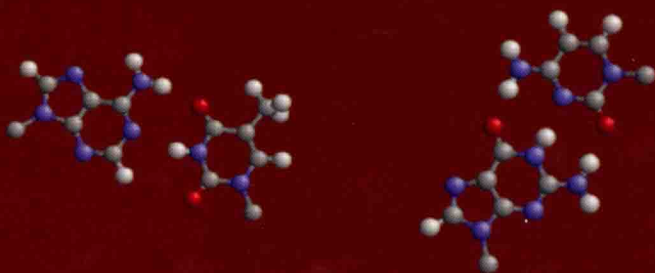
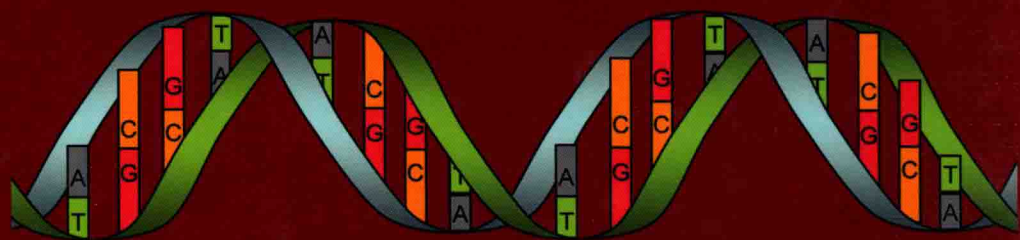


Chapman & Hall/CRC
Mathematical and Computational Biology Series

ALGORITHMS IN BIOINFORMATICS

A PRACTICAL INTRODUCTION



WING-KIN SUNG



CRC Press

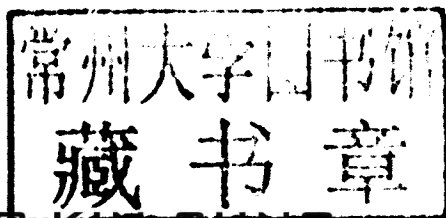
Taylor & Francis Group

A CHAPMAN & HALL BOOK

Chapman & Hall/CRC Mathematical and Computational Biology Series

ALGORITHMS IN BIOINFORMATICS

A PRACTICAL INTRODUCTION



WING KIN SONG



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business

A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2010 by Taylor and Francis Group, LLC
Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-7033-0 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Sung, Wing-Kin.

Algorithms in bioinformatics : a practical introduction / Wing-Kin Sung.

p. cm. -- (CHAPMAN & HALL/CRC mathematical and computational biology series)

Includes bibliographical references and index.

ISBN 978-1-4200-7033-0 (hardcover : alk. paper)

1. Bioinformatics. 2. Genetic algorithms. I. Title. II. Series.

QH324.2.S86 2009

572.80285--dc22

2009030738

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

ALGORITHMS IN BIOINFORMATICS

A PRACTICAL INTRODUCTION

CHAPMAN & HALL/CRC

Mathematical and Computational Biology Series

Aims and scope:

This series aims to capture new developments and summarize what is known over the whole spectrum of mathematical and computational biology and medicine. It seeks to encourage the integration of mathematical, statistical and computational methods into biology by publishing a broad range of textbooks, reference works and handbooks. The titles included in the series are meant to appeal to students, researchers and professionals in the mathematical, statistical and computational sciences, fundamental biology and bioengineering, as well as interdisciplinary researchers involved in the field. The inclusion of concrete examples and applications, and programming techniques and examples, is highly encouraged.

Series Editors

Alison M. Etheridge
Department of Statistics
University of Oxford

Louis J. Gross
Department of Ecology and Evolutionary Biology
University of Tennessee

Suzanne Lenhart
Department of Mathematics
University of Tennessee

Philip K. Maini
Mathematical Institute
University of Oxford

Shoba Ranganathan
Research Institute of Biotechnology
Macquarie University

Hershel M. Safer
Weizmann Institute of Science
Bioinformatics & Bio Computing

Eberhard O. Voit
The Wallace H. Couter Department of Biomedical Engineering
Georgia Tech and Emory University

Proposals for the series should be submitted to one of the series editors above or directly to:

CRC Press, Taylor & Francis Group

4th, Floor, Albert House
1-4 Singer Street
London EC2A 4BQ
UK

Published Titles

Algorithms in Bioinformatics: A Practical Introduction

Wing-Kin Sung

Bioinformatics: A Practical Approach

Shui Qing Ye

Cancer Modelling and Simulation

Luigi Preziosi

Combinatorial Pattern Matching Algorithms in Computational Biology Using Perl and R

Gabriel Valiente

Computational Biology: A Statistical Mechanics Perspective

Ralf Blossey

Computational Neuroscience: A Comprehensive Approach

Jianfeng Feng

Data Analysis Tools for DNA Microarrays

Sorin Draghici

Differential Equations and Mathematical Biology

D.S. Jones and B.D. Sleeman

Engineering Genetic Circuits

Chris J. Myers

Exactly Solvable Models of Biological Invasion

Sergei V. Petrovskii and Bai-Lian Li

Gene Expression Studies Using Affymetrix Microarrays

Hinrich Göhlmann and Willem Talloen

Glycome Informatics: Methods and Applications

Kiyoko F. Aoki-Kinoshita

Handbook of Hidden Markov Models in Bioinformatics

Martin Gollery

Introduction to Bioinformatics

Anna Tramontano

An Introduction to Systems Biology: Design Principles of Biological Circuits

Uri Alon

Kinetic Modelling in Systems Biology

Oleg Demin and Igor Goryanin

Knowledge Discovery in Proteomics

Igor Jurisica and Dennis Wigle

Meta-analysis and Combining Information in Genetics and Genomics

Rudy Guerra and Darlene R. Goldstein

Modeling and Simulation of Capsules and Biological Cells

C. Pozrikidis

Niche Modeling: Predictions from Statistical Distributions

David Stockwell

Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems

Qiang Cui and Ivet Bahar

Optimal Control Applied to Biological Models

Suzanne Lenhart and John T. Workman

Pattern Discovery in Bioinformatics: Theory & Algorithms

Laxmi Parida

Python for Bioinformatics

Sebastian Bassi

Spatial Ecology

*Stephen Cantrell, Chris Cosner, and
Shigui Ruan*

Spatiotemporal Patterns in Ecology and Epidemiology: Theory, Models, and Simulation

*Horst Malchow, Sergei V. Petrovskii, and
Ezio Venturino*

Stochastic Modelling for Systems Biology

Darren J. Wilkinson

Structural Bioinformatics: An Algorithmic Approach

Forbes J. Burkowski

The Ten Most Wanted Solutions in Protein Bioinformatics

Anna Tramontano

Preface

Bioinformatics is the study of biology through computer modeling or analysis. It is a multi-discipline research involving biology, statistics, data-mining, machine learning, and algorithms. This book is intended to give an in-depth introduction of the algorithmic techniques applied in bioinformatics.

The primary audiences of this book is advanced undergraduate students and graduate students who are from mathematics or computer science departments. We assume no prior knowledge of molecular biology beyond the high school level. In fact, the first chapter gives a brief introduction to molecular biology. Moreover, we do assume the reader has some training in college-level discrete mathematics and algorithms.

This book was developed from the teaching material for the course on “Combinatorial Methods in Bioinformatics”, which I taught at the National University of Singapore, Singapore. The chapters in this book is classified based on the biology application domains. For each topic, an in-depth biological motivation is given and the corresponding computation problems are precisely defined. Different methods and the corresponding algorithms are also provided. Furthermore, the book gives detailed examples to illustrate each algorithm. At the end of each chapter, a set of exercises is provided.

Below, we give a brief overview of the chapters in the book.

Chapter 1 introduces the basic concepts in molecular biology. It describes the building blocks of our cells which include DNA, RNA, and protein. Then, it describes the mechanism in the cell and some basic biotechnologies. It also briefly describes the history of bioinformatics.

Chapter 2 describes methods to measure sequence similarity, which is fundamental for comparing DNA, RNA, and protein sequences. We discuss various alignment methods, including global alignment, local alignment, and semi-global alignment. We also study gap penalty and scoring function.

Chapter 3 introduces the suffix tree and gives its simple applications. We also present Farach’s algorithm for constructing a suffix tree. Furthermore, we study variants of the suffix tree including suffix array and FM-index. We also discuss how to use suffix array for approximate matching.

Chapter 4 discusses methods for aligning whole genomes. We discuss MUMmer and Mutation Sensitive Alignment. Both methods apply the suffix tree and the longest common subsequence algorithm.

Chapter 5 considers the problem of searching a sequence database. Due to the advance of biotechnology, sequence data (including DNA, RNA, and protein) increases exponentially. Hence, it is important to have methods which

allow efficient database searching. In this chapter, we discuss various biological database searching methods including FASTA, BLAST, BLAT, QUASAR, BWT-SW, etc.

Chapter 6 introduces methods for aligning multiple biological sequences. The chapter describes four algorithms: an exact solution based on dynamic programming, an approximation algorithm based on star alignment, and two heuristics. The two heuristics are ClustalW (a progressive alignment method) and MUSCLE (an iterative method).

Chapter 7 first describes a phylogenetic tree and its applications. Then, we discuss how to construct a phylogenetic tree given a character-based dataset or a distance-based dataset. We cover the methods: maximum parsimony, compatibility, maximum likelihood, UPGMA, and neighbor-joining. Lastly, we discuss whether the character-based methods and the distance-based methods can reconstruct the correct phylogenetic tree.

Chapter 8 covers the methods for comparing phylogenetic trees. We discuss methods for computing similarity and distance. For similarity, we consider maximum agreement subtree (MAST). For distance, we consider Robinson-Foulds distance, nearest neighbor interchange (NNI) distance, subtree transfer (STT) distance, and quartet distance. Furthermore, we discuss methods for finding consensus of a set of trees. We consider strict consensus tree, majority rule consensus tree, median tree, greedy consensus tree, and R^* consensus tree.

Chapter 9 investigates the problem of genome rearrangement. We discuss various possible genome rearrangements including reversal, transposition, etc. Since reversal can simulate other types of genome rearrangements, this chapter focuses on reversal distance. For computing the unsigned reversal distance, the problem is NP-hard. We describe a 2-approximation algorithm for this problem. For computing the signed reversal distance, we present the Hannenhalli-Pevzner theorem and Bergeron's algorithm.

Chapter 10 introduces the problem of motif finding. We discuss a number of de novo motif finding methods including Gibb Sampler, MEME, SP-star, YMF, and suffix tree-based methods like Weeder. Since multiple motif finders exist, we also discuss ensemble methods like MotifVoter, which combines information from multiple motif finders. All the above methods perform de novo motif finding with no extra information. Last, two motif finding methods which utilize additional information are described. The first method RE-DUCE improves motif finding by combining microarray and sequence data. The second method uses phylogenetic information to improve motif finding.

Chapter 11 discusses methods for predicting the secondary structure of RNA. When there is no pseudoknot, we discuss the Nussinov algorithm and the ZUKER algorithm. When pseudoknot is allowed, we discuss Akutsu's algorithm.

Chapter 12 covers methods for reconstructing the peptide sequence using a mass spectrometer. We discuss both de novo peptide sequencing and

database search methods. For de novo peptide sequencing, we discuss Peaks and Sherenga. For database searching, we discuss SEQUEST.

Chapter 13 covers the computational problems related to population genetics. We discuss Hardy-Weinberg equilibrium and linkage disequilibrium. Then, we discuss algorithms for genotype phasing, tag SNP selection, and association study.

Supplementary material can be found at

http://www.comp.nus.edu.sg/~ksung/algo_in_bioinfo/.

I would like to thank the students who took my classes. I thank them for their efforts in writing the lecture scripts. I would also like to thank Xu Han, Caroline Lee Guat Lay, Charlie Lee, and Guoliang Li for helping me to proofread some of the chapters.

I would like to thank my PhD supervisors Tak-Wah Lam and Hing-Fung Ting and my collaborators Francis Y. L. Chin, Kwok Pui Choi, Edwin Chung, Wing Kai Hon, Jansson Jesper, Ming-Yang Kao, Hon Wai Leong, See-Kiong Ng, Franco P. Preparata, Yijun Ruan, Kunihiko Sadakane, Chialin Wei, Limsoon Wong, Siu-Ming Yiu, and Louxin Zhang. My knowledge of bioinformatics was enriched through numerous discussions with them. I would also like to thank my parents Kang Fai Sung and Siu King Wong, my three brothers Wing Hong Sung, Wing Keung Sung, and Wing Fu Sung, my wife Lily Or, and my two daughters Kelly and Kathleen for their support.

Finally, if you have any suggestions for improvement or if you identify any errors in the book, please send an email to me at ksung@comp.nus.edu.sg. I thank you in advance for your helpful comments in improving the book.

Wing-Kin Sung

Contents

| | |
|---|-----------|
| Preface | xv |
| 1 Introduction to Molecular Biology | 1 |
| 1.1 DNA, RNA, and Protein | 1 |
| 1.1.1 Proteins | 1 |
| 1.1.2 DNA | 4 |
| 1.1.3 RNA | 9 |
| 1.2 Genome, Chromosome, and Gene | 10 |
| 1.2.1 Genome | 10 |
| 1.2.2 Chromosome | 10 |
| 1.2.3 Gene | 11 |
| 1.2.4 Complexity of the Organism versus Genome Size | 11 |
| 1.2.5 Number of Genes versus Genome Size | 11 |
| 1.3 Replication and Mutation of DNA | 12 |
| 1.4 Central Dogma (from DNA to Protein) | 13 |
| 1.4.1 Transcription (Prokaryotes) | 13 |
| 1.4.2 Transcription (Eukaryotes) | 14 |
| 1.4.3 Translation | 15 |
| 1.5 Post-Translation Modification (PTM) | 17 |
| 1.6 Population Genetics | 18 |
| 1.7 Basic Biotechnological Tools | 18 |
| 1.7.1 Restriction Enzymes | 19 |
| 1.7.2 Sonication | 19 |
| 1.7.3 Cloning | 19 |
| 1.7.4 PCR | 20 |
| 1.7.5 Gel Electrophoresis | 22 |
| 1.7.6 Hybridization | 23 |
| 1.7.7 Next Generation DNA Sequencing | 24 |
| 1.8 Brief History of Bioinformatics | 26 |
| 1.9 Exercises | 27 |
| 2 Sequence Similarity | 29 |
| 2.1 Introduction | 29 |
| 2.2 Global Alignment Problem | 30 |
| 2.2.1 Needleman-Wunsch Algorithm | 32 |
| 2.2.2 Running Time Issue | 34 |
| 2.2.3 Space Efficiency Issue | 35 |

| | | |
|----------|---|-----------|
| 2.2.4 | More on Global Alignment | 38 |
| 2.3 | Local Alignment | 39 |
| 2.4 | Semi-Global Alignment | 41 |
| 2.5 | Gap Penalty | 42 |
| 2.5.1 | General Gap Penalty Model | 43 |
| 2.5.2 | Affine Gap Penalty Model | 43 |
| 2.5.3 | Convex Gap Model | 45 |
| 2.6 | Scoring Function | 50 |
| 2.6.1 | Scoring Function for DNA | 50 |
| 2.6.2 | Scoring Function for Protein | 51 |
| 2.7 | Exercises | 53 |
| 3 | Suffix Tree | 57 |
| 3.1 | Introduction | 57 |
| 3.2 | Suffix Tree | 57 |
| 3.3 | Simple Applications of a Suffix Tree | 59 |
| 3.3.1 | Exact String Matching Problem | 59 |
| 3.3.2 | Longest Repeated Substring Problem | 60 |
| 3.3.3 | Longest Common Substring Problem | 60 |
| 3.3.4 | Longest Common Prefix (LCP) | 61 |
| 3.3.5 | Finding a Palindrome | 62 |
| 3.3.6 | Extracting the Embedded Suffix Tree of a String from the Generalized Suffix Tree | 63 |
| 3.3.7 | Common Substring of 2 or More Strings | 64 |
| 3.4 | Construction of a Suffix Tree | 65 |
| 3.4.1 | Step 1: Construct the Odd Suffix Tree | 68 |
| 3.4.2 | Step 2: Construct the Even Suffix Tree | 69 |
| 3.4.3 | Step 3: Merge the Odd and the Even Suffix Trees | 70 |
| 3.5 | Suffix Array | 72 |
| 3.5.1 | Construction of a Suffix Array | 73 |
| 3.5.2 | Exact String Matching Using a Suffix Array | 73 |
| 3.6 | FM-Index | 76 |
| 3.6.1 | Definition | 77 |
| 3.6.2 | The <i>occ</i> Data Structure | 78 |
| 3.6.3 | Exact String Matching Using the FM-Index | 79 |
| 3.7 | Approximate Searching Problem | 81 |
| 3.8 | Exercises | 82 |
| 4 | Genome Alignment | 87 |
| 4.1 | Introduction | 87 |
| 4.2 | Maximum Unique Match (MUM) | 88 |
| 4.2.1 | How to Find MUMs | 89 |
| 4.3 | MUMmer1: LCS | 92 |
| 4.3.1 | Dynamic Programming Algorithm in $O(n^2)$ Time | 93 |
| 4.3.2 | An $O(n \log n)$ -Time Algorithm | 93 |

| | | |
|----------|---|------------|
| 4.4 | MUMmer2 and MUMmer3 | 96 |
| 4.4.1 | Reducing Memory Usage | 97 |
| 4.4.2 | Employing a New Alternative Algorithm for Finding MUMs | 97 |
| 4.4.3 | Clustering Matches | 97 |
| 4.4.4 | Extension of the Definition of MUM | 98 |
| 4.5 | Mutation Sensitive Alignment | 99 |
| 4.5.1 | Concepts and Definitions | 99 |
| 4.5.2 | The Idea of the Heuristic Algorithm | 100 |
| 4.5.3 | Experimental Results | 102 |
| 4.6 | Dot Plot for Visualizing the Alignment | 103 |
| 4.7 | Further Reading | 105 |
| 4.8 | Exercises | 105 |
| 5 | Database Search | 109 |
| 5.1 | Introduction | 109 |
| 5.1.1 | Biological Database | 109 |
| 5.1.2 | Database Searching | 109 |
| 5.1.3 | Types of Algorithms | 110 |
| 5.2 | Smith-Waterman Algorithm | 111 |
| 5.3 | FastA | 111 |
| 5.3.1 | FastP Algorithm | 112 |
| 5.3.2 | FastA Algorithm | 113 |
| 5.4 | BLAST | 114 |
| 5.4.1 | BLAST1 | 115 |
| 5.4.2 | BLAST2 | 116 |
| 5.4.3 | BLAST1 versus BLAST2 | 118 |
| 5.4.4 | BLAST versus FastA | 118 |
| 5.4.5 | Statistics for Local Alignment | 119 |
| 5.5 | Variations of the BLAST Algorithm | 120 |
| 5.5.1 | MegaBLAST | 120 |
| 5.5.2 | BLAT | 121 |
| 5.5.3 | PatternHunter | 121 |
| 5.5.4 | PSI-BLAST (Position-Specific Iterated BLAST) . . | 123 |
| 5.6 | Q-gram Alignment based on Suffix ARrays (QUASAR) | 124 |
| 5.6.1 | Algorithm | 124 |
| 5.6.2 | Speeding Up and Reducing the Space for QUASAR | 127 |
| 5.6.3 | Time Analysis | 127 |
| 5.7 | Locality-Sensitive Hashing | 128 |
| 5.8 | BWT-SW | 130 |
| 5.8.1 | Aligning Query Sequence to Suffix Tree | 130 |
| 5.8.2 | Meaningful Alignment | 133 |
| 5.9 | Are Existing Database Searching Methods Sensitive Enough? | 136 |
| 5.10 | Exercises | 136 |

| | | |
|----------|---|------------|
| 6 | Multiple Sequence Alignment | 139 |
| 6.1 | Introduction | 139 |
| 6.2 | Formal Definition of the Multiple Sequence Alignment Problem | 139 |
| 6.3 | Methods for Solving the MSA Problem | 141 |
| 6.4 | Dynamic Programming Method | 142 |
| 6.5 | Center Star Method | 143 |
| 6.6 | Progressive Alignment Method | 146 |
| 6.6.1 | ClustalW | 147 |
| 6.6.2 | Profile-Profile Alignment | 147 |
| 6.6.3 | Limitation of Progressive Alignment Construction | 149 |
| 6.7 | Iterative Method | 149 |
| 6.7.1 | MUSCLE | 150 |
| 6.7.2 | Log-Expectation (LE) Score | 151 |
| 6.8 | Further Reading | 151 |
| 6.9 | Exercises | 152 |
| 7 | Phylogeny Reconstruction | 155 |
| 7.1 | Introduction | 155 |
| 7.1.1 | Mitochondrial DNA and Inheritance | 155 |
| 7.1.2 | The Constant Molecular Clock | 155 |
| 7.1.3 | Phylogeny | 156 |
| 7.1.4 | Applications of Phylogeny | 157 |
| 7.1.5 | Phylogenetic Tree Reconstruction | 158 |
| 7.2 | Character-Based Phylogeny Reconstruction Algorithm | 159 |
| 7.2.1 | Maximum Parsimony | 159 |
| 7.2.2 | Compatibility | 165 |
| 7.2.3 | Maximum Likelihood Problem | 172 |
| 7.3 | Distance-Based Phylogeny Reconstruction Algorithm | 178 |
| 7.3.1 | Additive Metric and Ultrametric | 179 |
| 7.3.2 | Unweighted Pair Group Method with Arithmetic Mean (UPGMA) | 184 |
| 7.3.3 | Additive Tree Reconstruction | 187 |
| 7.3.4 | Nearly Additive Tree Reconstruction | 189 |
| 7.3.5 | Can We Apply Distance-Based Methods Given a Character-State Matrix? | 190 |
| 7.4 | Bootstrapping | 191 |
| 7.5 | Can Tree Reconstruction Methods Infer the Correct Tree? | 192 |
| 7.6 | Exercises | 193 |
| 8 | Phylogeny Comparison | 199 |
| 8.1 | Introduction | 199 |
| 8.2 | Similarity Measurement | 200 |
| 8.2.1 | Computing MAST by Dynamic Programming | 201 |
| 8.2.2 | MAST for Unrooted Trees | 202 |

| | | |
|-----------|--|------------|
| 8.3 | Dissimilarity Measurements | 203 |
| 8.3.1 | Robinson-Foulds Distance | 204 |
| 8.3.2 | Nearest Neighbor Interchange Distance (NNI) | 209 |
| 8.3.3 | Subtree Transfer Distance (STT) | 210 |
| 8.3.4 | Quartet Distance | 211 |
| 8.4 | Consensus Tree Problem | 214 |
| 8.4.1 | Strict Consensus Tree | 215 |
| 8.4.2 | Majority Rule Consensus Tree | 216 |
| 8.4.3 | Median Consensus Tree | 218 |
| 8.4.4 | Greedy Consensus Tree | 218 |
| 8.4.5 | R^* Tree | 219 |
| 8.5 | Further Reading | 220 |
| 8.6 | Exercises | 222 |
| 9 | Genome Rearrangement | 225 |
| 9.1 | Introduction | 225 |
| 9.2 | Types of Genome Rearrangements | 225 |
| 9.3 | Computational Problems | 227 |
| 9.4 | Sorting an Unsigned Permutation by Reversals | 227 |
| 9.4.1 | Upper and Lower Bound on an Unsigned Reversal Dis- tance | 228 |
| 9.4.2 | 4-Approximation Algorithm for Sorting an Unsigned Permutation | 229 |
| 9.4.3 | 2-Approximation Algorithm for Sorting an Unsigned Permutation | 230 |
| 9.5 | Sorting a Signed Permutation by Reversals | 232 |
| 9.5.1 | Upper Bound on Signed Reversal Distance | 232 |
| 9.5.2 | Elementary Intervals, Cycles, and Components | 233 |
| 9.5.3 | The Hannenhalli-Pevzner Theorem | 238 |
| 9.6 | Further Reading | 243 |
| 9.7 | Exercises | 244 |
| 10 | Motif Finding | 247 |
| 10.1 | Introduction | 247 |
| 10.2 | Identifying Binding Regions of TFs | 248 |
| 10.3 | Motif Model | 250 |
| 10.4 | The Motif Finding Problem | 252 |
| 10.5 | Scanning for Known Motifs | 253 |
| 10.6 | Statistical Approaches | 254 |
| 10.6.1 | Gibbs Motif Sampler | 255 |
| 10.6.2 | MEME | 257 |
| 10.7 | Combinatorial Approaches | 260 |
| 10.7.1 | Exhaustive Pattern-Driven Algorithm | 261 |
| 10.7.2 | Sample-Driven Approach | 262 |
| 10.7.3 | Suffix Tree-Based Algorithm | 263 |

| | | |
|-----------|---|------------|
| 10.7.4 | Graph-Based Method | 265 |
| 10.8 | Scoring Function | 266 |
| 10.9 | Motif Ensemble Methods | 267 |
| 10.9.1 | Approach of MotifVoter | 268 |
| 10.9.2 | Motif Filtering by the Discriminative and Consensus Criteria | 268 |
| 10.9.3 | Sites Extraction and Motif Generation | 270 |
| 10.10 | Can Motif Finders Discover the Correct Motifs? | 271 |
| 10.11 | Motif Finding Utilizing Additional Information | 274 |
| 10.11.1 | Regulatory Element Detection Using Correlation with Expression | 274 |
| 10.11.2 | Discovery of Regulatory Elements by Phylogenetic Footprinting | 277 |
| 10.12 | Exercises | 279 |
| 11 | RNA Secondary Structure Prediction | 281 |
| 11.1 | Introduction | 281 |
| 11.1.1 | Base Interactions in RNA | 282 |
| 11.1.2 | RNA Structures | 282 |
| 11.2 | Obtaining RNA Secondary Structure Experimentally | 285 |
| 11.3 | RNA Structure Prediction Based on Sequence Only | 286 |
| 11.4 | Structure Prediction with the Assumption That There is No Pseudoknot | 286 |
| 11.5 | Nussinov Folding Algorithm | 288 |
| 11.6 | ZUKER Algorithm | 290 |
| 11.6.1 | Time Analysis | 292 |
| 11.6.2 | Speeding up Multi-Loops | 292 |
| 11.6.3 | Speeding up Internal Loops | 294 |
| 11.7 | Structure Prediction with Pseudoknots | 296 |
| 11.7.1 | Definition of a Simple Pseudoknot | 296 |
| 11.7.2 | Akutsu's Algorithm for Predicting an RNA Secondary Structure with Simple Pseudoknots | 297 |
| 11.8 | Exercises | 300 |
| 12 | Peptide Sequencing | 305 |
| 12.1 | Introduction | 305 |
| 12.2 | Obtaining the Mass Spectrum of a Peptide | 306 |
| 12.3 | Modeling the Mass Spectrum of a Fragmented Peptide | 310 |
| 12.3.1 | Amino Acid Residue Mass | 310 |
| 12.3.2 | Fragment Ion Mass | 310 |
| 12.4 | De Novo Peptide Sequencing Using Dynamic Programming | 312 |
| 12.4.1 | Scoring by Considering y-Ions | 313 |
| 12.4.2 | Scoring by Considering y-Ions and b-Ions | 314 |
| 12.5 | De Novo Sequencing Using Graph-Based Approach | 317 |
| 12.6 | Peptide Sequencing via Database Search | 319 |

| | | |
|-----------|--|------------|
| 12.7 | Further Reading | 320 |
| 12.8 | Exercises | 321 |
| 13 | Population Genetics | 323 |
| 13.1 | Introduction | 323 |
| 13.1.1 | Locus, Genotype, Allele, and SNP | 323 |
| 13.1.2 | Genotype Frequency and Allele Frequency | 324 |
| 13.1.3 | Haplotype and Phenotype | 325 |
| 13.1.4 | Technologies for Studying the Human Population | 325 |
| 13.1.5 | Bioinformatics Problems | 325 |
| 13.2 | Hardy-Weinberg Equilibrium | 326 |
| 13.3 | Linkage Disequilibrium | 327 |
| 13.3.1 | D and D' | 328 |
| 13.3.2 | r^2 | 328 |
| 13.4 | Genotype Phasing | 328 |
| 13.4.1 | Clark's Algorithm | 329 |
| 13.4.2 | Perfect Phylogeny Haplotyping Problem | 330 |
| 13.4.3 | Maximum Likelihood Approach | 334 |
| 13.4.4 | Phase Algorithm | 336 |
| 13.5 | Tag SNP Selection | 337 |
| 13.5.1 | Zhang et al.'s Algorithm | 338 |
| 13.5.2 | IdSelect | 339 |
| 13.6 | Association Study | 339 |
| 13.6.1 | Categorical Data Analysis | 340 |
| 13.6.2 | Relative Risk and Odds Ratio | 341 |
| 13.6.3 | Linear Regression | 342 |
| 13.6.4 | Logistic Regression | 343 |
| 13.7 | Exercises | 344 |
| | References | 349 |
| | Index | 375 |

Chapter 1

Introduction to Molecular Biology

1.1 DNA, RNA, and Protein

Our bodies consist of a number of organs. Each organ is composed of a number of tissues, and each tissue is a collection of similar cells that group together to perform specialized functions. The individual cell is the minimal self-reproducing unit in all living species. It performs two types of functions: (1) stores and passes the genetic information for maintaining life from generation to generation; and (2) performs chemical reactions necessary to maintain our life.

For function (1), our cells store the genetic information in the form of double-stranded DNA. For function (2), portions of the DNA called genes are transcribed into closely related molecules called RNAs. RNAs guide the synthesis of protein molecules. The resultant proteins are the main catalysts for almost all the chemical reactions in the cell. In addition to catalysis, proteins are involved in transportation, signaling, cell-membrane formation, etc.

Below, we discuss these three main molecules in our cells, namely, protein, DNA, and RNA.

1.1.1 Proteins

Proteins constitute most of a cell's dry mass. They are not only the building blocks from which cells are built, but also execute nearly all cell functions. Understanding proteins can guide us to understand how our bodies function and other biological processes.

A protein is made from a long chain of amino acids, each linking to its neighbor through a covalent peptide bond. Therefore, proteins are also known as polypeptides. There are 20 types of amino acids and each amino acid carries different chemical properties. The length of a protein is in the range of 20 to more than 5000 amino acids. On average, a protein contains around 350 amino acids.

In order to perform their chemical functions, proteins need to fold into certain 3 dimensional shapes. The folding of the proteins is caused by the weak interactions among amino acid residues. The weak interactions include