

Computational Biology

Editors-in-Chief

Andreas Dress

University of Bielefeld (Germany)

Martin Vingron

Max Planck Institute for Molecular Genetics (Germany)

Editorial Board

Gene Myers, Janelia Farm Research Campus, Howard Hughes Medical Institute (USA)

Robert Giegerich, University of Bielefeld (Germany)

Walter Fitch, University of California, Irvine (USA)

Pavel A. Pevzner, University of California, San Diego (USA)

Advisory Board

Gordon Grippen, University of Michigan

Joe Felsenstein, University of Washington (USA)

Dan Gusfield, University of California, Davis (USA)

Sorin Istrail, Brown University, Providence (USA)

Samuel Karlin, Stanford University

Thomas Lengauer, Max Planck Institut Informatik (Germany)

Marcella McClure, Montana State University (USA)

Martin Nowak, Harvard University (USA)

David Sankoff, University of Ottawa (Canada)

Ron Shamir, Tel Aviv University (Israel)

Mike Steel, University of Canterbury (New Zealand)

Gary Stormo, Washington University Medical School (USA)

Simon Tavaré, University of Southern California (USA)

Tandy Warnow, University of Texas, Austin (USA)

The *Computational Biology* series publishes the very latest, high-quality research devoted to specific issues in computer-assisted analysis of biological data. The main emphasis is on current scientific developments and innovative techniques in computational biology (bioinformatics), bringing to light methods from mathematics, statistics and computer science that directly address biological problems currently under investigation.

The series offers publications that present the state-of-the-art regarding the problems in question; show computational biology/bioinformatics methods at work; and finally discuss anticipated demands regarding developments in future methodology. Titles can range from focused monographs, to undergraduate and graduate textbooks, and professional text/reference works.

Author guidelines: [springer.com](http://www.springer.com) > Authors > Author Guidelines

For other titles published in this series, go to
<http://www.springer.com/series/5769>

Kun-Mao Chao • Louxin Zhang

Sequence Comparison

Theory and Methods



Springer

Kun-Mao Chao, BS, MS, PhD
Department of Computer Science and
Information Engineering,
National Taiwan University,
Taiwan

Louxin Zhang, BSc, MSc, PhD
Department of Mathematics,
National University of Singapore,
Singapore

Computational Biology Series ISSN 1568-2684

ISBN: 978-1-84800-319-4

e-ISBN: 978-1-84800-320-0

DOI 10.1007/978-1-84800-320-0

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2008928295

© Springer-Verlag London Limited 2009

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

The use of 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 laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Printed on acid-free paper

Springer Science+Business Media
springer.com

KMC:

To Daddy, Mommy, Pei-Pei and Liang

LXZ:

To my parents

Foreword

My first thought when I saw a preliminary version of this book was: Too bad there was nothing like this book when I really needed it.

Around 20 years ago, I decided it was time to change my research directions. After exploring a number of possibilities, I decided that the area of overlap between molecular biology and computer science (which later came to be called "bioinformatics") was my best bet for an exciting career. The next decision was to select a specific class of problems to work on, and the main criterion for me was that algorithmic methods would be the main key to success. I decided to work on sequence analysis. A book like this could have, so to speak, straightened my learning curve.

It is amazing to me that those two conclusions still apply: bioinformatics is a tremendously vibrant and rewarding field to be in, and sequence comparison is (arguably, at least) the subfield of bioinformatics where algorithmic techniques play the largest role in achieving success. The importance of sequence-analysis methods in bioinformatics can be measured objectively, simply by looking at the numbers of citations in the scientific literature for papers that describe successful developments; a high percentage of the most heavily cited scientific publications in the past 30 years are from this new field. Continued growth and importance of sequence analysis is guaranteed by the explosive development of new technologies for generating sequence data, where the cost has dropped 1000-fold in the past few years, and this fantastic decrease in cost means that sequencing and sequence analysis are taking over jobs that were previously handled another way.

Careful study of this book will be valuable for a wide range of readers, from students wanting to enter the field of bioinformatics, to experienced users of bioinformatic tools wanting to use tool options more intelligently, to bioinformatic specialists looking for the killer algorithm that will yield the next tool to sweep the field. I predict that you will need more than just mastery of this material to reach stardom in bioinformatics – there is also a huge amount of biology to be learned, together with a regular investment of time to keep up with the latest in data-generation technology and its applications. However, the material herein will remain useful for years, as new sequencing technologies and biological applications come and go.

I invite you to study this book carefully and apply ideas from it to one of the most exciting areas of science. And be grateful that two professionals with a combined 30 years of experience have taken the time to open the door for you.

State College, Pennsylvania

Webb Miller

June 2008

Preface

Biomolecular sequence comparison is the origin of bioinformatics. It has been extensively studied by biologists, computer scientists, and mathematicians for almost 40 years due to its numerous applications in biological sciences. Today, homology search is already a part of modern molecular biology. This book is a monograph on the state-of-the-art study of sequence alignment and homology search.

Sequence alignment, as a major topic of bioinformatics, is covered in most bioinformatics books. However, these books often tell one part of the story. The field is evolving. The BLAST program, a pearl of pearls, computes local alignments quickly and evaluates the statistical significance of any alignments that it finds. Although BLAST homology search is done more than 100,000 times per day, the statistical calculations used in this program are not widely understood by its users. In fact, these calculations keep on changing with advancement of alignment score statistics. Simply using BLAST without a reasonable understanding of its key ideas is not very different from using a PCR without knowing how PCR works. This is one of the motivations for us to write this book. It is intended for covering in depth a full spectrum of the field from alignment methods to the theory of scoring matrices and to alignment score statistics.

Sequence alignment deals with basic problems arising from processing DNA and protein sequence information. In the study of these problems, many powerful techniques have been invented. For instance, the filtration technique, powered with spaced seeds, is shown to be extremely efficient for comparing large genomes and for searching huge sequence databases. Local alignment score statistics have made homology search become a reliable method for annotating newly sequenced genomes. Without doubt, the ideas behind these outstanding techniques will enable new approaches in mining and processing structural information in biology. This is another motivation for us to write this book.

This book is composed of eight chapters and three appendixes. Chapter 1 works as a tutorial to help all levels of readers understand the connection among the other chapters. It discusses informally why biomolecular sequences are compared through alignment and how sequence alignment is done efficiently.

Chapters 2 to 5 form the method part. This part covers the basic algorithms and methods for sequence alignment. Chapter 2 introduces basic algorithmic techniques that are often used for solving various problems in sequence comparison.

In Chapter 3, we present the Needleman-Wunsch and Smith-Waterman algorithms, which, respectively, align a pair of sequences globally and locally, and their variants for coping with various gap penalty costs. For analysis of long genomic sequences, the space restriction is more critical than the time constraint. We therefore introduce an efficient space-saving strategy for sequence alignment. Finally, we discuss a few advanced topics of sequence alignment.

Chapter 4 introduces four popular homology search programs: FASTA, BLAST family, BLAT, and PatternHunter. We also discuss how to implement the filtration idea used in these programs with efficient data structures such as hash tables, suffix trees, and suffix arrays.

Chapter 5 covers briefly multiple sequence alignment. We discuss how a multiple sequence alignment is scored, and then show why the exact method based on a dynamic-programming approach is not feasible. Finally, we introduce the progressive alignment approach, which is adopted by ClustalW, MUSCLE, YAMA, and other popular programs for multiple sequence alignment.

Chapters 6 to 8 form the theory part. Chapter 6 covers the theoretic aspects of the seeding technique. PatternHunter demonstrates that an optimized spaced seed improves sensitivity substantially. Accordingly, elucidating the mechanism that confers power to spaced seeds and identifying good spaced seeds become new issues in homology search. This chapter presents a framework of studying these two issues by relating them to the probability of a spaced seed hitting a random alignment. We address why spaced seeds improve homology search sensitivity and discuss how to design good spaced seeds.

The Karlin-Altschul statistics of optimal local alignment scores are covered in Chapter 7. Optimal segment scores are shown to follow an extreme value distribution in asymptotic limit. The Karlin-Altschul sum statistic is also introduced. In the case of gapped local alignment, we describe how the statistical parameters of the distribution of the optimal alignment scores are estimated through empirical approach and discuss the edge-effect and multiple testing issues. We also relate theory to the calculations of the Expect and P-values in BLAST program.

Chapter 8 is about the substitution matrices. We start with the reconstruction of popular PAM and BLOSUM matrices. We then present Altschul's theoretic-information approach to scoring matrix selection and recent work on compositional adjustment of scoring matrices for aligning sequences with biased letter frequencies. Finally, we discuss gap penalty costs.

This text is targeted to a reader with a general scientific background. Little or no prior knowledge of biology, algorithms, and probability is expected or assumed. The basic notions from molecular biology that are useful for understanding the topics covered in this text are outlined in Appendix A. Appendix B provides a brief introduction to probability theory. Appendix C lists popular software packages for pairwise alignment, homology search, and multiple alignment.

This book is a general and rigorous text on the algorithmic techniques and mathematical foundations of sequence alignment and homology search. But, it is by no means comprehensive. It is impossible to give a complete introduction to this field because it is evolving too quickly. Accordingly, each chapter concludes with the bibliographic notes that report related work and recent progress. The reader may ultimately turn to the research articles published in scientific journals for more information and new progress.

Most of the text is written at a level that is suitable for undergraduates. It is based on lectures given to the students in the courses in bioinformatics and mathematical genomics at the National University of Singapore and the National Taiwan University each year during 2002 – 2008. These courses were offered to students from biology, computer science, electrical engineering, statistics, and mathematics majors. Here, we thank our students in the courses we have taught for their comments on the material, which are often incorporated into this text.

Despite our best efforts, this book may contain errors. It is our responsibility to correct any errors and omissions. A list of errata will be compiled and made available at <http://www.math.nus.edu.sg/~matzlx/sequencebook>.

Taiwan & Singapore
June 2008

Kun-Mao Chao
Louxin Zhang

Acknowledgments

We are extremely grateful to our mentor Webb Miller for kindly writing the foreword for this book. The first author particularly wants to thank Webb for introducing him to the emerging field of computational molecular biology and guiding him from the basics nearly two decades ago.

The second author is particularly thankful to Ming Li for guiding and encouraging him since his student time in Waterloo. He also thanks his collaborators Kwok Pui Choi, Aaron Darling, Minmei Hou, Yong Kong, Jian Ma, Bin Ma, and Franco Preparata, with whom he worked on the topics covered in this book. In addition, he would like to thank Kal Yen Kaow Ng and Jialiang Yang for reading sections of the text and catching some nasty bugs.

We also thank the following people for their inspiring conversations, suggestions, and pointers: Stephen F. Altschul, Vineet Bafna, Louis H.Y. Chen, Ross C. Hardison, Xiaohu Huang, Tao Jiang, Jim Kent, Pavel Pevzner, David Sankoff, Scott Schwartz, Nikola Stojanovic, Lusheng Wang, Von Bing Yap, and Zheng Zhang.

Finally, it has been a pleasure to work with Springer in the development of this book. We especially thank our editor Wayne Wheeler and Catherine Brett for patiently shepherding this project and constantly reminding us of the deadline, which eventually made us survive. We also thank the copy editor C. Curioli for valuable comments and the production editor Frank Ganz for assistance with formatting.

About the Authors

Kun-Mao Chao was born in Tou-Liu, Taiwan, in 1963. He received the B.S. and M.S. degrees in computer engineering from National Chiao-Tung University, Taiwan, in 1985 and 1987, respectively, and the Ph.D. degree in computer science from The Pennsylvania State University, University Park, in 1993. He is currently a professor of bioinformatics at National Taiwan University, Taipei, Taiwan. From 1987 to 1989, he served in the ROC Air Force Headquarters as a system engineer. From 1993 to 1994, he worked as a postdoctoral fellow at Penn State's Center for Computational Biology. In 1994, he was a visiting research scientist at the National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland. Before joining the faculty of National Taiwan University, he taught in the Department of Computer Science and Information Management, Providence University, Taichung, Taiwan, from 1994 to 1999, and the Department of Life Science, National Yang-Ming University, Taipei, Taiwan, from 1999 to 2002. He was a teaching award recipient of both Providence University and National Taiwan University. His current research interests include algorithms and bioinformatics. He is a member of Phi Tau Phi and Phi Kappa Phi.

Louxin Zhang studied mathematics at Lanzhou University, earning his B.S. and M.S. degrees, and studied computer science at the University of Waterloo, where he received his Ph.D. He has been a researcher and teacher in bioinformatics and computational biology at National University of Singapore (NUS) since 1996. His current research interests include genomic sequence analysis and phylogenetic analysis. His research interests also include applied combinatorics, algorithms, and theoretical computer science. In 1997, he received a Lee Kuan Yew Postdoctoral Research Fellowship to further his research. Currently, he is an associate professor of computational biology at NUS.

Contents

Foreword	vii
Preface	ix
Acknowledgments	xiii
About the Authors	xv
1 Introduction	1
1.1 Biological Motivations	1
1.2 Alignment: A Model for Sequence Comparison	2
1.2.1 Definition	2
1.2.2 Alignment Graph	3
1.3 Scoring Alignment	7
1.4 Computing Sequence Alignment	8
1.4.1 Global Alignment Problem	9
1.4.2 Local Alignment Problem	10
1.5 Multiple Alignment	11
1.6 What Alignments Are Meaningful?	12
1.7 Overview of the Book	12
1.8 Bibliographic Notes and Further Reading	13
Part I. Algorithms and Techniques	15
2 Basic Algorithmic Techniques	17
2.1 Algorithms and Their Complexity	18
2.2 Greedy Algorithms	18
2.2.1 Huffman Codes	19
2.3 Divide-and-Conquer Strategies	21
2.3.1 Mergesort	21
2.4 Dynamic Programming	23
2.4.1 Fibonacci Numbers	24

2.4.2	The Maximum-Sum Segment Problem	25
2.4.3	Longest Increasing Subsequences	27
2.4.4	Longest Common Subsequences	29
2.5	Bibliographic Notes and Further Reading	32
3	Pairwise Sequence Alignment	35
3.1	Introduction	36
3.2	Dot Matrix	37
3.3	Global Alignment	37
3.4	Local Alignment	42
3.5	Various Scoring Schemes	46
3.5.1	Affine Gap Penalties	46
3.5.2	Constant Gap Penalties	48
3.5.3	Restricted Affine Gap Penalties	48
3.6	Space-Saving Strategies	49
3.7	Other Advanced Topics	54
3.7.1	Constrained Sequence Alignment	54
3.7.2	Similar Sequence Alignment	56
3.7.3	Suboptimal Alignment	57
3.7.4	Robustness Measurement	59
3.8	Bibliographic Notes and Further Reading	60
4	Homology Search Tools	63
4.1	Finding Exact Word Matches	64
4.1.1	Hash Tables	64
4.1.2	Suffix Trees	66
4.1.3	Suffix Arrays	67
4.2	FASTA	68
4.3	BLAST	69
4.3.1	Ungapped BLAST	69
4.3.2	Gapped BLAST	72
4.3.3	PSI-BLAST	73
4.4	BLAT	74
4.5	PatternHunter	75
4.6	Bibliographic Notes and Further Reading	78
5	Multiple Sequence Alignment	81
5.1	Aligning Multiple Sequences	81
5.2	Scoring Multiple Sequence Alignment	82
5.3	An Exact Method for Aligning Three Sequences	84
5.4	Progressive Alignment	85
5.5	Bibliographic Notes and Further Reading	86

Part II. Theory	89
6 Anatomy of Spaced Seeds	91
6.1 Filtration Technique in Homology Search	92
6.1.1 Spaced Seed	92
6.1.2 Sensitivity and Specificity	92
6.2 Basic Formulas on Hit Probability	93
6.2.1 A Recurrence System for Hit Probability	95
6.2.2 Computing Non-Hit Probability	97
6.2.3 Two Inequalities	98
6.3 Distance between Non-Overlapping Hits	99
6.3.1 A Formula for μ_π	100
6.3.2 An Upper Bound for μ_π	101
6.3.3 Why Do Spaced Seeds Have More Hits?	103
6.4 Asymptotic Analysis of Hit Probability	104
6.4.1 Consecutive Seeds	104
6.4.2 Spaced Seeds	107
6.5 Spaced Seed Selection	110
6.5.1 Selection Methods	110
6.5.2 Good Spaced Seeds	111
6.6 Generalizations of Spaced Seeds	112
6.6.1 Transition Seeds	112
6.6.2 Multiple Spaced Seeds	114
6.6.3 Vector Seed	115
6.7 Bibliographic Notes and Further Reading	115
7 Local Alignment Statistics	119
7.1 Introduction	120
7.2 Ungapped Local Alignment Scores	122
7.2.1 Maximum Segment Scores	123
7.2.2 E-value and P-value Estimation	128
7.2.3 The Number of High-Scoring Segments	130
7.2.4 Karlin-Altschul Sum Statistic	131
7.2.5 Local Ungapped Alignment	132
7.2.6 Edge Effects	133
7.3 Gapped Local Alignment Scores	134
7.3.1 Effects of Gap Penalty	134
7.3.2 Estimation of Statistical Parameters	135
7.3.3 Statistical Parameters for BLOSUM and PAM Matrices	138
7.4 BLAST Database Search	139
7.4.1 Calculation of P-values and E-values	140
7.4.2 BLAST Printouts	142
7.5 Bibliographic Notes and Further Reading	146

8	Scoring Matrices	149
8.1	The PAM Scoring Matrices	150
8.2	The BLOSUM Scoring Matrices	153
8.3	General Form of the Scoring Matrices	155
8.4	How to Select a Scoring Matrix?	157
8.5	Compositional Adjustment of Scoring Matrices	158
8.6	DNA Scoring Matrices	161
8.7	Gap Cost in Gapped Alignments	163
8.8	Bibliographic Notes and Further Reading	164
A	Basic Concepts in Molecular Biology	173
A.1	The Nucleic Acids: DNA and RNA	173
A.2	Proteins	174
A.3	Genes	175
A.4	The Genomes	175
B	Elementary Probability Theory	177
B.1	Events and Probabilities	177
B.2	Random Variables	178
B.3	Major Discrete Distributions	179
B.3.1	Bernoulli Distribution	179
B.3.2	Binomial Distribution	180
B.3.3	Geometric and Geometric-like Distributions	180
B.3.4	The Poisson Distribution	180
B.3.5	Probability Generating Function	181
B.4	Major Continuous Distributions	182
B.4.1	Uniform Distribution	182
B.4.2	Exponential Distribution	182
B.4.3	Normal Distribution	183
B.5	Mean, Variance, and Moments	183
B.5.1	The Mean of a Random Variable	183
B.5.2	The Variance of a Random Variable	185
B.5.3	The Moment-Generating Function	186
B.6	Relative Entropy of Probability Distributions	187
B.7	Discrete-time Finite Markov Chains	188
B.7.1	Basic Definitions	188
B.7.2	Markov Chains with No Absorbing States	189
B.7.3	Markov Chains with Absorbing States	190
B.7.4	Random Walks	191
B.7.5	High-Order Markov Chains	191
B.8	Recurrent Events and the Renewal Theorem	191
C	Software Packages for Sequence Alignment	195
	References	197
	Index	207