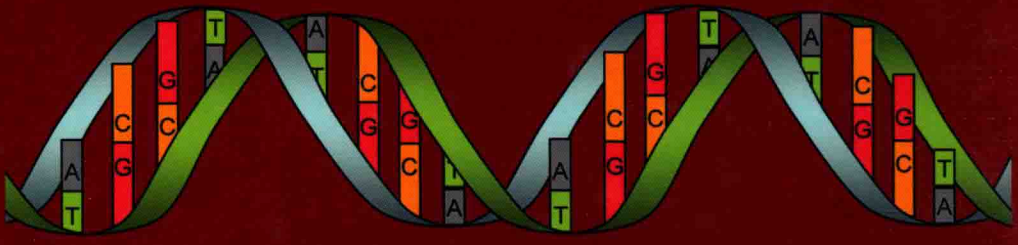
Chapman & Hall/CRC
Mathematical and Computational Biology Series

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 SUNG

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

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)

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.

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 REDUCE 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

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