AN INTRODUCTION TO
BIOINFORMATICS ALGORITHMS

NEIL C. JONES AND PAVEL A. PEVZNER

An Introduction to Bioinformatics Algorithms

Neil C. Jones
Pavel A. Pevzner

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

Typeset in 10/13 Lucida Bright by the authors using LaTeX 2_ε.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Jones, Neil C.
An introduction to bioinformatics algorithms/ by Neil C. Jones and Pavel A. Pevzner.
p. cm.—(computational molecular biology series)
"A Bradford book."
Includes bibliographical references and index (p.).
ISBN 0-262-10106-8 (hc : alk. paper)
1. Bioinformatics. 2. Algorithms. I. Pevzner, Pavel. II. Title
QH324.2.J66 2004
570'.285—dc22 2004048289
CIP

10 9 8 7 6 5 4 3

An Introduction to Bioinformatics Algorithms

Computational Molecular Biology

Sorin Istrail, Pavel Pevzner, and Michael Waterman, editors

Computational molecular biology is a new discipline, bringing together computational, statistical, experimental, and technological methods, which is energizing and dramatically accelerating the discovery of new technologies and tools for molecular biology. The MIT Press Series on Computational Molecular Biology is intended to provide a unique and effective venue for the rapid publication of monographs, textbooks, edited collections, reference works, and lecture notes of the highest quality.

Computational Molecular Biology: An Algorithmic Approach
Pavel A. Pevzner, 2000

Computational Methods for Modeling Biochemical Networks
James M. Bower and Hamid Bolouri, editors, 2001

Current Topics in Computational Molecular Biology
Tao Jiang, Ying Xu, and Michael Q. Zhang, editors, 2002

Gene Regulation and Metabolism: Postgenomic Computation Approaches
Julio Collado-Vides, editor, 2002

Microarrays for an Integrative Genomics
Isaac S. Kohane, Alvin Kho, and Atul J. Butte, 2002

Kernel Methods in Computational Biology
Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert, 2004

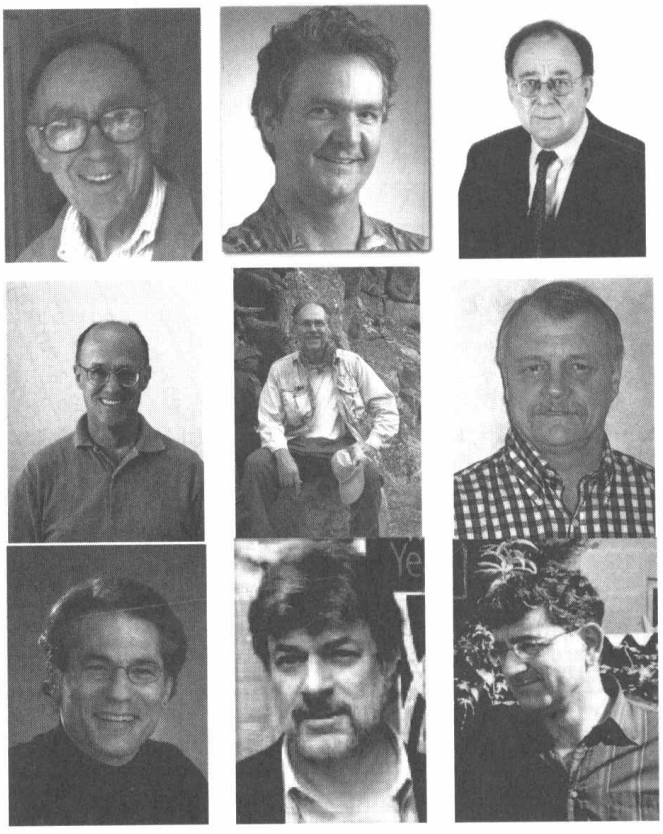
An Introduction to Bioinformatics Algorithms
Neil C. Jones and Pavel A. Pevzner, 2004

For my Grandfather, who lived to write and vice versa.
—NCJ

To Chop-up and Manny.
—PAP

Contents in Brief

Subject	4 Exhaustive Search	5 Greedy Algorithms	6 Dynamic Programming Algorithms	7 Divide-and-Conquer Algorithms	8 Graph Algorithms	9 Combinatorial Pattern Matching	10 Clustering and Trees	11 Hidden Markov Models	12 Randomized Algorithms
Mapping DNA	○								
Sequencing DNA					○				
Comparing Sequences			○	○		○			
Predicting Genes			○						
Finding Signals	○	○						○	○
Identifying Proteins					○				
Repeat Analysis						○			
DNA Arrays					○		○		
Genome Rearrangements		○							
Molecular Evolution							○		



Featuring historical perspectives from:

Russell Doolittle	Chapter 3
David Haussler	Chapter 11
Richard Karp	Chapter 2
Webb Miller	Chapter 7
Gene Myers	Chapter 9
David Sankoff	Chapter 5
Ron Shamir	Chapter 10
Gary Stormo	Chapter 4
Michael Waterman	Chapter 6

Preface

In the early 1990s when one of us was teaching his first bioinformatics class, he was not sure that there would be enough students to teach. Although the Smith-Waterman and BLAST algorithms had already been developed they had not become the household names among biologists that they are today. Even the term "bioinformatics" had not yet been coined. DNA arrays were viewed by most as intellectual toys with dubious practical application, except for a handful of enthusiasts who saw a vast potential in the technology. A few bioinformaticians were developing new algorithmic ideas for nonexistent data sets: David Sankoff laid the foundations of genome rearrangement studies at a time when there was practically no gene order data, Michael Waterman and Gary Stormo were developing motif finding algorithms when there were very few promoter samples available, Gene Myers was developing sophisticated fragment assembly tools when no bacterial genome has been assembled yet, and Webb Miller was dreaming about comparing billion-nucleotide-long DNA sequences when the 172,282-nucleotide Epstein-Barr virus was the longest GenBank entry. GenBank itself just recently made a transition from a series of bound (paper!) volumes to an electronic database on magnetic tape that could be sent to scientists worldwide.

One has to go back to the mid-1980s and early 1990s to fully appreciate the revolution in biology that has taken place in the last decade. However, bioinformatics has affected more than just biology—it has also had a profound impact on the computational sciences. Biology has rapidly become a large source of new algorithmic and statistical problems, and has arguably been the target for more algorithms than any of the other fundamental sciences. This link between computer science and biology has important educational implications that change the way we teach computational ideas to biologists, as well as how applied algorithmics is taught to computer scientists.

For many years computer science was taught to only computer scientists, and only rarely to students from other disciplines. A biology student in an algorithms class would be a surprising and unlikely (though entirely welcome) guest in the early 1990s. But these things change; many biology students now take some sort of Algorithms 101. At the same time, curious computer science students often take Genetics 101 and Bioinformatics 101. Although these students are still relatively rare, keep in mind that the number of bioinformatics classes in the early 1990s was so small as to be considered nonexistent. But that number is not so small now. We envision that undergraduate bioinformatics classes will become a permanent component at every major university. This is a feature, not a bug.

This is an introductory textbook on bioinformatics algorithms and the computational ideas that have driven them through the last twenty years. There are many important probabilistic and statistical techniques that we do not cover, nor do we cover many important research questions that bioinformaticians are currently trying to answer. We deliberately do not cover all areas of computational biology; for example, important topics like protein folding are not even discussed. The very first bioinformatics textbooks were Waterman, 1995 (108), which contains excellent coverage of DNA statistics and Gusfield, 1997 (44) which includes an encyclopedia of string algorithms. Durbin et al., 1998 (31) and Baldi and Brunak, 1997 (7) emphasize Hidden Markov Models and machine learning techniques; Baxevanis and Ouellette, 1998 (10) is an excellent practical guide to bioinformatics; Mount, 2001 (76) excels in showing the connections between biological problems and bioinformatics techniques; and Bourne and Weissig, 2002 (15) focuses on protein bioinformatics. There are also excellent web-based lecture notes for many bioinformatics courses and we learned a lot about the pedagogy of bioinformatics from materials on the World Wide Web by Serafim Batzoglou, Dick Karp, Ron Shamir, Martin Tompa, and others.

Website

We have created an extensive website to accompany this book at

`http://www.bioalgorithms.info`

This website contains a number of features that complement the book. For example, though this book does not contain a glossary, we provide this service, a searchable index, and a set of community message boards, at the above web address. Technically savvy students can also download practical

bioinformatics exercises, sample implementations of the algorithms in this book, and sample data to test them with. Instructors and students may find the prepackaged lecture notes on the website to be especially helpful. It is our hope that this website be used as a repository of information that will help introduce students to the diverse world of bioinformatics.

Acknowledgements

We are indebted to those who kindly agreed to be featured in the biographical sketches scattered throughout the book. Their insightful and heartfelt responses definitely made these the most interesting part of this book. Their life stories and views of the challenges that lay ahead will undoubtedly inspire students in the exploration of the unknown. There are many more scientists whose bioboxes we would like to have in this book and it is only the page limit (which turned out to be 200 pages too small) that prevented us from commissioning more of them. Special thanks go to Ethan Bier who inspired us to include biographical sketches in this book.

This book would not have been possible without the diligent teaching assistants in bioinformatics courses taught during the winter and fall of 2003 and 2004: Derren Barken, Bryant Forsgren, Eugene Ke, Coleman Mosley, and Degui Zhi all helped find technical errors, refine practical exercises, and design problems in the book. Helen Wu and John Allison spent many hours making technical figures, which is a thankless task like no other. We are also grateful to Vagisha Sharma who was kind enough to read the book from cover to cover and provide insightful comments and, unfortunately, bugs in the pseudocode. Steve Wasserman provided us with invaluable comments from a biologist's point of view that eventually led to new sections in the book. Alkes Price and Haixu Tang pointed out ambiguities and helped clarify the chapters on graphs and clustering. Ben Raphael and Patricia Jones provided feedback on the early chapters and helped avoid some potential misunderstandings. Dan Gilbert, of Dan Gilbert Art Group, Inc. kindly provided us with Triazzles to illustrate the problems of DNA sequence assembly.

Our special thanks go to Randall Christopher, the artist behind the website `www.kleemanandmike.com`. Randall illustrated the book and designed many unique graphical representations of some bioinformatics algorithms.

It has been a pleasure to work with Robert Prior of The MIT Press. With sufficient patience and prodding, he managed to keep us on track. We also appreciate the meticulous copyediting of G. W. Helfrich.

Finally, we thank the many students in different undergraduate and graduate bioinformatics classes at UCSD who provided comments on earlier versions of this book.

PAP would like to thank several people who taught him different aspects of computational molecular biology. Andrey Mironov taught him that common sense is perhaps the most important ingredient of any applied research. Mike Waterman was a terrific teacher at the time PAP moved from Moscow to Los Angeles, both in science and in life. PAP also thanks Alexander Karzanov, who taught him combinatorial optimization, which, surprisingly, remains the most useful set of skills in his computational biology research. He especially thanks Mark Borodovsky who convinced him to switch into the field of bioinformatics in 1985, when it was an obscure discipline with an uncertain future.

PAP also thanks his former students, postdocs, and lab members who taught him most of what he knows: Vineet Bafna, Guillaume Bourque, Sridhar Hannenhalli, Steffen Heber, Earl Hubbell, Uri Keich, Zufar Mulyukov, Alkes Price, Ben Raphael, Sing-Hoi Sze, Haixu Tang, and Glenn Tesler.

NCJ would like to thank his mentors during undergraduate school— Toshihiko Takeuchi, Harry Gray, John Baldeschwieler, and Schubert Soares—for patiently but firmly teaching him that persistence is one of the more important ingredients in research. Also, he thanks the admissions committee at the University of California, San Diego who gambled on a chemist-turned-programmer, hopefully for the best.

Neil Jones and Pavel Pevzner
La Jolla, California, 2004

Contents