Computational Biology

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 > 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

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 that 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 *Kun-Mao Chao*
June 2008 *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, Xiaoqiu 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