Shiyi Shen
Jack A. Tuszynski

BIOLOGICAL AND MEDICAL PHYSICS, BIOMEDICAL ENGINEERING

Theory and Mathematical Methods for Bioinformatics

Springer

Shiyi Shen · Jack A. Tuszynski

Theory and Mathematical Methods for Bioinformatics

With 47 Figures and 59 Tables

江苏工业学院图书馆
藏书章

Springer

Prof. Shiyi Shen
Nankai University
College of Mathematical Sciences
Tianjin 300071
China
syshen@nankai.edu.cn

Prof. Jack A. Tuszynski
University of Alberta
Department of Physics
Edmonton T6G 2G7, Alberta
Canada
jtus@phys.ualberta.ca

ISBN 978-3-540-74890-8 e-ISBN 978-3-540-74891-5

DOI 10.1007/978-3-540-74891-5

Biological and Medical Physics, Biomedical Engineering ISSN 1618-7210

Library of Congress Control Number: 2007939206

© 2008 Springer-Verlag Berlin Heidelberg

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable for prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting and production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig, Germany
Cover design: eStudioCalamar S.L., F. Steinen-Broo, Girona, Spain

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

BIOLOGICAL AND MEDICAL PHYSICS, BIOMEDICAL ENGINEERING

BIOLOGICAL AND MEDICAL PHYSICS, BIOMEDICAL ENGINEERING

The fields of biological and medical physics and biomedical engineering are broad, multidisciplinary and dynamic. They lie at the crossroads of frontier research in physics, biology, chemistry, and medicine. The Biological and Medical Physics, Biomedical Engineering Series is intended to be comprehensive, covering a broad range of topics important to the study of the physical, chemical and biological sciences. Its goal is to provide scientists and engineers with textbooks, monographs, and reference works to address the growing need for information.

Books in the series emphasize established and emergent areas of science including molecular, membrane, and mathematical biophysics; photosynthetic energy harvesting and conversion; information processing; physical principles of genetics; sensory communications; automata networks, neural networks, and cellular automata. Equally important will be coverage of applied aspects of biological and medical physics and biomedical engineering such as molecular electronic components and devices, biosensors, medicine, imaging, physical principles of renewable energy production, advanced prostheses, and environmental control and engineering.

Editor-in-Chief:

Elias Greenbaum, Oak Ridge National Laboratory,
Oak Ridge, Tennessee, USA

Editorial Board:

Masuo Aizawa, Department of Bioengineering,
Tokyo Institute of Technology, Yokohama, Japan

Olaf S. Andersen, Department of Physiology,
Biophysics & Molecular Medicine,
Cornell University, New York, USA

Robert H. Austin, Department of Physics,
Princeton University, Princeton, New Jersey, USA

James Barber, Department of Biochemistry,
Imperial College of Science, Technology
and Medicine, London, England

Howard C. Berg, Department of Molecular
and Cellular Biology, Harvard University, Cambridge,
Massachusetts, USA

Victor Bloomfield, Department of Biochemistry,
University of Minnesota, St. Paul, Minnesota, USA

Robert Callender, Department of Biochemistry,
Albert Einstein College of Medicine,
Bronx, New York, USA

Britton Chance, Department of Biochemistry/
Biophysics, University of Pennsylvania,
Philadelphia, Pennsylvania, USA

Steven Chu, Department of Physics,
Stanford University, Stanford, California, USA

Louis J. DeFelice, Department of Pharmacology,
Vanderbilt University, Nashville, Tennessee, USA

Johann Deisenhofer, Howard Hughes Medical Institute,
The University of Texas, Dallas, Texas, USA

George Feher, Department of Physics,
University of California, San Diego, La Jolla,
California, USA

Hans Frauenfelder, CNLS, MS B258,
Los Alamos National Laboratory, Los Alamos,
New Mexico, USA

Ivar Giaever, Rensselaer Polytechnic Institute,
Troy, New York, USA

Sol M. Gruner, Department of Physics,
Princeton University, Princeton, New Jersey, USA

Judith Herzfeld, Department of Chemistry,
Brandeis University, Waltham, Massachusetts, USA

Mark S. Humayun, Doheny Eye Institute,
Los Angeles, California, USA

Pierre Joliot, Institut de Biologie Physico-Chimique,
Fondation Edmond de Rothschild, Paris, France

Lajos Keszthelyi, Institute of Biophysics,
Hungarian Academy of Sciences, Szeged, Hungary

Robert S. Knox, Department of Physics and Astronomy,
University of Rochester, Rochester, New York, USA

Aaron Lewis, Department of Applied Physics,
Hebrew University, Jerusalem, Israel

Stuart M. Lindsay, Department of Physics
and Astronomy, Arizona State University, Tempe, Arizona,
USA

David Mauzerall, Rockefeller University,
New York, New York, USA

Eugenie V. Mielczarek, Department of Physics
and Astronomy, George Mason University, Fairfax,
Virginia, USA

Markolf Niemz, Medical Faculty Mannheim,
University of Heidelberg, Mannheim, Germany

V. Adrian Parsegian, Physical Science Laboratory,
National Institutes of Health, Bethesda, Maryland, USA

Linda S. Powers, NCDMF: Electrical Engineering,
Utah State University, Logan, Utah, USA

Earl W. Prohofsky, Department of Physics,
Purdue University, West Lafayette, Indiana, USA

Andrew Rubin, Department of Biophysics,
Moscow State University, Moscow, Russia

Michael Seibert, National Renewable Energy Laboratory,
Golden, Colorado, USA

David Thomas, Department of Biochemistry,
University of Minnesota Medical School, Minneapolis,
Minnesota, USA

Samuel J. Williamson, Department of Physics,
New York University, New York, New York, USA

此为试读，需要完整PDF请访问：www.ertongbook.com

Preface

Bioinformatics is an interdisciplinary science which involves molecular biology, molecular chemistry, physics, mathematics, computational sciences, etc. Most of the books on biomathematics published within the past ten years have consisted of collections of standard bioinformatics problems and informational methods, and focus mainly on the logistics of implementing and making use of various websites, databases, software packages and serving platforms. While these types of books do introduce some mathematical and computational methods alongside the software packages, they are lacking in a systematic and professional treatment of the mathematics behind these methods.

It is significant in the field of bioinformatics that not only is the amount of data increasing exponentially, but collaboration is also both widening and deepening among biologists, chemists, physicists, mathematicians, and computer scientists. The sheer volume of problems and databases requires researchers to continually develop software packages in order to process the huge amounts of data, utilizing the latest mathematical methods. The intent of this book is to provide a professional and in-depth treatment of the mathematical topics necessary in the study of bioinformatics.

Although there has been great progress in bioinformatics research, many difficult problems are still to be solved. Some of the most well-known include: multiple sequence alignment, prediction of 3D structures and recognition of the eukaryote genes. Since the Human Genome Project (HGP) was developed, the problems of the network structures of the genomes and proteomes, as well as regulation of the cellular systems are of great interest. Although there is still much work to be done before these problems are solved, it is our hope that the key to solving these problems lies in an increased collaboration among the different disciplines to make use of the mathematical methods in bioinformatics.

This book is divided into two parts: the first part introduces the mutation and alignment theory of sequences, which includes the general models and theory to describe the structure of mutation and alignment, the fast algorithms for pairwise and multiple alignment, as well as discussion based on the output given by fast multiple alignment. Part I contains a fairly advanced treatment, and it demonstrates how mathematics may be successfully used in bioinformatics. The success achieved using fast algorithms of multiple alignment illustrates the important role of mathematics.

Part II analyzes the protein structures, which includes the semantic and cluster analysis based on the primary structure and the analysis of the 3D structure for main chains and side chains of proteins. The wiggly angle (dihedral angle) was used when analyzing the configuration of proteins, making the description of the configuration more exact. Analyzing the configuration differs from predicting the secondary or 3D structures. We collect all pockets, grooves, and channels in a protein as configuration characteristics, analyze the structure of these characteristics, and give the algorithms to compute them.

Parts I and II offer independent treatments of biology and mathematics. This division is convenient, as the reader may study both separately. In each part we include results and references from our own research experiences. We propose some novel concepts, for example, the modulus structures, alignment space, semantic analysis for protein sequences, and the geometrical methods to compute configuration characteristics of proteins, etc. Study of these concepts is still in its infancy and so there is much to still be explored. It is our hope that these issues continue to be examined mathematically so that they remain at the forefront, both in mathematics and bioinformatics.

We recognize the importance of considering the computational aspect while introducing mathematical theories. A collection of computational examples have been included in this book so that our theoretical results may be tested, and so that the reader may see the corresponding theories illustrated. Additionally, some of these examples have implications which may be applied to biology directly, and may be downloaded from the website [99] as the data is too large to include in this book. (As an example, when examining the alignment of the HIV gene, the size is $m = 405, n = 10{,}000$ bp.)

An understanding of the fundamentals of probability, statistics, combinatorial graph theory, and molecular biology is assumed, as well as programming ability. In order that the reader may solidify their understanding, problems have been added at the end of each chapter, with hints to help get started, It is our hope that this book will be useful in the field of bioinformatics both as a textbook and as a reference book.

Acknowledgements

This book is the product of a collaboration between the Chinese and Canadian groups. In addition to the two authors, many members of both groups made important contributions. Shiyi Shen's assistants Gang Hu, Kui Wang, Liuhuan Dong, Hua Zhang (doctoral candidate), Zhonghua Wu (doctoral candidate), and Guogen Shan helped him to collect and analyze the data, along with other works in order to create the Chinese edition of this work. Dr. Jishou

Ruan and his graduate students provided the initial translation of the Chinese edition into non-native-speaker English, and later checked the native-speaker English edition to confirm that the meaning of the original was retained. These students are Guangyue Hu, Tuo Zhang, Guoxiang Lu, Jianzhao Gao, Yubing Shang, Shengqi Li, and Hanzhe Chen. The manuscript was rendered into native-speaker English by Jack Tuszynski's assistant, Michelle Hanlon. The authors acknowledge and are grateful to all involved for the hard work that went into this project.

Shen's group would like to thank Chun-Ting Zhang (Fellow of the Chinese Academy of Sciences) who encouraged Shiyi Shen's work on bioinformatics. The group also wishes to thank the Liuhui Center for Applied Mathematics, the Tianjin University and Nankai University joint program, the NSFC (grant NSFC10271061), and the Sciences and Technology program of Tianjin (STPT 04318511-17) for providing funding for this project. We would also like to thank the Chern Institute of Mathematics, which offers much assistance, including computers and offices.

Both authors would like to thank Angela Lahee of Springer-Verlag for her guidance and encouragement.

Tianjin, China
Edmonton, Canada

Shiyi Shen
Jack A. Tuszynski
June 2007

Contents

Outline 1

Part I Mutations and Alignments

1 Introduction 5

1.1 Mutation and Alignment 5

1.1.1 Classification of Biological Sequences 5

1.1.2 Definition of Mutations and Alignments 6

1.1.3 Progress on Alignment Algorithms and Problems to Be Solved 8

1.1.4 Mathematical Problems Driven by Alignment and Structural Analysis 12

1.2 Basic Concepts in Alignment and Mathematical Models 13

1.2.1 Mutation and Alignment 13

1.3 Dynamic Programming-Based Algorithms for Pairwise Alignment 17

1.3.1 Introduction to Dynamic Programming-Based Algorithms 17

1.3.2 The Needleman–Wunsch Algorithm, the Global Alignment Algorithm 18

1.3.3 The Smith–Waterman Algorithm 21

1.4 Other Notations 24

1.4.1 Correlation Functions of Local Sequences 24

1.4.2 Pairwise Alignment Matrices Among Multiple Sequences 25

1.5 Remarks 26

1.6 Exercises, Analyses, and Computation 27

2 Stochastic Models of Mutations and Structural Analysis . . . 29
2.1 Stochastic Sequences and Independent Sequence Pairs 29
2.1.1 Definitions and Notations of Stochastic Sequences 29
2.1.2 Independently and Identically Distributed Sequences . . . 31
2.1.3 Independent Stochastic Sequence Pairs 33
2.1.4 Local Penalty Function and Limit Properties of 2-Dimensional Stochastic Sequences 36
2.2 Stochastic Models of Flow Raised by Sequence Mutations 37
2.2.1 Bernoulli Processes 37
2.2.2 Poisson Flow 40
2.2.3 Mutated Flows Resulting from the Four Mutation Types 43
2.3 Stochastic Models of Type-I Mutated Sequences 45
2.3.1 Description of Type-I Mutation 45
2.3.2 Properties of Type-I Mutations 47
2.4 Type-II Mutated Sequences 50
2.4.1 Description of Type-II Mutated Sequences 51
2.4.2 Stochastic Models of Type-II Mutated Sequences 51
2.4.3 Error Analysis of Type-II Mutated Sequences 54
2.4.4 The Mixed Stochastic Models Caused by Type-I and Type-II Mutations 57
2.5 Mutated Sequences Resulting from Type-III and Type-IV Mutations 58
2.5.1 Stochastic Models of Type-III and Type-IV Mutated Sequences 58
2.5.2 Estimation of the Errors Caused by Type-III and Type-IV Mutations 59
2.5.3 Stochastic Models of Mixed Mutations 61
2.6 Exercises 64

3 Modulus Structure Theory 67
3.1 Modulus Structure of Expanded and Compressed Sequences . . 67
3.1.1 The Modulus Structures of Expanded Sequences and Compressed Sequences 67
3.1.2 The Order Relation and the Binary Operators on the Set of Expanded Modes 71
3.1.3 Operators Induced by Modes 73
3.2 Modulus Structure of Sequence Alignment 76
3.2.1 Modulus Structures Resulting from Multiple Alignment 76
3.2.2 Structure Analysis of Pairwise Alignment 78
3.2.3 Properties of Pairwise Alignment 81
3.2.4 The Order Relation and the Operator Induced by Modulus Structure 84

3.3 Analysis of Modulus Structures Resulting from Sequence Mutations 85
3.3.1 Mixed Sequence Mutations 85
3.3.2 Structure Analysis of Purely Shifting Mutations 87
3.3.3 Structural Representation of Mixed Mutation 93
3.4 The Binary Operations of Sequence Mutations 93
3.4.1 The Order Relationship Among the Modes of Shifting Mutations 93
3.4.2 Operators Induced by Modes of Shifting Mutations 96
3.5 Error Analysis for Pairwise Alignment 100
3.5.1 Uniform Alignment of Mutation Sequences 100
3.5.2 Optimal Alignment and Uniform Alignment 102
3.5.3 Error Analysis of Uniform Alignment 104
3.5.4 Local Modification of Sequence Alignment 106
3.6 Exercises 107

4 Super Pairwise Alignment 109
4.1 Principle of Statistical Decision-Based Algorithms for Pairwise Sequences 109
4.1.1 Uniform Alignment and Parameter Estimation for Pairwise Sequences 109
4.1.2 The Locally Uniform Alignment Resulting from Local Mutation 111
4.1.3 The Estimations of Mutation Position and Length 113
4.2 Operation Steps of the SPA and Its Improvement 115
4.2.1 Operation Steps of the SPA 115
4.2.2 Some Unsolved Problems and Discussions of SPA 118
4.2.3 Local Modifications for Sequence Alignment 121
4.3 Index Analysis of SPA 122
4.3.1 The Statistical Features of Estimations 122
4.3.2 Improvement of the Algorithm to Estimate $\hat{s}^*$ 128
4.3.3 The Computational Complexity of the SPA 131
4.3.4 Estimation for the Error of Uniform Alignment Induced by a Hybrid Stochastic Mutation Sequence 132
4.4 Applications of Sequence Alignment and Examples 135
4.4.1 Several Applications of Sequence Alignment 135
4.4.2 Examples of Pairwise Alignment 137
4.5 Exercises 146

5 Multiple Sequence Alignment 149
5.1 Pairwise Alignment Among Multiple Sequences 149
5.1.1 Using Pairwise Alignment to Process Multiple Sequences 149
5.1.2 Topological Space Induced by Pairwise Alignment of Multiple Sequences 150

此为试读,需要完整PDF请访问: www.ertongbook.com

5.2 Optimization Criteria of MA 156
5.2.1 The Definition of MA 156
5.2.2 Uniform Alignment Criteria and SP-Optimization Criteria for Multiple Sequences 156
5.2.3 Discussion of the Optimization Criterion of MA 160
5.2.4 Optimization Problem Based on Shannon Entropy 164
5.2.5 The Similarity Rate and the Rate of Virtual Symbols 170
5.3 Super Multiple Alignment 172
5.3.1 The Situation for MA 172
5.3.2 Algorithm of SMA 174
5.3.3 Comparison Among Several Algorithms 179
5.4 Exercises, Analyses, and Computation 180

6 Network Structures of Multiple Sequences Induced by Mutation 183
6.1 General Method of Constructing the Phylogenetic Tree 183
6.1.1 Summary 183
6.1.2 Distance-Based Methods 184
6.1.3 Feature-Based (Maximum Parsimony) Methods 188
6.1.4 Maximum-Likelihood Method and the Bayes Method 191
6.2 Network Structure Generated by MA 197
6.2.1 Graph and Tree Generated by MA 197
6.2.2 Network System Generated by Mutations of Multiple Sequences 200
6.3 The Application of Mutation Network Analysis 206
6.3.1 Selection of the Data Sample 206
6.3.2 The Basic Steps to Analyze the Sequences 208
6.3.3 Remarks on the Alignment and Output Analysis 210
6.4 Exercises, Analyses, and Computation 216

7 Alignment Space 219
7.1 Construction of Alignment Space and Its Basic Theorems 219
7.1.1 What Is Alignment Space? 219
7.1.2 The Alignment Space Under General Metric 221
7.2 The Analysis of Data Structures in Alignment Spaces 224
7.2.1 Maximum Score Alignment and Minimum Penalty Alignment 224
7.2.2 The Structure Mode of the Envelope of Pairwise Sequences 226
7.2.3 Uniqueness of the Maximum Core and Minimum Envelope of Pairwise Sequences 230
7.2.4 The Envelope and Core of Pairwise Sequences 231
7.2.5 The Envelope of Pairwise Sequences and Its Alignment Sequences 233

7.3 The Counting Theorem of the Optimal Alignment and Alignment Spheroid . . . 237
7.3.1 The Counting Theorem of the Optimal Alignment . . . 237
7.3.2 Alignment Spheroid . . . 238
7.4 The Virtual Symbol Operation in the Alignment Space . . . 241
7.4.1 The Definition of the Virtual Symbol Operator . . . 241
7.4.2 The Modulus Structure of the Virtual Symbol Operator . . . 243
7.4.3 The Isometric Operation and Micro-Adapted Operation of Virtual Symbols . . . 247
7.5 Exercises, Analyses, and Computation . . . 250

Part II Protein Configuration Analysis

8 Background Information Concerning the Properties of Proteins . . . 255
8.1 Amino Acids and Peptide Chains . . . 255
8.1.1 Amino Acids . . . 255
8.1.2 Basic Structure of Peptide Chains . . . 257
8.2 Brief Introduction of Protein Configuration Analysis . . . 259
8.2.1 Protein Structure Database . . . 259
8.2.2 Brief Introduction to Protein Structure Analysis . . . 260
8.3 Analysis and Exploration . . . 263

9 Informational and Statistical Iterative Analysis of Protein Secondary Structure Prediction . . . 265
9.1 Protein Secondary Structure Prediction and Informational and Statistical Iterative Analysis . . . 265
9.1.1 Protein Secondary Structure Prediction . . . 265
9.1.2 Data Source and Informational Statistical Model of PSSP . . . 267
9.1.3 Informational and Statistical Characteristic Analysis on Protein Secondary Structure . . . 269
9.2 Informational and Statistical Calculation Algorithms for PSSP . . . 271
9.2.1 Informational and Statistical Calculation for PSSP . . . 271
9.2.2 Informational and Statistical Calculation Algorithm for PSSP . . . 273
9.2.3 Discussion of the Results . . . 275
9.3 Exercises, Analyses, and Computation . . . 277

10 Three-Dimensional Structure Analysis of the Protein Backbone and Side Chains . . . 279
10.1 Space Conformation Theory of Four-Atom Points . . . 279
10.1.1 Conformation Parameter System of Four-Atom Space Points . . . 279
10.1.2 Phase Analysis on Four-Atom Space Points . . . 283
10.1.3 Four-Atom Construction of Protein 3D Structure . . . 286
10.2 Triangle Splicing Structure of Protein Backbones . . . 288
10.2.1 Triangle Splicing Belts of Protein Backbones . . . 288
10.2.2 Characteristic Analysis of the Protein Backbone Triangle Splicing Belts . . . 291
10.3 Structure Analysis of Protein Side Chains . . . 292
10.3.1 The Setting of Oxygen Atom O and Atom C_B on the Backbones . . . 293
10.3.2 Statistical Analysis of the Structures of Tetrahedrons V_O, V_B . . . 295
10.4 Exercises, Analyses, and Computation . . . 297

11 Alignment of Primary and Three-Dimensional Structures of Proteins . . . 299
11.1 Structure Analysis for Protein Sequences . . . 299
11.1.1 Alignment of Protein Sequences . . . 299
11.1.2 Differences and Similarities Between the Alignment of Protein Sequences and of DNA Sequences . . . 301
11.1.3 The Penalty Functions for the Alignment of Protein Sequences . . . 302
11.1.4 Key Points of the Alignment Algorithms of Protein Sequences . . . 306
11.2 Alignment of Protein Three-Dimensional Structures . . . 307
11.2.1 Basic Idea and Method of the Alignment of Three-Dimensional Structures . . . 307
11.2.2 Example of Computation in the Discrete Case . . . 310
11.2.3 Example of Computation in Consecutive Case . . . 314
11.3 Exercises, Analyses, and Computation . . . 323

12 Depth Analysis of Protein Spatial Structure . . . 325
12.1 Depth Analysis of Amino Acids in Proteins . . . 325
12.1.1 Introduction to the Concept of Depth . . . 325
12.1.2 Definition and Calculation of Depth . . . 327
12.1.3 Proof of Theorem 40 . . . 329
12.1.4 Proof of Theorem 41 . . . 334
12.2 Statistical Depth Analysis of Protein Spatial Particles . . . 335
12.2.1 Calculation for Depth Tendency Factor of Amino Acid . 335
12.2.2 Analysis of Depth Tendency Factor of Amino Acid . . . 338

12.2.3 Prediction for Depth of Multiple Peptide Chains in Protein Spatial Structure . . . 342
12.2.4 The Level Function in Spatial Particle System . . . 347
12.2.5 An Example . . . 347
12.3 Exercises . . . 353

13 Analysis of the Morphological Features of Protein Spatial Structure . . . 355
13.1 Introduction . . . 355
13.1.1 Morphological Features of Protein Spatial Structure . . . 355
13.1.2 Several Basic Definitions and Symbols . . . 357
13.1.3 Preliminary Analysis of the Morphology of Spatial Particle System . . . 360
13.1.4 Example . . . 362
13.2 Structural Analysis of Cavities and Channels in a Particle System . . . 366
13.2.1 Definition, Classification and Calculation of Cavity . . . 366
13.2.2 Relationship Between Cavities . . . 368
13.2.3 Example . . . 371
13.3 Analysis of γ-Accessible Radius in Spatial Particle System . . . 371
13.3.1 Structural Analysis of a Directed Polyhedron . . . 371
13.3.2 Definition and Basic Properties of γ-Accessible Radius . 377
13.3.3 Basic Principles and Methods of γ-Accessible Radius . . . 378
13.4 Recursive Algorithm of γ-Function . . . 382
13.4.1 Calculation of the γ-Function Generated by 0-Level Convex Hull . . . 382
13.4.2 Recursive Calculation of γ-Function . . . 384
13.4.3 Example . . . 387
13.5 Proof of Relative Theorems and Reasoning of Computational Formulas . . . 387
13.5.1 Proofs of Several Theorems . . . 387
13.5.2 Reasoning of Several Formulas . . . 390
13.6 Exercises . . . 394

14 Semantic Analysis for Protein Primary Structure . . . 395
14.1 Semantic Analysis for Protein Primary Structure . . . 395
14.1.1 The Definition of Semantic Analysis for Protein Primary Structure . . . 395
14.1.2 Information-Based Stochastic Models in Semantic Analysis . . . 397
14.1.3 Determination of Local Words Using Informational and Statistical Means and the Relative Entropy Density Function for the Second and Third Ranked Words . . . 400
14.1.4 Semantic Analysis for Protein Primary Structure . . . 402

14.2 Permutation and Combination Methods for Semantic Structures . . . 412
14.2.1 Notation Used in Combinatorial Graph Theory . . . 416
14.2.2 The Complexity of Databases . . . 420
14.2.3 Key Words and Core Words in a Database . . . 421
14.2.4 Applications of Combinatorial Analysis . . . 425
14.3 Exercises, Analyses, and Computation . . . 429

Epilogue . . . 431

References . . . 433

Index . . . 441

Outline

This book discusses several important issues including mathematical and computational methods and their applications to bioinformatics.

This book contains two parts. Part I introduces sequence mutations and the theory of data structure used in alignment. Stochastic models of alignment and algebraic theory are introduced as a means to describe data structure. Dynamic programming algorithms and statistical decision algorithms for pairwise sequence alignment, fast alignment algorithms, and the analysis and application of multiple sequence alignment output are also introduced. Part II includes the introduction of frequency analysis, cluster analysis and semantic analysis of the primary sequences, the introduction of the 3D-structure analysis of the main chains and the side chains, and the introduction of configuration analysis.

Many mathematical theories are presented in this monograph, such as stochastic analysis, combinational graph theory, geometry and algebra, informatics, intelligent computation, etc. A large number of algorithms and corresponding software packages make use of these theories, which have been developed to deal with the large amounts of biological and clinical data that play such an important role in the fields of bioinformatics, biomedicine, and so on.

This book has three main goals. The first is to introduce these classical mathematical theories and methods within the context of the current state of mathematics, as they are used in the fields of molecular biology and bioinformatics. The second is to discuss the potential mathematical requirements in the study of molecular biology and bioinformatics, which will drive the development of new theories and methods. Our third goal is to propose a framework within which bioinformatics may be combined with mathematics.

Within each chapter, we have included results and references from our own research experience, to illustrate our points. It is our hope that this book will be useful as a textbook for both undergraduate and graduate students, as well as a reference for teachers or researchers in the field of bioinformatics or mathematics, or those interested in understanding the relation between the two.

此为试读，需要完整PDF请访问：www.ertongbook.com