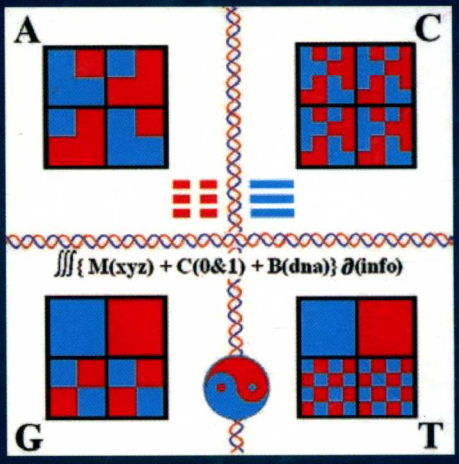
Wiley Series on Bioinformatics • Yi Pan and Albert Y. Zomaya, Series Editors

Mathematics of Bioinformatics

Theory, Practice, and Applications



MATTHEW HE
SERGEY PETOUKHOV

WILEY

MATHEMATICS OF BIOINFORMATICS

Theory, Practice, and Applications

Matthew He

Sergey Petoukhov

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data Is Available

He, Matthew
Mathematics of bioinformatics: theory, practice, and applications / Matthew He, Sergey Petoukhov

Includes bibliographical references and index.

ISBN 978-0-470-40443-0 (cloth)

Printed in Singapore

10 9 8 7 6 5 4 3 2 1

MATHEMATICS OF BIOINFORMATICS

Wiley Series on

Bioinformatics: Computational Techniques and Engineering

A complete list of the titles in this series appears at the end of this volume.

PREFACE

Recent progress in the determination of genomic sequences has yielded many millions of gene sequences. But what do these sequences tell us, and what generalities and rules are governed by them? There is more to life than the genomic blueprint of each organism. Life functions within the natural laws that we know and those we do not know. It appears that we understand very little about genetic contexts required to "read" these sequences. Mathematics can be used to understand life from the molecular level to the level of the biosphere. This book is intended to further integrate the mathematics and biological sciences. The reader will gain valuable knowledge about mathematical methods and tools, phenomenological results, and interdisciplinary connections in the fields of molecular genetics, bioinformatics, and informatics.

Historically, mathematics, probability, and statistics have been widely used in the biological sciences. Science is challenged to understand the system organization of the molecular genetics ensemble, with its unique properties of reliability and productivity. Disclosing key aspects of this organization constitutes a big step in science about nature as a whole and in creating the most productive biotechnologies. Knowledge of this structural organization should become a part of mathematical natural science.

Advances in mathematical methods and techniques in bioinformatics have been growing rapidly. Mathematics has a fundamental role in describing the complexities of biological structures, patterns, and processes. Mathematical analysis of structures of molecular systems has essential meaning for bioinformatics, biomathematics, and biotechnology. Mathematics is used to elucidate trends, patterns, connections, and relationships in a quantitative manner that can lead to important discoveries in biology. This book is devoted to drawing a closer connection and better integration between mathematical methods and biological codes, sequences, structures, networks, and systems biology. It is intended for researchers and graduate students who want an overview of the field and who want information on the possibilities (and challenges) of the interface between mathematics and bioinformatics. In short, the book provides a broad overview of the interfaces between mathematics and bioinformatics.

ORGANIZATION OF THE BOOK

To reach a broad spectrum of readers, this book does not require a deep knowledge of mathematics or biology. The reader will learn fundamental

concepts and methods from mathematics and biology. The book is organized into 10 chapters covering mathematical topics in relation to genetic code systems, biological sequences, structures and functions, networks and biological systems, matrix genetics, cognitive informatics, and the central dogma of informatics. Three appendixes, on bioinformatics notations, a historical time line of bioinformatics, and a bioinformatics glossary, are included for easy reference.

Chapter 1 provides an overview of bioinformatics history, genetic code and mathematics, background mathematics for bioinformatics, and the big picture of bioinformatics–informatics.

Chapter 2 is devoted to symmetrical analysis for genetic systems. Genetic coding possesses noise immunity. Mathematical theories of noise-immunity coding and discrete signal processing are based on matrix methods of representation and analysis of information. These matrix methods, which are connected closely to relations of symmetry, are borrowed for a matrix analysis of ensembles of molecular elements of the genetic code. A uniform representation of ensembles of genetic multiplets in the form of matrices of a cumulative Kronecker family is described. The analysis of molecular peculiarities of the system of nitrogenous bases reveals the first significant relations of symmetry in these genetic matrices. It permits one to introduce a natural numbering of the multiplets in each of the genetic matrices and to provide a basis for further analysis of genetic structures. Connections of the numerated genetic matrices with famous matrices of dyadic shifts and with the golden section are demonstrated.

In Chapter 3 we define biological, mathematical, and binary sequences in theoretical computer science. We describe pairwise, multiple, and optimal sequence alignment. We discuss the scoring system used to rank alignments, the algorithms used to find optimal (or good) scoring alignments, and statistical methods used to evaluate the significance of an alignment score.

Chapter 4 provides an introduction to the structures of DNA, key elements of knot theory, such as links, tangles, and knot polynomials, and applications of knot theory to the study of closed circular DNA. The physical and chemical properties of this type of DNA can be explained in terms of basic characteristics of a linking number which is invariant under continuous deformation of the DNA structure and is the sum of two geometric quantities, twist and writhing.

In Chapter 5 we introduce protein primary, secondary, tertiary, and quaternary structure by geometric means. We also discuss the classification of proteins, physical forces in proteins, protein motion (folding and unfolding), and basic methods for secondary and tertiary structure prediction.

Chapter 6 covers the topics of network approaches in biological systems. These approaches offer the tools to analyze and understand a host of biological systems. In particular, within the cell the variety of interactions among genes, proteins, and metabolites are captured by network representations. In

this chapter we focus our discussion on biological applications of the theory of graphs and networks.

Chapter 7 covers the topics of biological systems and genetic code systems. We explain how the presence of fractal geometry can be used in an analytical way to study genetic code systems and predict outcomes in systems, to generate hypotheses, and to help design experiments. At the end of the chapter we discuss the emerging field of systems biology, as well as challenges and perspectives in biological systems.

Chapter 8 continues the discussion introduced in Chapter 2 on genetic matrices and their symmetries and algebraic properties. The algebraic theory of coding is one of the modern fields of applications of algebra and uses matrix algebra intensively. This chapter is devoted to matrix forms of presentations of the genetic code for algebraic analysis of a basic scheme of degeneracy of the genetic code. Similar matrix forms are utilized in the theory of signal processing and encoding. The Kronecker family of the genetic matrices is investigated, which is based on the genetic matrix [C A; U G], where C, A, U, and G are the letters of the genetic alphabet. This matrix in the third Kronecker power is the 8×8 matrix, which contains all 64 genetic triplets in a strict order with a natural binary numeration of the triplets by numbers from 0 to 63. Peculiarities of the basic scheme of the genetic code degeneracy are reflected in the symmetrical black-and-white mosaic of this genetic 8×8 matrix. Unexpectedly, this mosaic matrix is connected algorithmically with Hadamard matrices, which are well known in the theory of signal processing and encoding, spectral analysis, quantum mechanics, and quantum computers. Furthermore, many types of cyclic permutations of genetic elements lead unexpectedly to reconstruction of initial Hadamard matrices into new Hadamard matrices. This demonstrates that matrix algebra is a promising instrument and adequate language in bioinformatics and algebraic biology.

In Chapter 9 we review briefly the intersections and connections between the two emerging fields of bioinformatics and cognitive informatics through a systems view of emerging pattern, dissipative structure, and evolving cognition of living systems. A new type of math-denotational mathematics for cognitive informatics is introduced. It is hoped that this brief review will encourage further exploration of our understanding of the biological basis of cognition, perception, learning, memory, thought, and mind.

In Chapter 10 we return to the big picture of informatics introduced in Chapter 1. We propose a general concept of data, information, and knowledge and then place the main focus on the process and transition from data to information and then to knowledge. We present the concept of the central dogma of informatics, in analogy to the central dogma of molecular biology.

Each chapter finishes with a summary of challenges and perspectives of corresponding topics. These summaries are structured to bridge the gaps among the interdisciplinary areas, which involve concepts and ideas from a

variety of sciences, including biology, biochemistry, physics, computer science, and mathematics.

THE CHALLENGES

The interface between mathematics and bioinformatics and computational biology presents challenges and opportunities for both mathematicians and biologists. Unique opportunities for research have surfaced within the last 10 to 20 years, both because of the explosion of biological data with the advent of new technologies and because of the availability of advanced and powerful computers that can organize the plethora of data. For biology, the possibilities range from the level of the cell and molecule to the level of the biosphere. For mathematics, the potential is great in traditional applied areas such as statistics and differential equations, as well as in such nontraditional areas as knot theory.

The primary purpose of encouraging biologists and mathematicians to work together is to investigate fundamental problems that cannot only be approached by biologists or by mathematicians. If this effort is successful, the future may produce individuals with both biological skills and mathematical insight and facility. At this time such people are rare; it is clear, however, that a greater percentage of the training of future biologists must be mathematically oriented. Both disciplines can expect to gain by this effort. Mathematics is the "lens through which to view the universe" and serves to identify important details of the biological data and suggest the next series of experiments. Mathematicians, on the other hand, can be challenged to develop new mathematics in order to perform this function.

In this book we explore some of the development and opportunities at the interface between biology and mathematics. To mathematicians, the book demonstrates that the stimulation of biological data and applications will enrich the discipline of mathematics for decades to come, as did applications in the past from the physical sciences. To biologists, the book presents the use of mathematical approaches to provide insights available for bioinformatics. To both communities, the book demonstrates the ferment and excitement of a rapidly evolving field—bioinformatics.

Acknowledgments

This book is part of the Wiley Series on Bioinformatics: Computational Techniques and Engineering. The authors would like to express our gratitude to the series editors, Yi Pan and Albert Zomaya, for giving us the opportunity to present our research interest as a book in this series. We would also like to thank many of our colleagues who worked with us in exploring topics relevant to this book. Their names can be found in the chapter references. Only literature closely related to our work is included in the references, and due to the

wide extent of subjects in the studies, the references cited are incomplete. The authors apologize deeply for any relevant omission.

We want to thank the Mechanical Engineering Institute of the Russian Academy of Sciences, Moscow, Russia and the Farquhar College of Arts and Sciences of Nova Southeastern University, Fort Lauderdale, Florida for their support. We are deeply indebted to our colleagues Diego Castano, Emily Schmitt, and Robin Sherman of Nova Southeastern University for offering suggestions and for reviewing the final version of the manuscript.

Special thanks also go to the publishing team at Wiley, whose contributions throughout the entire process from initial proposal to final publication have been invaluable: particular to the Wiley assistant development editing team, who continuously provided prompt guidance and support throughout the book editing process.

Finally, we would like to give our special thanks to our families for their patient love, which enabled us to complete this work.

Nova Southeastern University
Fort Lauderdale, Florida

MATTHEW HE

Russian Academy of Sciences
Moscow, Russia

SERGEY PETOUKHOV

March 16, 2010

ABOUT THE AUTHORS

Matthew He, Ph.D., is a full professor and director of the Division of Mathematics, Science, and Technology of Nova Southeastern University in Florida. He has been a full professor and grand Ph.D. of the World Information Distributed University since 2004, as well as an academician of the European Academy of Informatization. He received a Ph.D. in mathematics from the University of South Florida in 1991. He was a research associate at the Department of Mathematics, Eldgenossische Technische Hochschule, Zurich, Switzerland, and the Department of Mathematics and Theoretical Physics, Cambridge University, Cambridge, England. He was also a visiting professor at the National Key Research Lab of Computational Mathematics of the Chinese Academy of Science and the University of Rome, Italy.

Dr. He has authored and edited eight books and published over 100 research papers in the areas of mathematics, bioinformatics, computer vision, information theory, mathematics, and engineering techniques in the medical and biological sciences. He is an editor of *International Journal of Software Science and Computational Intelligence*, *International Journal of Cognitive Informatics and Natural Intelligence*, *International Journal of Biological Systems*, *and International Journal of Integrative Biology*. He is an invited series editor of Henry Stewart Talk "Using Bioinformatics in Exploration in Genetic Diversity" in Biomedical and Life Sciences Series. He received the World Academy of Sciences Achievement Award in recognition of his research contributions in the field of computing in 2003. He is chairman of the International Society of Symmetry in Bioinformatics and a member of International Advisory Board of the International Symmetry Association. He is a member of the American Mathematical Society, the Association of Computing Machinery, the IEEE Computer Society, the World Association of Science Engineering, and an international advisory board member of the bioinformatics group of the International Federation for Information Processing. He was an international scientific committee co-chair of the International Conference of Bioinformatics and Its Applications in 2004 and a general co-chair of the International Conference of Bioinformatics Research and Applications in 2009, and has been a keynote speaker at many international conferences in the areas of mathematics, bioinformatics, and information science and engineering.

Sergey Petoukhov, Ph.D., is a chief scientist of the Department of Biomechanics, Mechanical Engineering Research Institute of the Russian Academy of

Sciences in Moscow. He has been a full professor and grand Ph.D. of the World Information Distributed University since 2004, as well as an academician of the European Academy of Informatization. He is a laureate of the state prize of the USSR (1986) for his achievements in biomechanics. Dr. Petoukhov graduated from the Moscow Physical-Technical Institute in 1970 and received a postgraduated from the institute in 1973 with a specialty in biophysics. He received a Golden Medal of the National Exhibition of Scientific Achievements in Moscow in 1973 for his physical model of human vestibular apparatus. He received his first scientific degree in the USSR in 1973: a Candidate of Biological Sciences degree with a in specialty in biophysics. He received his second scientific degree in the USSR in 1988: Doctor of Physical-Mathematical Sciences in two specialties, biomechanics and crystallography and crystallophysics. He was an academic foreign stager of the Technical University of Nova Scotia, Halifax, Canada in 1988. He was elected an academician of Academy of Quality Problems (Russia) in 2000. Dr. Petoukhov is a director of the Department of Biophysics and chairman of the Scientific-Technical Council at the Scientific-Technical Center of Information Technologies and Systems in Moscow. He was vice-president of the International Society for the Interdisciplinary Study of Symmetry from 1989 to 2000 and chairman of the international advisory board of the International Symmetry Association (with headquarters in Budapest, Hungary; http://symmetry.hu/) from 2000 to the present. Dr. Petoukhov has been honorary chairman of the board of directors of the International Society of Symmetry in Bioinformatics since 2000 and vice-president and academician of the National Academy of Intellectual and Social Technologies of Russia since 2003. Dr. Petoukhov is academician of the International Diplomatic Academy (Belgium; www.bridgeworld.org). He is Russian chairman (chief) of an official scientific cooperative body of the Russian and Hungarian Academies of Sciences on the theme "Nonlinear Models in Biomechanics, Bioinformatics, and the Theory of Self-organizing Systems."

Dr. Petoukhov has published over 150 research papers (including seven books) in biomechanics, bioinformatics, mathematical and theoretical biology, theory of symmetries and its applications, and mathematics. He is a member of the editorial board of two international journals: *Journal of Biological Systems* and *Symmetry: Culture and Science*. He was a guest editor of special issues (on bioinformatics) of the international journal *Journal of Biological Systems* in 2004. Dr. Petoukhov is the book editor of *Symmetries in Genetic Informatics* (2001), *Advances in Bioinformatics and Its Applications* (2004), and a Russian edition (2006) of a book by Canadian professor R. V. Jean, *Phyllotaxis: A Systemic Study in Plant Morphogenesis* (Cambridge University Press, Cambridge, UK, 1994). He is a co-organizer of international conferences on the theory of symmetries and its applications (Budapest, Hungary, 1989; Hirosima, Japan, 1992; Washington, D.C., 1995; Haifa, Izrael, 1998; Budapest, Hungary, 2003, 2006, and 2009; Moscow, Russia, 2006). He was chairman of the international program committee of the International Conference on

Bioinformatics and Its Applications in Fort Lauderdale, Florida, in 2004. He was co-chairman of the organizing committees of international conferences on "Modern Science and Ancient Chinese 'The Book of Changes' (*I Ching*)" in Moscow in 2003, 2004, 2005, and 2006. He teaches a course on biophysics and bioinformatics at the Moscow Physical-Technical Institute and a course in architectural bionics at the Peoples' Friendship University of Russia. He is actively involved in promoting science, education, and technology.

CONTENTS

1 Bioinformatics and Mathematics

Traditionally, the study of biology is from morphology to cytology and then to the atomic and molecular level, from physiology to microscopic regulation, and from phenotype to genotype. The recent development of bioinformatics begins with research on genes and moves to the molecular sequence, then to molecular conformation, from structure to function, from systems biology to network biology, and further investigates the interactions and relationships among, genes, proteins, and structures. This new reverse paradigm sets a theoretical starting point for a biological investigation. It sets a new line of investigation with a unifying principle and uses mathematical tools extensively to clarify the ever-changing phenomena of life quantitatively and analytically.

It is well known that there is more to life than the genomic blueprint of each organism. Life functions within the natural laws that we know and those that we do not know. Life is founded on mathematical patterns of the physical world. Genetics exploits and organizes these patterns. Mathematical regularities are exploited by the organic world at every level of form, structure, pattern, behavior, interaction, and evolution. Essentially all knowledge is intrinsically unified and relies on a small number of natural laws. Mathematics helps us understand how monomers become polymers necessary for the assembly of cells. Mathematics can be used to understand life from the molecular to the biosphere levels, including the origin and evolution of organisms, the nature of genomic blueprints, and the universal genetic code as well as ecological relationships.

Mathematics and biological data have a synergistic relationship. Biological information creates interesting problems, mathematical theory and methods provide models for understanding them, and biology validates the mathematical models. A model is a representation of a real system. Real systems are too complicated, and observation may change the real system. A good system model should be simple, yet powerful enough to capture the behavior of the real system. Models are especially useful in bioinformatics. In this chapter we provide an overview of bioinformatics history, genetic code and mathematics, background mathematics for bioinformatics, and the big picture of bioinformatics–informatics.

Mathematics of Bioinformatics: Theory, Practice, and Applications, By Matthew He and Sergey Petoukhov