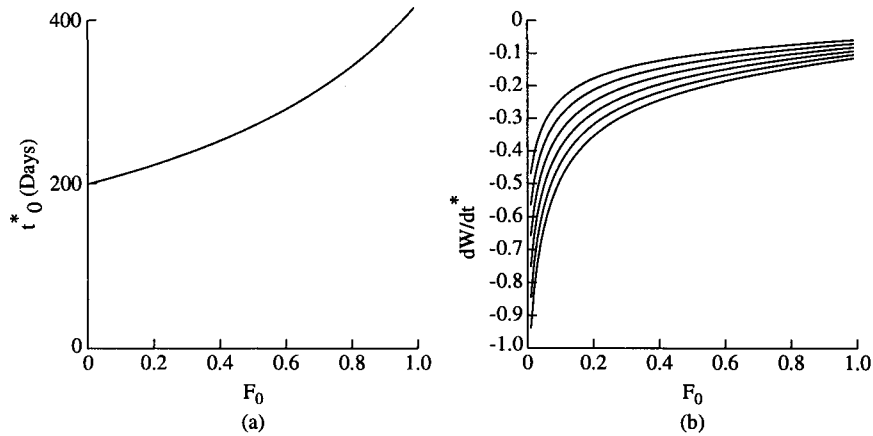


Fig. 9. (a) The characteristic time, t_0^* in days versus F_0 and (b) The maximum rate of change of relative weight versus initial body fat percentage for 50%, 55%, 60%, 65%, 70% and 75% Calorie Diet (top to bottom) (reprinted from [2]).

times yields Fig. 10. The rates of weight loss for the high body fat cases have slowed compared to Fig. 8 because of the higher characteristic times (lower E_0^*/W_0^* ratio). The rates are nearly the same until the amount of weight loss is significant enough to induce the metabolic reduction.

If we further introduce the fat/nonfat effects embodied in λ_Φ, we get the results displayed in Fig. 11. The rate of weight loss has increased significantly for the low fat (high β_2) cases because more low density lean mass is providing a higher percentage of the energy deficit compared to the high fat cases. Again, when the weight loss is enough to trigger metabolic reduction, the energy deficit is lowered and the rate is reduced; more for the low fat than the high fat cases.

5.3. *Comparing the models in dynamics*

We start by solving Eq. (21) for (1) the Harris-Benedict case for nominal energy density ($\beta_1 = 1, \beta_2 = 0, \lambda_\Phi = 1$), and (2) the setpoint model adjusted for body composition. In both case t_0^* is calculated from Eq. (26). Figure 12 shows the comparative results for various values of F_0 as indicated, for 25% calorie reduction ($C = 0.75$). For low values of F_0, the HB model under estimates the rate of weight loss in the early days because it over estimating the calories from fat and the effects of metabolic adaption are small until there is some level of weight loss. As time goes on

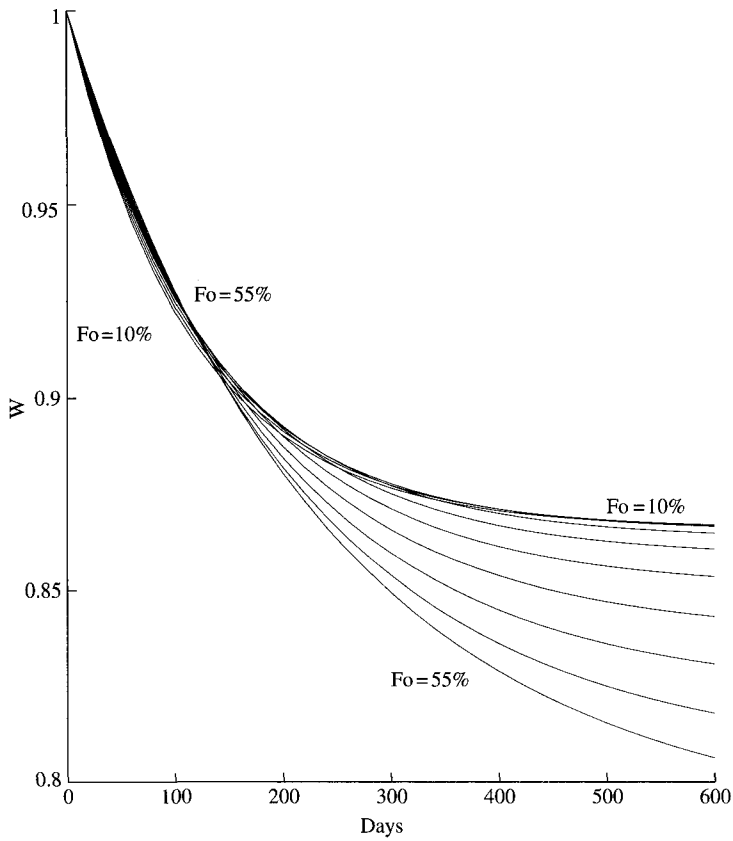


Fig. 10. Weight ratio for $F_0 = 10\%$ to $F_0 = 55\%$ in 5% steps ($\beta_2 = 1.0 \rightarrow 0.2$) versus time for 0.75 calorie ratio diet with individual characteristic time and constant energy density ($\lambda_\Phi = 1$).

the metabolic reduction dominates the setpoint model and the HB over estimate the weight loss as shown in [23].

As F_0 increases, the energy density increases coming closer to the nominal value while the level of adaption decreases (Fig. 3). The time to crossover, the time when the setpoint weight is greater than the HB weight, increases while the maximum difference between the two models gets smaller. For $F_0 = 0.55$ there is no perceptible difference for the first 300 days.

Further comparing the models for more restrictive diets (50%–70%) shows that all cases exhibit the initial under estimate by the HB model

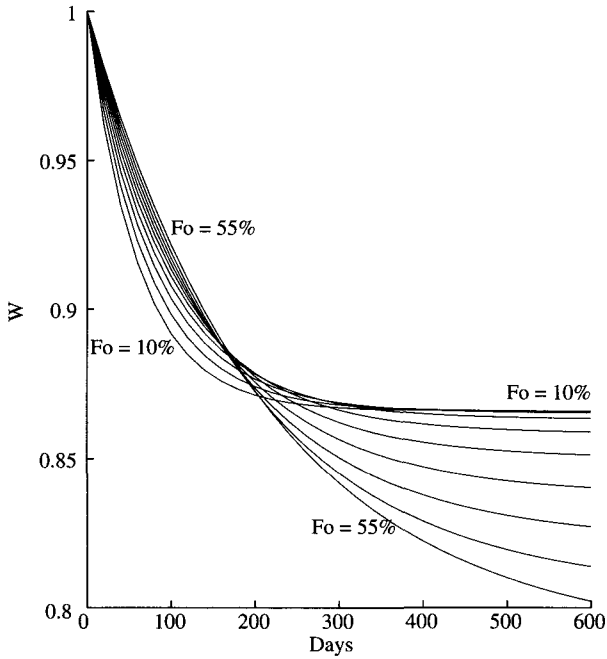


Fig. 11. Weight ratio for $F_0 = 10\%$ to $F_0 = 55\%$ in 5% steps ($\beta_2 = 1.0 \rightarrow 0.2$) versus Time for 0.75 calorie ratio diet corrected for characteristic time and energy density.

where the cross over time is more sensitive to the initial body fat (F_0) than the caloric intake (Fig. 13(a)). The minimum cross over time occurs for $F_0 \approx 0.3$ where the metabolic adaption (β_2, Fig. 3) is high and the fat density ratio is approaching 1.0 (Fig. 6(b)). To the left of 0.3, the metabolic adaption slows the weight decrease while higher consumption of lean body weight increases the rate of weight loss. As F_0 increases cross over time increases but the divergence between the two models get smaller (Fig. 13(b)), again showing the concurrence of the two model for the very obese subjects.

5.4. *Comparing the models in equilibrium*

Equation (21) shows that the selection of fat versus nonfat does not change the equilibrium $\left(\frac{dW}{dt} = 0\right)$ values of W, for a given C:

$$W_\infty = \frac{\lambda - \beta_1}{2\beta_2}(Setpoint), \qquad W_\infty = C(HB). \tag{28}$$

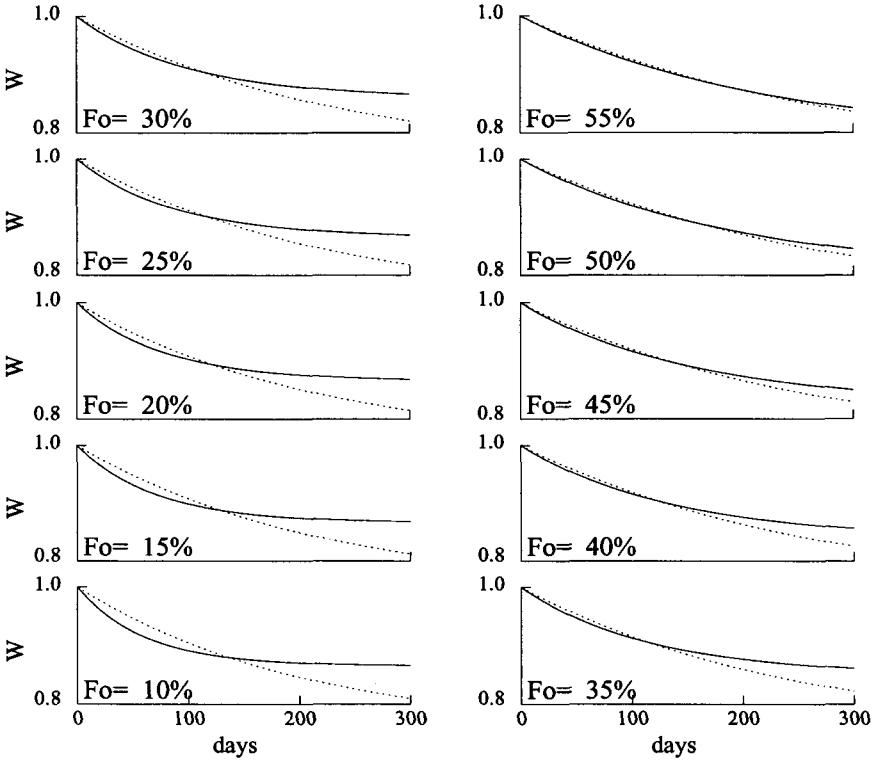


Fig. 12. Weight ratio for Harris–Benedict Model (dashed) with energy density of 3500 Kcal/Lb and Setpoint Model (solid) (adjusted for body composition) versus time for various initial fat ratios as indicated ($C = 0.75$) (reprinted from [2]).

At $\beta_2 = 1$, $W_\infty = \sqrt{C}$. For $\beta_2 = 0$ Eq. (28) reduces to $W_\infty = C$. Then $C \leq W_\infty \leq \sqrt{C}$ according to β_2. Figure (14) compares the final weight to the initial weight $\left(\frac{W^*_\infty}{W^*_0}\right)$, the final percentage body fat to the initial value $\left(\frac{F^*_\infty/W^*_\infty}{F^*_0/W^*_0}\right)$ and the final percentage of nonfat body mass to the initial value $\left(\frac{N^*_\infty/W^*_\infty}{N^*_0/W^*_0}\right)$ for a $C = 0.75$ diet. The setpoint model predicts a much higher final weight as the body defends the setpoint weight. Both models closely predict that the percentage of lean body mass will increase, this should be expected since $\Phi = \frac{dN^*}{dW^*} < 1$ (Fig. 6(a)). The setpoint models values of percent fat remain much higher, because of the higher body weight. Then the defense of setpoint weight is also a defense of setpoint body fat percentage.

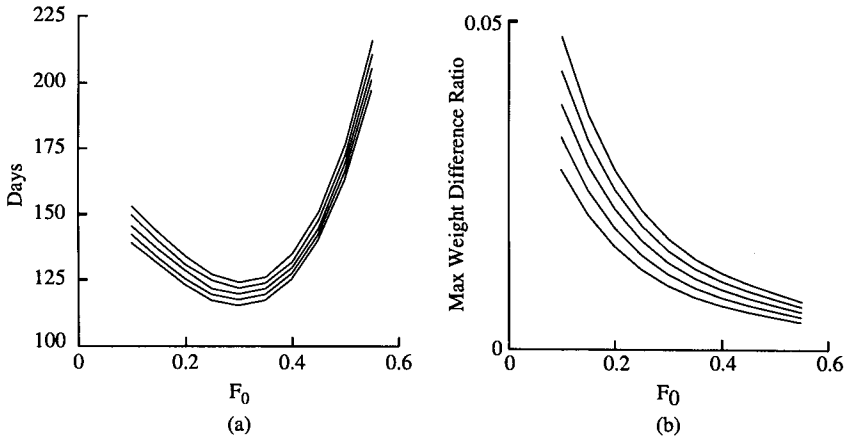


Fig. 13. (a) The time in days until the Harris–Benedict Model predicted weight is less than the body composition adjusted Setpoint predicted weight versus initial body fat ratio (F_0) and (b) The ratio of the maximum difference between the two models before cross over divided by the initial weight versus initial body fat ratio (F_0). (Both figures for 50%, 55%, 60%, 65% and 70% Calorie Diet (top to bottom)) (reprinted from [2]).

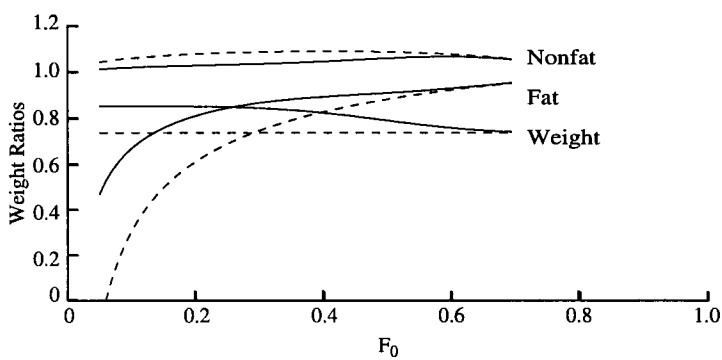


Fig. 14. Ratios of equilibrium values for 75% Calorie diet versus original (Setpoint) values: Lean — % percent lean body mass, Fat — % percent body fat mass and Weight — final weight over initial weight. Harris–Benedict Model (dashed) and Setpoint Model (solid) (reprinted from [2]).

6. Discussion and Conclusions

Models have been presented for calculating a metabolic reduction factor (β_2) and an energy density ratio (λ_Φ) which effect weight dynamics during a reduced calorie diet. These have both been modeled as dependent on the

initial (setpoint) percent body fat (F_0). Of course, conversion to dependence on percent nonfat mass is easily accomplished. A characteristic time (t_0^*) which also effects the rate of weight loss was defined and modeled as dependent on F_0. Expressing t_0^* in terms of F_0 simplified the analysis. However t_0^* can be calculated directly from the setpoint values E_0^* and W_0^*, if known. The models predict that the obese will loose weight and fat/nonfat more closely to the nominal/equilibrium models while the lean will experience significant departure. The difference of weight loss dynamics between lean and obese individuals is compactly displayed in Fig. 11.

The shaping of Fig. 3 is from nominal values found in the literature. Future efforts will be to further refine the β_2 form. The present model will also be expanded to metabolic adaption to weight gain (an increase in the relative energy needs).

References

[1] S. S. Alpert, Growth, thermogenesis, and hyperphagia, *Amer. J. Clin. Nutr.* **52** (1990) 784–792.

[2] F. P. Kozusko, The effects of body composition on setpoint based weight loss, *Math. Computer Modelling* (2001).

[3] E. Ravussin and B. A. Swinburn, Energy metabolism, in *Obesity: Theory and Therapy*, 2nd edn, eds. A. J. Stunkard and T. A. Walden (Raven Press, 1993), pp. 97–98.

[4] J. A. Harris and F. G. Benedict, *A Biometric Study of Basal Metabolism in Man* (Carnegie Institute of Washington, Washington, DC, 1919).

[5] W. A. Rowe, D. C. Frankenfield and E. R. Muth, The harris-bennedict studies of human basal metabolism: History and limitations, *J. Am. Diet. Assoc.* **98**(9) (1998) 970–971.

[6] M. Kleiber, *The Fire of Life: An Introduction to Animal Energetics* (Robert E. Kreiger Co., 1975).

[7] W. H. Van der Walt and C. H. Wyndham, An equation for prediction of energy expenditure of walking and running, *J. Appl. Phys.* **34** (1973) 559–563.

[8] R. L. Leibel, M. Rosenbaum and J. Hirsch, Changes in energy expenditure resulting from altered body weight, *N. Engl. Med.* **332** (1995) 621–628.

[9] G. D. Foster *et al.*, The energy cost of walking before and after significant weight loss, *Med. Sci. Sports Exerc.* **27**(6) (1995) 888–894.

[10] D. S. Weigle, K. J. Sande, P. H. Iverius, E. R. Monsen and J. D. Brunzell, Weight loss leads to a marked decrease in nonresting energy expenditure in ambulatory human subjects, *Metabolism* **37** (1988) 930–936.

[11] J. S. Stern, S. W. Corbett and R. E. Kersey, Energy expenditure in rats with diet induced obesity, *Amer. J. Clin. Nutr.* **44** (1986) 173–180.

[12] R. L. Leibel and J. Hirsch, Diminished energy requirements in reduced obesepatients, *Metabolism* **33** (1984) 164–170.

[13] J. O. der Boer *et al.*, Adaption of energy metabolism of overweight women to low-energy intake, studied with whole-body calorimeters, *Amer. J. Clin. Nutr.* **44** (1986) 5485–5495.

[14] E. S. Horton, Introduction: An overview of the assessment and regulation of energy balance in humans, *Amer. J. Clin. Nutr.* **38** (1983) 972–977.

[15] R. L. Leibel, Is obesity due to a heritable difference in "set point" for adiposity? *West J. Med.* **153** (1990) 429–431.

[16] D. M. Garner and P. E. Garfinkel, *Handbook of Treatments for Eating Disorders* (Guilford Press, 1997).

[17] R. L. Leibel, J. Hirsch, L. C. Hudgins and M. Rosenbaum, Diet composition and energy balance in humans, *Amer. J. Clin. Nutr.* **67** (1998) 551S–555S.

[18] A. Luke and D. A. Schoeller, Basal metabollic rate, fat-free mass, and body cellmass during energy restriction, *Metabolism* **41**(4) (1992) 450–456.

[19] A. C. Segal, A linear diet model, *The College Math. J.* **18** (1987) 44–45.

[20] S. S. Alpert, The energy density of weight loss in semistarvation, *Int. J. Obesity* **12** (1988) 533–542.

[21] S. S. Alpert, A two-reservoir energy model of the human body, *Amer. J. Clin. Nutr.* **32** (1979) 1710–1718.

[22] V. W. Antonetti, The equations governing weight change in human beings, *Amer. J. Clin. Nutr.* **26** (1973) 64–71.

[23] F. P. Kozusko, A setpoint based dieting model, *Math Computer Modelling* **29** (1999) 1–7.

[24] D. E. Matthews, Proteins and amino acids, in *Modern Nutrition in Health and Disease*, 9th edn., eds. J. A. O. Maurice, E. Shils and A. C. Ross (Lippincott Williams and Wilkins, 1999), pp. 11–48.

[25] S. B. Heymsfield, R. N. Baumgartner and S.-F. Pan, Nutritional assessment of malnutrition by anthropometric measures, in *Modern Nutrition in Health and Disease*, 9th edn., eds. J. A. O. Maurice, E. Shils and A. C. Ross (Lippincott, Williams and Wilkins, 1999), pp. 903–921.

[26] S. N. Kreitzman, Factors influencing body composition during very-low-calorie diets, *Amer. J. Clin. Nutr.* **56** (1992) 217S–223S.

[27] G. B. Forbes, Body fat content influences the body composition response to nutrition and exercise, *Ann. N Y Acad. Sci.* **904** (2000) 359–365.

[28] G. B. Forbes, *Human Body Composition: Growing, Aging, Nutrition and Activity* (Springer-Verlag, 1987).

[29] A. Keys, *The Biology of Human Starvation* (University of Minn Press, 1950).

[30] F. P. Kozusko, Body weight setpoint, metabolic adaption and human starvation, *Bull. Math. Biology* **63**(2) (2001) 393–404.

CHAPTER 15

MATHEMATICAL MODELS IN POPULATION DYNAMICS AND ECOLOGY

RUI DILÃO

Non-Linear Dynamics Group, Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisbon, Portugal

and

Institut des Hautes Études Scientifiques
Le Bois-Marie, 35, route de Chartres, F-91440 Bures-sur-Yvette, France
rui@sd.ist.utl.pt

We introduce the most common quantitative approaches to population dynamics and ecology, emphasizing the different theoretical foundations and assumptions. These populations can be aggregates of cells, simple unicellular organisms, plants or animals.

The basic types of biological interactions are analysed: consumer-resource, prey-predation, competition and mutualism. Some of the modern developments associated with the concepts of chaos, quasi-periodicity, and structural stability are discussed. To describe short- and long-range population dispersal, the integral equation approach is derived, and some of its consequences are analysed. We derive the standard McKendrick age-structured density dependent model, and a particular solution of the McKendrick equation is obtained by elementary methods. The existence of demography growth cycles is discussed, and the differences between mitotic and sexual reproduction types are analysed.

1. Introduction

In a region or territory, the number of individuals of a species or a community of species changes along the time. This variation is due to the mechanisms of reproduction and to the physiology of individuals, to the resources supplied by the environment and to the interactions or absence of interactions between individuals of the same or of different species.

Biology is concerned with the architecture of living organisms, its physiology and the mechanisms that originated life from the natural elements. Ecology studies the relations between living organisms and the environment, and, in a first approach, detailed physiological mechanisms of individuals have a secondary importance.

As a whole, understanding the phenomena of life and the interplay between living systems and the environment make biological sciences, biology together with ecology, a complex science. To handle the difficulties inherently associated to the study of living systems, the input of chemistry, physics and mathematics is fundamental for the development of an integrative view of life phenomena.

In the quantitative description of the growth of a population, several interactions are involved. There are intrinsic interactions between each organism and its environment and biotic interactions between individuals of the same or of different species. These interactions have specific characteristic times or time scales, and affect the growth and fate of a species or a community of species. Population dynamics deals with the population growth within a short time scale, where evolutionary changes and mutations do not affect significantly the growth of the population, and the population is physiologically stable. In this short time scale, the variation of the number of individuals in the population is determined by reproduction and death rates, food supply, climate changes and biotic interactions, like predation, competition, mutualism, parasitism, disease and social context.

There exists an intrinsic difficulty in analysing the factors influencing the growth and death of a species. There are species that are in the middle of a trophic web, being simultaneously preys and predators, and the trophic web exhibits a large number of interactions. For example, the food web of Little Rock Lake, Wisconsin, shows thousands of inter-specific connections between the top levels predators down to the phytoplankton, [1]. In this context, organisms in batch cultures and the human population are the simplest populations. In batch cultures, organisms interact with their resources for reproduction and growth. The human population is at the top of a trophic chain. Even for each of these simple cases, we can have different modelling approaches and strategies.

From the observational point of view, one of the best-known populations is the human population for which we have more than 50 years of relatively accurate census, and some estimates of population numbers over larger time intervals. Observations of population growth of micro-organisms in batch cultures are important to validate models and to test growth projections based on mathematical models.

Mathematical models give an important contribution to ecological studies. They propose quantities that can be measured, define concepts enabling to quantify biological interactions, and even propose different modelling strategies with different assumptions to describe particular features of the populations.

In population dynamics, and from the mathematical point of view, there are essentially two major modelling strategies: (i) The continuous time approach using techniques of ordinary differential equations; (ii) The discrete time approach which is more closely related with the structure of the census of a population. Both approaches use extensively techniques of the qualitative theory of dynamical systems.

In the continuous time approach, the number of individuals of a population varies continuously in time and the most common modelling framework applies to the description of the types of biotic inter-specific interactions and to the interactions of one species with the environment. They are useful for the determination of the fate of a single population or of a small number of interacting species. These models have been pioneered by Pierre-Françoise Verhulst, in the 19th century, with the introduction of the logistic model, and by Vito Volterra, in the first quarter of the 20th century, with the introduction of a model to describe qualitatively the cycling behavior of communities of carnivore and herbivore fishes.

In the discrete time approach, models are built in order to describe the census data of populations. They are discontinuous in time, and are closer to the way population growth data are obtained. These models are useful for short time prediction, and their parameters can be easily estimated from census data.

Modern ecology relies strongly on the concepts of carrying capacity (of the environment) and growth rate of a population, introduced by the discrete and the continuous models. In the 20th century, the works of McKendrick and Leslie gave an important contribution to modern ecology and demography.

The usefulness of population dynamics to predictability and resource management depends on the underlying assumptions of the theoretical models. Our goal here is to introduce in a single text the most common quantitative approaches to population dynamics, emphasizing the different theoretical foundations and assumptions.

In the next two sections, we introduce the continuous and the discrete age-structured approach to quantitative ecology. These are essentially two review sections, where we emphasise on the assumptions made in the derivation of the models, and whenever possible, we present case studies taken from real data. As the reader has not necessarily a background on the techniques of the qualitative theory of dynamical systems, we introduce some of its geometric tools and the concept of structural stability. In modelling situations where there exists some arbitrariness, structural stability is a useful tool to infer about the qualitative aspects of the solutions

of ordinary differential equations upon small variations of its functional forms.

In Sec. 4, and in the sequence of the Leslie type age-structured discrete models (Sec. 3), we make the mathematical analysis of the Portuguese population based on the census data for the second half of the 20th century. Here we introduce a very simple model in order to interpret data and make demographic projections, to analyse migrations and the change of socio-economic factors. This is a very simple example that shows the importance of mathematical modelling and analysis in population studies. In Sec. 5 we introduce discrete time models with population dependent growth rates, and we analyse the phenomenon of chaos. In Sec. 6, the consumer-resources interaction is introduced, and we discuss the two types of randomness found in dynamical systems: quasi-periodicity and chaos. In Sec. 7, we derive a general approach to the study of population dispersal (short- and long-range), and we derive a simple integro-difference equation to analyse the dispersion of a population.

In Sec. 8, we introduce the standard continuous model for age-structured density dependent populations, the McKendrick model, showing the existence of time periodic solutions by elementary techniques. In this context, we discuss demography cycles and the concept of growth rate. In Sec. 9, we derive a modified McKendrick model for populations with mitotic type reproduction, and compare the growth rates between populations with sexual and mitotic reproduction types. In the final section, we resume the main conclusions derived along the text and we compare the different properties of the analysed models.

2. Biotic Interactions

2.1. *One species interaction with the environment*

We consider a population of a single species in a territory with a well-defined boundary. Let $x(t)$ be the number of individuals at some time t. The growth rate of the population (by individual) is,

$$\frac{1}{x}\frac{dx}{dt} = r. \tag{1}$$

If the growth rate r is a constant, independently of the number of individuals of the population, then Eq. (1) has the exponential solution $x(t) = x(t_0)e^{r(t-t_0)}$, where $x(t_0)$ is the number of individuals in the population at time t_0. If, $r > 0$, $x(t) \to \infty$, as $t \to \infty$. In a realistic situation, such a population will exhaust resources and will die out in finite time.

Equation (1) with r constant is the Malthusian law of population growth. Exponential growth is in general observed in batch cultures of micro-organisms with a large amount of available resources and fast reproduction times, [2] and [3]. From the solution of Eq. (1) follows that the doubling time of the initial population (t_d) is related with the growth rate by $t_d = \ln 2/r$. For example, with the data of the world population, [4], we can determine the variation of the doubling time or the growth rate of the human population along historical times, Fig. 1. The curve in Fig. 1 suggests that, for human populations and at a large time scale, the growth rate r cannot be taken constant as in the Malthusian growth law (1), but must depend on other factors, as, for example, large-scale diseases, migrations, *etc.*

For large population densities, and in order to avoid unrealistic situations of exponential growth or explosion of population numbers, it is expected that the growth rate becomes population dependent. Assuming that, for large population numbers, $r \equiv r(x(t)) < 0$, for $x > K$, and, $r \equiv r(x(t)) > 0$, for $x < K$, where K is some arbitrary constant, the simplest form for the growth rate $r(x)$ is, $r(x) = r_0(K - x)$. Substitution of this population dependent growth rate into Eq. (1) gives,

$$\frac{dx}{dt} = r_0 x(K - x) := rx(1 - x/K) \tag{2}$$

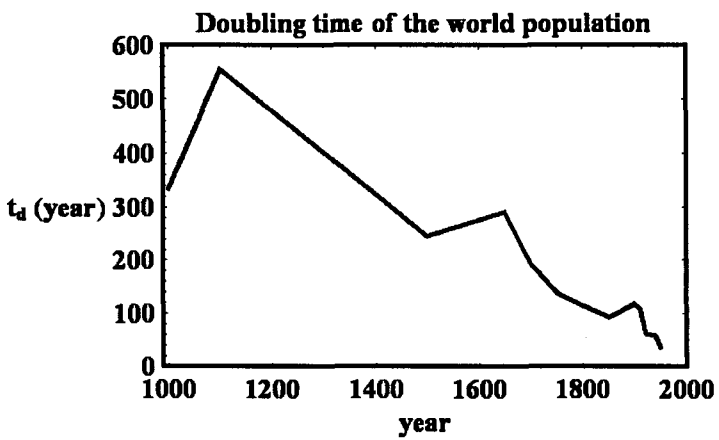


Fig. 1. Evolution of the doubling time of the world population. The doubling time has been calculated according to the formula $t_d = \ln 2/r$, where $r = (N(t + h) - N(t))/(hN(t))$, and $N(t)$ is the world population at year t. The data set is from reference [4].

where r_0 is a rate constant. Equation (2) is the logistic or Verhulst equation for one-species populations. For a population with $x(t_0) > 0$ at some time t_0, the general solution of (2) is,

$$x(t) = \frac{x(t_0)Ke^{r_0(t-t_0)}}{x(t_0)(e^{r_0(t-t_0)} - 1) + K} \tag{3}$$

and, in the limit $t \to \infty$, $x(t) \to K$. The constant K is called the *carrying capacity* of the environment and is defined as the maximum number of individuals of a species that the territory can support. For the same species, larger territories and bigger renewable resources correspond to larger values of K.

The logistic Eq. (2) describes qualitatively the growth of single colonies of micro-organisms in batch experiments, [3]. For example, in a batch experiment, Gause, [5], fitted the measured growth curve of the protozoa *Paramecium caudatum*, finding a good agreement with the solution (3) of the logistic equation. For the human population, the agreement is not so good, being dependent on technological developments, sociological trends and other factors, [2]. Depending on the data set, and from country to country, some authors find a good fit between the solutions of the logistic equation and demography data (see for example [4]), and others propose empirical models based on the delayed logistic equation, $x'(t) = rx^{\alpha}(1 - x(t-T)/K)$, [4].

In the derivation of the logistic equation, the plausibility of the mathematical form of the growth rate is assumed, without any assumptions about the relations between population growth and environmental support, or about the mechanisms of interaction between individuals and the environment. It is simply supposed that, for each species, the environment ensures enough resources. The carrying capacity constant can only be measured *a posteriori* through the asymptotic solution, $x(t) \to K$, as $t \to \infty$.

A possible mechanism for the derivation of the logistic equation is based on the mass action law of chemical kinetics, [6] and [7, pp. 295–300]. To be more specific, we represent species and resources by, A_j, with $j = 1, \ldots, m$. The interactions between species or between species and resources can be represented by n collision diagrams,

$$\nu_{i1}A_1 + \cdots + \nu_{im}A_m \xrightarrow{r_i} \mu_{i1}A_1 + \cdots + \mu_{im}A_m, \quad i = 1, \ldots, n \tag{4}$$

where r_i measure the rate at which the interactions occur, and the constants ν_{ij} and μ_{ij} are positive parameters measuring the number of individuals or units of resources that are consumed or produced in a collision. The mass

action law asserts that the time evolution of the (mean) concentration of A_j is given by,

$$\frac{dA_j}{dt} = \sum_{i=1}^{n} r_i(\mu_{ij} - \nu_{ij}) A_1^{\nu_{i1}} \cdots A_m^{\nu_{im}}, \quad j = 1, \ldots, m. \tag{5}$$

As we have in general n interaction diagrams and m species or resources, the system of Eq. (5) are not independent. In general, by simple inspection of the m Eq. (5), it is possible to derive the associated conservation laws, that is, a set of linear relations between the concentrations A_j. With these conservations laws, we obtain a system of $s \leq m$ linearly independent differential equations.

In this framework, reproduction in the presence of resources can be seen as the collision of the members of a population with the resources. In the case of the logistic equation, the collisions between individuals and the resource is represented by the diagram,

$$A + x \xrightarrow{r_0} (1 + e)x \tag{6}$$

where A represents resources, x is the number of individuals in the population, collisions occur at the rate r_0, and the inequality $e > 0$ expresses the increase in the number of individuals. By (4) and (5), to the diagram (6) is associated the logistic equation (2), together with the conservation law $x(t) + eA(t) = x(t_0) + eA(t_0)$, where the carrying capacity is given by $K = x(t_0) + eA(t_0)$. As, in the limit $t \to \infty$, $x(t) \to K$, then, in the same limit, $A(t) \to 0$. In this interpretative framework, when the population attains the equilibrium value K, resources are exhausted. In realistic situations, after reaching the equilibrium, the number of individuals of the population decreases and, some time afterwards, the population disappears, [3]. However, this asymptotic behavior is not predicted by Eq. (2). To further include this effect, we can add to the collision mechanism (6) a new death rate diagram, $x \xrightarrow{d}$. In this case, by the mass action law, (4) and (5), the time variation of the number of individuals of the population is not of logistic type anymore, obeying to the equations,

$$\begin{cases} x'(t) = r_0 e x A - dx \\ A'(t) = -r_0 x A \end{cases} \tag{7}$$

without any conservation law and, consequently, without a carrying capacity parameter. Numerical integration of the system of Eq. (7) leads to the conclusion that, for a small initial population, a fast exponential growing phase is followed by a decrease in the number of individuals of the

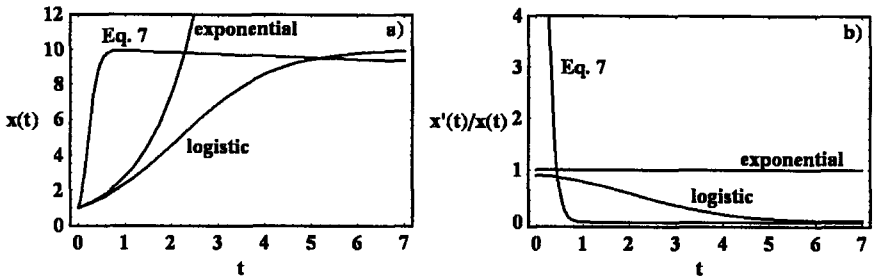


Fig. 2. (a) Comparison between the solutions of the exponential (1), logistic (2) and equation (7), for the initial conditions $x(0) = 1$, $A(0) = 9$, and the parameters $r = 1$, $r_0 = 1$, $K = 10$, $d = 0.01$ and $e = 1$. (b) Growth rates as a function of time for Eq. (1), (2) and (7).

population, and extinction occurs when $t \to \infty$, Fig. 2(a). The growth behavior predicted by Eq. (7) is in qualitative agreement with the growth curves observed in generic microbiological batch experiments, [3].

In Fig. 2(a), we compare the solutions of the three Eq. (1), (2) and (7) for the growth of one-species. In the growing phase, the solutions of the three growth models show qualitatively the same type of exponential behavior. For Eq. (7), the concept of carrying capacity is lost but the growth maximum is approximated by the value of the carrying capacity of the logistic equation. In these models, and for the same data set, it is possible to obtain different values for the fitted growth rates, as it is clearly seen in Fig. 2(b).

The approach developed so far introduces into the language of population dynamics the concepts of exponential or Malthusian growth, growth rate, doubling time and carrying capacity. The agreement between the models and data from laboratory experiments is, in some situations, very good, but in others deviates from observations. In the situations where no agreement with observations is found, it is believed that other relevant factors besides reproduction are not included in the modelling process. In modern ecology, the modelling concepts introduced here enable a rough estimate of the population growth and are the starting point for more specific and specialized approaches. For a more extensive study and applications of the exponential and logistic models see [8–10].

2.2. *Two interacting species*

Here, we introduce the basic models for the different types of biotic interactions between the populations of two different species. As models become

non-linear, and no general methods for the determination of solutions of non-linear differential equations exist, in parallel, we introduce some of the techniques of the qualitative theory of differential equations (dynamical systems theory).

We consider two interacting species in the same territory, and we denote by $x(t)$ and $y(t)$, their total population numbers at time t. The growth rates by individual of both interacting species are,

$$\begin{cases} \dfrac{1}{x}\dfrac{dx}{dt} = f(x,y) \\ \dfrac{1}{y}\dfrac{dy}{dt} = g(x,y) \end{cases}$$

defining the two-dimensional system of differential equations, or vector field,

$$\begin{cases} \dfrac{dx}{dt} = xf(x,y) \\ \dfrac{dy}{dt} = yg(x,y). \end{cases} \tag{8}$$

The particular form of the system of Eq. (8) ensures that the coordinate axes of the (x,y) phase space are invariant for the flow defined by the vector field (8), in the sense that, any initial condition within any one of the coordinate axis is transported by the phase flow along that axis. Due to this particular invariant property, in the literature of ecology, Eq. (8) are said to have the Kolmogoroff form, [8, p. 62].

In general, the system of differential Eq. (8) is non-linear and there are no general methods to integrate it explicitly. We can overcome this problem by looking at Eq. (8) as defining a flow or vector field in the first quadrant of the two-dimensional phase space ($x \geq 0, y \geq 0$). Adopting this point of view, the flow lines are the images of the solutions of the differential equation in the phase space, Fig. 3. At each point in phase space, the flow lines have a tangent vector whose coordinates are $xf(x,y)$ and $yg(x,y)$, and the flow lines can be visualised through the graph of the vector field. In fact, given a set of points in phase space, we can calculate the x- and y-coordinates of the vector field components, $xf(x,y)$ and $yg(x,y)$, and draw the directions of the tangent vectors to the flow lines. The solutions of the differential Eq. (8) are tangent to the vector field.

The phase space points for which we have simultaneously, $xf(x,y) = 0$ and $yg(x,y) = 0$, are the fixed points of the flow. The fixed points are stationary solutions of the ordinary differential Eq. (8). In dimension two,

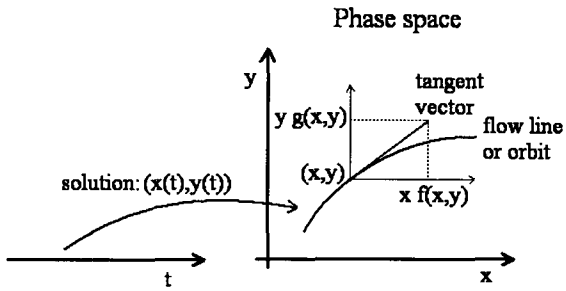


Fig. 3. The differential Eq. (8) defines a vector field or phase flow in the two-dimensional phase space. The flow lines are the images in phase space of the solutions of the differential equation. The flow lines are parameterised by the time t. At each point (x, y) in phase space, the tangent vector to the flow line or orbit has local coordinates $xf(x, y)$ and $yg(x, y)$.

the knowledge of these stationary solutions determines the overall topology of the flow lines in phase space. With the additional knowledge of the two nullclines, defined by equations $xf(x, y) = 0$ and $yg(x, y) = 0$, we can qualitatively draw in phase space the flow lines of the differential equation and to determine the asymptotic states of the dynamics, which, in generic cases, are isolated fixed points.

The (isolated) fixed points of a differential equation can be (Lyapunov-) stable or unstable. They are stable if, for any initial condition sufficiently close to the fixed point, and for each $t > 0$, the solution of the equation remain at a finite distance from the fixed point. If in addition, in the limit $t \to \infty$, the solution converges to the fixed point, we say that the fixed point is asymptotically stable. A fixed point is unstable if it is not stable.

Around a fixed point, the stability properties of the solutions of a differential equation can be easily analysed. Let (x^*, y^*) be a fixed point of Eq. (8), and let $(x(t) = x^* + \bar{x}(t),\ y(t) = y^* + \bar{y}(t))$ be a solution defined locally around (x^*, y^*). Introducing this solution into (8), we obtain, up to the first order in $\bar{x}$ and $\bar{y}$,

$$\begin{pmatrix} \frac{d\bar{x}}{dt} \\ \frac{d\bar{y}}{dt} \end{pmatrix} = \begin{pmatrix} f(x^*, y^*) + x^* \frac{\partial f}{\partial x}(x^*, y^*) & x^* \frac{\partial f}{\partial y}(x^*, y^*) \\ y^* \frac{\partial g}{\partial x}(x^*, y^*) & g(x^*, y^*) + y^* \frac{\partial g}{\partial y}(x^*, y^*) \end{pmatrix} \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$$

$$=: DF \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \tag{9}$$

where DF is the Jacobian matrix of the vector field (8) evaluated at (x^*, y^*). In the conditions of the theorem below, the solutions of the linear differential Eq. (9) are equivalent to the solutions of the nonlinear Eq. (8) near (x^*, y^*).

Theorem 1 (Hartman-Grobman, [11]). *If none of the eigenvalues of the Jacobian matrix DF rest on the imaginary axis of the complex plane, then, near the fixed point (x^*, y^*), the phase flows of Eqs.* (8) *and* (9) *are similar or topologically equivalent.*

Under the conditions of the Hartman–Grobman theorem (Theorem 1), by a simple linear analysis, it is possible to determine the stability of the fixed points of the non-linear Eq. (8), and, therefore, to determine the asymptotic behavior of the solutions of the non-linear Eq. (8). The global flow in the first quadrant of phase space is conditioned by the fixed points with non-negative coordinates. This approach is geometrically intuitive and is one of the most powerful tools of the theory of dynamical systems, [11] and [12]. As will see now, this enables the analysis of models for biotic interactions with a minimum of technicalities.

We now introduce the most common types of two-species interactions. There are essentially three basic two-species interactions: prey-predator, competition and mutualism. In the prey-predator interaction, for large predator numbers, the growth rate of the prey becomes negative, but in the absence of predators, the growth rate of the prey is positive. If the prey is not the only resource for predators, the growth rate of the predators is always positive. In competition, and in the presence of both species, both growth rates decrease. In mutualistic interactions, the growth rates of both species increase.

Adopting the same empirical formalism as in the case of the logistic equation (Sec. 2.1), we assume that the growth rates f and g are sufficiently well behaved functions, and the above ecology definitions can be stated into the mathematical form:

$$\begin{aligned} &\text{Prey-predator:} \quad \frac{\partial f}{\partial y} < 0, \quad \frac{\partial g}{\partial x} > 0 \\ &\text{Competition:} \quad \frac{\partial f}{\partial y} < 0, \quad \frac{\partial g}{\partial x} < 0 \\ &\text{Mutualism :} \quad \frac{\partial f}{\partial y} > 0, \quad \frac{\partial g}{\partial x} > 0. \end{aligned} \tag{10}$$

In the simplest situation where f and g are affine functions, $f(x, y) = d_1 + d_2 x + d_3 y$ and $g(x, y) = d_4 + d_5 x + d_6 y$, and further assuming that

in the absence of one of the species the growth of the other species is of logistic type, by (10), we obtain for the growth rates,

$$\begin{array}{lll} \text{Prey-predator:} & f = r_x(1 - x/K_x - c_1 y) \quad \text{and} & g = r_y(1 + c_2 x - y/K_y) \\ \text{Competition:} & f = r_x(1 - x/K_x - c_1 y) \quad \text{and} & g = r_y(1 - c_2 x - y/K_y) \\ \text{Mutualism:} & f = r_x(1 - x/K_x + c_1 y) \quad \text{and} & g = r_y(1 + c_2 x - y/K_y) \end{array} \tag{11}$$

where c_1, c_2, K_x and K_y are positive constants. The constants in the growth rate functional forms (11) have been chosen in such a way that, in the absence of any one of the species, we obtain the logistic equation (2). Introducing (11) into (8), we obtain three systems of non-linear ordinary differential equations for prey-predation, competition and mutualism. The topological structure in phase space of the solutions of these equations can be easily analysed by the qualitative methods just described above.

The generic differential Eq. (8) defines a flow in the first quadrant of the two-dimensional phase space, and the simplest solutions are the fixed points of the flow. These fixed points are obtained by solving simultaneously the equations, $xf(x,y) = 0$ and $yg(x,y) = 0$. For any of the values of the parameters in (11), and for the three biotic interaction types, we have the fixed points $(x_0, y_0) = (0,0)$, $(x_1, y_1) = (K_x, 0)$ and $(x_2, y_2) = (0, K_y)$, which correspond to the absence of one or both species. The fixed points $(K_x, 0)$ and $(0, K_y)$ are the asymptotic solutions associated to any non-zero initial condition on the phase space axis x and y, respectively. The zero fixed point corresponds to the absence of both species. For a particular choice of the parameters, a forth fixed point can exists:

$$\begin{aligned} \text{Prey-predator: } (x_3, y_3) &= \left(K_x \frac{1 - c_1 K_y}{1 + c_1 c_2 K_x K_y}, K_y \frac{1 + c_2 K_x}{1 + c_1 c_2 K_x K_y} \right) \\ &\qquad \text{if } c_1 K_y < 1 \\ \text{Competition: } (x_3, y_3) &= \left(K_x \frac{c_1 K_y - 1}{c_1 c_2 K_x K_y - 1}, K_y \frac{c_2 K_x - 1}{c_1 c_2 K_x K_y - 1} \right) \\ &\qquad \text{if } c_1 K_y > 1,\ c_2 K_x > 1 \\ \text{Mutualism: } (x_3, y_3) &= \left(K_x \frac{1 + c_1 K_y}{1 - c_1 c_2 K_x K_y}, K_y \frac{1 + c_2 K_x}{1 - c_1 c_2 K_x K_y} \right) \\ &\qquad \text{if } c_1 c_2 K_x K_y < 1. \end{aligned} \tag{12}$$

In Fig. 4, we show, for the differential Eq. (8) and the three growth rate functions (11), all the qualitative structures of the flows in phase space. The fixed points with non-zero coordinates (12) correspond to cases (a)–(c), and are marked with a square.

To determine qualitatively the structure of the solutions of Eq. (8) for the different cases depicted in Fig. 4, we have analysed the signs of the components of the vector field along the nullclines. The arrows in Fig. 4 represent the directions of the flow in phase space. Except the case of Fig. 4(a), the vector field directs the flow towards the fixed points, and the limiting behavior of the solutions as $t \to \infty$ is easily derived.

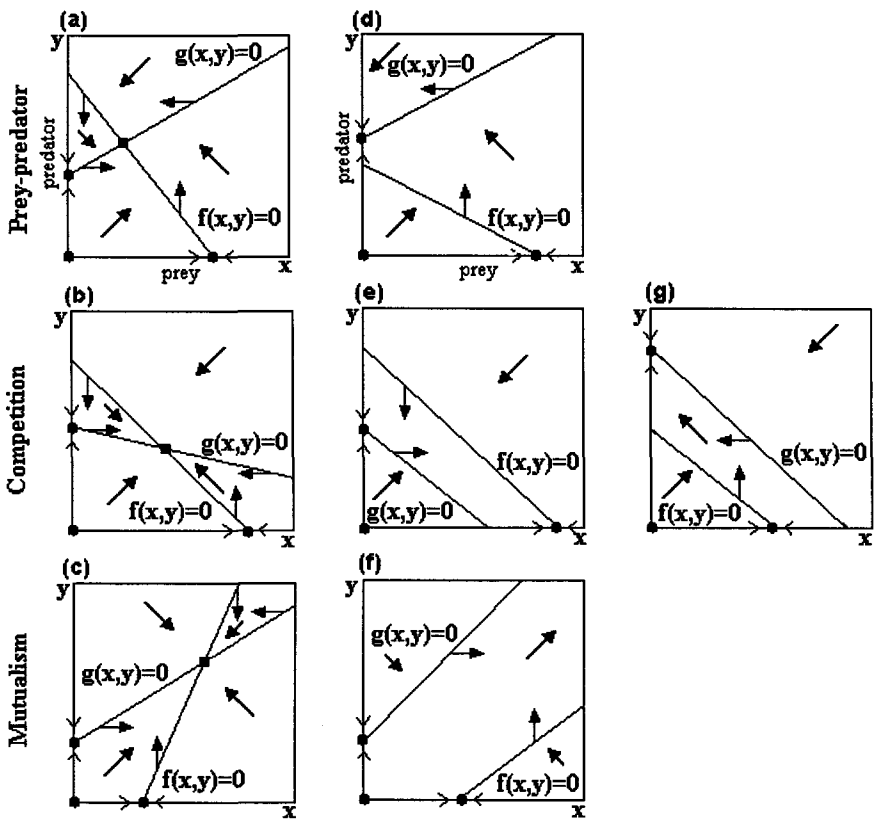


Fig. 4. Qualitative structures of the flow in phase space of the differential equation (8), for the growth rate functions (11). Bullets and squares represent fixed points. In cases (a)–(c), a non-zero fixed point exists if the conditions in (12) are verified. Cases (d)–(g) correspond to different arrangements of nullclines. The arrows represent the directions of the vector fields, and the solutions of Eq. (8) are tangent to the vector field. The sign of the vector field is calculated from the sign of the functions f and g at each point in phase space.

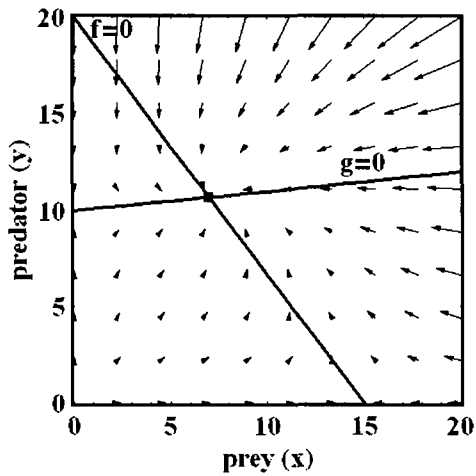


Fig. 5. Vector field and nullclines for the prey-predator equation of Fig. 4(a), with parameter values $c_1 = 0.05$, $c_2 = 0.01$, $K_x = 15$, $K_y = 10$ and $r_x = r_y = 1$. As it is clearly seen, the vector field directs the flow to the fixed point (x_3, y_3). This fixed point is asymptotically stable.

To analyse the prey-predator case of Fig. 4(a), we have calculated the directions of the vector field near the fixed point (x_3, y_3), Fig. 5. In this case, the flow turns around the fixed point (x_3, y_3), and to determine the local structure of the flow, we use the technique provided by the Hartman–Grobman theorem. Linearising Eq. (8) around (x_3, y_3), by (11) and (12), we obtain the linear system of differential equations,

$$\begin{pmatrix} \frac{d\bar{x}}{dt} \\ \frac{d\bar{y}}{dt} \end{pmatrix} = \begin{pmatrix} -r_x x_3/K_x & -r_x x_3 c_1 K_x \\ r_y c_2 y_3 & -r_y y_3/K_y \end{pmatrix} \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = DF \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \tag{13}$$

where $(x, y) = (x_3 + \bar{x}, y_3 + \bar{y})$. The stability near the fixed point (x_3, y_3) is determined by the eigenvalues of the matrix DF, provided that they are not on the imaginary axis of the complex plane. As, $Trace(DF) = \lambda_1 + \lambda_2 < 0$ and $Det(DF) = \lambda_1 \lambda_2 > 0$, the eigenvalues λ_1 and λ_2 of DF are both real and negative or, complex conjugate with negative real parts. As the solution of the linear system of Eq. (13) is a linear combination of terms of the form $e^{\lambda_i t}$, and the eigenvalues have negative real parts, this implies that $\bar{x}(t)$ and $\bar{y}(t)$ converge to zero as $t \to \infty$. Therefore, for non-zero initial

conditions, the solutions of the prey-predator system of Fig. 4(a) converge to the stable fixed point (x_3, y_3), Fig. 5.

In the prey-predator case of Fig. 4(d), $c_1K_y > 1$, the effect of the predator on the prey is so strong that asymptotically predators consume all the preys, and, as $t \to \infty$, the solutions converge for the fixed point $(x_2, y_2) = (0, K_y)$. In this case, we do not need to make the linear analysis near the fixed points because the directions of the vector field show clearly the convergence of the solutions to the asymptotically stable fixed point.

For the competitive and mutualistic interactions of Figs. 4(b) and 4(c), Eq. (8) has always a stable fixed point which is also an asymptotic solution for non-zero initial conditions. In cases (e) and (g), we have $c_1K_y < 1$ or $c_2K_x < 1$, and, asymptotically in time, only one of the species survives. For the mutualistic interaction (f), we have $c_1c_2K_xK_y > 1$, and, asymptotically in time, both population numbers explode to infinity. (Note however that, in this last case, it is possible that the solutions go to infinity in finite time due to the non-Lipschitz nature of the right-hand side of (8).)

From the models for the prey-predator, competition and mutualistic interactions, it is possible to derive some ecological consequences. In the prey-predator system, the prey brings advantage to the predator in the sense that its presence increases the number of predators at equilibrium, but the presence of predator decreases the equilibrium population of the prey. If the effect of the predator on the prey is too strong, predators consume all the preys, and, in the long time scale, predators lose advantage.

For competition, the asymptotic equilibrium between the two species assumes lower values for both species when compared to the cases where they are isolated.

In the mutualistic interaction, the situation is opposed to the competition case, where the equilibrium between the two species assumes higher values. However, for strong mutualistic interactions, we can have overcrowding as in the Malthusian model (1), leading to the death of the species by over consumption of resources. These conclusions, derived from the mathematical models (8) and (11), are in agreement with the biological knowledge about predation, competition and mutualism, [13] and [14].

Another model for the prey-predator interaction that has a conceptual and historical importance is the Lotka–Volterra model. This model has been used as an explanation to justify the resumption of carnivore fishes, after the cessation of fishing in the North Adriatic Sea after the First World

War, [8]. To be more specific, the prey-predator Lotka–Volterra interaction model is,

$$\begin{cases} \dfrac{dx}{dt} = r_x x(1 - c_1 y) \\ \dfrac{dy}{dt} = r_y y(c_2 x - 1) \end{cases} \tag{14}$$

where c_1, c_2, r_x and r_y are constants.

This model obeys the prey-predator conditions in (10), but assumes that predators have an intrinsic negative growth rate and do not survive without preys. For preys alone, it assumes that they have exponential growth as in model (1).

The Lotka–Voltrerra model (14) has one horizontal and one vertical nullcline in phase space, Fig. 6, and one non-zero fixed point with coordinates $(x, y) = (1/c_2, 1/c_1)$. One of the eigenvalues of the Jacobian matrix of (14) calculated at the fixed point is zero, and as the conditions on the Hartman–Grobman theorem fail: the local structure of the flow cannot be characterized by the linear analysis around the fixed point. It can be shown that, in the first quadrant of phase space, the solution orbits of (14) are closed curves around the fixed point, Fig. 6, corresponding to oscillatory

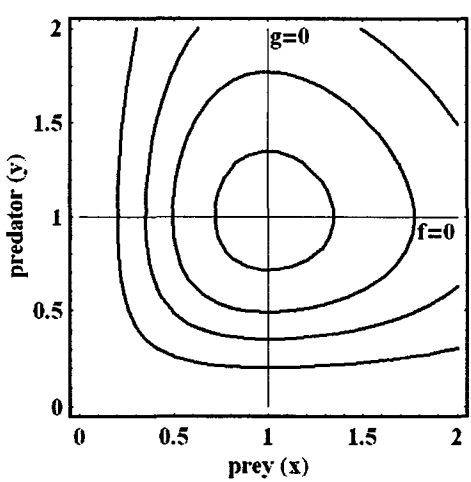


Fig. 6. Qualitative structure of the flow in phase space of the Lotka–Volterra system of Eqs. (14), for parameter values $c_1 = c_2 = r_x = r_y = 1$. Away from the non-zero fixed point, the solutions are periodic in time, suggesting a simple explanation for the oscillatory behavior observed in prey-predator real systems. It can be shown that the orbits of the system of Eqs. (14) are the level sets of the function, $H(x, y) = r_y \log x - r_y c_2 x + r_x \log y - r_x c_1 y$.

motion in the prey and predator time series (for a proof see [12]). Moreover, along each phase space cycle, the temporal means of prey and predators are independent of the amplitude of the cycles, being given by, $\langle x \rangle = 1/c_2$ and $\langle y \rangle = 1/c_1$, respectively. This property of the solutions of Eqs. (14) has been used to assert that fluctuations in fisheries are periodic but the time average during each cycle is conserved, [8, p. 93].

One of the important issues in the Lotka–Volterra model is to suggest the possibility of existence of time oscillations in prey-predator systems. A long-term observation of prey-predator oscillations was provided by the hare-lynx catches data during 90 years, from the Hudson Bay Company, [14] and [15]. The catches of lynx and hare are in principle proportional to the abundances of these animals in nature, and the time series shows an out of phase oscillatory abundance, with the lynx maximum preceding the hare maximum. Making a naïf analogy between the solutions of the Lotka–Volterra model and the oscillations found in the lynx-hare interaction, it turns out that the maximum number of preys is observed before the maximum numbers of predator. This is in clear disagreement with the Lotka–Volterra model where the prey maximum precedes in time the predator maximum, Fig. 6. Several attempts were made to explain this out of phase behavior but no consistent explanations have been found, [15].

One possible meaningful argument against the plausibility of the Lotka–Volterra model (14) to describe the prey-predator interaction is based on the property that any perturbation of the right-hand side of Eq. (14) destroys the periodic orbits in phase space. In mathematical terms, it means that the Lotka–Volterra system (14) is not structurally stable or robust. In general, a two-dimensional dynamical system is structurally stable if all its fixed points obey the conditions of the Hartman–Grobman theorem, and there are no phase space orbits connecting unstable fixed points (saddle points), [11] and [12]. The only types of structurally stable two-dimensional differential equations with periodic orbit in phase space are equations with isolated periodic orbits or limit cycles. In this case, the growth rate functions f and g must be at most quadratic, and several models with this property appeared in the literature, [12] and [15]. However, all these models show the same wrong out of phase effect as in the hare-lynx data.

In modern theoretical ecology, the development of more specialized models relies on the conditions (10) and on further assumptions on the functional behavior of the growth rate functions, [13], [12] and [14]. In some cases, the assumptions are introduced in analogy with some mechanisms derived from chemical kinetics, [6], [16] and [17]. For example,

the mechanisms,

$$\begin{aligned} &A + x \xrightarrow{r_1} (1 + e_1)x \\ &x + y \xrightarrow{r_2} (1 + e_2)y + cx \\ &x \xrightarrow{d_1} \\ &y \xrightarrow{d_2} \end{aligned} \tag{15}$$

with $e_1 > 0$, $e_2 > 0$, $r_1 > 0$, $r_2 > 0$ and $c < 1$, and,

$$\begin{aligned} &A + x \xrightarrow{r_1} (1 + e_1)x \\ &B + y \xrightarrow{r_2} (1 + e_2)y \\ &x + y \xrightarrow{r_3} c_1 x + c_2 y \\ &x \xrightarrow{d_1} \\ &y \xrightarrow{d_2} \end{aligned} \tag{16}$$

with $e_1 > 0$, $e_2 > 0$, $r_1 > 0$, $r_2 > 0$, $r_3 > 0$, $c_1 > 0$ and $c_2 > 0$, are examples of possible mechanisms for the prey-predator and generic biotic interactions. The phase space structure of the orbits of the Lotka–Volterra system (14) and the model (8)–(11) are different from the ones derived from the model equations associated to (15) and (16). However, the mechanistic interpretation of models (15) and (16) are closer to the biological situations. A detailed account of models for predation and parasitism is analysed in [18] and [19].

3. Discrete Models for Single Populations. Age-Structured Models

One important fact about the individuals of a species is the existence of age classes and life stages. Within each age class, the individuals of a species behave differently, have different types of dependencies on the environment, have different resource needs, *etc.* For example, in insets, three stages are generally identified: egg, larval and the adult. In mammals, in the childhood phase, reproduction is not possible, neither hunting nor predation.

To describe a population with age classes or stages, we can adopt a discrete formalism, where the transition between different age classes or stages is described in matrix form. One of the advantages of this type of models is that they can be naturally related with field data. One discrete model that accounts for age or stage classes has been proposed by Leslie in 1945, [20].

The Leslie model considers that, at time n, a population is described by a vector of population numbers, $(N^n)^T = (N_1^n, \ldots, N_m^n)$, where N_i^n is

the number of individuals with age class i (or in life stage i). The time transition between age classes is described by the map,

$$N^{n+1} = AN^n \tag{17}$$

where A is the Leslie time transition matrix. Under the hypothesis that from time n to time $n+1$, the individuals die out or change between consecutive age classes, the matrix A has the form,

$$A = \begin{pmatrix} 0 & e_2 & e_3 & \cdots & e_{k-1} & e_k \\ \alpha_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \alpha_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha_{k-1} & 0 \end{pmatrix} \tag{18}$$

where the e_i are fertility coefficients, and the α_i are the fraction of individuals that survive in the transition from age class $i-1$ to i. Clearly, $e_i \geq 0$ and $0 < \alpha_i \leq 1$. We consider always that $e_k > 0$, where e_k is the last reproductive age class. If $e_k = 0$, the determinant of matrix A is zero. Obviously, we can have populations with age classes such that $e_p = 0$ and $\alpha_p > 0$, for $p > k$. In this case, if $e_k > 0$, the solutions N_p^n, with $p > k$, are determined from the solutions obtained from the discrete difference Eq. (17) in dimension k. For example, if $e_{k+1} = 0$, then $N_{k+1}^n = \alpha_k N_k^{n-1}$. Therefore, without loss of generality, we always consider that $e_k > 0$ and $e_p = 0$, for $p > k$.

In Fig. 7(a), we show the distribution of age classes for the Portuguese population obtained from the census of 1991, 1992 and 1999, [21]. As it is

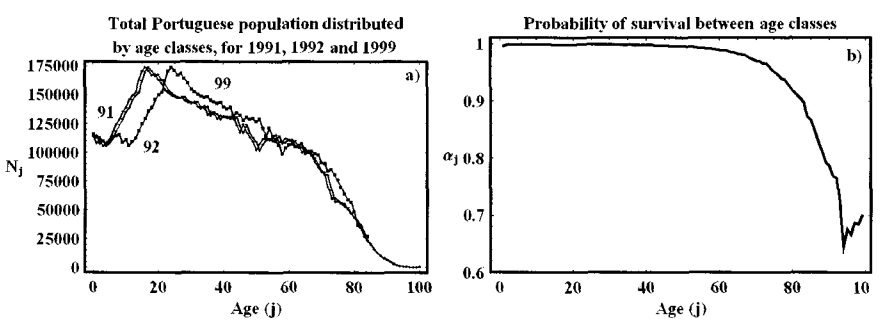


Fig. 7. (a) Total Portuguese population distributed by age classes for 1991, 1992 and 1999, [21]. (b) Probability of survival between age classes calculated with the population data of 1991 and 1992. The fraction of individuals that survive in the transition from age class $j-1$ to j is given by, $\alpha_j = N_{j+1}^{n+1}/N_j^n$.

clearly seen, the population in 1992 and 1999 is approximately obtained from the population in 1991 by a translation along the age axis, a property shared by the Leslie transition matrix (18). In Fig. 7(b), we show the values of the survival probability α_j as a function of the age classes, calculated from the census of 1991–92. For age classes $j < 45$, the values of α_j are close to 1.

The solution of the Leslie discrete map model (17)–(18) is easily determined. As the discrete Eq. (17) is linear, its general solution is, [22],

$$N_i^n = \sum\nolimits_{j=1}^{k} c_{ij} \lambda_j^n \tag{19}$$

where the c_{ij} are constants determined by the initial conditions and the coefficients of the matrix A, and the λ_j are the eigenvalues of (18). To simplify, we assume that the eigenvalues of A have multiplicity 1. As A is a non-negative matrix with non-zero determinant ($e_k > 0$), by the Frobenius–Perron theory, [23], its dominant eigenvalue λ is positive with multiplicity 1, implying the existence of a non-zero steady state if, and only if, $\lambda = 1$. If, $\lambda < 1$, the solutions (19) go to zero, as $n \to \infty$. If, $\lambda > 1$, the solutions (19) go to infinity.

Calculating the characteristic polynomial of A, we obtain (by induction in k),

$$P(\lambda) = (-1)^k \left(\lambda^k - \sum_{i=2}^{k} e_i \lambda^{k-i} \prod_{j=2}^{i} \alpha_{j-1} \right). \tag{20}$$

Imposing the condition that $\lambda = 1$ is a root of the polynomial (20), from the condition $P(1) = 0$, we define the constant,

$$G := \sum_{i=2}^{k} e_i \prod_{j=2}^{i} \alpha_{j-1} \tag{21}$$

where G is the *inherent net reproductive number of the population.* If we make the approximation, $\alpha_j \approx 1$, we have the *net fertility number* $\bar{G} := \sum_{i=2}^{k} e_i$.

The condition for the existence of asymptotic stable population numbers, given by the dominant eigenvalue of A, can be stated through the inherent net reproductive number of the population. So, in a population where $G < 1$, any initial condition leads to extinction. If, $G > 1$, we have unbounded growth. If, $G = 1$, in the limit $n \to \infty$, the population attains a stable age distribution.

The Leslie model is important to describe populations where there exists a complete knowledge of the life cycle of the species, including survival probabilities and fertilities by age classes. For example, in the Leslie paper [20], it is described a laboratory observation of the growth of *Rattus norvegicus*. For a period of 30 days, the projected total population number was overestimated with an error of 0.06% of the total population.

For human populations, survival probabilities are easily estimated from census data, Fig. 7. However, data from the fertility coefficients are difficult to estimate due to sex distinction and to the distribution of fertility across age classes. For an exhaustive account about Leslie type models, its modifications, and several case studies we refer to [23], [24] and [13]. For tables of the world population by country and the measured parameters of the Leslie matrix, we refer to [25].

Comparing the discrete and the continuous time approaches, the Leslie population growth model presents exactly the same type of unbounded growth as the exponential model (1). To overcome the exponential type of growth, we can adopt two different points of view. One approach is to introduce a dependence of the growth rates on the population numbers, as it has been done in Sec. 2, in the derivation of the logistic equation from the Malthusian growth equation. Another alternative is to introduce a limitation on the growth rates through the resource consumption of the population. These two types of development of the Leslie model lead to the introduction of the concepts of chaos and randomness and will be developed below.

4. A Case Study with a Simple Linear Discrete Model

Here, we introduce a simplified discrete linear model enabling to make projections about human population growth based on census data. With this simple model, we avoid the difficulty associated with the choice of the fertility coefficients by age classes, a characteristic of the Leslie model.

We characterize a population in a finite territory at time n by a two-dimensional vector $(B^n, N^n)^T$, where B^n represents the age class of newborns, individuals with less than 1 year, and N^n represents the total number of individuals with one or more years. By analogy with the Leslie approach of the previous section, the time evolution equations are now,

$$\begin{pmatrix} B^{n+1} \\ N^{n+1} \end{pmatrix} = \begin{pmatrix} 0 & e \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} B^n \\ N^n \end{pmatrix} \tag{22}$$

where e is the (mean) fertility coefficient of the population, α is the probability of survival of the total population between consecutive years, and β is the probability of survival of new-borns. In census data, the fertility coefficient is given in number of new-borns per thousand, but here we use the convention that the fertility coefficient is given in number of new-borns by individual.

Following the same approach as in the previous section, the solution of the discrete Eq. (22) is,

$$\begin{aligned} B^n &= c_1\lambda_1^n + c_2\lambda_2^n \\ N^n &= c_3\lambda_1^n + c_4\lambda_2^n \end{aligned} \tag{23}$$

where λ_1 and λ_2 are the eigenvalues of the matrix defined in (22), and the c_i are constants to be determined from the initial data taken at some initial census time n_0. If the dominant eigenvalue of the matrix in (22) is $\lambda = 1$, in the limit $n \to \infty$, the solution (23) converges to a non-zero constant solution, from any non-zero initial data. As the characteristic polynomial of the matrix in (22) is,

$$P(\lambda) = \lambda^2 - \alpha\lambda - e\beta$$

the condition of existence of a non-zero steady state is,

$$I = \alpha + e\beta = 1. \tag{24}$$

As in (21), we say that $I = \alpha + e\beta$ is the inherent net reproductive number of the population. If, $I > 1$, then $\lambda > 1$, and the solution (23) diverges to infinity as $n \to \infty$. If, $I < 1$, then the solution (23) goes to zero.

In order to calibrate the simple model (22), we take the census data for the Portuguese population in the period 1941–1999, Fig. 8.

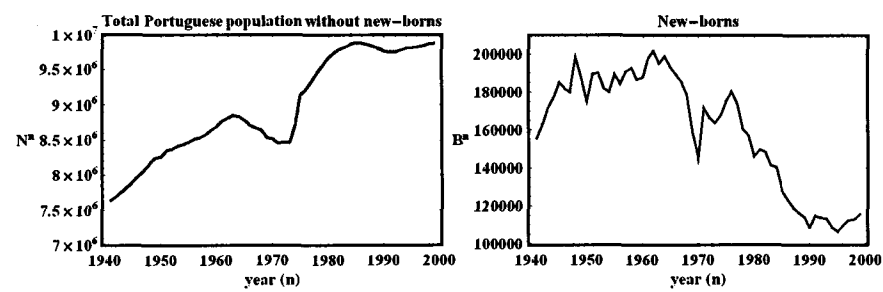


Fig. 8. Portuguese population and new-borns for the years 1941–1999, from reference [21].

As we see from Fig. 8, the total Portuguese population without new-borns shows strong variations, sometimes with a negative growth rate. This negative growth rate is due to emigration, decrease of population fertility and other social factors. The data for new-borns also shows negative growth rates. Therefore, the growth behavior shown in Fig. 8 is influenced by other factors that are necessary to quantify.

The values of the parameters α, β and e, are calculated from the census data and are shown in Fig. 9. The probability of survival of the population is approximately constant with mean $\alpha = 0.9891$, and a standard deviation of the order of 10^{-6}. The coefficient of fertility e and the new-borns survival probability β vary along the years. The last two coefficients are very sensitive to socio-economic and technological factors, suggesting that, for growth predictions, we must introduce into model (22) their time variation.

From the data of Fig. 9, the net reproductive number can be estimated. In Fig. 10, we show the variation of $I = \alpha + e\beta$ for the period 1960–1999. For 1960, we have $I = 1.01137$ and, for 1999, $I = 1.00074$, both very close to the steady state condition (24). Therefore, during this period of time, the Portuguese population is growing with a net reproductive number $I > 1$, but very close to 1. The decrease in the population number in the period 1960–1974 is essentially due to emigration.

To make population growth projections, we consider that α is constant, Fig. 9, with the mean value $\alpha = 0.9891$, and we consider that e and β are time varying functions. Due to the form of the curves in Fig. 9, the functions,

$$\begin{aligned} e(t) &= c_1 + \frac{c_2}{c_3 + (t - 1945)^{c_4}} \\ \beta(t) &= 1 - c_5 e^{-c_6(t-1960)} \end{aligned} \tag{25}$$

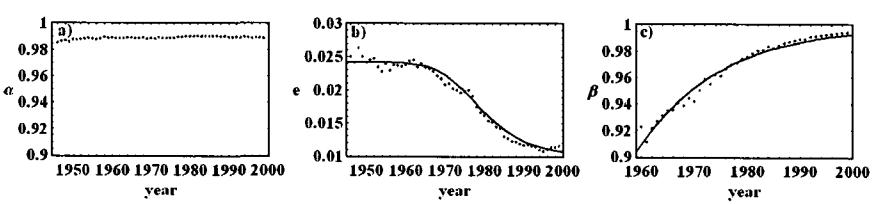


Fig. 9. (a) Probability of survival α, and (b) fertility coefficient e for the period 1945–1999. (c) Probability of survival of new-borns β in the period 1960–1999. The probabilities of survival α and β have been calculated with the death rate data by thousand habitants, [21]. In (b) and (c), we show the fitting functions (25), for the parameter values (26).

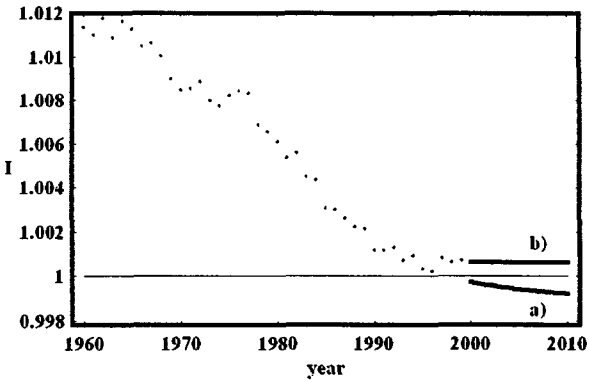


Fig. 10. Dots represent the inherent net reproductive number of the Portuguese population, calculated from the data of Fig. 9 (1960–1999). The two lines correspond to the two possible projections for the net reproductive number I, for the period 2000–2010. In estimate (a), we have considered that the time dependence of β and e is given by (25), for the parameter values (26). In estimate (b), we have taken constant values for β and e, obtained with the 1999 census values, [21].

are reasonable choices, with fitting constants,

$$\begin{aligned} &c_1 = 0.00979232, \quad c_2 = 5.65086 \times 10^6, \quad c_3 = 3.93899 \times 10^8, \\ &c_4 = 5.63644, \quad c_5 = 0.09058, \quad c_6 = 0.06355. \end{aligned} \tag{26}$$

In Fig. 9, we show the fitting functions (25), for the parameter values (26). In the limit $t \to \infty$, the new-borns survival probability converges to 1 and the mean fertility coefficient converges to $c_1 \approx 0.0097$, which corresponds roughly to 10 new-borns per thousand individuals in the population. The census value of e for 1999 corresponds to 11.6 new-borns per thousand.

To estimate the population growth for the period 2000–2010, we adopt two strategies for the iteration of map (22). In the first case, we iterate (22) with the time dependent functions (25), and we introduce as initial conditions the census data for 1999, Fig. 11. In the second case, we take for β and e the 1999 values. We also apply these two strategies to estimate the net reproductive number (24) as a function of time, Fig. 10.

With the simplified model (22), it is possible to make a short time projection of population numbers. However, for a good calibration and greater accuracy, emigration and immigration factors must be taken into account.

From the data and the fits in Figs. 10 and 11, we can derive several conclusions. The projections for the period 2000–2010 show two different growth behaviors: In case (a) of Fig. 10 and Fig. 11, we have,

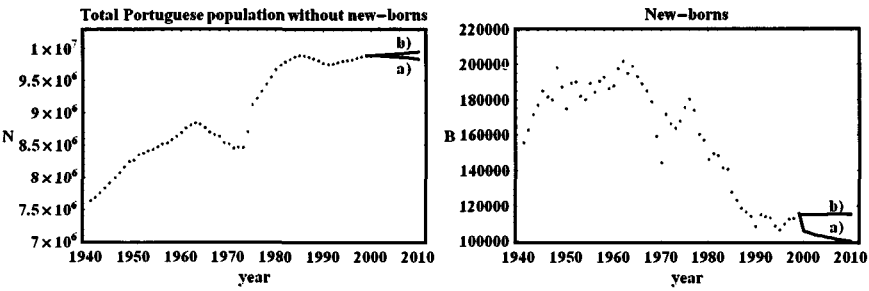


Fig. 11. Projections of the population numbers for the period 2000–2010, from the initial data of the year 1999. In estimate (a) we have considered that the time dependence of β and e is given by (25)–(26). In the estimate (b), we took β and e with the 1999 census values. These projections do not take into account emigration or immigration factors.

$B^{2010} = 99,966$ and $N^{2010} = 9{,}835{,}840$, with $B^{1999} = 115,440$ and $N^{1999} = 9{,}882{,}150$, implying a negative growth with an inherent net reproductive number $I < 1$. In case (b), we have, $B^{2010} = 115,251$ and $N^{2010} = 9{,}940{,}660$, corresponding to a positive growth of the population but with I close to the transition value $I = 1$.

Emigration and immigration are strongly dependent on several social factors. However, even in this simplified model, we can make an estimate of the balance between emigration and immigration. Iterating map (22) for all the initial conditions between 1960 and 1998, we can compare with the census data the projected value for the next year. The differences between these values is an estimate of the balance between emigration and immigration, Fig. 12. In this case, we have used the mean value $\alpha = 0.9891$, which does not change too much during the period under analysis.

The period 1960–1974 is characterized by a strong emigration, reflected in the negative growth of the population and new-borns. During the period 1974–1982, immigrants outnumber emigrants, introducing a larger growth in the population and in the newborns. For the period 1983–1999, we have an oscillatory balance. In the period 1960–1974, the emigration-immigration balance is of the order of 0.68 millions habitants, implying that emigration was stronger than immigration. In the period 1975–1999, the external income of population dominates, and the emigration-immigration balance is of the order of -0.18 millions habitants.

The example presented here shows how this type of simple discrete models can help us to predict the overall growth of a population, the impact of historical events and the impact of policies of social protection and medical

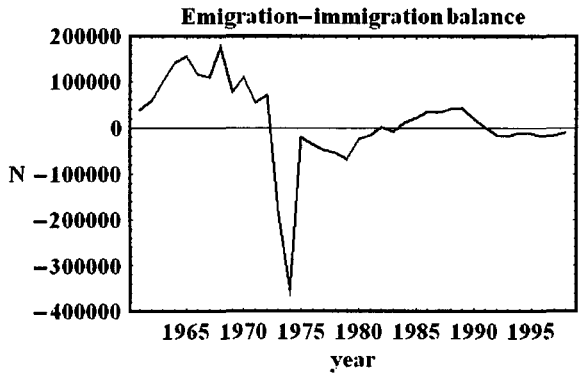


Fig. 12. Emigration-immigration balance for the period 1961–1999 calculated with the census data and model Eq. (22). Positive values correspond to larger emigration when compared to immigration. Negative values mean that immigration is larger than emigration.

care. In fact, the main features presented in the figures reflect important social transformations that occurred in Portugal in the last 40 years. This approach can be further extended in order to introduce emigration and immigration factors and age classes.

In fact, in demography studies, the Leslie discrete model of Sec. 3 is nowadays the basic tool for demographic projections in human populations, [25]. In microbiology, most of modelling approaches are based on the exponential and logistic models, [3].

5. Discrete Time Models with Population Dependent Parameters

In discrete time models with population dependent parameters, we introduce the same kind of reasoning as developed in the continuous models of Sec. 2: In a bounded territory, the growth rate of a population shows sensitivity to population numbers, Fig. 1. As we have seen, this choice has been used in the derivation of the logistic model and, in some sense, has been validated by the predictions of growth of *Paramecium caudatum* in a batch culture.

The simplest discrete population dependent model is described by the Ricker map [26],

$$N_{n+1} = rN_n e^{-N_n} \tag{27}$$

where r is the growth rate and N_n represents the number of individuals of a population at time n. This map has been used for many years in the analysis of fisheries.

The overall dynamics of map (27) is similar to the discrete logistic model,

$$N_{n+1} = aN_n(1 - N_n) \tag{28}$$

used in the modern theory of dynamical systems as a prototype of a chaotic systems, [27]. In particular, the map (28) is a finite differences approximation to the logistic Eq. (2). Applying a finite difference approximation for the derivative in (2), we obtain, $x_{n+1} = x_n(1 + r\Delta t - x_n r\Delta t/K)$. With the linear change of coordinates, $N_n = r\Delta t x_n/K(r\Delta t + 1)$ and $a = (r\Delta t + 1)$, we get the discrete logistic map (28). These types of models introduce a population dependent growth rate in the form of a decreasing function of the number of individuals in the population. The right-hand side of both maps (27) and (28) have a local maximum at $N_n = 1$ and $N_n = 1/2$, respectively.

The dynamic behavior of maps (27) and (28) introduce into the language of population dynamics and ecology the concept of chaos, [27]. The motivation for this approach is based on some observations that, during time, some populations show erratic variations in the population numbers, apparently without external causes, as for example the diminishing of environmental resources.

From the mathematical point of view, there are essentially two types of random behavior in dynamical systems. One type of random behavior, called quasi-periodicity, is associated with the non-periodicity of a temporal time series, as in the iteration of circle maps, [11]. The other type of erratic behavior, called chaoticity, is associated with the existence of an infinite number of unstable periodic orbits in phase space. The Ricker and the logistic maps have chaotic behavior for several parameter values.

To analyse the type of random behavior of the Ricker map (27), we construct a bifurcation diagram: For each value of the parameter r, we iterate the Ricker map from a given initial condition, say 1000 times, and we plot the last 500 iterates. Then, we repeat this procedure for other parameter values. The graph obtained gives information about the asymptotic states of the trajectories of the map, as a control parameter is varied, Fig. 13. For simple enough maps, as one-dimensional maps with one maximum, the information obtained by this method is independent of the initial conditions.

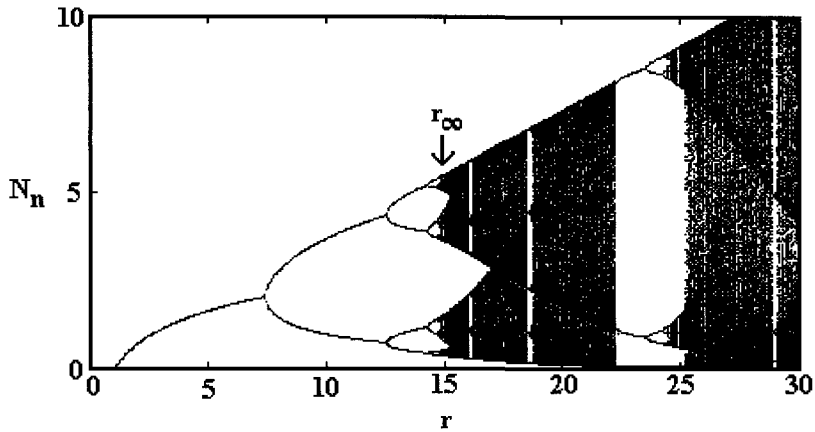


Fig. 13. Bifurcation diagram for the Ricker map (27). Chaos occurs for $r > r_\infty$.

As we see from Fig. 13, for small values of the parameter r, the map (27) has an asymptotically stable steady state — the iterates converge to a stable period-1 fixed point or stable steady state. Increasing the value of the parameter r, the period-1 stable steady state disappears and a new stable steady state with period-2 appears. The parameter value of the transition is a bifurcation in the dynamics of the map. For the parameter values where the period-2 orbit is stable, there exists an unstable period-1 orbit, which obviously does not appear in the bifurcation diagram. For increasing values of the parameter r, a sequence of period doubling bifurcations appears. This sequence accumulates at $r = r_\infty$. For $r > r_\infty$, we say that the dynamics of map (27) is chaotic, [11].

One of the characteristics of the chaotic region in one-dimensional maps ($r > r_\infty$) is the existence of an infinite number of unstable periodic orbits in phase space, even in the regions where the asymptotic states are stable fixed points. The typical time series of a chaotic map is represented in Fig. 14.

It is generally believed that populations can have chaotic behavior in time [27]. In a laboratory experiment with a flour beetle, Costantino *et al.* [28] have shown that, by manipulating the adult mortality, the number of individuals of the population can have erratic behavior in time. In this case, the experimental system shows qualitatively the same type of bifurcation behavior as in a non-linear three-dimensional discrete model for the time evolution of the feeding larvae, the large larvae and the mature adults. However, there is no clear evidence that such erratic behavior shares the same

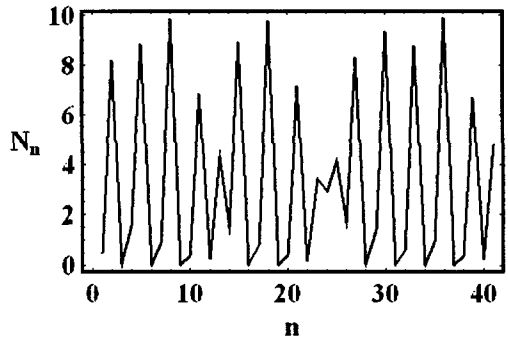


Fig. 14. Chaotic time series obtained with the Ricker map (27), for $r = 26$.

dynamic properties of maps with chaotic behavior, despite the similarities between bifurcation diagrams.

Based on observational data, some authors claim that, in the time series of some populations, the observed erratic behavior has quasi-periodic characteristics, [29]. In the next section, we present a consumer-resource approach model with a bifurcation diagram with some characteristics that are similar to the one obtained with the Ricker map.

6. Resource Dependent Discrete Models

The consumer-resource interaction is a fundamental issue in ecology, [30] and [31]: Without resources, living organisms do not survive or reproduce. In a rich environment, it seams natural to assume that the effect of resources on a small population is not an important limiting factor for growth. However, if resources are scarce, we can expect an increase of death rates and an increase in mobility for the search for other territories.

In ecological modelling, the effect of resources is sometimes introduced as external forcing factors. A typical example is the modelling based on the logistic equation with a time varying carrying capacity. In this case, the response of the population numbers to the external forcing follows the time varying characteristics of the forcing function. Here, we adopt a more generic approach and introduce directly the dynamics of the resources into the models, [31].

To maintain some degree of generality in the derivation of resource dependent models, we adopt a Leslie age-structured approach. The compromise between simplicity and generality is to consider that, in age-structured

populations, resources only affect the probability of survival of the reproductive age classes. Under these conditions, we can write a resource dependent model in the generic form, [31],

$$\begin{pmatrix} N_1^{n+1} \\ N_2^{n+1} \\ N_3^{n+1} \\ \vdots \\ N_k^{n+1} \end{pmatrix} = \begin{pmatrix} 0 & e_2 & e_3 & \cdots & e_{k-1} & e_k \\ \alpha_1(R^n) & 0 & 0 & \cdots & 0 & 0 \\ 0 & \alpha_2(R^n) & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha_{k-1}(R^n) & 0 \end{pmatrix} \begin{pmatrix} N_1^n \\ N_2^n \\ N_3^n \\ \vdots \\ N_k^n \end{pmatrix} \tag{29}$$

$$R^{n+1} = f(R^n)\phi(N^n)$$

where we have introduced a dynamic for the resources, and $N^n = N_1^n + \cdots + N_k^n$ is the total number of individuals in the population at time n. We also assume that the fertility coefficients are resource independent, which must be understood as an oversimplification.

To analyse the dynamical properties of map (29), we make some assumptions on the form of the functions $\alpha_i(R)$, $f(R)$ and $\phi(N)$. In order to derive general properties about the asymptotic states of the population numbers, we establish plausible limiting behaviors for the model functions, without specifying any particular functional forms for $\alpha_i(R)$, $f(R)$ and $\phi(N)$. The function $f(R)$ describes the dynamics of the resources alone through the iteration $R^{n+1} = f(R^n)$. We further assume that the resource map, $R^{n+1} = f(R^n)$, has a stable fixed point for $R^n = K$ and an unstable fixed point for $R^n = 0$.

We further assume that both $\alpha_i(R)$ and $f(R)$ are non-negative and monotonic increasing functions of R, and $\phi(N)$ is non-negative and monotonic decreasing function of N. Therefore, we have the following limiting values,

$$\begin{aligned} &\alpha_i(R) \to 0, \quad \text{as } R \to 0 \\ &\alpha_i(R) \to 1, \quad \text{as } R \to \infty \\ &f(R) \to 0, \quad \text{as } R \to 0 \\ &\phi(N) \to 1, \quad \text{as } N \to 0 \\ &\phi(N) \to 0, \quad \text{as } N \to \infty \end{aligned} \tag{30}$$

Under these conditions, it can be proved that ([31]):

Theorem 2. *The map* (29) *together with conditions* (30) *is a diffeomorphism in the interior of the set* $B = (N_1 \geq 0, \ldots, N_m \geq 0, R > 0)$. *The resource dependent inherent net reproductive number*

$G(K) = \sum_{i=2}^{k} e_i \prod_{j=2}^{i} \alpha_{j-1}(K)$ *is a bifurcation parameter for map* (29). *If* $G(K) > 1$, *but* $G(K)$ *is close to the value* 1, *then the map* (29) *has a non-zero stable fixed and is structurally stable in the interior of* B.

The importance of Theorem 2 relies on the statement that the resources control the structural stability of model map (29), in the sense that any small perturbations of the map will not destroy the stability of the non-zero fixed point. Moreover, the structural stability result is independent of the functional form of $\alpha_i(R)$, $f(R)$ and $\phi(N)$.

In order to better understand the dynamic properties of map (29), we take a prototype model with three age classes, and we choose plausible functions $\alpha_i(R)$, $f(R)$ and $\phi(N)$. For the resources, we choose a logistic growth function,

$$f(R^n) = \frac{K\beta R^n}{R^n(\beta - 1) + K} \tag{31}$$

where $\beta \geq 1$ is the discrete time intrinsic growth rate, and K is the carrying capacity. Function (31) is a logistic type growth function for the resources and follows from (3), with $t = \Delta t$ and $\beta = e^{r_0 \Delta t}$. For the probability of survival between age classes, we assume that it has the form of a birth-and-death (stochastic) process, [32],

$$\alpha_i(R^n) = \frac{\gamma_i R^n}{\gamma_i R^n + 1} \tag{32}$$

where the $\gamma_i > 0$ are parameters. As γ_i becomes large, the probability of survival in the transition between consecutive age classes becomes close to 1, and $\alpha_i(R^n)$ becomes sensitive to the variations of resources only for small values of R^n (few available resources). The function $\phi(N)$ is assumed to have the Poisson form,

$$\phi(N^n) = e^{-N^n}. \tag{33}$$

With $k = 3$, and introducing (31)–(33) into (29), we obtain the resource dependent map,

$$\begin{pmatrix} N_1^{n+1} \\ N_2^{n+1} \\ N_3^{n+1} \end{pmatrix} = \begin{pmatrix} 0 & e_2 & e_3 \\ \dfrac{\gamma_1 R^n}{\gamma_1 R^n + 1} & 0 & 0 \\ 0 & \dfrac{\gamma_2 R^n}{\gamma_2 R^n + 1} & 0 \end{pmatrix} \begin{pmatrix} N_1^n \\ N_2^n \\ N_3^n \end{pmatrix} \tag{34}$$

$$R^{n+1} = \frac{K\beta R^n}{R^n(\beta - 1) + K} \exp\left(-(N_1^n + N_2^n + N_3^n)\right).$$

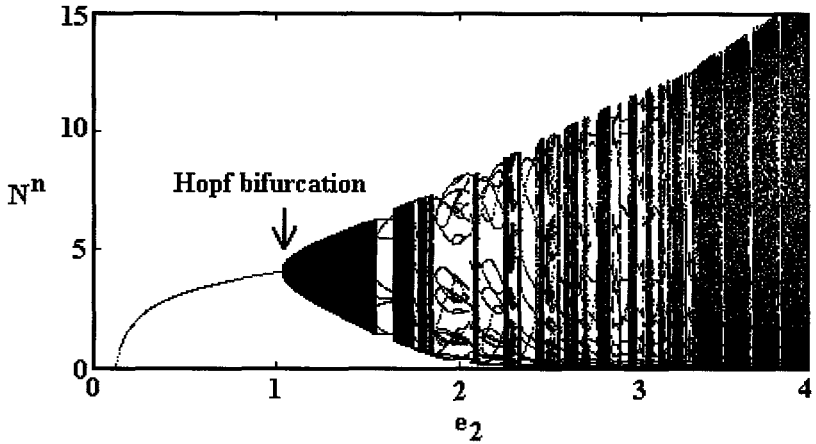


Fig. 15. Bifurcation diagram of map (34) for the parameter values $e_3 = 0.8$, $\gamma_1 = \gamma_2 = 1$, $K = 100$ and $\beta = 1000$, [31].

The phase space of map (34) has dimension 4. However, to analyze the bifurcation diagrams of the number of individuals, we simply plot the total number of individuals of the population, $N^n = N_1^n + N_2^n + N_3^n$, as a function of the control parameters. In Fig. 15, we show the bifurcation diagram for the total number of individuals calculated from map (34), as a function of the control parameter e_2. The other parameters have been fixed to the values $e_3 = 0.8$, $\gamma_1 = \gamma_2 = 1$, $K = 100$ and $\beta = 1000$.

For $0.11 < e_2 < 1.04$, the map (34) has a non-zero stable fixed point. Increasing e_2, this fixed point becomes unstable. The instability of the fixed point is due to a discrete Hopf bifurcation, and an invariant circle in phase space appears, [33], Fig. 15. The discrete Hopf bifurcation occurs when two complex conjugate eigenvalues of the Jacobian matrix of (34) evaluated at the period-1 fixed point cross the unit circle in the complex plane. On the invariant circle, the time evolution is not periodic anymore, and any time series or orbit becomes quasi-periodic. Increasing further the control parameter, there is a continuous region in the parameter space where regions of invariant circles in phase space and regions of periodic behavior appear. These regions are separated by bifurcations from quasi-periodic to periodic attractors (saddle-node discrete bifurcations) characteristic of circle maps, [33].

In Fig. 16, we show the attractors and the time series for two different values of the parameter e_2. In Fig. 16(a), the invariant trajectory in phase

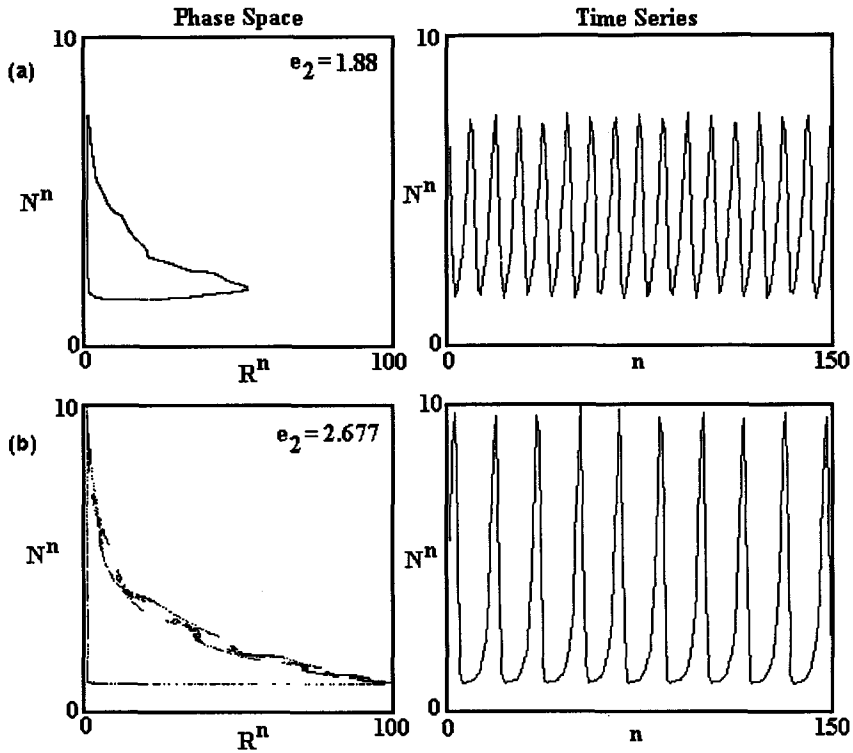


Fig. 16. Invariant sets in phase space and time series for map (34), for several values of the control parameter e_2. In (a), the invariant set in phase space is homeomorphic to a circle, and in (b) it is similar to a fractal set. In both cases, the corresponding time series are quasi-periodic. The time series of these quasi-periodic motions should be compared with the chaotic time series of Fig. 14.

space is homeomorphic to a circle and has a quasi-periodic time series. In Fig. 16(b), the invariant circle is destroyed and an invariant set appears, apparently, with a fractal structure. In this case, the quasi-periodicity of the time series is maintained. Further numerical analysis for other parameter values leads the conclusion that the random behavior found in this map has a different characteristic than the one found in the chaotic case of the previous section.

Quasi-periodic time series have random behavior. In fact, there exists a continuous probability distribution characterizing the permanence time of the iterates of the map on the attractor in phase space, [34]. This probability distribution also exists in the chaotic maps (27) and (28). The difference

between map (34) and maps (27) and (28) is that, in map (34), the phase trajectories on the invariant set have no unstable periodic orbits, whereas in chaotic systems invariant sets contain an infinite number of unstable periodic orbits. In both cases, the trajectories on the invariant sets are random because they are ergodic, leaving invariant the above-mentioned probability distribution, with support on the attractor. The time series of the chaotic system is clearly irregular, contrary to the quasi-periodic case where it is almost regular, despite the apparent similarities of bifurcation diagrams.

There is one more important distinction between the map (34) and the maps (27) and (28). The map (34) is a diffeomorphism in the positive part of phase space, and maps (27) and (28) are not invertible. Prototype models of chaotic dynamics are in general non-invertible. The map (34) becomes non-invertible in the limit $\beta \to \infty$. In this limit, we obtain, for the resource dynamics, $R^{n+1} = Ke^{-(N_1^n+N_2^n+N_3^n)}$. This corresponds to the introduction of a fast recovery time of the resources when compared with the time scale of the population. In this case, the structure of the invariant sets in phase space becomes more complex than the ones depicted in Fig. 16, and we observe regions of quasi-periodic behavior mixed with regions with more complex time series, [31].

From the comparison between the models (27), (28) and (34), the main conclusion we want to point out is that, for hypothetical populations following these dynamics, the information provided by the bifurcation diagrams is not enough to decide about the degree of complexity of a time series. The structure of the attractors in phase space has to be taken into account. In principle, any analytical strategy in order to calibrate and validate models with chaotic or quasi-periodic behavior in time must elucidate about the topological structure of the attractors and about fixed points in phase space.

7. Spatial Effects

In general, in the search of resources or simply to avoid overcrowding, populations spread along space. In some species, the spreading is short-range and in others it has a long-range effect. In order to describe both effects in a simple formalism, we take the simplest case where the population at time $n+1$ relates to the population at time n, through the difference equation,

$$N^{n+1} = f(N^n) \tag{35}$$

where f is some arbitrary function, and N^n is the number of individuals in the population at time n. To introduce the effect of spatial spreading

into (35), we let $N^n \equiv N^n(x)$, and the total population is,

$$\int_a^b N^n(x)dx = N_{tot}^n \tag{36}$$

where a and b define the limits of the territory. The limiting cases $a = -\infty$ and $b = +\infty$ are allowed. Now, $N(x)$ is the number of individuals of the population per unit of length or area. The spatial spreading effect is introduced into (35) by a dispersion kernel $k(y-x)$, and the local dynamics becomes,

$$N^{n+1}(x) = \int_a^b k(y-x)f(N^n(y))dy \tag{37}$$

where, to avoid spatial asymmetries, we assume that $k(\cdot)$ is an even function of its argument. We impose further that,

$$\int_a^b k(z)dz = 1. \tag{38}$$

With condition (38), the kernel function $k(z)$ can be understood as a probability distribution, and the term $k(y-x)f(N^n(y))$ in (37) is the frequency of individuals that were at y at time n and will be at x at time $n+1$. Introducing, $z = y - x$ into (37), we obtain,

$$N^{n+1}(x) = \int_a^b k(z)f(N^n(z+x))dz. \tag{39}$$

The integro-difference Eq. (39) describes the time evolution of the density of individuals in the population and, depending of the form of the kernel function, it accounts for short- and long-range spreading effects.

In order to describe only short-range effects, we develop $f(N^n(z+x))$ in Taylor series around $z = 0$, and from (39), we obtain,

$$N^{n+1}(x) = f(N^n(x)) + D\frac{\partial^2}{\partial x^2}(f(N^n(x))) + \cdots \tag{40}$$

where we have used the normalization condition (38), and,

$$D = \int_a^b k(z)z^2dz \tag{41}$$

is the diffusion coefficient or second moment of the kernel function $k(z)$. The kernel function $k(z)$ is specific to the species under consideration, and, in principle, is related with the mobility of the population. Obviously, D is also species dependent. In [15] and [35], some examples of kernel functions used in ecological modelling are discussed.

We apply now this formalism to the study of the dispersal of a hypothetical population that follows a Ricker type dynamics (Sec. 5). Introducing the Ricker map (27) into the integro-difference equation (39), we obtain,

$$N^{n+1}(x) = r \int_{-\infty}^{+\infty} k(z) N^n(z+x) e^{-N^n(z+x)} dz \tag{42}$$

and we chose as dispersion kernel the normalized Gaussian function,

$$k(z) = \frac{1}{2\sqrt{\pi D}} e^{-z^2/4D} \tag{43}$$

where D is the diffusion coefficient.

To follow in time and space the growth of this hypothetical population, we consider a one-dimensional infinite domain with the initial density distribution,

$$N^0(x) = \begin{cases} 2 & \text{if } |x| \leq 1 \\ 0 & \text{if } |x| > 1. \end{cases} \tag{44}$$

By (36), the initial total population described by (44) has four individuals.

To follow the space and time evolution of the number of individuals in the population, we introduce (44) into (42), and we iterate (42) for several values of the growth rate parameter r. In Fig. 17, we show the first iterates of the integro-difference equation (42), for the initial condition (44), and parameter values: $D = 0.1$ and $r = 5$; $D = 0.1$ and $r = 17$. For these parameter values, the Ricker map (27) has periodic and chaotic behavior (Fig. 13), respectively. After four iterations, the total population numbers are: $N^4_{tot} = 9.6$, for $r = 5$, and $N^4_{tot} = 21.2$, for $r = 17$. In both cases, the initial condition corresponds to $N^0_{tot} = 4$.

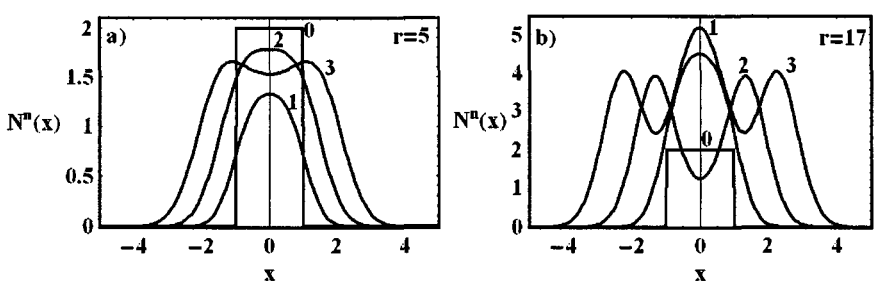


Fig. 17. First iterates of the integro-difference Eq. (42) associated to the Ricker map, for $D = 0.1$, (a) $r = 5$ and (b) $r = 17$. $N^n(x)$ is the density of individuals at the spatial region x and time n. The initial condition at time zero is given by (44). The numbers in the graphs represent the iteration time n.

The numerical simulations in Fig. 17(a) show the formation of a dispersal front. Initially, the front amplitude has small oscillations. After some time, the front amplitude stabilizes and its value equals the value of the fixed point of the Ricker map (27). The front propagates in space and the population number increase in time. When $n \to \infty$, then $N^n \to \infty$. In Fig. 17(b), the Ricker map has chaotic behavior, and a dispersion front appears as time increases. In this case, at each point of the spatial region, the oscillations of the population density during time are chaotic, Fig. 18.

In Fig. 18(a), we show, for the Ricker map (27), the time evolution of $N^n(0)$, calculated with the integro-difference equation (42). The time series at a given point of the extended system has the same type of chaotic behavior as the time series obtained with Ricker map. In Fig. 18(b), we show the graph of the points $(N^{n+1}(0), N^n(0))$. Also, in the extended system, the chaotic behavior of the local map persists. In real extended systems, this effect gives information about the chaoticity or periodicity of the local dynamics.

This simple example shows that the dispersal effect strongly increases the equilibrium values of the population, implying that the dimension of the territory of a population is a constraining factor for the population

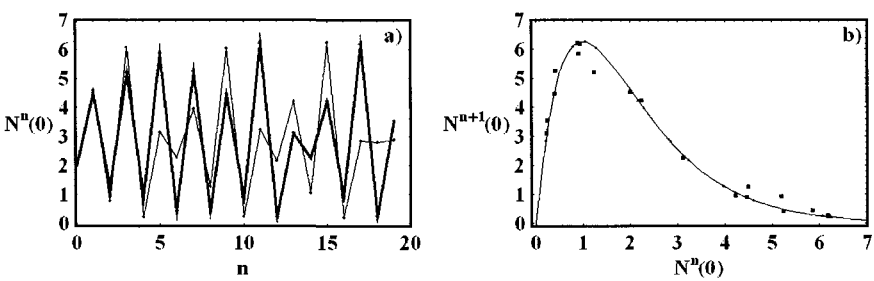


Fig. 18. (a) The heavy line is the time series of the first iterates of the density of individuals $N^n(0)$, calculated with the integro-difference equation (42), for the Ricker map (27). The parameters values are, $D = 0.1$, $r = 17$, and the initial condition is given by (44). The thin line is the iterate of the Ricker map, for the same parameter values, and the initial condition $N^0 = 2$. For these parameter values, the Ricker map is chaotic. As it is clearly seen from the time series, the chaotic behavior is still present in the extended system. In (b), we show the graph of the function $N^{n+1} = f(N^n) = rN^n e^{-N^n}$ (thin line), and the points $(N^{n+1}(0), N^n(0))$ from the time series in (a). The iterates of the Ricker map are on the graph of the function $f(N^n)$ (triangles). The points $(N^{n+1}(0), N^n(0))$ obtained from the time series of the extended system are near to the graph of the function $f(N^n)$ (squares).

growth. Note that, the Ricker map only admits finite values for the number of individuals of a population.

Suppose now that we have a population evolving according to the logistic equation (2), and we want to take into account dispersal effects. Applying to the logistic equation the same reasoning leading to the integro-difference Eq. (39), we obtain,

$$\frac{\partial N}{\partial t} = \int_a^b k(z) r N(z+x,t)(1 - N(z+x,t)/K) dz \tag{45}$$

together with the normalization condition (38). In this case, we obtain an integro-differential time evolution equation. To analyze short-range dispersal effects, by (40), we have,

$$\frac{\partial N}{\partial t} = rN(1 - N/K) + rD\frac{\partial^2}{\partial x^2}(N(1 - N/K)) \tag{46}$$

which is a parabolic partial differential equation.

We take the simple equilibrium solution of the logistic equation, $N(x) = K$. Introducing this solution into (46), $N(x) = K$ is a stationary solution of the parabolic Eq. (46). Therefore, by (36), the logistic model with dispersal admits an infinite population in an infinite territory. Once more, it is clear that population numbers are dependent of the dimensions of the territory.

In the literature of ecology, dispersion effects are analysed through the parabolic equation,

$$\frac{\partial N}{\partial t} = rN(1 - N/K) + D\frac{\partial^2 N}{\partial x^2} \tag{47}$$

which is known as the Kolmogoroff–Petrovskii–Piskunov or Fisher equation. This equation is in general derived under the assumption of Fick laws that asserts that migration occurs from regions with higher densities to regions with lower densities. One of the properties of the solutions of the equation (47) is the possibility of having a propagating front along space, [36] and [37], analogous to the fronts of Fig. 17(a).

To incorporate dispersion effects into continuous models of population dynamics, we can follow the reasoning leading to equations (45) and (46), or adopt the view leading to Eq. (47), introduced by Kolmogoroff–Petrovskii–Piskunov, [36]. In the last case, this corresponds to add a diffusive term, transforming the differential equations into a quasi-linear parabolic partial differential equation. Most of these spatial models are non-linear and their space and time solutions are found numerically. In references [15] and [35], several models with spatial effects are analysed.

8. Age-Structured Density Dependent Models

In the literature of ecology, density dependent models appear in several contexts. In the Leslie approach, the effects on population density can be introduced through the dependence of the probability of survival between age classes on the population numbers. In this case, the survival probabilities between the age classes are of the form $\alpha_i(N^n)$, where N^n is the total population at time n, [38], and the general non-linear map obtained falls in the class of non-linear maps of Sec. 5. Also, in the previous section, we found a density dependent model in the sense that the state variable of the population has the generic form $N(x)$, representing the number of individuals by unit of area or length. All these models are, in a certain sense, density dependent. Here, we are interested in density dependent models where the age and time variables have a continuous nature, [39], [23], [40] and [41].

We consider a population age density function $n(a,t)$ such that,

$$N(t) = \int_0^{+\infty} n(a,t)da \tag{48}$$

where $N(t)$ is the total population at time t, and a represents age. The function $n(a,t)$ is the density of the individuals of the population with age a at time t. The births are described by the fertility function by age class $b(a,t)$, and are given by,

$$n(0,t) = \int_0^{+\infty} b(a,t)n(a,t)da. \tag{49}$$

Assuming that the normalised death rate of the density of individuals (mortality modulus) of the population is constant (μ), we have,

$$\frac{dn}{dt} = -\mu n. \tag{50}$$

As ageing is time dependent, $a \equiv a(t)$, and is measured in the same time scale of time, $\frac{da}{dt} = 1$, by (50), the function $n(a(t),t)$ obeys the first order linear partial differential equation,

$$\frac{\partial n(a,t)}{\partial t} + \frac{\partial n(a,t)}{\partial a} = -\mu n(a,t) \tag{51}$$

together with the boundary condition (49). Equation (51) is the McKendrick equation [39].

Let us find now a solution of the McKendrick partial differential Eq. (51) by elementary geometric methods. Writing Eq. (51) in the form, $\frac{dn(a,t)}{dt} = -\mu n(a,t)$, where a is also a function of t, the solutions of (51)

are obtained through the solutions of the system of ordinary differential equations,

$$\begin{cases} \dfrac{dn}{dt} = -\mu n \\ \dfrac{da}{dt} = 1. \end{cases} \tag{52}$$

These two independent equations have the general solution,

$$\begin{cases} n(a,t) = n(a_0,t_0)e^{-\mu(t-t_0)} \\ a - a_0 = t - t_0 \end{cases} \tag{53}$$

where a_0 is the continuous age variable for $t = t_0$. The second equation in (53) is the equation of the characteristic curves of the partial differential Eq. (51), Fig. 19. Introducing the second equation in (53) into the first one, we obtain the solution of the McKendrick equation,

$$n(a,t) = n(a-t,0)e^{-\mu t}, \quad \text{for } t < a \tag{54}$$

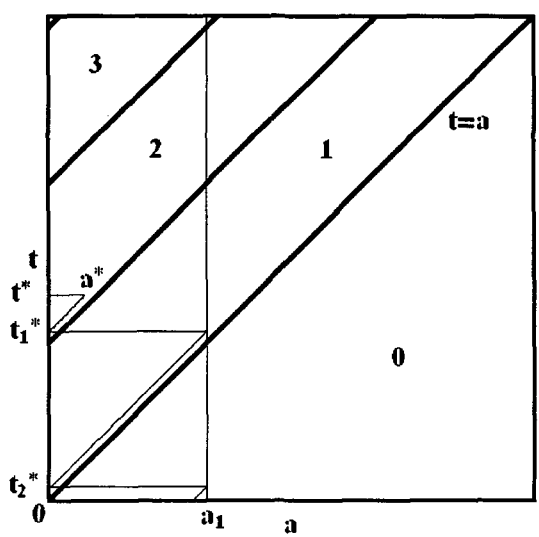


Fig. 19. Characteristic curves $a - a_0 = t - t_0$ for the McKendrick equation (51). The graph of the characteristic curves lies in the domain of the solution $n(a,t)$. Given an arbitrary point (a^*, t^*) in the domain of the partial differential equation, the solution $n(a^*, t^*)$ is easily obtained following the thin line backwards in time down to $t = 0$. The heavy lines are characteristic curves defined by the equations $t = a + ma_1$, for $m = 0, 1, \ldots$.

where $n(a,0)$ is an initial density distribution of the population at $t_0 = 0$. For $t < a$, the solution (54) is independent on the boundary condition (49). The domain of the solution (54) is the region labeled with a zero in Fig. 19.

To extend the solution (54) for $t \geq a$, the boundary condition must be introduced, as well as some additional simplifications. We suppose that births occur at some unique fixed age $a = a_1$, Fig. 19. Then, the fertility function is necessarily concentrated at the point $a = a_1$. Therefore, as fertility function by age class, we take,

$$b(a,t) = b\delta(a - a_1)$$

where $\delta(\cdot)$ is the Dirac delta function and b is a (mean) fertility parameter. Under these conditions, the boundary condition (49) simplifies to,

$$n(0,t) = bn(a_1,t). \tag{55}$$

We extend now the solution (54) to $t = a$, with the boundary condition (55). By (53) and (55), the solution of the McKendrick equation (51) is,

$$n(a,t) = n(0,0)e^{-\mu t} = bn(a_1,0)e^{-\mu t}, \quad \text{for } a = t. \tag{56}$$

With a simple geometric construction, we calculate now the solution for $t > a$. We take the point (a^*,t^*) on the line $t = t^*$, Fig. 19. This point is in the region labelled 2 in Fig. 19. By (53), the characteristic line that passes by (a^*,t^*) crosses the line $a = 0$ at some time $t = t_1^*$, and $n(a^*,t^*) = n(0,t_1^*)e^{-\mu(t^*-t_1^*)} = n(0,t^* - a^*)e^{-\mu(t^*-t_1^*)}$, where $t_1^* = t^* - a^*$. Imposing on this solution the boundary condition (55), we obtain, $n(a^*,t^*) = bn(a_1,t^* - a^*)e^{-\mu(t^*-t_1^*)}$. As $n(a_1,t^* - a^*)$ is the solution of the McKendrick equation at the point $(a_1,t^* - a^*)$, we repeat the above construction, by drawing the horizontal line connecting the points $(0,t_1^*)$ with (a_1,t_1^*), Fig. 19. Iterating this procedure backwards in time, we obtain,

$$\begin{aligned} n(a^*,t^*) &= b^2 n(a_1,t^* - a^* - a_1)e^{-\mu(t^*-t_1^*+t_1^*-t_2^*)} \\ &= b^2 n(2a_1 + a^* - t^*,0)e^{-\mu(t^*-t_1^*+t_1^*-t_2^*+t_2^*-0)} \\ &= b^2 n(2a_1 + a^* - t^*,0)e^{-\mu t^*} \end{aligned} \tag{57}$$

where $t_2^* = t^* - a_1^* - a^*$, and we systematically have used the equation of the characteristic curves, $a - a_0 = t - t_0$. In Fig. 19, for any initial condition inside the region labelled with an integer, say m, an easy induction argument shows that the solution of the McKendrick equation (51) has the form (57), where the factor 2 is substituted by m, where $m = [(t-a)/a_1 + 1]$ and

$[x]$ stands for the integer part of x. Then, by simple geometric arguments, we have proved:

Theorem 3. *Let $n(a,0)$ be an initial data function for the McKendrick partial differential equation* (51), *with $a \geq 0$, $t \geq 0$ and $\mu > 0$. Then, for the boundary condition* (55), *with a_1 and b positive constants, the general solution of the McKendrick partial differential equation is,*

$$\begin{aligned} n(a,t) &= n(a-t,0)e^{-\mu t}, \quad \text{for } t < a \\ n(a,t) &= b^{[(t-a)/a_1+1]} n([(t-a)/a_1+1]a_1 + a - t, 0)e^{-\mu t}, \quad \text{for } t \geq a \end{aligned} \tag{58}$$

where $[x]$ stands for the integer part of x.

We analyse now the stability of the solution (58) of the McKendrick partial differential equation. As $[(t-a)/a_1]+1 = m$, where m is a positive integer, we have, $(t-a)/a_1 + 1 = m + \varepsilon(a,t)$, and, for fixed a, with $t \geq a$, the function $\varepsilon(a,t)$ is time periodic with period a_1. Therefore,

$$b^{[(t-a)/a_1+1]}e^{-\mu t} = e^{[(t-a)/a_1](\ln b - \mu a_1)} b e^{-\mu a} e^{-\mu a_1 \varepsilon(a,t)}.$$

By (58), and for $t \geq a$, we have,

$$\begin{aligned} n(a,t) = {} & e^{[(t-a)/a_1](\ln b - \mu a_1)} \\ & \times n([(t-a)/a_1+1]a_1 + a - t, 0) b e^{-\mu a} e^{-\mu a_1 \varepsilon(a,t)}. \end{aligned} \tag{59}$$

Then, if $(\ln b - \mu a_1) = 0$, the asymptotic solution of the McKendrick partial differential equation is periodic in time. If, $b > e^{\mu a_1}$, asymptotically in time, the population density goes to infinity, and if, $b < e^{\mu a_1}$, the population density goes to zero. Hence, we have:

Corollary 4. *Let $n(a,0)$ be a differentiable and bounded initial data function for the McKendrick partial differential equation* (51), *with boundary condition* (55). *Suppose in addition that $\ln b = \mu a_1$. Then, the asymptotic solution of the McKendrick equation is bounded and periodic in time, with period a_1:*

$$n(a,t) = n([(t-a)/a_1+1]a_1 + a - t, 0) b e^{-\mu a} e^{-\mu a_1 \varepsilon(a,t)}.$$

In Fig. 20, we show the time evolution of an initial population with an uniform age distribution, with a maximal age class, and obeying to the stability condition $b = e^{\mu a_1}$. We have chosen for initial conditions the density function: $n(a,0) = 2$, for $a \leq 100$, and $n(a,0) = 0$, for $a > 100$. By (48), this corresponds to the initial population, $N(0) = 200$. The calculated population numbers for $t = 30$ and $t = 100$ are $N(30) = 110.7$ and $N(100) = 70.9$, respectively. As it is clearly seen in Fig. 20, after a transient time, the population density becomes periodic in time, as asserted in Corollary 4.

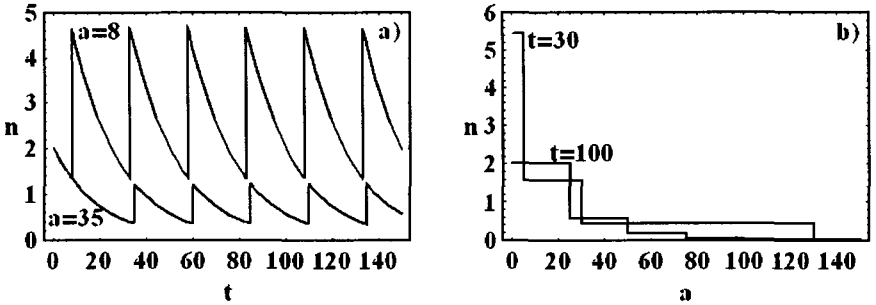


Fig. 20. (a) Time evolution of the solution of the McKendrick equation (51), for the age classes $a = 8$ and $a = 35$. (b) Distribution of the number of individuals by age class, for $t = 30$ and $t = 100$. All these solutions have been calculated with the initial data condition, $n(a, 0) = 2$, for $a \leq 100$, and $n(a, 0) = 0$, for $a > 100$, and parameter values, $a_1 = 25$ (unique reproductive age class), $\mu = 0.05$ (death rate) and $b = e^{\mu a_1} = 3.49$ (mean fertility).

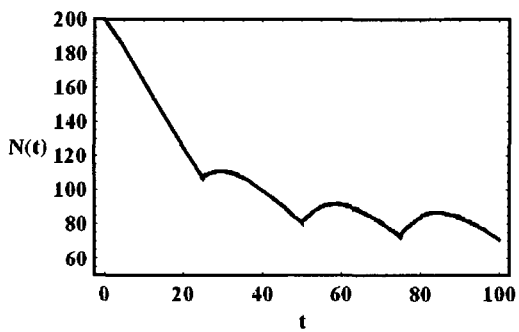


Fig. 21. Total population as a function of time, for the same parameter values of Fig. 20. After a transient time, the period of oscillations is a_1, the age of the unique reproductive age class.

In Fig. 21, we show the total population as a function of time, calculated from Theorem 3 and (48), with the initial condition and parameters from Fig. 20. In the McKendrick continuous age-structured approach, the asymptotic stable state of the dynamics are not fixed points as in the case of the Leslie type maps (Sec. 3), but bounded time periodic function. Moreover, by Theorem 3, the amplitude of oscillations depends on the initial data function $n(a, 0)$.

If, $b > e^{\mu a_1}$, by (59), it can be shown that the growth curve is modulated by a time periodic function with period a_1, [42]. For two reproductive age classes and boundary condition $n(0, t) = b_1 n(a_1, t) + b_2 n(a_2, t)$, the growth curves are always modulated by two time periodic functions with periods

a_1 and a_2. If these periods are not rationally related ($n_1a_1+n_2a_2 = 0$, has no integer solutions in n_1 and n_2), the modulation function is quasi-periodic. In the population dynamics literature the quasi-periodic modulations of the growth curves are called demography cycles. For a detailed discussion see [42].

In the more general case of age dependent fertility function and mortality modulus, $b(a)$ and $\mu(a)$, the qualitative behavior and stability properties of the solutions of the McKendrick equation (51) are determined by the Lotka growth rate, [24] and [42],

$$r = \int_{a_1}^{a_2} b(c)e^{-\int_0^c \mu(s)ds}dc \tag{60}$$

where a_1 and a_2 are the ages of the first and the last reproductive age classes, and $b(a)$ and $\mu(a)$ are determined from demography data.

To estimate the growth of populations, the Lotka growth rate is an important demography tool, [25]. A detailed analysis shows that, the discrete approximation of (60) coincides with the inherent net reproductive number of a population (21), [42], introduced in the discrete time and age Leslie formalism of Sec. 3.

We can now compare the solution of the McKendrick equation derived in Theorem 3 with the exponential solution of the Malthusian growth model (1). Assuming that $t \geq a$, we have, $n(a, t + sa_1) = n(a, t)r^s$, where s is an integer and $r = be^{-\mu a_1}$ is the Lotka growth rate. With $t = a_1$ and $\tau = a_1 + sa_1$, we obtain, $n(a, \tau) = n(a, a_1)r^{(\tau - a_1)/a_1}$. Integrating $n(a, \tau)$ in a, we have for the total population,

$$N(\tau) = N(a_1)r^{(\tau - a_1)/a_1} \tag{61}$$

which is a Malthusian growth function with Lotka growth rate r. Therefore, with a time step equal to the age of the only reproductive age class (a_1), the solution of the McKendrick equation behaves like the exponential growth model of Sec. 2. In fact, taking the derivative of (61) in order to τ, the total population N obeys to the differential equation,

$$\frac{dN}{d\tau} = \frac{\log r}{a_1} N$$

which is the Malthusian growth model (1). This shows that Malthusian type models describe population growth at a larger time scale when compared with age-dependent McKendrick type models.

9. Growth by Mitosis

When reproduction occurs by mitosis as in cells and some micro-organisms, the growth model of the previous section must be modified. We consider a population of micro-organisms or cells in a media with enough resources, eventually infinite. We assume that the micro-organisms replicate by mitosis, and the time of the mitotic processes can be neglected. We represent by $n(a,t)$ the density of organisms in the media with age a at time t.

We consider that the probability of dying depends only on the age. We denote by $\mu(a)$ the probability density of death with age a, and $b(a)$ is the probability density of undergoing mitosis with age a . If an organism initiates mitosis, then, after some time, the organism transforms into two new ones with age zero. Therefore, the density of newborns at time t is,

$$n(0,t) = 2\int_0^\infty b(a)n(a,t)da \tag{62}$$

where the factor 2 accounts for the mitotic process.

Hence, the time evolution of the colony of micro organisms is described by the modified McKendrick equation,

$$\frac{\partial n(a,t)}{\partial t} + \frac{\partial n(a,t)}{\partial a} = -(\mu(a) + b(a))n(a,t) \tag{63}$$

together with the boundary condition (62).

We consider a population where all the individuals initiate mitosis at some fixed age $a = \alpha$, such that $b(a) = \delta(a-\alpha)$. To simplify even further, we suppose that $\mu(a) = \mu$ is constant, independently of the age. So, by (60), (62) and (63), the Lotka growth rate is now,

$$r = 2\int_0^\infty \delta(a-\alpha)e^{-\int_0^a (\mu(s)+\delta(s-\alpha))ds}da. \tag{64}$$

As,

$$\int_0^a \delta(s-\alpha)ds = \begin{cases} 0, & \text{if } a < \alpha \\ 1/2, & \text{if } a = \alpha \\ 1, & \text{if } a > \alpha \end{cases} \tag{65}$$

by (65) and (64), for a population with mitotic reproduction type, the Lotka growth rate is,

$$r = 2e^{-\mu\alpha}e^{-1/2}. \tag{66}$$

Making $\mu = 0$ in (66), for the Lotka growth rate, we obtain the constant value,

$$r = 2e^{-1/2} = 1.21306.$$

In the case of sexual reproduction and in the limit case of zero mortality, by (60), we have,

$$r = \int_{a_1}^{a_2} b(c)dc \tag{67}$$

which can be larger or smaller than 1, depending on the intrinsic fertility of the species. In mitosis, in the limiting case of zero mortality, the Lotka growth rate is always greater than 1, ensuring an exponential growth process. On the other hand, by (66), in organisms that reproduce by mitosis, r cannot take large values, but in the case of organisms with sexual reproduction, by (67), r can be arbitrary large.

The application of the McKendrick approach to the growth of colonies of micro-organisms is described in detail in Rubinow [43].

10. Conclusions

Along this review we have derived the most common models used in ecology and population dynamics. Differential and difference equation models, describe a population as whole, in the sense that they do not distinguish either intra-specific characteristics of the individuals, or their spatial distribution. The calibration of these models with real biological situations is not always successful and, in the context of growth projections, they must be considered as toy models. However, the analysis of their dynamic behavior introduced in the language of ecology new concepts that later were generalised to the age-structure approach. This is the case of the concept of growth rate, introducing a quantitative measure of the stability or instability of a population.

The dependence of the carrying capacity and of the growth rate parameters on the dimensions of the territories and on the available resources, is also an important issue. Experimental measurements of growth of micro-organisms as a function of resources has been done by Monod, [44].

The class of age-structured models gives us a more detailed insight on the dynamics of a population. In demography studies, the Leslie and the McKendrick approaches are nowadays the reference models. In the context of microbiology, the same structure and theoretical setting holds as it has been shown in Rubinow, [43]. One of the consequences of the age-structured McKendrick approach to the population growth is that, in a time scale of the order of the mean age of the reproductive classes, populations have a Malthusian type growth pattern with a Lotka growth rate. The Malthusian growth pattern is perturbed by periodic functions which globally induce quasi-periodic demography cycles.

The distinction between chaos and quasi-periodicity in real ecological systems, is not a simple subject. We have presented models based on the difference equations approach that are non-invertible and have chaotic behavior. However their calibration with real biological systems presents some difficulties. Models showing quasi-periodicity are invertible, and the restriction to the phase space attractors make them similar to circle maps. Technically, both chaotic and quasi-periodic systems are ergodic, leaving invariant a probability measure. This property implies that chaotic and quasi-periodic systems have the same type of randomness.

In generic terms, the dynamical properties of chaotic and quasi-periodic systems are strongly dependent of the adopted mathematical models. The question whereas a basal population or a population at the top of a trophic chain follow any of these choices can only be verified by observation.

One possible way of distinguishing between chaos and quasi-periodicity in populations is through the analysis of the population dispersal. As it has been seen in Sec. 7 for the dispersal of a population, the characteristics of the temporal dynamics is imprinted in the density profile along space. The analysis of this effect gives information about chaoticity or periodicity of the local dynamics.

Some of these models are sensitive to perturbations of the functional forms of the growth rates, others are not. For populations at the top or at the bottom of a trophic chain, it is difficult to conceive that growth models could be too sensitive to arbitrary small external factors. In this case, the structural stability property should be a mandatory property for the choice of any mathematical model for the population growth.

We have compared the dynamic properties of the models derived along this review. We have analysed the models according the possibility of describing population extinction and explosion, non-zero equilibrium states, oscillations, chaos and quasi-periodicity. The structural stability or robustness of the models in phase space is, in simple terms, related with the conditions of the Hartman-Grobman theorem (Theorem 1), and with the non-existence of phase space trajectories connecting unstable fixed points. For this global comparison, we have considered that growth rates are always positive. The properties of the analysed models are summarised in Table 1.

For all the models analysed along this review, it is evident that, the same biological assumptions, but different technical options, lead to different models with different properties, Table 1. The choice of the appropriate model to describe a specific living system must rely on the calibration and validation of the model results with the growth projections.

Table 1. Comparison between the properties of the population growth models analysed along the text. In the case of partial differential equation models, the concept of structural stability needs a different approach from the one explored here. Notes: (1) Structural stability refers to persistence upon perturbations of the non-zero equilibrium in phase space. (2) In mutualistic interactions we can have explosion of population numbers. This case has not been considered in this classification. (3) The zero equilibrium state is always unstable. (4) Technically, if the rotation number around the closed trajectories is irrational, the time series is quasi-periodic. (5) In general, the Leslie matrix has non dominant complex eigenvalues, introducing quasi-periodic modulations in the growth curves. (6) If the Leslie matrix has no eigenvalues on the unit circle of the complex plane, then the Leslie map is structuraly stable. (7) Chaos occurs for some of these models when the recovery time of the resources goes to zero, [31]. (8) The McKendrick model has quasi-periodic time behavior in the sense that growth curves are modulated by periodic functions, with periods equal to the ages of the reproductive age classes, [42]. (9) In general, the concept of structural stability is not defined for partial differential equations.

Population growth model	Non-zero equilibrium	Population goes to ∞	Population dies out	Oscillations	Chaos	Quasi-periodicity	Structural stability (1)
Malthusian	no	yes	no	no	no	no	yes
Logistic	yes	no	no	no	no	no	yes
Logistic controlled by resources	no	no	yes	no	no	no	yes
Biotic interactions (Sec. 2.2)	yes	no (2)	no (3)	no	no	no	yes
Lotka-Volterra	yes	no	no	yes	no	yes (4)	no
Leslie	yes	yes	yes	yes (5)	no	yes (5)	yes (6)
Leslie controlled by resources	yes	no	yes	yes	yes (7)	yes	yes
Ricker	yes	no	yes	yes	yes	no	yes
McKendrick	no	yes	yes	yes	no	yes (8)	(9)

Some other modelling approaches to ecology and population dynamics were not focused in this review. This is the case of the dynamic energy budget approach, [45], topics related with resource exploration and eco-economics, [46], harvesting, [47], [46] and [48], epidemics and dispersal of diseases, [15], [30] and [49]. In the dynamic energy budget approach, the main objective is to make an integrative view of the different levels of organization of biological systems, from simple micro-organisms to ecosystems. In the resource exploration aspects of ecology and eco-economics, concepts of economic theory are introduced in the framework of ecology. Harvesting models are important to analyse problems of control and over consumption of natural resources, [46]. Also, harvesting models are an alternative approach to describe predation in species that are incorporated in trophic webs, [48]. Epidemics and dispersal of diseases are important subjects due to its immediate application to health prevention issues. For a recent account on more specialised mathematical models in biological sciences see [50].

Acknowledgments

I would like to thank Professor J. C. Misra for his kind invitation to write an introductory review on mathematical modelling in ecology and population dynamics. This work has been partially supported by the PRAXIS XXI Project P/FIS/13161/1998 (Portugal), and by the Institut des Hautes Études Scientifiques (Bures-sur-Yvette, France).

References

[1] S. H. Strogatz, Exploring complex networks, *Nature* **410** (2001) 268–276.
[2] M. Braun, Single species population models, in *Differential Equations Models*, Vol. 1, eds. M. Braun, C. S. Coleman and D. A. Drew (Springer-Verlag, New York, 1978).
[3] H. G. Schlegel, *General Microbiology* (Cambridge University Press, Cambridge, 1993).
[4] H. Haberl and H. P. Aubauer, Simulations of human population dynamics by a hyperlogistic time-delay equation, *J. Theo. Biol.* **156** (1992) 499–511.
[5] G. F. Gause, *The Struggle for Existence* (Williams and Wilkins, New York, 1934).
[6] R. Dilão and T. Domingos, A general approach to the modelling of trophic chains, *Ecological Modelling* **132** (2000) 191–202.
[7] A. I. Volpert, V. A. Volpert and V. A. Volpert, *Travelling Wave Solutions of Parabolic Systems* (American Mathematical Society, Providence, 1994).
[8] F. M. Scudo and J. R. Ziegler, *The Golden Age of Theoretical Ecology: 1923–1940*, Lecture Notes in Biomathematics, No. 22 (Springer-Verlag, Berlin, 1978).

[9] V. Volterra, *Leçons sur la Théorie Mathématique de la Lutte pour la Vie* (Gauthier-Villars, Paris, 1931).
[10] R. L. Coren, *The Evolutionary Trajectory* (Gordon and Breach, New York, 1998).
[11] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (Springer-Verlag, Berlin, 1983).
[12] M. W. Hirsh and S. Smale, *Differential Equations, Dynamical Systems and Linear Algebra* (Academic Press, 1974).
[13] N. J. Gotteli, *A Primer of Ecology* (Sinauer Associates, Sunderland, 1995).
[14] R. M. May, *Stability and Complexity in Model Ecosystems* (Princeton University Press, Princeton, 1974).
[15] J. D. Murray, *Mathematical Biology* (Springer-Verlag, New York, 1989).
[16] C. S. Holling, The functional response of predators to prey density and its role in mimicry and population regulation, *Mem. Entomol. Soc. Canada* **45** (1965) 5–60.
[17] L. A. Real, The kinetics of functional response, *The American Naturalist* **111** (1977) 289–300.
[18] A. A. Berryman, The origins and evolution of predator-prey theory, *Ecology* **73** (1992) 1530–1535.
[19] T. Royama, A comparative study of models for predation and parasitism, *Res. Popul. Ecol. Supp.* **1** (1971) 1–90.
[20] P. H. Leslie, On the use of matrices in certain population mathematics, *Biometrika* **33** (1945) 183–212.
[21] Instituto Nacional de Estatística-Portugal. Idicadores demográficos séries 1960–1999; Resultado definitivos, censos 1940, 1960, 1970, 1980 e 1991; I-Movimento geral da população.
[22] R. E. Mickens, *Difference Equations, Theory and Applications* (van Nostrand, New York, 1990).
[23] H. Caswell, *Matrix Population Models* (Inc. Publishers, Sunderland, 1989).
[24] J. M. Cushing, *An Introduction to Structured Population Dynamics* (SIAM, Philadelphia, 1998).
[25] N. Keyfitz and W. Flieger, *World Population Growth and Aging* (University of Chicago Press, Chicago, 1990).
[26] W. Ricker, Stock and recruitment, *J. Fish. Res. Board Canada* **11** (1954) 559–663.
[27] R. M. May, Simple mathematical models with very complicated dynamics, *Nature* **261** (1976) 459–467.
[28] R. F. Costantino, R. A. Desharnais and J. M. Cushing, Chaotic dynamics in an insect population, *Science* **275** (1997) 389–391.
[29] C. Zimmer, Life after chaos, *Science* **284** (1999) 83–86.
[30] R. H. MacArthur, *Geographical Ecology* (Princeton University Press, Princeton, 1972).
[31] R. Dilão and T. Domingos, Periodic and quasi-periodic behavior in resource-dependent age structured population models, *Bull. Math. Biology* **63** (2001) 207–230.

[32] N. T. J. Bailey, *The Elements of Stochastic Processes* (Wiley Interscience Publications, New York, 1964).
[33] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory* (Springer-Verlag, New York, 1995).
[34] V. I. Arnold and A. Avez, *Problèmes Ergodiques de la Mécanique Classique* (Gauthier-Villars, Paris, 1967).
[35] M. Kot, M. A. Lewis and P. Driessche, Dispersal data and the spread of invading organisms, *Ecology* **77** (1996) 2027–2042.
[36] A. Kolmogoroff, I. Petrovskii and N. Piskunov, A study of the equation of diffusion with increase in the quantity of matter, and its application to a biological problem, Moscow Uni. Bull. Math. 1 (1937) 1–25. Reprinted in P. Pelcé, *Dynamics of Curved Fronts* (Academic Press, Boston, 1988), pp. 105–130.
[37] P. Grindrod, *Patterns and Waves* (Clarendon Press, Oxford, 1991).
[38] A. Wikan, Four-periodicity in Leslie matrix models with density dependent survival probabilities, *Theoretical Population Biology* **53** (1998) 85–97.
[39] A. G. McKendrick, Applications of mathematics to medical problems, *Pro. Edinburgh Math. Soc.* **44** (1926) 98–130.
[40] G. Oster and J. Guckenheimer, Bifurcation phenomena in population models, in *The Hopf Bifurcation and its Applications*, eds. J. E. Marsden and M. McCracken (Springer-Verlag, New York, 1976).
[41] A. M. Roos, A gentle introduction to physiologically structured population models, in *Structured-Population Models in Marine, Terrestrial, and Freshwater Systems*, eds. S. Tuljapurkar and H. Caswell (Chapman & Hall, New York, 1996).
[42] R. Dilão and A. Lakmeche, On the weak solutions of the McKendrick equation with time and age dependent birth rate, in preparation.
[43] S. I. Rubinow, *Mathematical Problems in the Biological Sciences* (SIAM, Philadelphia, 1973).
[44] J. Monod, *Recherches sur la Croissance des Cultures Bactériennes* (Hermann & Cie., Paris, 1942).
[45] S. A. L. M. Kooijman, *Dynamic and Energy Mass Budgets in Biological Systems* (Cambridge University Press, Cambridge, 2000).
[46] C. W. Clark, *Mathematical Bioeconomics* (Wiley InterScience, New York, 1990).
[47] J. R. Beddington and R. M. May, Harvesting natural populations in a randomly fluctuating environment, *Science* **197** (1977) 463–465.
[48] R. Dilão, T. Domingos and E. Shahverdiev, Harvesting in a resource dependent age structured Leslie type population model, *Mathematical Biosciences* **189** (2004) 141–151.
[49] N. M. Ferguson, M. J. Keeling, W. J. Edmunds, R. Gani, B. T. Grenfell, R. M. Anderson and S. Leach, Planning for smallpox outbreaks, *Nature* **425** (2003) 681–685.
[50] L. Chen, Y. Kuang, S. Ruan and G. Webb, Advances in Mathematical Biology, *Discrete and Continuous Dynamical Systems — Series B* **4** (2004) 501–866.

This page is intentionally left blank

CHAPTER 16

MODELLING IN BONE BIOMECHANICS

J. C. MISRA* and S. SAMANTA
Department of Mathematics, Indian Institute of Technology
Kharagpur — 721302, India
**jcm@maths.iitkgp.ernet.in*

This chapter gives a brief account of various mathematical models developed by different investigators in connection with various studies in Biomechanics of Bones. A mention of some relevant experimental investigations has also been made. For the convenience of readers some basic concepts and some associated topics useful for a better comprehension of this interesting interdisciplinary area of study are also included.

1. Introduction

As early as 1638 Galileo propounded the idea that the form of the bones depends on the load they carry (Ascenzi, 1993). But it was not easy to readily apply the laws of mechanics, that are used to discuss the statics and dynamics of inanimate objects subjected to loads, to a biological material like bone which is capable of "self-repair" or "self-organization". Before renaissance period people used to think that the world of living beings has nothing in common with that of the non-living bodies. Nobody even thought of explaining biological events in physical terms. But from the time of Descartes, when people with liberal views were questioning the past habits and lines of thought, they were trying to give a systematic interpretation of every phenomenon. Serious attempts were made by them to investigate the mechanical behavior of biological elements. These studies helped develop the discipline of Biomechanics. As the application of the concepts of biomechanics to study different aspects of bone, from a structural point of view and also as a system, this subject area of study grew as a subdiscipline of Biomechanics, known as Bone Biomechanics. The growth of bone biomechanics that took place till 1973 has been summarized by Evans (1973). This treatise contains a review of most of all the important researches on the mechanical properties of bones. Subsequently Cowin and

his associates (1976–1998) made significant contribution to mathematical modelling of the functional adaptation of bones under load. Roesler (1987) presented the historical development of the fundamental concepts of bone biomechanics and the important milestones in the history of researches in this interesting field of study.

During the last three decades, a tremendous growth of bone biomechanics has taken place. Basing upon adaptive elasticity theory and computational mechanics, Cowin and his co-workers investigated extensively different aspects of the process of bone remodelling. They also dealt with problems associated with bone implants. Lakes and Katz (1974–1979) made a series of investigations on the material behavior of bones using wave propagation technique, analyzed and compared their results with those reported by other workers in this field. They also tried to explain the discrepancies between experimental results and theoretical results estimated on the basis of the classical theory of elasticity in several cases, using the concept of Cosserat elasticity. Subsequently several other researchers also contributed significantly to the growth of bone biomechanics. A comprehensive discussion on the Mechanics of Head Injury and the Fracture and Remodelling Mechanics of bones are available in the recent book published by Misra (2005).

The aim of the present chapter is to provide some useful information on mathematical modelling in Bone Biomechanics along with experimental observations on different physical properties and mechanical behavior of bones. It is believed that the comprehensive material presented in this chapter would stimulate further research in the important domain of biomechanics. Sections 3 and 4 deal with a brief account of the physical properties of ideal solids and the relevant constitutive relations. The subsequent sections include discussions on the properties and relations that are adapted in case of bones. Theoretical formulations of the control mechanism for internal and surface bone remodelling are also presented. The last section gives an idea of the current state-of-the-art of mathematical modelling of some important problems in Bone Biomechanics.

2. Bone Biomechanics and Its Mathematical Analogues

Biomechanics, according to Hatze (1974), may be defined as the study of the structure and function of biological systems by means of the methods of mechanics. The particular subdiscipline which is concerned with studies pertaining to the mechanics of bones is called the Bone Biomechanics. The importance of researches in this important area lies in ascertaining the

mechanical properties of bone tissue with an aim to determine the pathological state of diseased bone, the fracture site etc., in understanding the remodelling processes that living bone continually undergoes in the course of our daily activities, by which bone adapts its histological structure to changes in long term loading and in constructing suitable biomechanical implants for replacements of skeleton joints e.g. vertebral, knee and hip joints.

In the realm of physical sciences, the motions of inanimate bodies, which are its objects of study, can be analyzed within the limits of practical importance but this is not exactly so for biological motions. The truth of this assertion lies in the fact that whereas a set of definite laws and almost a definite knowledge of the forces governing the motions of non-living objects, even in molecular and atomic levels exist, such basic information regarding the forces and the laws governing the biological motions is yet to take a definite shape. Although Medical Physics, in particular and Biomedical Engineering, in general, have undergone an unprecedented development in recent years, because of the lack of accurate knowledge of the basic principles governing the motions of biological elements, our knowledge of the human skeleton as a load carrying system which is so vital for mankind, has been in no more than a prenatal state as compared to our comprehension of the mechanical behavior of technological structures. Even today this assertion continues to be largely true. So it is of utmost importance to analyze the material response of bone, the most important constituent of the skeletal system. Only through a continual comparison of the theoretical predictions derived by using suitable mathematical models for bones with the experimental data obtained from experiments on real bone specimen, the exact principles governing the biological motions may be developed — and finally an exact mathematical model simulating the osseous medium can be constructed. Punjabee (1979) pointed out that such a model combined with experimental data for physical properties of osseous tissue structure or system form a mathematical analogue which even without validation, has the possibility of representing the reality. Some recent experimental studies indicate success to actualize this possibility.

At early stages of researches on bone, it was modelled as an isotropic, homogeneous and linear elastic solid and the mathematical methods of stress analysis used to be employed, based on the simplified rules of Strength of Materials. With the profusion of experimental studies on bone, it has been revealed that bone is rate-sensitive and it should be modelled as a viscoelastic solid. Time-dependence of the mechanical properties or the viscoelastic behavior of bones actually owes its origin to various on-going

physical processes e.g. thermomechanical coupling, piezoelectric coupling, etc. (Lakes *et al.*, 1979) inside the bone. With passage of time, more complicated mathematical models have been proposed to account for the results of experimental observations on the material behavior of bones, carried out in a continuous manner by many researchers. However, it is always possible to work on simplified bone models which provide useful information for many practical situations. On the basis of such mathematical models, it has already been possible to replace the hip and knee joints by artificial structures and attempts are now going on to design and improve many other replacement joints.

It should, however, be borne in mind that bone is not exactly an engineering material — it is a living organ which continually undergoes the processes of growth, re-inforcement and resorption which are collectively referred to as bone remodelling due to which a living bone adapts its histological structure to changes in the case of long term loading. What follows from this is that in formulating a mathematical analogue of a real bone specimen, which can be directly used in the design and construction of prosthetic devices, one must pay due attention to the mechanism of bone remodelling.

It follows from the above discussion that one can mathematically analyze the problems on bone, including its remodelling mechanism, by using the principles of mechanics but utmost importance should be paid to the choice or construction of a suitable mathematical model with due cognizance of experimental data available for different physical properties of bone tissues obtained from the latest experiments on bone specimens. Once this is done, such an analysis can provide information which should be of significant importance for further biomechanical and clinical investigations. Before we proceed further, let us discuss briefly some relevant matters associated with the response of structural solids subjected to external loading conditions. This will be followed by a brief discussion on human skeletal system, composition of bone as well as its mechanical and physical properties. All these are deemed necessary for a better comprehension of the current state-of-the-art of the highly interesting interdisciplinary field of bone biomechanics.

3. Material Response of Structural Solids to External Excitations

In the classical formulation of the theory of elasticity it is generally held that the mechanical energy stored in a solid continuum during a deformation

process initiated by an external mechanical load, is completely recoverable. This implies that the classical theory totally ignores the possibility of energy loss. Thus in the case of a simple conservative system, it is possible to fit the classical elasticity theory exactly into the purview of reversible thermodynamics. But in reality, dissipation of energy is a common phenomenon exhibited in deformable bodies; this is quite apparent from the subsidence of vibrations set up in them due to excitation. This subsidence, as interpreted by Love (1927), is due to the loss of work done against viscous resistances offered by solid continua. So, in a deformation process, the work has to be done not only against the elastic forces with which the molecules in solids are bound together but also against the time and rate-dependent viscous resistances which render regular molar motions into molecular agitations. While the work done against elastic forces is completely recoverable upon withdrawal of the external load (in the classical theory), the work done against viscous resistances is totally dissipated as heat.

There are, however, other interesting effects of viscous resistances on the material response behavior of solids when subjected to mechanical loads. If the specimen of a perfectly elastic solid is subjected to a sudden loading state held constant thereafter, response is supposed to be in the form of an instantaneous deformation which should remain constant. But the common experience is that a real solid, under such a loading state, exhibits an instantaneous deformation followed by a flow process which may or may not remain limited as time progresses. This particular response behavior is termed as creep in solids. Again, if the specimen is subjected to a constant deformation state, the stress developed in it continuously decreases with time through a process which is known as stress-relaxation. Both creep and stress-relaxation in real solids can be explained by assuming induced viscous resistances which are time- as well as rate-dependent. Indeed, such material response behavior, though unnatural from the standpoint of classical elasticity theory, is of common occurrence to people working with so called elastic metals at high temperatures and pressures, and with polymers and biological materials even at ordinary physical conditions. In fact, this type of material response may be described as the inherent property of all solids; it is observable only under favorable physical conditions.

To account for the instantaneous deformation and the steady flow following it, which are respectively the characteristics of a perfectly elastic solid and a Newtonian viscous fluid occurring simultaneously in a solid specimen, an entirely new theory is essential. Such a theory, which combines both the features of material response of real solids subjected to external excitations,

is known as the theory of viscoelasticity. The materials whose mechanical response characteristics are represented through this theory are called as viscoelastic.

It is quite interesting to note that due to the occurrence of such a flow process, the deformation field is dependent on the history of loading state. This is clearly observed when we consider the application of the external load to the specimen in two or more stages, in succession. It is understood that the deformation in the specimen, just after the last stage of loading, is the result of superposition of all the deformations induced at the current time by each increment of the load applied at different times in the stages preceding the last one, together with the instantaneous deformation produced by the last increment. This means that the specimen experiences not only the instantaneous response to the last stage of loading but also the continuing time responses of the other incremental loadings prior to the last one. Thus in order to study the net material response, one must keep an eye on the past history in addition to the current state of loading. Such materials are said to possess memory. The modern theory of viscoelasticity has been built upon the basis of the memory hypothesis. It may, however, be mentioned here that there exist other theories on the mechanical behavior of materials which have a memory of deformation. For example, incremental theory of plasticity accounts for the dissimilar material behavior in loading and unloading conditions in a specimen loaded beyond the elastic limit. This difference in material responses in loading-unloading programmes may be attributed to the memory of deformation in the specimen. In the plasticity theory, the time-scale involved in such programmes is deemed practically unimportant while the theory of viscoelasticity considers specific time- or rate-dependence (Boley and Weiner, 1967).

It is clear from the above discussion that besides the deformation field, in a real solid specimen subjected to a loading state, a temperature field may also be induced due to heat dissipation in it. Of course, a temperature field is also induced in ideal elastic solids due to thermo-mechanical coupling, which is applicable only to the dynamical conditions of loading. But in a viscoelastic specimen the temperature field is present even under statical conditions of loading due to the loss of work done against viscous resistances. It is true that except in cyclical loadings such dissipation is really too small in most solids to effect any appreciable temperature-rise. But theoretically speaking such a temperature-rise, however small, can be accounted for by solving the appropriate energy equation conforming to the physical conditions of a particular thermo-viscoelastic boundary value problem.

In the mechanically loaded specimen of some crystalline solids, the deformation- and temperature-fields are associated, in general, together with an electric field. In such materials, known as piezoelectric solids, the electric field is induced by the deformation field. This thermodynamically reversible phenomenon is known as piezoelectric effect. A temperature field, instead of the deformation field, may also cause a similar effect, which is known as pyroelectric effect. But under isothermal or nearly isothermal conditions, only the former effect is of much importance. Elastic and viscoelastic solids, belonging to some particular classes of crystals exhibit these effects. A dynamical loading induces an electric field accompanied by a magnetic field in elastic solids and an electromagnetic field in viscoelastic solids, even in static loading situations. For many viscoelastic materials of common use, e.g. polymers, some biological elements, etc., both the temperature and the piezoelectric effects induced by the mechanical load, are not too small to be neglected; of course, the induced electrical polarization depends entirely on the degree of crystallization.

So far, we have discussed how the deformation, temperature, electric and magnetic fields are induced in a specimen of real solid subjected to a mechanical excitation. But irrespective of the method of excitation — mechanical, thermal or electrical, all such fields are simultaneously induced in a continuum (particularly in a crystalline solid) and are usually coupled together. Besides these thermodynamically reversible or quasi-reversible processes, some other irreversible processes like electrostriction may also take place in the medium.

It is now apparent that an exact theory of solids should account for all such material responses. Such a theory has been developed by Nowinski (1978) from thermodynamical considerations for elastic solids. One may develop a similar theory for viscoelastic solids from similar thermodynamic reasoning. With reference to an experimental study carried out by Satter *et al.* (1999) for the treatment of bone fracture by pulse electromagnetic fields, Eringen (2004) has developed a general electromagnetic theory of microstretch elasticity considering different material properties for determining different aspects of remodelling in bone, modelled as an elastic solid having interconnected voids, microcracks or stretchable micro elements. However, the mathematical analysis of a particular problem, by considering such a general theory, is hardly tractable. So we shall now present the theories which are somewhat simple and at the same time do not cause much loss to the generality of the results derived therefrom.

4. Deformable Solids

4.1. *Basic concepts*

For the convenience of readers, before describing the theories regarding the material response behavior of solids subjected to external excitations, it is pertinent to present some fundamental concepts relevant to the central theme of our present discussion. Let the coordinates of a material point in a specimen of a solid be X_i and x_i $(i = 1, 2, 3)$ with reference to the rectangular cartesian coordinates X and x respectively. The system X is assumed to be fixed in space, whereas x is attached with the body, both being coincident initially. Then

$$x_i(\tau) = x_i(X_i, \tau), \quad -\alpha \leq \tau \leq t \tag{1}$$

where τ is the time variable and t is the current time. The displacement vector of the point may be given by

$$u_i(\tau) = x_i(\tau) - X_i(\tau). \tag{2}$$

Differentiating (2) with respect to X_j, one obtains

$$\frac{\partial u_i}{\partial X_j} = \frac{\partial x_i}{\partial X_j} - \delta_{ij}. \tag{3}$$

If $\left\| \frac{\partial u_i}{\partial X_j} \right\| \ll 1$, the deformation is said to be infinitesimal. In the infinitesimal deformation theory of solids, usually no distinction is made between differentiation with respect to X_i and that w.r. to x_i. Bearing this in mind, the strain components may be defined as

$$S_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i}) \tag{4}$$

in which a comma before an index denotes differentiation with respect to the corresponding space coordinate.

Let us now consider an element of area, δA within the body or on its bounding surface, with its orientation defined by the unit positive normal $\bar{n}$. If the total force acting on this area be F_i, the stress at the central point of the elementary area δA may be defined as

$$T_i = \lim_{\delta A \to 0} \frac{F_i}{\delta A}. \tag{5}$$

For each orientation of the area δA, there would be a different stress vector. Hence the stress tensor T_{ji} is defined through the following transformation

which relates the components of the stress vectors to the orientation of the elementary surface.

$$T_i = T_{ji} n_i. \tag{6}$$

By using the principle of conservation of linear momentum in a small tetrahedron, one can derive the transformation (6). Similarly the conservation of angular momentum in an element of volume of the specimen gives rise to

$$T_{ij} = T_{ji}. \tag{7}$$

4.2. *Equilibrium equation*

The principles of Newtonian mechanics assert that irrespective of the nature of the continuum and the type of excitations, the total force acting on a body is always zero; this means

$$\int_V \rho \left(f_i - \frac{\partial^2 u_i}{\partial t^2} \right) dV + \int_B T_{ij}\, n_j dA = 0 \tag{8}$$

where f_i is the body force density, $\frac{\partial^2 u}{\partial t^2}$ the acceleration at a point, ρ the density of the material, V the volume of the body and B the area of the surface bounding it. By using Gauss divergence theorem in (8) and considering the equilibrium of each element of the volume, one finds

$$T_{ij,j} = \rho \ddot{u}_i \tag{9}$$

in which a dot denotes differentiation with respect to time. Equation (9) represents the equilibrium of a dynamical system. Under statical or quasi-statical conditions the inertia term in (9) may be neglected, so that the condition of equilibrium reduces to

$$T_{ij,j} = 0. \tag{10}$$

4.3. *Linear viscoelastic constitutive relations: non-piezoelectric materials*

The phenomenological development of the relations among stress, strain and other fields is based on the memory hypothesis (explained in Sec. 3), together with the observation that different material bodies having identical mass and geometrical configuration, respond quite differently to the same excitation. This individual material response behavior is actually due to the difference in internal constitution, which is different for different materials. As a result, the functional form of the relation, commonly known as the

constitutive relation, is the same for all materials, though its constituents change from one material to the other.

By employing the memory hypothesis and using relevant theorems of mathematical analysis, the following linear constitutive relations for viscoelastic materials which do not exhibit piezoelectric effects, have been developed in the form (cf. Christensen, 1971)

$$T_{ij}(x,t) = \int_o^t G_{ijkl}(t-\tau)\frac{\partial S_{kl}}{\partial \tau}d\tau, \tag{11}$$

in which the relaxation functions $G_{ijkl}(\tau)$ characterizing individual material response behavior have the following symmetry and analytical properties:

$$\begin{aligned} &G_{ijkl}(t) = G_{jikl}(t) = G_{ijlk}(t) \\ &G_{ijkl}(t) = 0, \quad -\infty < t < 0; \end{aligned} \tag{12}$$

$G_{ijkl}(t)$ and $\dot{G}_{ijkl}(t)$ being assumed to be defined in $0 \leq t \leq \infty$ and dots denote time-derivatives.

Equation (11) represents the integral form of the linear constitutive relations for viscoelastic solids with most general type of anisotropic material behavior, under isothermal conditions. The integral equations of the type (11) may be rewritten in an abridged form as

$$T_{ij}(x,t) = G_{ijkl}(t) * \dot{S}_{kl}(x,t). \tag{13}$$

By inverting the relation (13), one may write

$$S_{ij}(x,t) = J_{ijkl}(t) * \dot{T}_{kl}(x,t), \tag{14}$$

in which the creep functions $J_{ijkl}(t)$ obey the same symmetry and analytical conditions (12).

For isotropic solids under isothermal conditions, the stress-strain relations (13) reduce to

$$t_{ij}(x,t) = 2G(t) * \dot{S}_{ij}(x,t) \tag{15}$$

and

$$T_{ii}(x,t) = 3K(t) * \dot{S}_{ii}(x,t) \tag{16}$$

where t_{ij} and s_{ij} represent deviatoric stress and strain defined respectively by

$$t_{ij} = T_{ij} - \frac{1}{3}T_{kk}\delta_{ij} \tag{17}$$

and

$$s_{ij} = S_{ij} - \frac{1}{3} S_{kk} \delta_{ij}. \tag{18}$$

$G(t)$ and $K(t)$ are termed as relaxation functions in shear and hydrostatic pressure respectively; they are also subjected to the analytical conditions (12). δ_{ij} is the Kronecker symbol.

At this point, it is very much instructive to observe a close correspondence of the viscoelastic constitutive relations (11) with those for elastic solids given by

$$T_{ij}(x,t) = C_{ijkl} S_{kl}(x,t) \tag{19}$$

in which C_{ijkl} are the elastic moduli. The Laplace transformations of the relations (11) and (19) with respect to time read

$$\bar{T}_{ij}(x,p) = p\bar{G}_{ijkl}(p)\bar{S}_{kl}(p) \tag{20}$$

$$\bar{T}_{ij}(x,p) = C_{ijkl}(p)\bar{S}_{kl}(p), \tag{21}$$

in which p is the Laplace transform variable and a bar over a function denotes its Laplace transform. From (20) and (21), one may conclude that in the Laplace space, the solution of a viscoelastic boundary value problem can be obtained from the transform solution of the corresponding elastic problem, if one replaces the elastic moduli in it by p times the transforms of the viscoelastic relaxation functions. Using the elastic-viscoelastic correspondence principle, the final solution can be obtained just by inverting the transformed solution.

Apart from the integral representations as above, the constitutive relations of a rate sensitive linear material may also be expressed as (cf. Moghe and Hsiao, 1966; Bulanowski and Yeh, 1971)

$$T_{ij}(x,t) = E_{ijkl}(D) S_{kl}(x,t) \tag{22}$$

in which $D \equiv \frac{\partial}{\partial t}$.

It is easy to note that

$$E_{ijkl}(p) = p\bar{G}_{ijkl}(p), \tag{23}$$

if $E_{ijkl}(D)$ are expressible as a power series in D or as the quotient of two such power series in D. In a similar fashion the linear constitutive relations

for isotropic materials, in differential form, are written as

$$P_1(D)t_{ij}(x,t) = Q_1(D)S_{ij}(x,t) \tag{24}$$

and

$$P_2(D)Tkk(x,t) = Q_2(D)Skk(x,t). \tag{25}$$

Here P's and Q's represent power series in D. It is needless to mention that the concept of elastic-viscoelastic correspondence can also be realised by using this differential formulation of the viscoelastic constitutive relations. But the principal advantage of this formulation lies in the fact that using the said principle one can obtain the time-dependent relaxation or creep functions for materials whose response behavior can be characterized by a mathematical model.

5. The Human Skeletal System

The two hundred and six bones, which are highly connective tissues, form the rigid frame-work of the human skeletal system. They are organs that consist of bone tissues, bone membranes and bone marrow. Among the biological materials, bone is the hardest of all. According to shape and size, bones in the skeletal system may be classified into five categories, viz. long, short, flat, seasamoid and irregular. In a long bone, the axial dimension is large compared to transverse one. Such bones are almost cylindrical in shape. They usually have two or more ends termed as epiphysis. In some bones the axial and transverse dimensions are comparable; such bones are termed as short bones. If the transverse dimension of a bone is larger than its axial dimension, it is called a flat bone. It consists of two layers of compact bones, separated by a layer of spongy bone. The short bones that develop in tendons resemble seasame seeds and are called seasamoid bones. Bones whose shapes are irregular and not included in any of the categories mentioned above, are called irregular bones. They include those contained in vertebrae, shoulders etc. In recent researches vertebral joints have been modelled as circular discs of trabecular bone.

Since some problems on long femur or tibia and inter-vertebral joints have been the subject of discussion in this chapter, brief discussions on these two types of bones will be made in the following two sections.

6. Long Bones

In femur and tibia, since the axial dimension is many times larger than the transverse one, they are long bones (cf. Fig. 1). Although microscopically

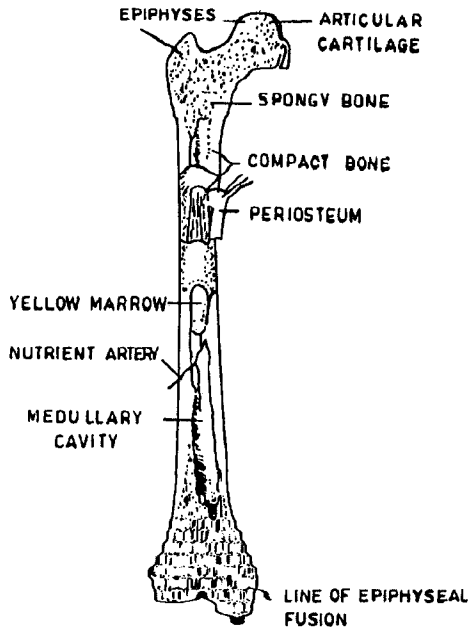


Fig. 1. A typical long bone.

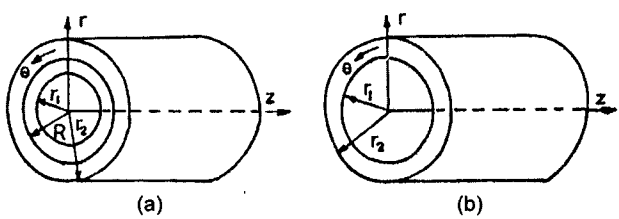


Fig. 2. Geometry of (a) two layered bone, (b) compact bone.

such bone is non-homogeneous and consists of osteons, interstitial tissues and other organic substances embedded in a viscoelastic matrix. A long bone can be described as a composite thick cylindrical shell of two types of materials (cf. Fig. 2) — compact in the outer layer and spongy in the inner layer (Vayo and Ghista, 1971). The outer surface or cortex of the femur is made of the hard tissue, the cortical bone. The cancellous or spongy bone inside the bone medium consists of a network of hard interconnected filaments called trabeculae filled with marrow and a larger number of blood vessels. Cancellous bone is structurally prominent near the joints, i.e. at

the epiphyses while cortical bone is structurally prominent in the middle portion of the femur, the bone diaphysis.

Carter *et al.* (1976) made an observation that adult bone has two or three distinct regions, which is in conformity to Vayo and Ghista's (1971) composite bone model.

The bone cross-section is non-uniform, in general. In the diaphysis region of human femur, it is almost uniform, but for human tibia it gradually decreases from the knee-joint, where the dimension of the cross section has its maximum value (Finlay *et al.*, 1982). For this reason, a human tibia of finite length may be modelled as a truncated conical bar [cf. Misra *et al.* (1989, 1992)].

7. Intervertebral Discs

There are 47 vertebral joints in the human spinal column. Each joint has three substructures (Fig. 3). The first substructure, called the intervertebral disc, lumps the cortical and trabecular bone regions with the end plates.

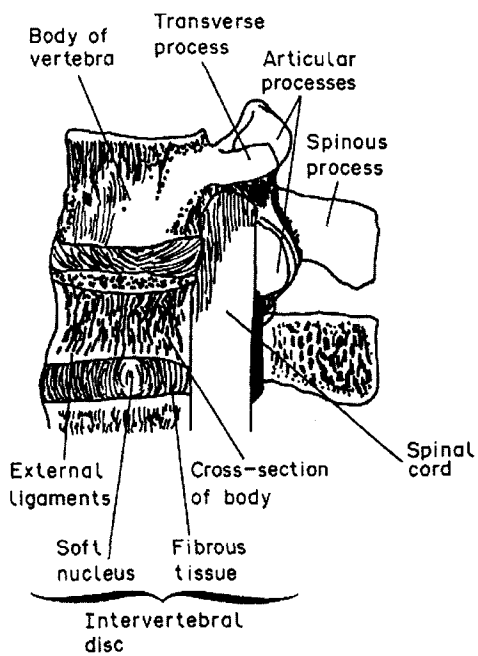


Fig. 3. Body of vertebra, intervertebral disc and the spinal cord.

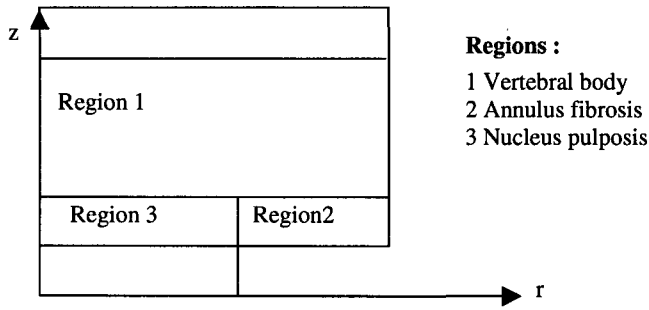


Fig. 4. Simplified model of the vertebral body/intervertebral disc.

The second structure presents the annulus fibrosis of the intervertebral disc and the last one represents the nucleus pulposis, an incompressible fluid.

In analytical studies, the vertebral discs have been supposed to possess circular shape with axial symmetry by some authors (Spilker, 1980; Spilker *et al.*, 1984) while the disc material has been considered to be isotropic, linear and viscoelastic by some researchers (Lu *et al.*, 2004; Lee *et al.*, 2004).

8. Composition of Bone

The composite structure of bone consists of crystalline mineral (hydroxyapatite), amorphous mineral, crystalline organic (collagen), amorphous organic (protein molecule) and liquid phases. The macro structural properties of bone depend on the properties and volume composition of different phases present in them. The major portion of the bone is made of hydroxyapatite and collagen; the rest consists of liquid in haversian canal, canaliculi and lacunae. It is supposed that collagen and organic amorphous phase lump together to form the fibre of the bone composite whereas lumping of hydroxyapatite crystals with the amorphous mineral provides the matrix in which the fibres are embedded. Spaces within the osseous media are interconnected and the flow of liquid through these pores can absorb large amount of energy, thereby increasing the toughness of the bone structure.

9. Microscopic Anatomy of Bone Tissue

The osteons or the haversian systems, which have an irregularly cylindrical and branching structure, form the primary histological units of bone structure. Each osteon is roughly a cylinder of 150 microns in diameter and one

to two centimeters in length. In cancellous bones, the osteons are arranged in a pattern along the line of average principal stress (Koch, 1917). In cortical bones, the osteons are parallel to the axis. The walls of the osteons are thick. There is a fluid-filled hollow or lumen, called the haversian canal, each of forty microns diameter, along the centre of its long axis. Each lumen is provided with a blood vessel required to supply nourishment to the bone cells inside the osteons. The blood vessels and lumina together constitute what is called the haversian system.

Osteoblasts, osteocytes and osteoclasts are the three cellular components of bones which actively participate in its remodelling mechanism. Osteoblasts are of cuboidal shape; they participate in the formation of bones and are attached to their surfaces. Osteoblasts that are present inside the bones are known as osteocytes; they help in the maintenance of bone as a living tissue. Osteoclasts, which are giant cells with a variable number of nuclei perform the function of resorption. Gottesman and Hasin (1980) investigated the mechanical properties of bone as a composite material, on the basis of a micromechanical model. They made an observation that the microscopic structure can be derived only when the exact mechanical properties of bulk bone are known.

10. Physical Properties of Bones

Being the major important constituent of the skeletal system, bone is always subjected to internal and external loads. The stresses and strains induced in it are determined by its mechanical as well as other physical properties. The first study on the mechanical properties of bone was carried out by Wertheim (1847). Rauber (1876) provided the first hint that elastic moduli of bone tissues are direction-dependent and also that they are history dependent (viscoelastic). As mentioned earlier, in the early researches bone was modelled as an isotropic elastic solid. Dempster and Liddicoat (1952) observed that a compact bone should be modelled as a non-isotropic material. Subsequent experiments confirmed the orientational characteristics of the physical properties of hard tissues. Most of the non-isotropic crystalline solids have a tendency to exhibit piezoelectric effects — electric field is induced in them when they are subjected to mechanical loads. Bone is not any exception to that rule. Fukada and Yasuda (1957) investigated quantitatively the piezoelectric effect in bone for the first time. Subsequently, intensive as well as extensive investigations were carried out in order to relate bone piezoelectricity with the mechanism of bone remodelling and search for piezoelectric effects in other biological tissues e.g. tendons, blood

vessels etc. Most relevant studies in this field were reviewed by Fukada (1968) as well as by Güzelsu and Demiray (1979). Like real solids, bone also exhibits viscoelastic material behavior. Early investigations on the viscoelastic material behavior of bone were carried out by Sedlin (1965) and McElhaney (1966). Sedlin, on the basis of his experimental study, proposed a mechanical model — standard linear solid for osseous media. Lakes and Katz (1979c) proposed that bone is a non-linear viscoelastic solid. Nowinski (1971) commented that bone is generally non-homogeneous and anisotropic. Nowinski (1974) put forward a mathematical analysis by assuming power law variation for the physical constants of bone tissues. Carter (1976) observed experimentally that the elastic moduli are connected with the apparent density of osseous tissues, in a non-linear manner. In an experimental study, Morgan *et al.* (2001) observed that bone possesses non-linear elastic behavior even at small strains.

Osseous tissues resemble a lattice structure, in which there are innumerable cavities in a solid skeleton. These cavities are interconnected and the liquid phase is distributed in such a fashion that it has access to a drainage path to the surface of the bone. The difference between the two major types of bone — compact and spongy is rather relative. The difference is owing to the variation in the proportion of the volume of the pores to the volume of the solid matter in a unit volume of bone. Owing to porosity in bone, Nowinski (1970, 1971) considered it to be a two-phase material, the solid skeleton being perfectly elastic (1970) or viscoelastic (1971) and the liquid phase being Newtonian.

It appears that bone has diverse physical properties. Basing upon the research studies carried out by Currey(1984) on the mechanical properties of bone, Roesler (1987) made an observation that "bone can have a wide variety of different mechanical properties, depending upon its function". So it is essential that each of these properties be considered separately, keeping an eye on the nature of the problem. The salient properties are discussed briefly in Secs. 11–15.

11. Bone Anisotropy

Rauber (1876) conjectured that the elastic moduli of bone are direction-dependent. The directional effect of loading on the bone compressive strength was clearly observed by Dempster *et al.* (1952, 1961) in their experimental studies on compact bone in different directions. They reported that the ultimate compressive strength is least in the tangentially loaded femoral bones and greatest in the axially loaded ones. These tests were

carried out by taking cubical specimens of femur. From an experiment on bovine femoral bones, fresh and dried, Lang (1970) determined all the five independent elastic moduli needed to characterize the material behavior of osseous medium. Lang (1970) also established experimentally the existence of hexagonal symmetry in cortical bones by using ultrasound technique for the first time. The theoretical predictions on cortical bones by Yoon and Katz (1976a) were in conformity to the experimental observations of Lang (1970). Yoon and Katz (1976b) also determined experimentally the said elastic moduli for human femoral bones. Reilly and Burstein (1974) conjectured that bone is transversely isotropic. In a subsequent study they (1975) confirmed their hypothesis. Katz (1980) also showed that Young's modulus of bone is direction-dependent. van Buskirk and Ashman (1981) modelled bone as an orthotropic material with nine elastic constants. Using ultrasonic wave technique they measured these constants and came to the conclusion that cortical bone is non-homogeneous.

The symmetry and consequently the anisotropic properties of osseous media result from the nature of its composition. The most important constituents of bone, hydroxyapatite and collagen are anisotropic in nature. Physical properties of bones are mostly derived from the relative orientation and amount of the organic and inorganic crystalline substances contained in them. In an experiment on bovine cortical bone, Lipson and Katz (1984) measured the elastic properties of bones having different histology along three different mutually orthogonal directions with an aim to investigate the influence of microstructure of bone on its mechanical properties. It was found that the plexiform bone is orthotropic while harversian bone is transversely isotropic. Ashman *et al.* (1984) also measured the corresponding elastic constants. They arrived at the conclusion that bone is both anisotropic and non-homogeneous and that the actual material behavior of human compact bone is orthotropic. Cowin and van Buskirk (1986) analysed the data reported by Ashman *et al.* (1984) and confirmed that their results were within the thermodynamic restrictions investigated by Mahanian and Pizialli (1985). In bone biomechanics the term "fabric" has been introduced to describe local anisotropy of a material microstructure. Based on the consideration of the orthotropy of bones, Cowin (1985, 1986) formulated elastic properties as a function of fabric and density. With an aim to analyze fabric dependence of orthotropic elastic constituents of bone, Turner *et al.* (1980), van Rietberger (1995, 1996) and Odgaard *et al.* (1997) conducted further studies in this direction using methods based on finite element techniques.

In order to take into account the anisotropic elastic behavior of osseous tissues, Eq. (19) may be taken as the constitutive relation in the case of bone.

12. Viscoelastic Properties of Osseous Tissues

The creep and relaxation phenomena in real solids which have been discussed in Section 3, are quite common in bones. Rauber (1876) studied creep as well as anisotropic material behavior of bones. On the basis of his experimental observations on bones under loads well below the fracture load, Sedlin (1965) proposed a viscoelastic model to explain the creep and relaxation behavior of bones in which a Hookean spring is in series with a Kelvin model; Kelvin model is that in which a Hookean spring is connected in parallel with a viscous dashpot. Smith *et al.* (1965) investigated viscoelastic properties of bones. Mc Elhaney (1966) also observed the strain dependence of the elastic moduli of bones experimentally and concluded that osseous tissues are viscoelastic.

The viscoelastic material behavior also owes its origin to the viscoelastic properties of the bone constituents — mainly the hydroxyapatite and collagen besides the ongoing physical processes, e.g. thermo-mechanical coupling, piezo-electric coupling etc. [cf. Lakes *et al.* (1979a)]. Gottesman and Hashin (1980) obtained the anisotropic relaxation functions by using the theory of composite materials. It is assumed that collagen together with the other organic phase constitutes the fibres whereas the mixture of the hydroxyapatite and another mineral phase acts as the matrix. In a theoretical study Nowinski (1974) used the generalized Sedlin model (1965) in which the anisotropic material behavior of bone was considered. Several researchers have been working on the stress analysis in the vertebral body by assuming it to be represented by a two-, three- or four-parameter Kelvin model [cf. Spilker (1982), Spilker *et al.* (1984), Furlong and Pulazotto (1983), Burns *et al.* (1980, 1984)].

The linear viscoelastic constitutive relation for one-dimensional stress analysis connecting the stress T with the strain S may be written as (Nowinski, 1971)

$$T = \left(\frac{E + \eta H D}{1 + \eta D} \right) S \tag{26}$$

in which $D = \frac{\partial}{\partial t}$ is the time operator, E, H and η are model parameters.

A relaxation process (mechanical, dielectric, magnetic or piezoelectric) is often described by means of a spectrum of relaxation times. A mechanical

system may possess a single characteristic rate at which it readjusts itself to equilibrium from its disturbed situation. Then the system is said to possess a single relaxation time. In this case its time dependent stiffness coefficient $C(t)$ may be given by Debye model (Lakes and Katz, 1979b), described by

$$C(t) = C_0 + C_1 e^{-t/T_0} \tag{27}$$

where T_0 is the single relaxation time and C_0 , C_1 are model parameters.

By taking both the viscoelastic and anisotropic material behavior of osseous tissues into account, the constitutive relation in the framework of a linear theory may be given by Eq. (11) presented above. In the operator-form the linear anisotropic viscoelastic constitutive relations for osseous media are represented by (23) (cf. Bulanowski and Yeh, 1971).

Lakes and Katz (1974) analyzed the experimental data obtained by Black and Korostoff (1973), Curry (1965), Lugassy and Korostoff (1969), Mc Elhaney (1965), Smith (1965) and Tennyson (1972) in order to compare them as well as to determine the relations among viscoelastic functions in anisotropic materials like bones.

A non-linear study on the viscoelastic properties of wet cortical bones was presented by Lakes and Katz (1979b). Sanjibee (1982) and Sanjibee *et al.* (1982b) observed non-linear viscoelastic effects in collagen, an important constituent of bone. In ordinary solids, as mentioned earlier, viscoelasticity arises due to relative motions of the different layers in solids subjected to load. Apart from this bone viscoelasticity may have other origins. A substantial part of recent research on bones aims at enlisting these causes and their contributions to the overall viscoelastic material behavior of bone. At the molecular level, investigation by Sasaki *et al.* (1995, 1997) suggests that an important constituent of bone, the collagen, a proteinaceous phase along with its water content may give rise to significant viscoelastic effects. Thermomechanical coupling giving rise to heat dissipation in elastic solids may also be a cause for damping of waves in bones, but this has not been well investigated. Lakes and Katz (1979b, c) made an observation that the piezoelectric coupling in hexagonal tissues may be a source of bone viscoelasticity; however, their study revealed that it has negligible contribution to the damping of waves in bones. Prior to these reports, it had been observed by Biot (1941) that fluid flow within bone induces viscoelasticity, while Katz (1980) and Lakes and Saha (1979) observed viscous-like cement line motion contributes significantly to bone viscoelasticity. Piekarski *et al.* (1977), Cowin (1993), Cowin *et al.* (1995) and Cowin (1998) observed that stress induced fluid flow leading to remodelling in bones is the

direct outcome of its viscoelastic properties. Misra and Samanta (1987) also studied the extent to which bone remodelling is influenced by viscoelasticity of osseous tissues.

Due to viscoelastic properties, waves through bones are attenuated and vibrations are damped. Misra and Samanta (1982) put forward mathematical analyses of several problems on waves and vibrations in bones and determined how wave and vibration parameters get modified, if bone viscoelasticity is taken into account. Based on wave propagation technique, Brodt *et al.* (1995) and Garner *et al.* (2000) performed experiments to correlate wave characteristics with bone viscoelasticity in different frequency ranges. While experimenting with compact bones in torsion and bending in the frequency range from 5 mHz to 5 KHz in bending and 5 KHz to 50 KHz in torsion, Garner *et al.* (2000) showed that wet bones are more viscoelastic than dry bones and the results obtained by them confirmed the earlier observation by Lakes *et al.* (1979a) and Sasaki *et al.* (1993). Solomonow *et al.* (2001) investigated lumbar viscoelastic creep due to prolonged cyclic loading. Lu *et al.* (2004) concluded on the basis of experimental observation on vertebral discs that repetitive lumbar loading at fast rates is indeed a risk factor as it induces large creep in the lumbar viscoelastic tissues, which in turn intensifies the resulting neuromuscular disorder. In a recent mathematical analysis based on a finite element model of lumbar interbody fusion under axial loading, Lee *et al.* (2004) have demonstrated that if the mechanical behavior of the intervertebral disc is considered to be viscoelastic, in addition to its composite characteristics, their theoretical predictions are in good agreement with the relevant experimental observations reported by previous researchers. All these studies confirm the validity of the consideration of viscoelasticity of intervertebral discs in an earlier study conducted by Misra and Samanta (1988).

13. Piezoelectric Effects in Bone

As mentioned earlier, in some materials possessing anisotropic material behavior, electric field is induced while they are subjected to mechanical loading. This is known as piezoelectric effect. Fukada and Yasuda (1957) observed and quantified such piezoelectric effects in bones, for the first time. The experimental studies of Anderson and Eriksson (1968, 1970) indicate that collagen in bone continues to exhibit piezoelectric behavior even when the bone specimen is fully hydrated. They further showed that the stress generated potentials, which are strongly related to the rate of deformation in moist bone are neutralised by the conductivity of physiological milieu.

It was reported that at low frequencies, the measurements of Anderson and Eriksson (1970) supported the Maxwell-Wagner type of polarization in bone tissues. Further studies (Gurdijian and Chem, 1974) in this area indicate that electric current even of the order of microampere possesses the potential to cause bone deformation. It was inferred that electric current originated due to bone piezoelectricity produces similar effects in the physiological state. Like other physical properties (e.g. anisotropy and viscoelasticity), bone piezoelectricity also owes its origin to its basic constituents. Osteon, the unit of bone histology, possesses hexagonal symmetry and exhibits piezoelectric effects. However, bioelectrical signals are always generated in some biological tissues, while performing their physiological functions. Intensive as well as extensive researches were undertaken to ascertain whether electricity induced in bones was due to piezoelectric effects or it had any other origin. Of the three possible origins of bioelectricity viz. piezoelectric effect, streaming potential effect and p-n junction effect, the piezoelectric effect has the primary contribution to bio-electricity in dry bones. Even in physiologically moist bones it is mainly of piezoelectric origin (cf. Güzelsu and Demray, 1979).

The piezoelectric effect exhibited by certain materials which lack a centre of symmetry, may be defined as the production of an electrical response due to mechanical excitation and vice versa (Cady, 1946). The coupling between the mechanical deformation and electric polarization in elastic solids, may be described by the relations, (cf. Cady, 1947; and Güzelsu, 1978)

$$T_m = C^E_{mn} S_n - e_{km} E_k \tag{28}$$

$$D_k = e_{km} S_m + \varepsilon^S_{ki} E_i, \quad (1 \leq m, n \leq 6; 1 \leq i, k \leq 3) \tag{29}$$

in which E_i is the electric field, D_i the electric displacement, C^E_{mn} the stiffness matrix at zero electric field, e_{km} the piezoelectric coefficient matrix and ε^S_{ki} the dielectric tensor at zero strain.

By considering the mechanical, dielectric and piezoelectric relaxations, in certain semi-crystalline polymers including those of biological origin, the above constitutive relations get modified to

$$T_m = \int_{-\infty}^{t} \left[C^E_{mn}(t-\tau) \frac{\partial S_n}{\partial \tau} d\tau - e_{km}(t-\tau) \frac{\partial E_k}{\partial \tau} d\tau \right] \tag{30}$$

and

$$D_k = \int_{-\infty}^{t} \left[e_{km}(t-\tau) \frac{\partial S_m}{\partial \tau} d\tau + e_{ki}(t-\tau) \frac{\partial E_k}{\partial \tau} d\tau \right] \tag{31}$$

in the convolution form. In operator formalism, they assume the form:

$$T_m = C^E_{mn}(D)S_n - e_{km}(D)E_k \tag{32}$$

and

$$D_k = e_{km}(D)S_m + \varepsilon_{ki}(D)E_k. \tag{33}$$

The frequency dependence of dielectric matrix is of common occurrence. The piezoelectric relaxations in polymers were studied by Furukawa and Fukada (1976). Pfeiffer (1977) invoked the idea of Thomson model to represent both the mechanical and piezoelectric relaxation in osseous tissues. In the problems of wave propagation and vibration one may take (cf. Bulanowski and Yeh, 1971)

$$C^E_{mn}(D) \equiv C^E_{mn}(i\omega)$$

when the medium is disturbed sinusoidally with a single frequency ω. One can also write

$$C^E_{mn}(i\omega) \equiv C^E_{mn}(\omega)[1 + i\delta(\omega)], \tag{34}$$

in which C^E_{mn} is the storage modulus and δ the loss tangent. It has been reported by Lakes and Katz (1979a) that for wave propagation problems in human cortical bones in the ultrasound frequency range, both storage modulus and loss tangent are independent of frequency. As a consequence of piezoelectric effects, mechanical strain waves propagating in a given specimen of bone along most directions are accompanied by induced electric fields which are in phase with the mechanical motion (if viscoelastic effects are disregarded). It is evident from the above constitutive relations that the electric field induced at high frequencies is accompanied by a magnetic field. The induced fields in the bone monitors effectively the characteristics of the traveling wave and can lift the equilibrium between osteogenic and osteoclastic activities (Park, 1979). If properly oriented, the fields can also induce calcium ion movement. Hence bone piezoelectricity can play an important role in fracture healing. Demiray (1983) put forward an analysis for studying the electromechanical remodelling of bone tissues, by considering wet bone to be composed of a charged fluid and the solid bone matrix which possesses the piezoelectric property. Saha and Lakes (1977b) designed an experiment to detect the induced magnetic field accompanying the travelling ultrasonic wave for the purpose of determining the physical properties by a completely invasive technique. For this purpose theoretical investigations were carried out by Güzelsu and Saha (1981), Misra and Samanta (1983a, 1983b, 1988) and Misra *et al.* (1988).

14. Bone Inhomogeneity

Nowinski (1971) emphasized that like most biological materials, a real bone is non-homogeneous. With this in mind he proposed a power law of variation for the physical properties of bone material, by using which Nowinski (1974) carried out stress analysis in a cylindrical bone specimen, treating bone tissue as anisotropic and viscoelastic. The relations he proposed in this study are as follows:

$$C_{ij}^{o} = C_{ij} r^{p} \quad \text{and} \quad \rho_o = \rho r^{p} \tag{35}$$

in which r is the radial co-ordinate of a representative point in the bone specimen, C_{ij}^{o} the elastic moduli, ρ_o the density, C_{ij}, p and ρ are material parameters.

Patel (1969) pointed out that for porous materials, the modulus C is related to the apparent density ρ_o (mass per unit volume of the bone specimen including voids) according to the law:

$$C = C_o \rho_o^{b} \tag{36}$$

where C_o and b are constants. From compression tests on cylindrical specimens of human and bovine trabecular bones, Carter and Hayes (1976) found that $b = 2$ for compressible strength. This is in agreement with the theoretical prediction made by Patel (1969) for cellular plastics and porous materials like bones. For shear strength, the value of the exponent obtained theoretically by Patel (for foams) is 1, whereas the experimental value of "b" for bones, as reported by Stone *et al.* (1982) is 1.65. It was also observed by Patel (1969) from his experimental studies on the shear strength of foams (which are porous material like bones), that the value of the exponent ranges from 1.0 to 1.5. The assumption of a linear variation of shear strength with density (i.e. $b = 1$) virtually leads to similar power laws for the shear modulus and density [see relations (35)].

15. Bone Remodelling

As indicated earlier, living bones cannot be treated just as an ordinary engineering material. The reason for this is that the microscopic structure of engineering materials remains unaltered for all time under the influence of external mechanical loads, whereas living bones undergo the continual processes of growth, reinforcement and resorption, which constitute the mechanism of bone remodelling by which the bone adapts its histological structure to changes in long term loading. Frost (1964) classified the remodelling processes as internal remodelling and surface remodelling. As early

as 1892, Julius Wolff, the German anatomist was the first scientist who carried out intensive research on this problem. The results of his experimental findings are embodied in what is called Wolff's law. According to this law, bone remodelling is dependent on strain or stress. The mechanism by which the bulk density of an osseous medium changes within fixed boundaries is termed as internal bone remodelling, while the process of bone deposition on the periosteal surface is known as surface remodelling.

Electrical and chemical properties of bones were found to influence the remodelling processes (Basset and Becker, 1962; Shamos, 1963; Becker and Murray, 1970; Basset, Pawlick, and Becker, 1964; Justus and Luft, 1970). Gjelsvik (1973a) postulated that the surface aspect of bone remodelling is governed, at least in part, by the piezoelectric polarization produced in a deformed bone. The internal remodelling, according to him, is a second mechanism aimed at making the material direction aligned with the primary stress distribution throughout the volume of a bone specimen, if there is any misalignment between the two directions. According to the Gjelsvik model of bone remodelling, the piezoelectric polarization (P_i) in the bone tissue due to a stress field (T_m) caused by the external loads may be determined from the simplified relation

$$P_i = d_{im}\, T_m \quad (i = 1, 2, 3;\ m = 1, 2, \ldots, 6) \tag{37}$$

in which d_{im} are the piezoelectric constants for bone tissues.

Apart from piezoelectric polarization, several mechanisms have been proposed for the transduction of mechanical loads to the remodelling response such as streaming potentials, alterations in mineral solubility due to stress, mechanical fatigue microdamage, extracellular fluid pressure effects on bone cells and direct load on bone membrane (Treharne, 1981). Each of these mechanisms is supported by experiments. Cowin and Hegedus (1976), and Hegedus and Cowin (1976) developed performed theoretical studies, by using the theory of adaptive elasticity, basing upon which Cowin *et al.* (1976–1981) explained the bone remodelling mechanism considering its chemical origin. They presented a porous and chemically reacting solid model for bone tissue to interpret its remodelling mechanism [cf. Cowin (1976), Cowin and Nachlinger (1978)]. They developed the models by using the concept of small strain elasticity under isothermal conditions. According to these models,

$$T_i = (\xi_0 + e) C_{ik} S_k, \quad (i, k = 1, 2, \ldots, 6) \tag{38}$$

in which ξ_0 is the reference volume fraction of bone matrix material, "e" measures an increment of this fraction from ξ_0 and all other symbols retain

their earlier meanings. Cowin and van Buskirk (1978) used this model to analyze the problem of internal remodelling induced by a medullary pin. In a later communication Cowin and Van Buskirk (1979) proposed a theory of surface remodelling and applied it to determine the changes in the endosteal and periosteal diameters of the bone specimen, modelled as a hollow cylinder. In this theory it was assumed that the normal speed of the remodelling surface u at a point Q, for small strain, is proportional to the strain tensor $S_{ik}(Q)$ measured from a reference value $S^0_{ik}(Q)$, so that

$$\vec{u} = C_{ik}(\vec{n}, Q)[S_{ik}(Q) - S^0_{ik}(Q)], \tag{39}$$

where $C_{ik}(\vec{n}, Q)$ are surface remodelling rate coefficients which are, in general, dependent upon the point Q and the normal $\vec{n}$ to the surface at Q.

Cowin and Firoozbaksh (1981) applied this theory of surface remodelling to predict the shape evolution of right hollow circular cylinders — the idealized models of the diaphyseal region of a long bone. Based on finite element method, Hurt *et al.* (1982, 1984) developed a computational model for the same situation. Cowin *et al.* (1985) determined the values of surface bone remodelling coefficients for combined axial loading and bending for actual bone cross-sections. D'Antoni (1987) developed a computer simulation model for a similar problem to the one mentioned above. Cowin (1987) proposed a theory of surface remodelling to include the effects of shearing strains as well as normal strains.

As discussed earlier, bone is viscoelastic and this material behavior is expected to have significant effect on bone remodelling dynamics. Misra *et al.* (1983, 1987) developed mathematical models and made pioneering contribution in this study of viscoelastic effects of osseous tissues on bone remodelling. Misra and Murthy (1983) analyzed the viscoelastic effect of osseous tissues on the physiological processes of internal bone remodelling induced by a medullary pin. The mathematical analysis they performed was, however, very much involved. Based on the remodelling equations (38) and (39), Misra and Samanta (1987) proposed an approximate method for solving remodelling problems by considering material damping behavior of bone tissues, in which undue mathematical complication could be avoided without diluting the physical relevance of the problem. It is worthwhile to mention that bone being dissipative, the condition of isothermality inherent in the remodelling theories of Cowin *et al.* is maintained. The study conducted by Misra and Samanta (1987) reveals that the effect of material damping behavior on the strain values is significant, particularly at intermediate times for smaller relaxation times.

Misra *et al.* (1989) also developed a mathematical model to study the internal bone remodelling mechanism in a specimen of long bone, by paying due attention to the non-isotropic elastic property of osseous tissues and the non-uniformity of the cross section of a long bone. The process of remodelling was considered to have been initiated by the force fitting of a metallic pin into the medulla of the bone specimen. The results of numerical simulation of this study clearly indicate that both the non-isotropy of bone tissues and the cross-sectional non-uniformity bear the potential to affect the remodelling mechanism in quantitative as well as qualitative terms, to a significant extent. Misra *et al.* (1992) constructed and analyzed another mathematical model for studying the remodelling of the diaphyseal surfaces of the specimen of a long tubular bone induced owing to the force fitting of a pin as a part of a surgical operation. This study reveals that at different transverse sections of the bone specimen, the nature of remodelling and its saturation were different. Using surface bone remodelling law, given by Eq. (39) and the boundary element method (BEM), Sadegh *et al.* (1993) studied bone remodelling along an implant interface. Martinez *et al.* (1998) reformulated this method by combining it with sensitivity and optimization method to efficiently model bone in-growth into a slot of an implant.

After reviewing many research studies on the functional adaptation of bone subjected to load and related problems, Ramtani and Zidi (2001) made an observation that throughout life, bone is continuously turning over by the well regulated process of bone formation and resorption. Bone damage occurring during daily activities of life is normally repaired in a continuous manner by means of remodelling processes. When an imbalance in this remodelling process occurs, bones may become more susceptible to fracture. With this observation, the authors made an attempt to formulate theoretically the competition between damage and internal remodelling in bones, under the purview of the general framework of continuum thermodynamics with reference to the theory of adaptive elasticity (cf. Cowin and Hegedus, 1976; Hegedus and Cowin, 1976). Their formulation was based upon a mathematical model developed by Prendergast and Taylor (1994) for prediction of bone adaptation by taking care of damage accumulation. They remarked that their formulation would provide a better understanding of bone remodelling induced by a medullary pin, if all the constants and the exact form of damage evolution during the adaptation process were known.

In another study, Hazelwood *et al.* (2001) developed a constitutive model for bone remodelling by incorporating several relevant mechanical and biological processes. They used this model to address differences in the remodelling behavior as a volume element of bone is placed in disuse

or overload. The authors claimed that it was a complete model for bone remodelling that had the potential to help studying bone diseases and their treatment.

It appears from the above discussion that the assumption of stress or strain dependence of the bone remodelling mechanism is common to all the models proposed so far. In determining experimentally the relationship between elastic properties and microstructure of bovine cortical bones, Lipson and Katz (1984) offered an indirect validation of this assumption. They observed that the level of osteonal remodelling is related to the pattern of mechanical stress.

16. Current State-of-the-Art

The concept of viscoelastic behavior of materials, in the realm of solid mechanics, is not of recent origin. The celebrated physicists like Maxwell, Kelvin and Voigt contributed much to the initial development of the theory of viscoelasticity more than a hundred years ago. But the general development and broader applications of the theory are of relatively recent occurrence. This is due to a rapid growth of research on polymers and materials of biological origin, which are mostly viscoelastic, during the last three decades. It is now widely accepted that polymers and metals at high temperature exhibit viscoelastic behavior, and a general theory for all such materials, linear or non-linear, are well-established. The excellent monograph of Bland (1960), Nowacki (1962), Boley and Weiner (1967), Ferry (1970), Christensen (1971), Findley *et al.* (1976) may be cited as the general references in this extensive field of research involving ordinary materials.

The researches on the viscoelastic properties of bones, however, are not so extensive. As remarked by Herman and Liebowitz (1972), the exploration of bone as a viscoelastic material is still in its infancy. After McElhaney (1961) and Sedlin (1965), Nowinski (1974) analyzed a viscoelastic problem involving a long bone by using the bone model proposed by Sedlin (1965). In a series of papers Lakes *et al.* carried out investigations on the viscoelastic cortical bone. In their first communication, Lakes *et al.* (1979) dealt with the dynamic measurements of relaxation properties in torsion. They observed that at high frequencies, in the ultrasound range, the material parameters are independent of frequency. In the second one, Lakes and Katz (1979a), discussed the possible contribution of the ongoing physical processes in bone to its relaxation mechanism. Non-linear viscoelastic constitutive relation for cortical bones was the subject of their third study (cf. Lakes and Katz, 1979b). Assuming bone to be a two-phase, fibre-reinforced

composite material, Gottesman and Hashin (1980) determined analytically the five relaxation functions by means of a micromechanical model. Pelker and Saha (1983) carried out an experiment on the propagation of stress waves in bones, modelled as a viscoelastic solid. By assuming that the material behavior of trabecular bone which is the chief constituent of the vertebral body at the intervertebral joints in the human skeletal system to be represented by two-, three- or four-parameter Kelvin model, a series of researches were made by various researchers [cf. Spilker (1980), Burns and Kaleps (1980), Furlong and Palazotto (1983), Spilker, Dangirda and Schultz (1984), Burns *et al.* (1984), Dangirda and Schultz (1984)]. These studies reveal that three- and four-parameter Kelvin models yield better results.

Sanjibee *et al.* (1982a, b) established the non-linear constitutive relations for a specimen of collagen, an important constituent of bone.

As mentioned earlier, the semi-crystalline polymers and materials having biological origin possess another interesting material behavior — they exhibit piezoelectric effects. Having discovered bone piezo-electricity, Fukada and Yasuda (1957) concluded that this was mainly due to collagen, the crystalline organic constituent present in bones. Following the observation made by Fukada and Yasuda (1957), many investigators tried to correlate bone piezoelectricity with the mechanism of bone remodelling [cf. Bassett *et al.* (1962), Cochran (1966), Cochran *et al.* (1968), Samos *et al.* (1963), Bassett (1965) and Marino *et al.* (1970)]. Bioelectricity in bones may have different origins, but Anderson and Ericksson (1970) tried to establish in a convincing manner that bone piezoelectricity is its primary source. Excellent reviews of the researches in this field of study are provided in Fukada (1968), Bassett (1971), Liboff *et al.* (1973, 1974) and Güzelsu and Demiray (1979).

As innumerable voids are present in bone, it is generally non-homogeneous. Nowinski (1974) described bone inhomogeneity by a power law variation. He assumed that elastic moduli are proportional to the apparent density of bone. But Patel (1969) theoretically proposed a power law of variation of elastic modulus with apparent density, for porous materials. Carter *et al.* (1976) experimentally obtained such power law of variation of bone compressive strength with its apparent density. Later on, Stone *et al.* (1982) experimentally confirmed the validity of this law for shear strength of osseous tissue.

Julius Wolff (1892) was the first to observe the strain-dependence of the mechanism of bone remodelling. Most of the experimental studies that

were aimed at relating bone remodelling with bone piezoelectricity have already been mentioned in Sec. 13. In his theoretical study, Gjelsvik (1973a) proposed a macrobone model by assuming that bone remodelling is effected mostly by piezoelectricity of osseous tissues. He used this model to explain the surface and internal bone remodelling and claimed that his results were justified from experimental and clinical observation on bone remodelling and the piezoelectric properties of bone. In a second study Gjelsvik (1973b) discussed the equilibrium and non-equilibrium forms of bone architecture by using his model.

Cowin *et al.* (1976–1981) explained the bone remodelling mechanism from a different standpoint. Having assumed its chemical origin, they presented a porous and chemically reacting elastic solid model for bone tissues to interpret its remodelling mechanism (Cowin, 1976; Cowin and Nachdinger, 1978). The concept of small strain elasticity under isothermal condition has been used in developing their model. Cowin and van Buskirk (1978) used this model to analyze the problem of internal remodelling induced by a medullary pin. In a later communication, Cowin and van Buskirk (1979) proposed a theory of surface remodelling of bone and applied it to obtain the changes in endosteal and periosteal diameters of a cylindrical bone specimen. In this study, remodelling was considered to be induced by an axial load or by a medullary pin. Cowin and Firoozbaksh (1981) made some theoretical predictions on the remodelling of diaphyseal surfaces under constant load.

The balance between local remodelling and accumulation of trabecular bone micro damage is believed to play an important role in the maintenance of skeletal integrity. However, the local mechanical parameters associated with micro damage initiation are not well understood. Trabecular bone micro damage and micro structural stresses under uniaxial compression were studied by Nagaraja *et al.* (2005).

Miller and Fuchs (2005) examined the effect of trabecular curvature on the stiffness of trabecular bone. They used simplified structural models of trabecular bone to model various forms of variability. The structural effects of variability of direction, length and thickness of the trabeculae have been studied using "lattice-type" finite element models.

For choosing intramedullary devices, the surgeon has to take care of the factors that may affect the functional outcome, viz. the strength of the inserted device, bone quality and fracture reduction. Sheer body weight and muscle contraction can result in nail failure, as well as bone penetration due to improper positioning or bone quality. Steinberg *et al.* (2005) studied

experimentally the biomechanical properties of the nail and peg as well as the influence of the peg's expansion upon cadaveric femoral heads.

The ability to accurately assess bone quality *in vivo* is essential for improving the diagnostic and therapeutic goals for bone loss from such varied etiologies as osteoporosis, micro gravity, bed rest, or stress-shielding from an implant. Early diagnostic ability is very important because the effectiveness of treatment diminishes with disease progressing, yet patients are rarely symptomatic before considerable bone loss has occurred and sometimes not until the first fracture has occurred (Davidson, 2003; Homminga *et al.*, 2004).

Stanczyk (2005) presented a study on modelling of PMMA bone cement polymerization. The model is constructed in such a way as to mimic the chemical processes taking place in the cement dough. On the basis of this study he could identify some important phenomena and put forward a mathematical formulation.

So far we have given a brief account of various earlier researches on bone, particularly on its anisotropic, viscoelastic, piezoelectric, non-homogeneous material behavior and its remodelling mechanism. Before we conclude this section, we present a brief review on previous researches devoted to waves and vibrations in bones.

It is known that studies on waves and vibrations in a continuum bear the potential to provide useful techniques for determining its elastomechanical, electromechanical, thermo-mechanical and damping characteristics. But for an osseous medium, which can adapt its shape according to loading conditions, such a study offers information regarding the pathological state, the site of fracture and the remodelling process in addition to its mechanical and electrical properties. Moreover, such information can be obtained for *in vivo* situations.

In fact due to its piezoelectric material behavior, all such information can be derived by studying the electromagnetic waves radiated by bone, when disturbed by travelling elastic waves, using some appropriate monitoring devices [cf. Saha and Lakes (1977a, b), Güzelsu and Saha (1981, 1984)]. Güzelsu and Saha (1983) made an investigation on the diagnostic capacities of flexural waves in wet bones. Chen and Saha (1987) developed a mathematical model for stress wave propagation in a long bone. Having used the data for human femurs for different age groups, reported by Martin and Atkinson (1977), they made an observation that osteoporosis sets in around the age of 55. They further suggested that the diagnostic methods based on wave propagation characteristics may be potentially useful in

detecting the onset of osteoporosis changes in the human skeletal system. Based on waves and vibration techniques, there have been different studies with a definite aim to assess bone pathology and monitor the rate of fracture healing [cf. Lewis (1975), Lewis and Goldsmith (1975), Wong *et al.* (1976), Wright *et al.* (1981), Pelker and Saha (1983, 1985), Rubin *et al.* (1984 a, b)]. Misra and Samanta (1984) put forward a mathematical analysis for a problem of wave propagation in tubular bones.

By considering the damping material behavior of the skull and brain tissues, the effects of various pulse shapes in vehicular impact situations were studied by Misra *et al.* (1977, 1978b). Misra (1978a) carried out a detailed study in order to examine the effect of triangular pulse, which is one of the most frequently occurring pulses during accidents caused due to collision of the head with a hard surface. Misra (1978c) put forward a theoretical study on the deformation of human head impacted by an external load. The study takes care of the eccentricity of the skull structure and also the viscoelastic properties of osseous tissues. Misra and Murty (1979) put forward a theoretical estimate of the intensified stress-field generated in the neighbourhood of a crack in a human-sized skull, while Misra and Mishra (1984) reported similar estimates for an exterior star-shaped crack in a bone medium. Misra (1986b) reported the results of his theoretical study on the distribution of stresses in a pre-cracked bone specimen.

Misra and Chakravarty (1984a) modelled the human skull as a poroelastic shell in their study of a problem concerned with head injury. In order to take care of the eccentricity of the human skull, Misra and Chakravarty (1984b) modelled the skull as a prolate spheroidal shell in their study of rotational brain injury.

Misra (1985) carried out a theoretical analysis on the distribution of stresses in a tubular bone exposed to heat radiation from a distant heat source. Misra (1986a) studied theoretically the stress-field in the human skull generated due to thermogenesis. In this study he made use of the temperature-distributions in the human head reported by Thron (1956) and the results presented by Richardson and Whitelaw (1968) for their problem of heat conduction in the head.

The three-layered structure of the skull bone was duly taken care of by Misra and Roy (1988) in their mathematical analysis of the free and forced vibrations of the cranial vault. They studied four different types of pulse shapes, viz. square pulse, half-sine pulse, triangular pulse and skewed pulse, which are reported to be encountered quite frequently in vehicular impact situations.

A study on local piezoelectric polarization of human cortical bone as a function of stress frequency was carried out by Pfeiffer (1977). Saha and Güzelsu (1981) derived the magnetic field induced by a travelling antisymmetric wave propagating in a single layered long bone. Pelker and Saha (1983) studied the viscoelastic effects on wave propagation in bones. Misra and Chakravarty (1982) analyzed the resonance spectrum of free vibrations of the human cranial system with due consideration to the damping material behavior of osseous tissues.

By modelling long human bone as a solid or hollow circular cylindrical shells of uniform cross-section, Jurist *et al.* (1970, 1973), Doherty and Wilson (1974) calculated the resonance frequencies assuming bone to be an elastic solid. However, no attention to different types of resonances and their attenuation with time was paid in these studies. Collier *et al.* (1982) studied the resonances with the objective of identifying various resonances in long human tibia *in vitro* with the consideration of constant cross-section of isosceles triangle. In this study too, damping of resonances was not considered.

Since materials with hexagonal symmetry are piezoelectric and since osseous media are viscoelastic, Misra and Samanta (1983a) examined the effects of these material properties on the wave propagation characteristics of a bone specimen. This study corresponds to a situation in which both the endosteal and periosteal surfaces of the long bone specimen are maintained at zero electric potential and are free from traction. The study demonstrated that though the piezoelectric properties do not have appreciable effect on wave propagation characteristics, they are useful to determine the induced radial polarization that affects bone remodelling, if one uses the method developed by Gjelsvik (1973). Misra *et al.* (1988) put forward theoretical estimates for the effects of inhomogeneity of bones on the wave propagation constant as well as on the remodelling processes. In a separate study Misra and Samanta (1983b) studied the torsional waves in long biphasic bones having viscoelastic and piezoelectric properties with material inhomogeneity. The authors determined the effects of material damping behavior on wave propagation constants and reported their theoretical estimates for the induced electric and magnetic fields. They discussed possible influence on bone remodelling processes. The study has been extended to a specimen of polymethyl methacrylate (bone cement) having geometry as that of the bone specimen. It was found that the effects of material inhomogenity and viscoelasticity are more prominent in the case of PMMA and dispersion of waves in it were also more significant, particularly at higher modes. It has been argued by Park *et al.* (1986) that owing to these differences in material properties between bone and PMMA (bone

cement), the two interfaces, viz. prosthesis-bone and bone-bone cement are vulnerable sites for loosening at the joints.

In order to derive the micro-structural composition of osseous media from bulk material characteristics, such as modulus of elasticity, piezoelectric coefficients etc. one needs the complete knowledge of all such material parameters. But it is not possible to determine all of them from a single experiment. With this end in view, Misra *et al.* (1989) analyzed the thickness vibration of long bone, considering different properties of osseous tissues as in their earlier studies on bones. This study bears the potential to devise an electromechanical transducer that could generate ultrasound waves within the body itself and is also useful for correlating certain material properties with vibration characteristics of bones. By considering the material non-homogeneity as well as the dissipative material behavior of osseous tissues, Misra *et al.* (1986) studied theoretically the vibration characteristics of a tubular bone in axial planes and conjectured that the stress wave propagation in bones results in induced electric and magnetic fields, the magnitudes of which depend upon the frequencies of the stress waves. They studied the frequency spectrum of a vibrating bone specimen by using experimental results on the variation of elastic constants with density of osseous tissues, reported by Patel (1969).

Since a real bone specimen is of finite length and non-uniform cross section, Collier *et al.* (1982) considered bone to have a non-uniform cross section. Finlay *et al.* (1982) assumed bone to have different circumferential dimensions at different locations. In analyzing the stress field inside the wall of carotid sinus von Maltzahn (1982) assumed it to be cone-shaped. Misra and Samanta (1988) treated human tibia as a truncated conical bar. This model took care of the finiteness of the length of bone specimen as well as non-uniformity of its cross section in a better manner. The geometrical model for bone specimen was used to analyze the lengthwise vibration of the bone specimen theoretically, with an aim to provide better correlation with the experimentally obtained vibration parameters. In a recent experiment, Garner *et al.* (2000) studied the viscoelastic dissipation in human compact bone, in dry and wet conditions, in torsion and bending in longitudinal and transverse directions.

References

[1] J. C. Anderson and C. Ericksson, *Nature* **227** (1970) 491.
[2] A. A. Armenakas, *J. Acous. Soc. Am.* **38** (1965) 439–446.
[3] A. Ascenzi, *J. Biomechanics* **26** (1993) 95–103.

[4] R. B. C. Ashman, S. C. Cowin, W. C. Van Buskirk and S. C. Riu, *J. Biomechanics* **17** (1984) 349–361.
[5] R. B. Ashman, S. C. Cowin, W. C. van Buskirk and J. C. Riu, *J. Biomechanics* **17** (1984) 349–361.
[6] C. A. L. Bassett, *The Biochemistry and Physiology of Bone*, ed. G. H. Broune (Iind Ed., New York, 1971).
[7] C. A. L. Bassett and R. O. Becker, *Science* **137** (1962) 1063–1064.
[8] C. A. L. Bassett, *Sci. Amer.* **213** (1965) 18–25.
[9] C. A. L. Bassett, R. J. Pawlick and R. O. Becker, *Nature* **204** (1964) 652–654.
[10] R. O. Becker and D. G. Murray, *Proc. Biodynamics Symposium* (Dayton, Ohio, 1970).
[11] M. A. Biot, *J. Appl. Phys.* **12** (1941) 155–164.
[12] J. Black and E. Korostoff, *J. Biomechanics* **16** (1973) 435.
[13] D. R. Bland, *Theory of Linear Viscoelasticity* (Pergamon Press, New York, 1960).
[14] B. A. Boley and R. H. Weiner, *Theory of Thermal Stress* (John Wiley and Sons, Inc., New York, 1967).
[15] E. A. Bulanowski, Jr. and H. Yeh, *J. Appl. Mech.* **38** (1971) 351–362.
[16] M. L. Burns and I. Maleps, *J. Biomechanics* **13** (1980) 959–964.
[17] M. L. Burns, I. Kaleps and L. E. Kazarian, *J. Biomechanics* **17** (1984) 113–130.
[18] W. C. Cady, *Piezoelectricity* (McGraw Hill Book Co., Inc. New York, London, 1946).
[19] D. R. Carter and W. C. Hayes, *Science* **194** (1976) 1174–1176.
[20] D. R. Carter, W. C. Hayes and D. J. Schruman, *J. Biomechanics* **9** (1976) 211–218.
[21] A. K. Chain, R. A. Sigelman and A. W. Guy, *IEEE Trans. Biomed. Engg.* **21** (1974) 280–285.
[22] R. M. Christensen, *Theory of Viscoelasticity* (Academic Press, New York, London, 1971).
[23] G. V. B. Cochran, Electromechanical properties of moist bones, Sc. D. (Med.) Columbia University (1966).
[24] G. V. B. Cochran, R. J. Pawlik and C. A. L. Bassett, *Clin. Orthop. Related Res.* **58** (1968) 249.
[25] R. J. Collier, O. Nadav and T. G. Thomas, *Biomechanics* **15** (1982) 545–553.
[26] G. A. Coquin and H. F. Tiersten, *J. Acous. Soc. Am.* **41** (1967) 921–939.
[27] S. C. Cowin, *Mech. Mater.* **4** (1985) 137–147.
[28] S. C. Cowin, *ASME, J. Biomech. Engg.* **108** (1986) 83–88.
[29] S. C. Cowin, *ASME, J. Biomech. Engg.* **115** (1993) 525–533.
[30] S. C. Cowin and Firoozbaksh, *J. Biomechanics* **14** (1981) 471–484.
[31] S. C. Cowin and D. M. Hegedus, *J. Elasticity* **6** (1976) 313–325.
[32] S. C. Cowin and R. R. Nachlinger, *J. Elasticity* **8** (1978) 285–295.
[33] S. C. Cowin and W. C. van Buskirk, *J. Biomechanics* **11** (1978) 269–275.
[34] S. C. Cowin and W. C. van Buskirk, *J. Biomechanics* **11** (1979) 269–276.

[35] S. C. Cowin and W. C. van Buskirk, *J. Biomechanics* **19** (1986) 85–87.
[36] S. C. Cowin, S. Weinbaum and N. Zeng, *J. Biomechanics* **28** (1995) 1281–1287.
[37] S. C. Cowin, *Int. J. Solids Structures* **35** (1998) 4981–4997.
[38] S. C. Cowin, R. T. Hart, J. R. Balser and D. H. Kohn, *J. Biomechanics* **18** (1985) 665–684.
[39] J. D. Currey, *Exp. Biol.* **43** (1965) 279.
[40] S. D. Currey, *The Mechanical Adaptations of Bones* (Princeton University Press, Princeton, 1984).
[41] J. D' Antoni, Bachelor's Honours Thesis, Biomedical Engineering Department, Tulane University, New Orleans, Louisiana (1987).
[42] M. R. Davidson, *J. Midwifery and Women's Health* **48** (2003) 39–52.
[43] H. J. Demiray, *Int. J. Engg. Sci.* **21** (1983) 1117–1126.
[44] W. T. Dempster and R. F. Coleman, *J. Appl. Physiology* **16** (1961) 355–360.
[45] W. T. Dempster and R. T. Liddicoat, *Am. J. Anatomy* **91** (1952) 331–362.
[46] W. P. Dohetry and B. Wilson, *J. Biomechanics* **7** (1974) 559.
[47] A. C. Eringen, *Int. J. Engg. Sci.* **42** (2004) 237–242.
[48] F. G. Evans, *Mechanical Properties of Bones* (Charles C. Thomas, Springfield, Illinois, 1973).
[49] F. G. Evans, *Stress and Strain in Bones* (Charles C. Thomas, Springfield, 1978).
[50] W. N. Findley, J. S. Lai and K. Onaran, *Creep and Relaxation of Non-linear Viscoelastic Solids* (North Holland Publishing Co., Amsterdam, 1976).
[51] J. B. Finlay and R. U. Repo, *J. Biomechanics* **11** (1978) 379–388.
[52] J. B. Finlay, R. B. Bourne and J. Melean, *J. Biomechanics* **15** (1982) 723–739.
[53] H. M. Frost, *Dynamics of Bone Remodelling in Bone Dynamics*, ed. H. M. Frost (Little and Brown, Boston, 1964a).
[54] H. M. Frost, *The Laws of Bone Structure* (Charles C. Thomas, Springfield, 1964b).
[55] E. Fukada, *Biorheology* **5** (1968) 199–208.
[56] E. Fukada and I. Yasuda, *J. Phys. Soc. Japan* **12** (1957) 1158–1162.
[57] D. R. Furlong and A. N. Pulazotto, *J. Biomechanics* **16** (1983) 785–795.
[58] T. Furukawa and E. Fukada, *Nature* **221** (1969) 1235–1236.
[59] T. Furukawa and E. Fukada, *J. Poly. Sci.* **14** (1976) 1979–2010.
[60] A. Gjelsvik, *J. Biomechanics* **6** (1973a) 69–77.
[61] A. Gjelsvik, *J. Biomechanics* **6** (1973b) 187–193.
[62] T. Gottesman and Z. Hashin, *J. Biomechanics* **13** (1980) 89–96.
[63] D. Gross and W. S. Williams, *J. Biomechanics* **15** (1982) 277–295.
[64] A. A. Gurdijian and H. L. Chem, *IEEE Trans. Biomed Engg.* **21** (1974) 177–182
[65] N. Güzelsu, *J. Biomechanics* **11** (1978) 257–267.
[66] N. Güzelsu and H. Demiray, *Int. J. Engg. Sci.* **17** (1979) 813–851.
[67] N. Güzelsu and S. Saha, *J. Biomechanics* **14** (1981) 19–33.
[68] N. Güzelsu and S. Saha, in *Biomechanics Symposium*, ed. S. L. Y. Woo, and R. E. Mates (1983), pp. 197–200.

[69] N. Güzelsu and S. Saha, *J. Biomech. Engg.* **106** (1984) 262–271.
[70] R. T. Hart, D. T. Davy and K. G. Heiple, *J. Biomech. Engg.* **106** (1984) 342–350.
[71] R. T. Hart, D. T. Davy and K. G. Heiple, in *Advances in Bioengineering*, (Am. Soc. Mech. Engrs., New York, 1982), pp. 123–126.
[72] H. Hatze, *J. Biomechanics* **7** (1974) 189–190.
[73] S. T. Hazelwood, R. B. Martin, M. M. Rashid and J. J. Radrigo, *J. Biomechanics* **34** (2001) 299–308.
[74] D. M. Hegedus and S. C. Cowin, *J. Elasticity* **6** (1976) 337–352.
[75] G. Herman and H. Liebowitz, *Fracture*, ed. H. Liebowitz VIII (AP, New York, 1972).
[76] J. Homminga, B. van Rietberger, E.-M. Lochmuller, H. Weinans, F. Eckstein and R. Huiskes, *Bone* **34** (2004) 510–516.
[77] T. C. Huang and C. C. Huang, *J. Appl. Mech.* **38** (1971) 515–521.
[78] R. J. Jendruko, Chang, Chao-Jan and W. A. Hyman, *J. Biomechanics* **10** (1977) 493–503.
[79] J. M. Jurist, *Phys. Med. Biol.* **15** (1970) 417–426.
[80] J. M. Jurist and K. Kianin, 19730, *J. Biomechanics* **6** (1973) 49–257.
[81] R. Justus and J. H. Lutt, *Calc. Tiss. Res.* **5** (1970) 222–235.
[82] J. L. Katz and V. C. Mow, *Biomat. Med. Dev. Art. Org.* **1** (1973) 575–638.
[83] J. C. Koch, *Am. J. Anatomy* **21** (1917) 177–293.
[84] R. S. Lakes and J. L. Katz, *J. Biomechanics* **7** (1974) 259–270.
[85] R. S. Lakes, J. L. Katz and S. S. Sternstein, *J. Biomechanics* **12** (1979a) 657–678.
[86] R. S. Lakes and S. Saha, *Science* **204** (1979) 501–503.
[87] R. S. Lakes and J. L. Katz, *J. Biomechanics* **12** (1979b) 679–688.
[88] R. S. Lakes and J. L. Katz, *J. Biomechanics* **12** (1979c) 689–698.
[89] S. B. Lang, *Science*, New York, **165** (1969) 287–288.
[90] S. B. Lang, *IEEE Trans. Biomed. Engg.* **17** (1970) 101–105.
[91] K. K. Lee, E. C. Teo, F. K. Fuss, V. Vanneuville, T. X. Qiu, H. W. Ng, K. Yang and R. J. Sabitizer, *IEEE Trans. Biomed. Engg.* **51** (2004) 393–400.
[92] J. L. Lewis, *J. Biomechanics* **8** (1975) 17–25.
[93] J. L. Lewis and W. Goldsmith, *J. Biomechanics* **8** (1975) 27–46.
[94] A. R. Liboff and M. H. Shamos, in *Biological Mineralization*, ed. I. Zipkin (Willey, New York, 1973).
[95] A. R. Liboff and R. A. Rinoldi (Eds.), Annals of the New York Acad. Sci., 238 (1974).
[96] S. R. Lipson and J. L. Katz, *J. Biomechanics* **17** (1984) 231–240.
[97] A. E. H. Love, *Theory of Elasticity* (Cambridge, 1927).
[98] D. Lu, M. Solomonow, B. Zhou, R. V. Baratta and L. Li, *J. Biomechanics* (2004) 845–855.
[99] A. A. Lugassy and E. Korogsoff, in *Research in Dental and Medical Materials* (Plenum, N. Y., 1969).
[100] S. Mahanian and R. Piziali, *J. Biomechanics* **18** (1985) 77–78.
[101] A. A. Marino and R. O. Becker, *Nature* **228** (1970) 473.
[102] R. B. Martin and P. J. Atkinson, *J. Biomechanics* **10** (1977) 223–231.

[103] M. Martinez, M. H. Aliabadi and H. Powl, *J. Biomechanics* **31** (1998) 1059–1062.
[104] W. P. Mason, *Piezoelectric Crystals and Their Applications to Ultrasonics* (von Nostrand, 1950).
[105] C. Mayer, *Trans. 29th Annual Meeting of the Orthopaedic Research Society* (Anaheim, California, March, 1983).
[106] J. H. Mc Elhaney, *J. Appl. Physiol.* **20** (1966) 1231–1236.
[107] Z. Miller and M. B. Fuchs, *J. Biomechanics* **38** (2005) 1855–1864.
[108] I. Mirsky, *J. Acous. Soc. Am.* **37** (1965) 1016–1026.
[109] J. C. Misra, *Medical and Life Sci. Engg.* **4** (1978a) 142–152.
[110] J. C. Misra, C. Hartung and O. Mahrenholtz, *Ing. Arch.* **47** (1978b) 329–337.
[111] J. C. Misra, *Ing. Archiv.* **47** (1978c) 11–19.
[112] J. C. Misra, *Rheol. Acta* **24** (1985) 520–533.
[113] J. C. Misra, *Rheol. Acta* **25** (1986a) 201–205.
[114] J. C. Misra, *Rheol. Acta* **25** (1986b) 485–490.
[115] J. C. Misra, *Math. Stud.* **55** (1987) 217–223.
[116] J. C. Misra, G. C. Bera and S. Samanta, *Mathematical Modelling* **7** (1986) 483–492.
[117] J. C. Misra, C. Hartung and O. Mahrenholtz, *Mech. Res. Comm.* **4** (1977) 297–302.
[118] J. C. Misra, C. Hartung and O. Mahrenholtz, *Ingenieur Archiv* **47** (1978) 329–337.
[119] J. C. Misra and S. Chakravarty, *J. Math. Anal. Applic.* **103** (1984a) 323–343.
[120] J. C. Misra and S. Chakravarty, *J. Biomechanics* **17** (1984b) 459–466.
[121] J. C. Misra and M. Mishra, *Engg. Fract. Mech.* **22** (1985) 65–76.
[122] J. C. Misra and S. Roy, *Computers Math. Applic.* **16** (1988) 247–266.
[123] J. C. Misra and S. Samanta, *Int. J. Solids Structures* **20** (1984) 55–62.
[124] J. C. Misra, G. C. Bera, S. C. Samanta and S. C. Misra, *Computer Math. Applic.* **24** (1992) 3–15.
[125] J. C. Misra and S. Chakravarty, *Acta Mechanica* **44** (1982) 159–168.
[126] J. C. Misra and S. Chakravarty, *Int. J. Engg. Sci.* **20** (1982) 445–454.
[127] J. C. Misra and S. B. Chakravarty, *J. Biomechanics* **15** (1982) 635–645.
[128] J. C. Misra and M. Mishra, *Engg. Fract. Mech.* **19** (1984) 101–112.
[129] J. C. Misra and V. V. Murthy, T. N. Froschung in *Ingenieurwesen* **47** (1981a) 37–40.
[130] J. C. Misra and V. V. T. N. Murthy, *Med. Life Sci. Engg.* **5** (1981b) 95–107.
[131] J. C. Misra and V. V. T. N. Murthy, *Medical and Life Sci. Engg.* **5** (1979) 95–107.
[132] J. C. Misra and V. V. T. N. Murthy, *Aeronautical Quarterly* **34** (1983) 303–313.
[133] J. C. Misra and V. V. T. N. Murthy, *Engg. Fract. Mech.* **18** (1983) 1075–1086.
[134] J. C. Misra and V. V. T. N. Murthy, *Bull. Tech. Univ. Istanbul* **36** (1983a) 67–80.

[135] J. C. Misra and V. V. T. N. Murthy, *Rheologica Acta* **22** (1983b) 512–518.
[136] J. C. Misra and S. Roy, *Computers Math. Applic.* **17** (1989) 1493–1502.
[137] J. C. Misra and S. Roy, *Modelling, Simulation and Control* **15** (1989) 41–63.
[138] J. C. Misra and S. Roy, *Computers Math. Applic.* **21** (1991) 141–147.
[139] J. C. Misra and S. Samanta, *Computers Math. Applic.* **15** (1988) 85–96.
[140] J. C. Misra and S. C. Samanta, *Int. J. Solids and Structures* **20** (1983a) 55–62.
[141] J. C. Misra and S. C. Samanta, *J. Math. Anal. Appl.* **96** (1983b) 313–329.
[142] J. C. Misra and S. C. Samanta, *J. Biomechanics* **20** (1987) 241–249.
[143] J. C. Misra and S. C. Samanta, *Computers Math. Applic.* **16** (1988) 1017–1026.
[144] J. C. Misra, G. C. Bera and S. C. Samanta, *Math. Comput. Modelling* **12** (1989) 611–624.
[145] J. C. Misra, G. C. Bera and S. Samanta, *J. Math. Biol.* **26** (1988) 105–120.
[146] J. C. Misra, G. C. Bera, S. C. Samanta and S. C. Misra, *J. Math. Anal. Applic.* **161** (1991) 474–507.
[147] J. C. Misra, G. C. Bera and S. Samanta, *Modelling Simulation Control* **17** (1989) 1–24.
[148] E. Mittra, C. Rubin and Y. X. Qin, *J. Biomechanics* **38** (2005) 1229–1237.
[149] A. Moon and R. Gatenby, *Trans. 25th Annual Meeting ACEMB* (Bal Harbour, Florida, October, 1972).
[150] E. E Morgan, O. C. Yeh, W. C. Chang and T. M. Keaveny, *J. Biomech. Engg.* **123** (2001) 1–9.
[151] S. Nagaraja, T. L. Couse and R. E. Guldberg, *J. Biomechanics* **38** (2005) 707–716.
[152] W. Nowacki, *Thermoelasticity* (Pergamon Press, Oxford, 1962).
[153] J. L. Nowinski, *Develoment of Applied Mechanics*, Vol. 5, ed. G. L. Rogres, Chapel Hill, The University of Carolina Press. J. L. Nowinski (1971), *AIAA J.* **9** (1970) 60–67.
[154] J. L. Nowinski, *Int. J. Mech. Sci.* **16** (1974) 285–288.
[155] J. L. Nowinski, *Theory of Thermoelasticity with Applications* (Sijthoff and Noordnoff, Alphen aan den Rej, 1978).
[156] A. Odgaard, J. Kabel, van Rietberger, M. Dalstra and R. Huiskes, *J. Biomech.* **30** (1997) 487–495.
[157] J. B. Park, *Biomaterials* (Plenum Press, New York, London, 1979).
[158] M. R. Patel, The deformation and fracture of rigid cellar plastics under multiaxial stress, Ph.D. Thesis, University of California, Barkley (1969).
[159] H. S. Paul, *J. Acous. Soc. Am.* **40** (1966) 1077–1080.
[160] R. Pelker and S. Saha, *J. Biomechanics* **16** (1983) 481–489.
[161] R. Pelker and S. Saha, *J. Biomechanics* **18** (1985) 745–753.
[162] B. H. Pfeiffer, *J. Biomechanics* **10** (1977) 53–57.
[163] K. Piekarski and M. Munro, *Nature (London)* **269** (1977) 80–82.
[164] P. J. Prendergast and D. Taylor, *J. Biomechanics* **27** (1994) 1067–1076.
[165] M. Punjabee, *J. Biomechanics* **12** (1979) 238.
[166] S. Ramtani and M. Zidi, *J. Biomechanics* **34** (2001) 471–479.
[167] A. A. Rauber, Wilhelm Engelmann, Leipzig (1876).

[168] D. T. Reilly and A. H. Burstein, *J. Biomechanics* **8** (1975) 393–405.
[169] D. T. Reilly and A. H. Burstein, *J. Bone Jt. Surg.* **56A** (1974) 1001–1022.
[170] P. D. Richardson and J. H. Whitelaw, *J. Franklin Inst.* **286** (1968) 169.
[171] H. J. Roesler, *Biomechanics* **20** (1987) 1025–1034.
[172] C. T. Rubin and L. E. Langon, *J. Bone Jt. Surg.* **66A** (1984a) 397–402.
[173] C. T. Rubin and L. E. Langon *J. Theor. Biol.* **107** (1984b) 321–327.
[174] A. M. Sadegh, G. M. Luo and S. C. Cowin, *J. Biomechanics* **26** (1993) 167–182.
[175] S. Saha and R. S. Pelker, *Proc. 28th ACEME* **17** (1975) 172.
[176] S. Saha and R. S. Lakes, *J. Biomechanics* **10** (1977a) 393–401.
[177] S. Saha and R. S. Lakes, *IEEE Trans. Biomed. Engg.* **24** (1977b) 508–512.
[178] R. Sanjibee, *J. Biomechanics* **15** (1982) 107–109.
[179] R. Sanjibee, N. Somanathan and D. Ramaswamy, *J. Biomechanics* **15** (1982) 181–183.
[180] J. Sapriel, *Acousto-Optics* (J. Wiley and Sons, 1979).
[181] S. A. Satter, M. J. Islam, K. S. Rabbani and M. S. Talukdar, *Bang. Med. Res. Bull.* (1999) 6–10.
[182] E. D. Sedlin, *Acta Ortho. Scand. Suppl.* 83, Munksgaurd, Copen. **1** (1965) 1–77.
[183] M. N. Shamos, L. S. Lavine and M. I. Shamos, *Nature* **197** (1963) 81.
[184] R. Smith and D. Keiper, *Am. J. Med. Electron* **4** (1965) 502.
[185] M. Solomonow, B. Zhow, R. V. Baralta, Y. Ln, M. Zhu and M. Harris, *Clinical Biomechanics* **15** (2000) 167–175.
[186] R. L. Spilker, *J. Biomechanics* **13** (1980) 395–901.
[187] R. L. Spilker, D. M. Dangirda and A. B. Schultz, *J. Biomechanics* **17** (1984) 103–112.
[188] M. Stanczyk, *J. Biomechanics* (2005) 1397–1403.
[189] E. L. Steinberg, N. Blumberg and S. Dekel, *J Biomechanics* **38** (2005) 63–68.
[190] J. L. Stone, G. S. Beaupre and W. C. Hayes, *J. Biomechanics* **15** (1982) 743–752.
[191] D. B. Swellow, P. Frasca, R. A. Harper and J. L. Katz, *Biomat. Mech. Dev. Art. Org.* **3** (1975) 121–153.
[192] R. C. Tennyson, R. Ewart and V. Niranjan, *Experim. Mech.* (1972) 502.
[193] G. A. Thompson, D. R. Young and D. Orne, *Med. and Biol. Engg.* **14** (1976) 253–262.
[194] H. L. Thron, *Pflügers Arch. Geo. Physiol.* **263** (1956) 107.
[195] R. W. Trehorne, *Orthopaedics Review* **10** (1981) 35–47.
[196] C. H. Turner, S. C. Cowin, J. Y. Rho, R. B. Ashman and J. C. Rice, *J. Biomechanics* **23** (1990) 549–561.
[197] W. C. van Buseirk and R. B. Ashman, *Mechanical Properties of Bone AMD*, Vol. 45, ed. S. C. Cowin (Am. Soc. Mech. Engrs., New York, 1981).
[198] B. van Rietberger, H. Weinans, R. Huiskes and A. Odgaard, *J. Biomechanics* **28** (1995) 69–78.
[199] H. W. Vayo and D. N. Ghista, *Bull. Math. Biophys.* **33** (1971) 463–479.
[200] W. W. von Maltzahn, *J. Biomechanics* **15** (1952) 757–765.

[201] Wertheim, *Ann. Chem. Phys.* **21** (1847) 385–414.
[202] J. Wolff, *Das Gesetz der Transformation der Knochen* (Hirschwald, Berlin, 1892).
[203] R. N. Wolosewick and S. Rayner, *J. Acous. Soc. Am.* **42** (1967) 417–421.
[204] A. T. C. Wong, W. Goldsmith and J. L. Sackman, *J. Biomechanics* **9** (1976) 813–825.
[205] T. M. Wright, F. Vosburgh and A. H. Burstein, *J. Biomechanics* **14** (1981) 405–410.
[206] H. S. Yoon and J. L. Katz, *J. Biomechanics* **9** (1976a) 401–412.
[207] H. S. Yoon and J. L. Katz, *J. Biomechanics* **9** (1976b) 459–464.

This page is intentionally left blank

INDEX

CONTRIBUTORS

R. Blauwkamp
Department of Mechanical and Industrial Engineering,
University of Illinois at Urbana-Champaign,
1206 West Green Street, Urbana,
IL 61801-2906, USA

J. Bentsman
Department of Mechanical and Industrial Engineering,
University of Illinois at Urbana-Champaign,
1206 West Green Street, Urbana,
IL 61801-2906, USA
jbentsma@uiuc.edu

A. K. Das
Department of Biotechnology,
Indian Institute of Technology,
Kharagpur 721302, India

I. V. Dardynskaia
Department of Environmental and Occupational Health Sciences,
University of Illinois at Chicago,
M/C 922,
2121 West Taylor Street,
Room 215, Chicago,
IL 60612, USA

Rui Dilão
Non-Linear Dynamics Group,
Instituto Superior Técnico,
Av. Rovisco Pais, 1049-001 Lisbon,
Portugal

Institut des Hautes Études Scientifiques,
Le Bois-Marie,
35, route de Chartres,
F-91440 Bures-sur-Yvette,
France
rui@sd.ist.utl.pt

P. D. Einziger
Department of Electrical Engineering, Technion,
Israel Institute of Technology,
Haifa 32000, Israel

Robert J. Gallop
Department of Mathematics and Applied Statistics,
West Chester University,
West Chester, PA 19383
rgallop@wcupa.edu

G. Gloushonok
Department of Radiation Chemistry,
Belarussian State University,
Minsk, Belarus

Donald Greenspan
Mathematics Department,
University of Texas at Arlington,
USA
greenspan@uta.edu

Sujoy K. Guha
School of Medical Science and Technology,
Indian Institute of Technology,
Kharagpur 721302, India
guha_sk@yahoo.com

Dirk Husmeier
Biomathematics and Statistics Scotland (BIOSS),
JCMB, The King's Buildings
Edinburgh EH9 3JZ,
United Kingdom
dirk@bioss.sari.ac.uk

Vladimir G. Ivancevic
Defence Science & Technology Organization,
Adelaide, Australia
vladimir.Ivancevic@dsto.defence.gov.au

Girija Jayaraman
Centre for Atmospheric Sciences,
Indian Institute of Technology,
New Delhi 110016, India
jgirija@cas.iitd.ernet.in

Frank P. Kozusko
Department of Mathematics,
Hampton University,
Hampton, Virginia, 23668, USA
frank.kozusko@hamptonu.edu

L. M. Livshitz
Department Biomedical Engineering, Technion,
Israel Institute of Technology,
Haifa 32000, Israel
jm@biomed.technion.ac.il

J. C. Misra
Department of Mathematics,
Indian Institute of Technology
Kharagpur 721302, India
jcm@maths.iitkgp.ernet.in

Charles J. Mode
Department of Mathematics and Computer Science,
Drexel University,
Philadelphia, PA 19104, USA

J. Mizrahi
Department Biomedical Engineering, Technion,
Israel Institute of Technology,
Haifa 32000, Israel
jm@biomed.technion.ac.il

S. Mukherjee
Department of Mathematics,
Indian Institute of Technology,
Kharagpur 721302, India

Jacob Oluwoye
The Built Environment Consultant
jacoboluwoye@hotmail.com

S. K. Pandey
Department of Mathematics,
Indian Institute of Technology
Kharagpur 721302, India

G. Pellegrinetti
Department of Mechanical and
Industrial Engineering,
University of Illinois at
Urbana-Champaign,
1206 West Green Street, Urbana,
IL 61801-2906, USA

G. Plank
Institut für Medizinische Physik
und Biophysik,
Karl Franzens Universitat Graz,
Austria
gernot.plank@meduni-graz.at

S. Ramtani
Université Paris 13 - IUT de
Saint-Denis,
Laboratoire des Propriétés
Mécaniques et Thermodynamiques
des Materiaux, CNRS-UPR9001,
Institut Galilée, Université Paris 13,
99, av. J.B. Clément,
93430 Villetaneuse, France
salah.ramtani@lpmtm.
univ-paris13.fr

S. Samanta
Department of Mathematics,
Indian Institute of Technology,
Kharagpur 721302, India

O. Shadyro
Department of Radiation
Chemistry,
Belarussian State University,
Minsk, Belarus

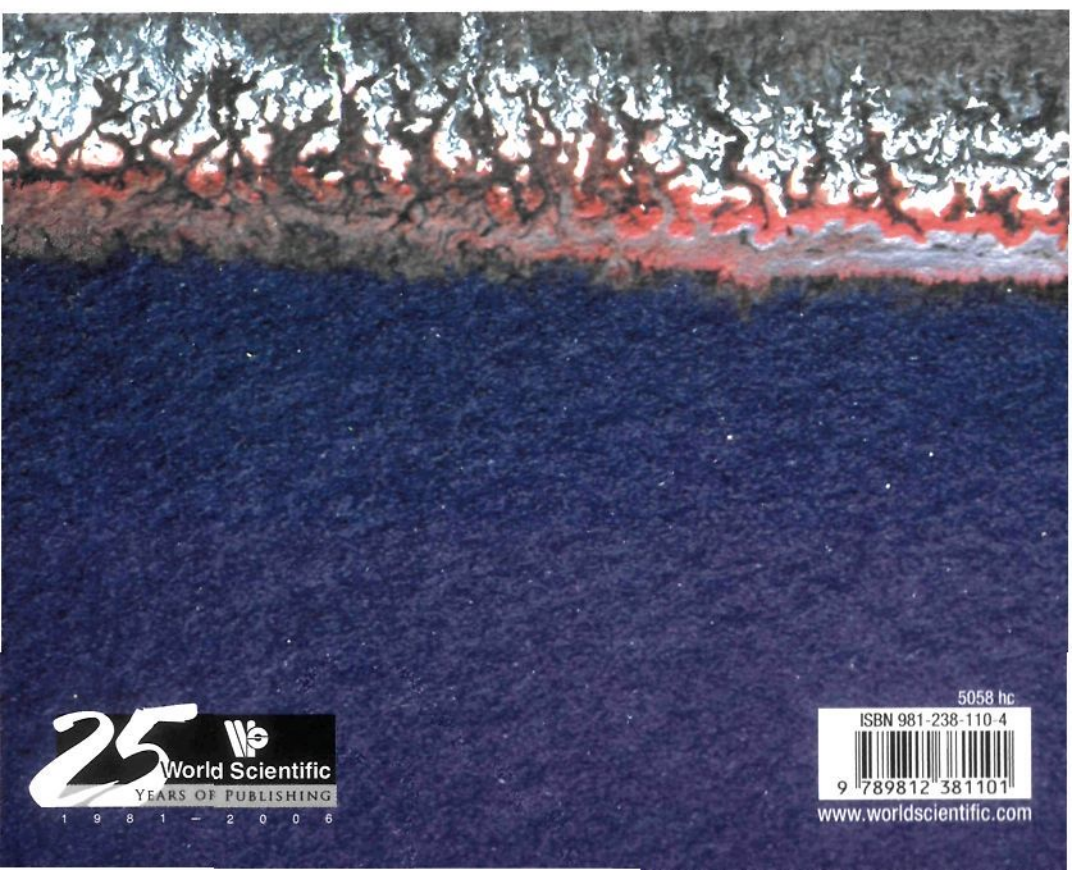
Shivani Sharma
Centre for Biomedical Engineering,
Indian Institute of Technology,
New Delhi 110016, India

Candace K. Sleeman
Department of Mathematics and
Computer Science,
Drexel University,
Philadelphia, PA 19104, USA

BIOMATHEMATICS

modelling and simulation

This book on modelling and simulation in biomathematics will be invaluable to researchers who are interested in the emerging areas of the field. Graduate students in related areas as well as lecturers will also find it beneficial. Some of the chapters have been written by distinguished experts in the field.



5058 hc
ISBN 981-238-110-4
9 789812 381101
www.worldscientific.com