NEW COMPREHENSIVE BIOCHEMISTRY VOLUME 32

general editor:
G. BERNARDI

Computational Methods in Molecular Biology

edited by
S.L. Salzberg
D.B. Searls
S. Kasif

ELSEVIER

Computational Methods in Molecular Biology

Editors

Steven L. Salzberg

The Institute for Genomic Research,
9712 Medical Center Drive, Rockville, MD 20850, USA

David B. Searls

SmithKline Beecham Pharmaceuticals, 709 Swedeland Road,
P.O. Box 1539, King of Prussia, PA 19406, USA

Simon Kasif

Department of Electrical Engineering and Computer Science,
University of Illinois at Chicago, Chicago, IL 60607-7053, USA

1998
ELSEVIER
Amsterdam · Lausanne · New York · Oxford · Shannon · Singapore · Tokyo

Elsevier Science B.V.
P.O. Box 211
1000 AE Amsterdam
The Netherlands

Library of Congress Cataloging-in-Publication Data

```
Computational methods in molecular biology / editors, Steven L.
  Salzberg, David B. Searls, Simon Kasif.
       p.   cm. -- (New comprehensive biochemistry ; v. 32)
    Includes bibliographical references and index.
    ISBN 0-444-82875-3 (alk. paper)
    1. Molecular biology--Mathematics.   I. Salzberg, Steven L., 1960-
  .  II. Searls, David B.   III. Kasif, Simon.   IV. Series.
  QD415.N48  vo. 32
  [QH506]
  572 s--dc21
  [572.8'01'51]                                              98-22957
                                                                 CIP
```

ISBN 0 444 82875 3
ISBN 0 444 80303 3 (series)

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of the rapid advances in the medical sciences, the publisher recommends that independent verification of diagnoses and drug dosages should be made.

∞ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Printed in the Netherlands

COMPUTATIONAL METHODS IN MOLECULAR BIOLOGY

New Comprehensive Biochemistry

Volume 32

General Editor

G. BERNARDI
Paris

ELSEVIER
Amsterdam · Lausanne · New York · Oxford · Shannon · Singapore · Tokyo

To Claudia and Annika – S.L.S.

To my parents – D.B.S.

To Orna and Mei – S.K.

Preface

The field of computational biology, or bioinformatics as it is often called, was born just a few years ago. It is difficult to pinpoint its exact beginnings, but it is easy to see that the field is currently undergoing rapid, exciting growth. This growth has been fueled by a revolution in DNA sequencing and mapping technology, which has been accompanied by rapid growth in many related areas of biology and biotechnology. No doubt many exciting breakthroughs are yet to come. All this new DNA and protein sequence data brings with it the tremendously exciting challenge of how to make sense of it: how to turn the raw sequences into information that will lead to new drugs, new advances in health care, and a better overall understanding of how living organisms function. One of the primary tools for making sense of this revolution in sequence data is the computer. Computational biology is all about how to use the power of computation to model and understand biological systems and especially biological sequence data.

This book is an attempt to bring together in one place some of the latest advances in computational biology. In assembling the book, we were particularly interested in creating a volume that would be accessible to biologists (as well as computer scientists and others). With this in mind, we have included tutorials on many of the key topics in the volume, designed to introduce biological scientists to some of the computational techniques that might otherwise be unfamiliar to them. Some of those tutorials appear as separate, complete chapters on their own, while others appear as sections within chapters. We also want to encourage more computer scientists to get involved in this new field, and with them in mind we included tutorial material on several topics in molecular biology as well. We hope the result is a volume that offers something valuable to a wide range of readers. The only required background is an interest in the exciting new field of computational biology.

The chapters that follow are broadly grouped into three sections. Loosely speaking, these can be described as an introductory section, a section on DNA sequence analysis, and a section on proteins. The introductory section begins with an overview by Searls of some of the main challenges facing computational biology today. This chapter contains a thought-provoking description of problems ranging from gene finding to protein folding, explaining the biological significance and hinting at many of the computational solutions that appear in later chapters. Searls' chapter should appeal to all readers. Next is Salzberg's tutorial on computation, designed primarily for biologists who do not have a formal background in computer science. After reading this chapter, biologists should find many of the later chapters much more accessible. The following chapter, by Fasman and Salzberg, provides a tutorial for the other main component of our audience, computational scientists (including computer scientists, mathematicians, physicists, and anyone else who might need some additional biological background) who want to understand the biology that underlies all the research problems described in later chapters. This tutorial introduces the non-biologist to many of the terms, concepts, and mechanisms of molecular biology and sequence analysis.

The second of the three major sections contains work primarily on DNA and RNA sequence analysis. Although the techniques covered here are not restricted to DNA sequences, most of the applications described here have been applied to DNA.

Krogh's chapter begins the section with a tutorial introduction to one of the hottest techniques in computational biology, hidden Markov models (HMMs). HMMs have been used for a wide range of problems, including gene finding, multiple sequence alignment, and the search for motifs. Krogh covers only one of these applications, gene finding, but he first gives a cleverly non-mathematical tutorial on this very mathematical topic.

The chapter by Overton and Haas describes a case-based reasoning approach to sequence annotation. They describe an informatics system for the study of gene expression in red blood cell differentiation. This type of specialized information resource is likely to become increasingly important as the amount of data in GenBank becomes ever larger and more diverse.

The chapter by States and Reisdorf describes how to use sequence similarity as the basis for sequence classification. The approach relies on clustering algorithms which can, in general, operate on whole sequences, partial sequences, or structures. The chapter includes a comprehensive current list of databases of sequence and structure classification.

Xu and Uberbacher describe many of the details of the latest version of GRAIL, which for years has been one of the leading gene-finding systems for eukaryotic data. GRAIL's latest modules include the ability to incorporate sequence similarity to the expressed sequence tag (EST) database and a nice technique for detecting potential frameshifts.

Burge gives a thorough description of how to model RNA splicing signals (donor and acceptor sites) using statistical patterns. He shows how to combine weight matrix methods with a new tree-based method called maximal dependence decomposition, resulting in a splice site recognizer that is state of the art. His technique is implemented in GENSCAN, currently the best-performing of all gene-finding systems.

Parsons' chapter includes tutorial material on genetic algorithms (GAs), a family of techniques that use the principles of mutation, crossover, and natural selection to "evolve" computational solutions to a problem. After the tutorial, the chapter goes on to describe a particular genetic algorithm for solving a problem in DNA sequence assembly. This description serves not only to illustrate how well the GA worked, but it also provides a case study in how to refine a GA in the context of a particular problem.

Salzberg's chapter includes a tutorial on decision trees, a type of classification algorithm that has a wide range of uses. The tutorial uses examples from the domain of eukaryotic gene finding to make the description more relevant. The chapter then moves on to a description of MORGAN, a gene-finding system that is a hybrid of decision trees and Markov chains. MORGAN's excellent performance proves that decision trees can be applied effectively to DNA sequence analysis problems.

Wei, Chang, and Altman's chapter describes statistical methods for protein structure analysis. They begin with a tutorial on statistical methods, and then go on to describe FEATURE, their system for statistical analysis of protein sequences. They describe several applications of FEATURE, including characterization of active sites, generation of substitution matrices, and protein threading.

Protein threading, or fold recognition, is essentially finding the best fit of a protein sequence to a set of candidate structures for that sequence. Lathrop, Rogers, Bienkowska, Bryant, Buturović, Gaitatzes, Nambudripad, White, and Smith begin their chapter with a tutorial section that describes what the problem is and why it is "hard" in the computer science sense of that word. This section should be of special interest to those who want to understand why protein folding is computationally difficult. They then describe their threading algorithm, which is an exhaustive search method that uses a branch-and-bound strategy to reduce the search space to a tractable (but still very large) size.

Jones' chapter describes THREADER, one of the leading systems for protein threading. He first introduces the general protein folding problem, reviews the literature on fold recognition, and then describes in detail the so-called double dynamic programming approach that THREADER employs. Jones makes it clear how this intriguing problem combines a wide range of issues, from combinatorial optimization to thermodynamics.

The chapter by Wolfson and Nussinov presents a novel application of geometric hashing for predicting the possibility of binding, docking and other forms of biomolecular interaction. Even when the individual structures of two molecules are accurately modeled, it remains computationally difficult to predict whether docking or binding are possible. Thus, this method naturally complements the work on structure prediction described in other chapters.

The chapter by Kasif and Delcher uses a probabilistic modeling approach similar to HMMs, but their formalism is known as probabilistic networks or Bayesian networks. These networks have slightly more expressive power and in some cases a more compact representation. For sequence analysis tasks, the probabilistic network approach allows one to model features such as motif lengths, gap lengths, long term dependencies, and the chemical properties of amino acids.

Finally, the end of the book contains some reference materials that all readers should find useful. The first appendix contains a list of Internet resources, including most of the software described in the book. This list is also available on a Web page whose address is given in the appendix. The Web page will be kept up to date long after the book's publication date. The second appendix contains an annotated bibliographical list for further reading on selected topics in computational biology. Some of these references, each of which contains a very short text description, point to more technical descriptions of the systems in the book. Others point to well-known or landmark papers in computational biology which would be of interest to anyone looking for a broader perspective on the field.

Steven Salzberg
David Searls
Simon Kasif
Baltimore, Maryland
October 1997

List of contributors*

Russ B. Altman 207
Section of Medical Informatics, 251 Campus Drive, Room x-215,
Stanford University School of Medicine, Stanford, CA 94305-5479, USA

Jadwiga Bienkowska 227
BioMolecular Engineering Research Center, Boston University, 36 Cummington Street,
Boston, MA 02215, USA

Barbara K.M. Bryant 227
Millennium Pharmaceuticals, Inc., 640 Memorial Drive, Cambridge, MA 02139, USA

Christopher B. Burge 129
Center for Cancer Research, Massachusetts Institute of Technology,
40 Ames Street, Room E17-526a, Cambridge, MA 02139-4307, USA

Ljubomir J. Buturović 227
Incyte Pharmaceuticals, Inc., 3174 Porter Drive, Palo Alto, CA 94304, USA

Jeffrey T. Chang 207
Section of Medical Informatics, 251 Campus Drive, Room x-215,
Stanford University School of Medicine, Stanford, CA 94305-5479, USA

Arthur L. Delcher 335
Computer Science Department, Loyola College in Maryland, Baltimore, MD 21210, USA

Kenneth H. Fasman 29
Whitehead Institute/MIT Center for Genome Research, 320 Charles Street,
Cambridge, MA 02141, USA

Chrysanthe Gaitatzes 227
BioMolecular Engineering Research Center, Boston University, 36 Cummington Street,
Boston, MA 02215, USA

Juergen Haas 65
Center for Bioinformatics, University of Pennsylvania, 13121 Blockley Hall,
418 Boulevard, Philadelphia, PA 19104, USA

* Authors' names are followed by the starting page number(s) of their contribution(s).

David Jones 285
Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, England, UK

Simon Kasif 335
Department of Electrical Engineering and Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7053, USA

Anders Krogh 45
Center for Biological Analysis, Technical University of Denmark, Building 208, 2800 Lyngby, Denmark

Richard H. Lathrop 227
Department of Information and Computer Science, 444 Computer Science Building, University of California, Irvine, CA 92697-3425, USA

Raman Nambudripad 227
Molecular Computing Facility, Beth Israel Hospital, 330 Brookline Avenue, Boston, MA 02215, USA

Ruth Nussinov 313
Sackler Inst. of Molecular Medicine, Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel; and
Laboratory of Experimental and Computational Biology, SAIC, NCI-FCRDC, Bldg. 469, rm. 151, Frederick, MD 21702, USA

G. Christian Overton 65
Center for Bioinformatics, University of Pennsylvania, 13121 Blockley Hall, 418 Boulevard, Philadelphia, PA 19104, USA

Rebecca J. Parsons 165
Department of Computer Science, University of Central Florida, P.O. Box 162362, Orlando, FL 32816-2362, USA

William C. Reisdorf, Jr. 87
Institute for Biomedical Computing, Washington University in St. Louis, 700 South Euclid Avenue, St. Louis, MO 63110, USA

Robert G. Rogers Jr. 227
BioMolecular Engineering Research Center, Boston University, 36 Cummington Street, Boston, MA 02215, USA

Steven Salzberg 11, 29, 187
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

David B. Searls 3
SmithKline Beecham Pharmaceuticals, 709 Swedeland Road, P.O. Box 1539,
King of Prussia, PA 19406, USA

Temple F. Smith 227
BioMolecular Engineering Research Center, Boston University, 36 Cummington Street,
Boston, MA 02215, USA

David J. States 87
Institute for Biomedical Computing, Washington University in St. Louis,
700 South Euclid Avenue, St. Louis, MO 63110, USA

Edward C. Uberbacher 109
Bldg. 1060 COM, MS 6480, Cumputational Biosciences Section,
Life Sciences Division, ORNL, Oak Ridge, TN 37831-6480, USA

Liping Wei 207
Section of Medical Informatics, 251 Campus Drive, Room x-215,
Stanford University School of Medicine, Stanford, CA 94305-5479, USA

James V. White 227
BioMolecular Engineering Research Center, Boston University, 36 Cummington Street,
Boston, MA 02215, USA; and *TASC, Inc., 55 Walkers Brook Drive, Reading, MA*
01867, USA

Haim Wolfson 313
Computer Science Department, Tel Aviv University,
Raymond and Beverly Sackler Faculty of Exact Sciences, Ramat Aviv 69978,
Tel Aviv, Israel

Ying Xu 109
Bldg. 1060 COM, MS 6480, Cumputational Biosciences Section,
Life Sciences Division, ORNL, Oak Ridge, TN 37831-6480, USA

Other volumes in the series

Volume 1. *Membrane Structure* (1982)
J.B. Finean and R.H. Michell (Eds.)

Volume 2. *Membrane Transport* (1982)
S.L. Bonting and J.J.H.H.M. de Pont (Eds.)

Volume 3. *Stereochemistry* (1982)
C. Tamm (Ed.)

Volume 4. *Phospholipids* (1982)
J.N. Hawthorne and G.B. Ansell (Eds.)

Volume 5. *Prostaglandins and Related Substances* (1983)
C. Pace-Asciak and E. Granström (Eds.)

Volume 6. *The Chemistry of Enzyme Action* (1984)
M.I. Page (Ed.)

Volume 7. *Fatty Acid Metabolism and its Regulation* (1984)
S. Numa (Ed.)

Volume 8. *Separation Methods* (1984)
Z. Deyl (Ed.)

Volume 9. *Bioenergetics* (1985)
L. Ernster (Ed.)

Volume 10. *Glycolipids* (1985)
H. Wiegandt (Ed.)

Volume 11a. *Modern Physical Methods in Biochemistry, Part A* (1985)
A. Neuberger and L.L.M. van Deenen (Eds.)

Volume 11b. *Modern Physical Methods in Biochemistry, Part B* (1988)
A. Neuberger and L.L.M. van Deenen (Eds.)

Volume 12. *Sterols and Bile Acids* (1985)
H. Danielsson and J. Sjövall (Eds.)

Volume 13. *Blood Coagulation* (1986)
R.F.A. Zwaal and H.C. Hemker (Eds.)

Volume 14. *Plasma Lipoproteins* (1987)
A.M. Gotto Jr. (Ed.)

Volume 16. *Hydrolytic Enzymes* (1987)
A. Neuberger and K. Brocklehurst (Eds.)

Volume 17. *Molecular Genetics of Immunoglobulin* (1987)
F. Calabi and M.S. Neuberger (Eds.)

Volume 18a. *Hormones and Their Actions, Part 1* (1988)
B.A. Cooke, R.J.B. King and H.J. van der Molen (Eds.)

Volume 18b. *Hormones and Their Actions, Part 2 – Specific Action of Protein Hormones* (1988)
B.A. Cooke, R.J.B. King and H.J. van der Molen (Eds.)

Volume 19. *Biosynthesis of Tetrapyrroles* (1991)
P.M. Jordan (Ed.)

Volume 20. *Biochemistry of Lipids, Lipoproteins and Membranes* (1991)
D.E. Vance and J. Vance (Eds.) – Please see Vol. 31 – revised edition

Volume 21. *Molecular Aspects of Transport Proteins* (1992)
J.J. de Pont (Ed.)

Volume 22. *Membrane Biogenesis and Protein Targeting* (1992)
W. Neupert and R. Lill (Eds.)

Volume 23. *Molecular Mechanisms in Bioenergetics* (1992)
L. Ernster (Ed.)

Volume 24. *Neurotransmitter Receptors* (1993)
F. Hucho (Ed.)

Volume 25. *Protein Lipid Interactions* (1993)
A. Watts (Ed.)

Volume 26. *The Biochemistry of Archaea* (1993)
M. Kates, D. Kushner and A. Matheson (Eds.)

Volume 27. *Bacterial Cell Wall* (1994)
J. Ghuysen and R. Hakenbeck (Eds.)

Volume 28. *Free Radical Damage and its Control* (1994)
C. Rice-Evans and R.H. Burdon (Eds.)

Volume 29a. *Glycoproteins* (1995)
J. Montreuil, J.F.G. Vliegenthart and H. Schachter (Eds.)

Volume 29b. *Glycoproteins II* (1997)
J. Montreuil, J.F.G. Vliegenthart and H. Schachter (Eds.)

Volume 30. *Glycoproteins and Disease* (1996)
J. Montreuil, J.F.G. Vliegenthart and H. Schachter (Eds.)

Volume 31. *Biochemistry of Lipids, Lipoproteins and Membranes* (1996)
D.E. Vance and J. Vance (Eds.)

Contents

I – Introduction and Tutorial Background

II – Learning and Pattern Discovery in Sequence Databases