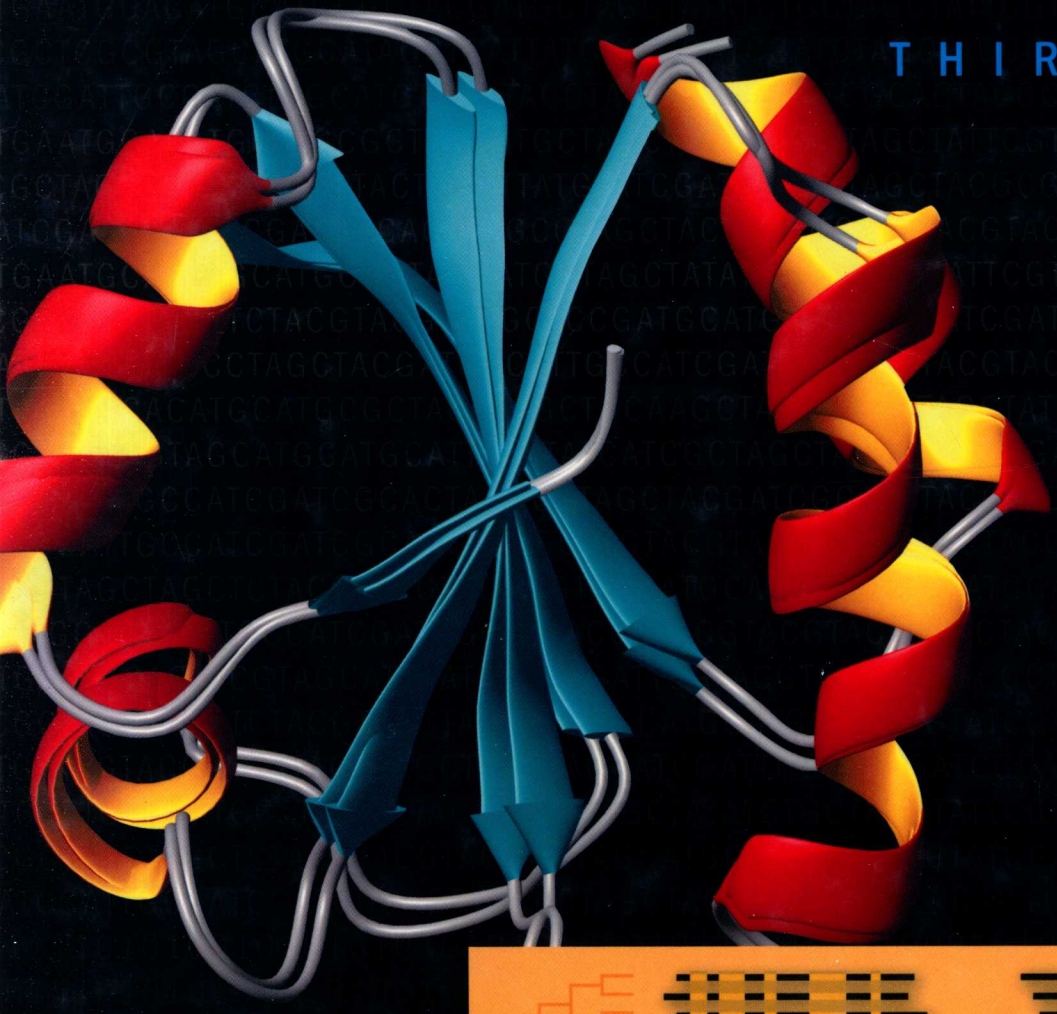
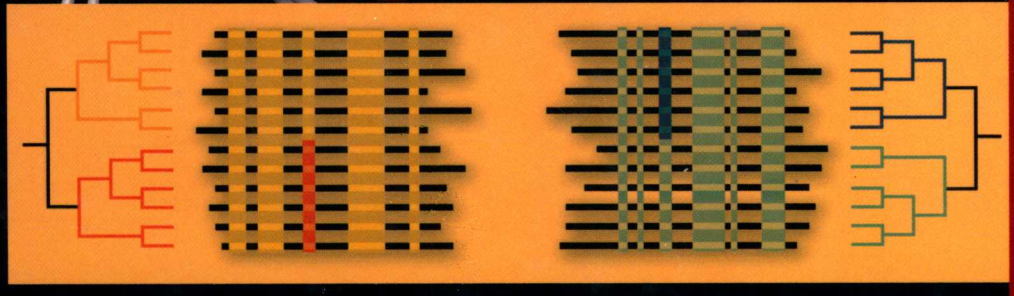
WILEY

BIOINFORMATICS

A PRACTICAL GUIDE TO THE ANALYSIS OF GENES AND PROTEINS

THIRD EDITION





EDITED BY

ANDREAS D. BAXEVANIS

B.F. FRANCIS OUELLETTE

THIRD EDITION

BIOINFORMATICS

A Practical Guide to the Analysis of Genes and Proteins

Edited By

Andreas D. Baxevanis

B. F. Francis Ouellette

WILEY-INTERSCIENCE

A John Wiley & Sons, Inc., Publication

This book is printed on acid-free paper.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Bioinformatics : a practical guide to the analysis of genes and proteins / edited by Andreas D. Baxevanis, B. F. Francis Ouellette.– 3rd ed.
p. ; cm.
Includes bibliographical references and index.
ISBN 0-471-47878-4 (cloth)
1. Bioinformatics. I. Baxevanis, Andreas D. II. Ouellette, B. F. Francis. [DNLM: 1. Base Sequence. 2. Sequence Analysis–methods. 3. Computational Biology–methods. 4. Databases, Genetic. QU 58 B61523 2005] I. Baxevanis, Andreas D. II. Ouellette, B. F. Francis.
QH324.2.B547 2005
572.8′633′0285–dc22

2004011822

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

THIRD EDITION

BIOINFORMATICS

A Practical Guide to the Analysis of Genes and Proteins

ADB dedicates this book to his best friend, Tom Young,
in celebration of a friendship full of laughter, loyalty, kindness, and all-around silliness.

BFFO dedicates this book to his daughter, Pascale.
Her smile and joy brings consonance to a complicated world.

Contents in Brief

Foreword

As we move into the 21st century, we stand at a grand inflection point in biology—how we view and practice biology has forever changed. This inflection point has been catalyzed by a number of events, perhaps the most important of which is the human genome project. It provided a genetics parts list and catalyzed the development of high throughput measurement tools (e.g., the high speed DNA sequences, DNA arrays, high throughput mass spectrometry, etc.) and high throughput measurement strategies (e.g., yeast two-hybrid technique for measuring protein/protein interactions and the genome-wide localization technique for delineating protein/DNA interactions), as well as stimulating the development of powerful new computational tools for acquiring, storing, and analyzing biological information.

The human genome project also changed how we view and practice biology in several other ways. First, it has catalyzed the view that biology is an informational science. There are two fundamental types of biological information, the digital genome and the environmental signals, intracellular, extracellular, or even from outside the organism, that impinge on the genome to facilitate the development of living organisms. These two types of biological information operate across three different time dimensions with regard to the lifetime of individual organisms—evolution, development and physiological responses. There are two major types of genomic (digital) information–the genes encoding proteins, which assemble to create the molecular machines and networks of life, and the cis-control elements that, through interactions with their cognate transcription factors, regulate the expression of their associated genes and establish the linkage relationships and architectures of the gene regulatory networks, those grand integrators of environmental signals, which then transduce the input information to the protein modules or protein networks that mediate the developmental and physiological responses of living organisms. Biological information is also hierarchical—as one moves from the genomes to ecologies, successively higher levels of biological information are created (DNA, RNA, protein machines, protein and gene regulatory networks, cells, etc). Since environmental signals change the information at each level, in order to understand systems one must collect and integrate information from as many different hierarchical levels as possible.

Second, biology has become increasingly cross-disciplinary as biologists, chemists, computer scientists, engineers, mathematicians, and physicists work together to develop the high throughput technologies and computational/mathematical tools required for this new biology—all driven by the contemporary needs of biology. Finally, all of these changes have enabled the emergence of systems biology—the idea that we can study the interactions of all the elements in a biological system and from these come to understand its systems or emergent properties. Systems approaches have been practiced for many years, but what is unique about today's systems biology is that it can make global measurements (e.g., all mRNA, all proteins, etc.) and can integrate the global measurements from different levels of biological information.

The world of biology is, accordingly, very different from what it was even ten years ago. There are many different categories of scientists that must be educated as to this new world—undergraduate, graduates, practicing biologists, and our cross-disciplinary colleagues. One of the biggest challenges with regard to this education is to bring an awareness and understanding of the central role that mathematics, computer science, and statistics plays in deciphering the complexities of this new world of biology. Indeed, one of the most interesting exercises we at the Institute for Systems Biology have undertaken in the past year is a series of institute-wide discussions concerning ten grand computational and mathematical challenges in biology (Table 1). I will not discuss these

challenges here, but will point out that they represent a very broad list focused on deciphering biological information—the digital information of the genome, the three- and four- dimensional information of proteins, the dynamic nature of protein and gene regulatory networks to the metrics and analyses necessary for global data sets—as well as their integration.

TABLE 1: ■ Grand Computational and Mathematical Challenges in Biology

1. How to fully decipher the (digital) information content of the genome
2. How to do all-vs.-all comparisons of thousands of genomes
3. How to extract protein and gene regulatory networks from 1 and 2
4. How to integrate multiple high-throughput data types dependably
5. How to visualize and explore large-scale, multi-dimensional data
6. How to convert static network maps into dynamic mathematical models
7. How to predict protein structure and function *ab initio*
8. How to identify signatures for cellular states (e.g., healthy vs. diseased)
9. How to build hierarchical models across multiple scales of time and space
10. How to reduce complex multi-dimensional models to underlying principles, e.g., the reduction of information dimensionality

This book attempts to bring this world of computer science and mathematics in biology to the entire spectrum of scientists with an interest in biology—advanced undergraduates, graduates students, practicing biologists, and our cross-disciplinary colleagues. The chapters represent serious attempts to deal with differing aspects of many of the challenges listed in Table 1. This book is particularly important because it represents a readable and concise approach to educating ordinary biologists and equipping them with the fundamental tools necessary for participating in the paradigm changes that have occurred as a consequence of the grand inflection point in biology. To all of you who are new to the worlds of genomics, proteomics, and systems biology—welcome and good reading.

Lee Hood

Preface

In the foreword to the Second Edition of *Bioinformatics*, Eric Lander conveyed the sentiment that modern biology had entered a new era with the official publication of the initial sequence and analysis of the human genome. Since that moment in time, in February of 2001, the impact of having the human sequence in hand has been nothing short of tremendous. In the last few years, we have witnessed the completion of human genome sequencing, the completion of numerous model organism genome sequences, the development of new genomic technologies and approaches, and a proliferation of innumerable databases attempting to catalog all of the information that has been learned about genes, proteins, structures, mutations, polymorphisms, and many other biological features of interest. The advent of the genomic era has also laid a strong foundation for the development of new areas of endeavor, such as proteomics and systems biology, fields that are still in their infancy but that have the potential to have an even greater impact on our understanding of basic biological processes and human disease. What has become obvious in these last few years is that, regardless of one's specific area of endeavor, one of the critical keys to being able to do cutting-edge biological research in this new era lies in the ability to combine both laboratory- and computationally-based approaches in a synergistic manner, allowing the investigator to better-design experiments (based on database searches and the like), as well as facilitating the analysis of larger and larger data sets generated through experimentation. Unfortunately, despite its great power and potential in solving biological problems, the realm of bioinformatics still remains *terra incognita* for many biologists. To address the need for training and education in this area, we have developed a new edition of this book as a resource for our scientific colleagues.

This new edition of *Bioinformatics* follows in the tradition of the last two editions in keeping up with the quick pace of change in this field. In this edition, tried-and-true concepts and approaches that have stood the test of time are featured, as well as new approaches and algorithms that have emerged since the publication of the First and Second Editions. In considering how to refine the focus and content of the book, a questionnaire was sent to a number of professors currently teaching bioinformatics courses, as well as to people who are actively called-upon to lecture on these topics. Based on these responses, published reviews of the Second Edition, and our own experience in the classroom, we have included a number of new features in the Third Edition.

Six chapters have been added on topics that have emerged as being important enough in their own right to warrant distinct and separate discussion: genomic databases, predictive techniques using RNA sequences, sequence polymorphisms, intermolecular (protein-protein) interactions, comparative genomics, and protein identification using proteomic techniques. The chapter on Internet basics has been retired, and the chapter on submitting sequence information to public databases has been folded into the chapter on sequence databases. We have supplemented many of the chapters with text boxes and appendices that highlight basic biological techniques or provide more advanced information that may be of interest to readers or useful to instructors. A more rigorous set of problem sets has been included, and we hope that the reader will work through these examples to reinforce the concepts presented throughout the book. The solutions to these problems are available through the book's Web site, at *http://www.wiley.com/bioinformatics*. We are also pleased that the current edition contains color figures throughout; this is in recognition of the way in which bioinformatic information is presented nowadays by many Web sites, using color much more than before to communicate basic biological information to the user. We are hopeful that the inclusion of all of these features, in response to the valuable feedback we have

received, will make this book much more useful both in the classroom and in the laboratory.

There are many people whom we need to thank, on many fronts, for all of their efforts in bringing this Third Edition to press. As always, our thanks go to all of the authors who took time out of their busy schedules to write the individual chapters in this book. Collectively, this stellar group of individuals from around the world has provided the kind of expertise and perspective that is, by far, the most important factor in making the content of this book as robust and insightful as it is. We also thank the authors for bearing with us through revisions, reminders, and a tight production schedule. We've thoroughly enjoyed the scientific discourse and the numerous chats about how to present all of this information in a way that our readers will be able to master it as easily as possible.

We also thank the numerous professors and instructors—whose identity is known to the folks at Wiley but not to us—for taking the time to respond to questionnaires and letting us know how to improve upon previous editions of the book. While we are quite pleased to see the number of course adoptions continue to grow well into the triple digits and translations of the book into languages ranging from Greek to Chinese, this kind of valuable feedback has helped us to continue to live up to our original goal in taking on this project: making bioinformatics accessible and useful to as broad of a group of scientists as possible, presented at a level that can successfully drive biological inquiry forward.

Of course, there are many people behind-the-scenes at Wiley who have worked tirelessly to actually produce the book and get it into our readers' hands. First, we thank our editor, Luna Han, for her continued support and confidence in this project, and we look forward to continuing our professional relationship with her well into the future. We also thank Kristen Hauser for taking care of all of the logistics that go into a project such as this one, a substantial undertaking with the large number of authors involved. Our thanks also go to Danielle Lacourciere and Camille Pecoul Carter for being part of this project once again, helping us with all of the production-related details and making the final product look as professional as it does. We thank Alexandra Anderson, our copyeditor, who has done a wonderful job in painstakingly proofreading the final text. Finally, we wholeheartedly thank Dr. Ann Boyle, our developmental editor, for joining us on conference calls (and putting up with our banter) week after week, giving us the benefit of her scientific expertise and, more importantly, for helping us achieve our goal of effectively communicating the concepts presented in this book to such a broad audience.

ADB would like to specifically thank Debbie Wilson for her help through rounds of editing, wading through a myriad of proofreading marks along the way, as well as her good-natured moral support during the last two editions of this book. I would also like to thank Drs. Shonda Leonard and Dan Davison, my colleagues at *Current Protocols in Bioinformatics*, for their helpful discussions and critical reading of a number of the chapters in this book. My hearfelt thanks also go to Darryl Leja, whose creativity has led to a very impressive and eye-catching cover design. I would also like to extend special thanks to both Eric Green and Tyra Wolfsberg at NHGRI for being incredibly supportive of this effort, letting me bend their ears on many occasions and providing much-appreciated insight and advice along the way.

BFFO would like to thank his colleagues at the University of British Columbia for their vision and continued support of bioinformatics. I also want to strongly acknowledge the uncompromising support provided by my spouse, Nancy Ryder. Nancy has made it possible for me to work on this book, enduring my long nights of work on this project and never-ending discussions about production issues. Since the last edition, we have been blessed with a new daughter, Pascale, and both she and our first daughter, Maya, have been very understanding of their Papa's spending long hours at the computer. Most of all, Andy Baxevanis is the one I owe the most gratitude to. His project management skills, in addition to his always keeping what was best for the reader first and foremost, has made this book something that I am very honored and proud to be able to share with him.

The field of bioinformatics continues to become more complex and diversified, requiring that all biologists have a firm understanding of the broad array of tools available to them. We truly hope that this book will help provide our students and colleagues with the kind of insight and vision needed for tackling their next big biological question.

Andreas D. Baxevanis
B. F. Francis Ouellette

Contributors

Rolf Apweiler, Ph.D. is Head of the Sequence Database Group at the European Bioinformatics Institute. He is currently a member of the editorial boards of the European Journal of Biochemistry, Proteomics, and Biochimica et Biophysica Acta.

Gary D. Bader, Ph.D. is currently a post-doctoral fellow in the lab of Dr. Chris Sander at the Computational Biology Center at Memorial Sloan-Kettering Cancer Center in New York City. Previously, Dr. Bader completed a Ph.D. in the lab of Dr. Christopher Hogue in the Department of Biochemistry at the University of Toronto and the Program in Proteomics and Bioinformatics at the Samuel Lunenfeld Research Institute at Mount Sinai Hospital in Toronto. His thesis was the development and research use of the Biomolecular Interaction Network Database (BIND).

Geoff Barton, Ph.D. is Professor of Bioinformatics at the School of Life Sciences, University of Dundee, and Co-Director of the Post Genomics and Molecular Interactions Centre. Dr. Barton obtained his doctorate in the Department of Crystallography, Birkbeck College, University of London, and then spent two years as an ICRF Fellow working with Chris Rawlings at the Imperial Cancer Research Fund Labs in London. In 1989 he was awarded a Royal Society University Research Fellowship to work in the Lab of Molecular Biophysics, University of Oxford. Since that time Dr. Barton has held posts as Head of Genome Informatics at the Wellcome Trust Centre for Human Genetics (1995-1997), Research and Development Team Leader at the EMBL European Bioinformatics Institute (1997-2001), and Head of the European Macromolecular Structure Database at EBI (1998-2001) before taking up his present appointments in 2001. Dr. Barton has published over 50 papers about computational protein sequence and structure analysis in refereed journals as well as around 20 book chapters and other contributions. Software developed by his group is widely distributed and in daily use in many research laboratories worldwide.

Andreas D. Baxevanis, Ph.D. is the Deputy Director for Intramural Research and the Director of the Computational Genomics Program at the National Human Genome Research Institute, National Institutes of Health. He is currently the editor-in-chief of Current Protocols in Bioinformatics, senior editor of Molecular Cancer Therapeutics, and associate editor of Proteins: Structure, Function, and Bioinformatics. His involvement in educational activities include teaching bioinformatics at The Johns Hopkins University, serving as adjunct faculty at Boston University, lecturing in numerous courses, and developing materials intended to facilitate the use of genomic sequence data. His current research focuses on better-understanding structure-function relationships in DNA-binding proteins and how mutations in these proteins contribute to human disease. He is the recipient of the Bodossaki Foundation's 2000 Academic Prize in Medicine and Biology, Greece's highest honor for young scientists of Greek heritage throughout the world.

Enrique Blanco is a computer engineer at Universitat Politècnica de Catalunya, Spain, finishing his Ph.D. thesis in the Genome Bioinformatics Laboratory from Institut Municipal d'Investigació Mèdica-Universitat Pompeu Fabra in Barcelona, Spain. He works in the design of new algorithms of sequence alignment for the detection and characterization of the gene promoter regions. He has been involved in many educational activities, including teaching bioinformatics in numerous courses, and developing materials to facilitate the understanding of bioinformatics for undergraduates, graduates, and Ph.D. students from other disciplines.

Gerard G. Bouffard, Ph.D. is the Director of the Bioinformatics Group of the National Institutes of Health Intramural Sequencing Center (NISC) where he oversees

data generation, management and analysis for this high-throughput DNA sequencing facility. His graduate work in mapping and sequencing in the *E. coli* genome evolved into post-doctoral research in physical mapping of human chromosome 7 and interest in comparative genomics.

Fiona S. L. Brinkman, Ph.D. is an Assistant Professor in Molecular Biology and Biochemistry at Simon Fraser University. She is also Research Director of the Genome Canada Pathogenomics Project, Coordinator of the Pseudomonas Community Genome Annotation Project, and a Core Faculty for the Canadian Bioinformatics Workshops. She has won numerous career awards for her evolutionary infectious disease and bioinformatics research, including being the only Canadian professor listed as one of the "Top 100 of the World's Young Innovators in Technology" in 2002 by the Massachusetts Institute of Technology, the 2003 Science Council of BC Young Innovator award, and Canada's "Top 40 Under 40" for 2003–2004. She has a strong interest in bioinformatics education through development of both graduate and undergraduate curricula and she developed the first undergraduate bioinformatics joint major program (computing science and molecular biology and biochemistry) in Canada.

Anton J. Enright, Ph.D. is a Research Group Leader at the Wellcome Trust Sanger Institute in Cambridge, United Kingdom. Previously he worked at the Computational Biology Center at Memorial Sloan-Kettering Cancer Center, New York and completed his EMBL and Cambridge University predoctoral fellowship at the European Bioinformatics Institute (EBI) in Cambridge, United Kingdom.

Morgan C. Giddings, Ph.D. is an Assistant Professor in the Departments of Microbiology and Immunology and Biomedical Engineering at the University of North Carolina at Chapel Hill. He is a founding member of the UNC-CH training program in bioinformatics and computational biology, computational advisor to the Michael Hooker Proteomics Core at UNC, and a member of the Carolina Center for Genome Sciences. He is the recipient of an NIH/NHGRI Genome Scholar Career Development award. His research applies both computational and laboratory science to examine how genomes encode proteomic diversity.

Roderic Guigó i Serra, Ph.D. leads the Genome Bioinformatics Laboratory at the Institut Municipal d'Investigació Mèdica in Barcelona. He is also Professor at the Universitat Pompeu Fabra, and coordinates the Bioinformatics and Genomics program within the recently created Center for Genomic Regulation. His research focuses on computational gene prediction. He is author of one of the first general purpose gene finding programs, and he has contributed to the development of standards for the evaluation of the accuracy of gene prediction programs. He has also participated in the analysis consortiums of numerous eukaryotic genomes.

Nancy Fisher Hansen, Ph.D. is a Member of the Bioinformatics group at the National Institutes of Health Intramural Sequencing Center. She designs, implements, and modifies software tools to facilitate the generation and analysis of Ordered and Oriented sequence data. She received her doctorate in physical chemistry from Stanford University, and subsequently worked as a software developer at the Stanford Genome Technology Center.

Mark Holmes is a Bioinformatics Developer in the Department of Microbiology & Immunology at the School of Medicine, University of North Carolina at Chapel Hill. His professional computing experience began in 1974 and includes early work automating his own experimental protocols at the Clinical Center of the U.S. National Institutes of Health. A former art historian and museum registrar, he worked for many years in the private sector as a senior computing consultant before returning to academic life in 2000.

David H. Mathews, M.D., Ph.D. is an Assistant Professor of Biochemistry and Biophysics in the Center for Human Genetics and Molecular Pediatric Disease at the University of Rochester Medical Center. He is the author of RNAstructure, a software package for analyzing RNA secondary structure on Windows.

Tara C. Matise, Ph.D. is Associate Professor in the Department of Genetics at Rutgers University in New Jersey. She runs the Laboratory of Computational Genomics, serves on the editorial board of Genome Research, has previously served on the Board of Scientific Counselors for the National Institute of Biotechnology Information (NCBI), National Institutes of Health (NIH), and was previously a HUGO editor for human chromosome 1.

James C. Mullikin, Ph.D. is an Associate Investigator within the Genome Technology Branch of the National Human Genome Research Institute, National Institutes of Health. Prior to joining NHGRI, Dr. Mullikin served as the Head of production software development from 1988–2002 at the Sanger Institute, where he also served as Acting Director of Informatics in 2002. He has been involved in polymorphism discovery and analysis methods since the beginning of The SNP Consortium project in 1999 while at the Sanger Institute.

Yanay Ofran, Ph.D. is a Research Fellow at Columbia University Bioinformatics Center (CUBIC) in the Department of Molecular Biophysics and Biochemistry. He is the developer of several tools and Web servers for sequence analysis and prediction. His educational activity includes teaching bioinformatics and computational biology at Columbia University. He is the recipient of the 2002 Freund Memorial Prize.

B. F. Francis Ouellette, M.Sc. is Director of the University of British Columbia (UBC) Bioinformatics Centre and Associate Professor in Medical Genetics and the Michael Smith Laboratories at UBC. He is also the Director for the Canadian Genetic Diseases Network (CGDN) Bioinformatics Core Facility, where he coordinates the Canadian Bioinformatics Workshop series. He works in comparative genomics and in building tools and databases for bioinformatics analyses. He has previously served as GenBank coordinator at the National Center for Biotechnology Information (NCBI), National Institutes of Health (NIH).

John Quackenbush, Ph.D. is an Investigator in Functional Genomics and Bioinformatics at The Institute for Genomic Research (TIGR) in Rockville, Maryland. He also holds appointments as Professor of Biochemistry at The George Washington University, as Adjunct Professor of Biostatistics at the Bloomberg School of Public Health at The Johns Hopkins University, and as Adjunct Professor of Chemical Engineering at the University of Maryland. Dr. Quackenbush has organized and taught many workshops and courses on DNA microarray analysis and has authored a book (with Helen Causton and Alvis Brazma) as well as numerous articles in the subject. Among other accomplishments, he and his group build the TIGR Gene Index databases, including RESOURCERER and are responsible for the freely-available, open-source TM4 software package for DNA microarray analysis. He is also actively involved in the Microarray Gene Expression Data society (MGED), which has been developing standards for microarray data reporting.

Kevin R. Ramkissoon, B.Sc. is a doctoral graduate student in the department of Microbiology & Immunology at the University of North Carolina at Chapel Hill. His current research interests include the development proteomic and bioinformatic methods to study viral and prokaryotic evolution with a focus on antimicrobial drug resistance, pathogenicity, and immune response evasion.

Burkhard Rost, Dr. rer. nat. is an Associate Professor in the Department of Biochemistry and Molecular Biophysics at Columbia University. Since obtaining his doctorate at the Institute of Theoretical Physics, Heidelberg, Dr. Rost held posts at EMBL Heidelberg (1990—1994), EBI Cambridge (1995), EMBL Heidelberg (1996—1998), and LION Biosciences (1998) before taking up his present appointment in 1999. His group focuses on methods for predicting protein structure and function from sequence, primarily in the context of entirely sequenced organisms. Dr. Rost has given 93 invited lectures in 16 countries, has had over 100 papers published that have been quoted over 5,000 times, and has been responsible for PredictProtein, one of the first Web servers in molecular biology.

Stephen T. Sherry, Ph.D. is Staff Scientist at the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), where he supervises the dbSNP database of genetic variation and directs the development of open-source software tools for quality assessment of DNA forensic data. He received the Library's Board of Regents Award for Scholarship or Technical Achievement in 2003 for his advisory role in applying computational forensic methods to help identify victims of the September 11, 2001 tragedy of the World Trade Center in New York City.

Lincoln D. Stein, M.D., Ph.D. is an Associate Professor at the Cold Spring Harbor Laboratory, where he works on genome databases. He teaches bioinformatics and genetics at the Watson School for Biomedical Sciences, and was recently awarded the Benjamin Franklin prize for service to bioinformatics.

Pamela Jacques Thomas, Ph.D. is a Member of the Bioinformatics Group at the National Institutes of Health Intramural Sequencing Center (NISC) where she focuses on the assembly and annotation of BAC-derived sequences from multiple vertebrates. She received her doctorate from Case Western Reserve University and previously worked as a GenBank Scientific Data Analyst at the National Center for Biotechnology Information (NCBI), National Institutes of Health (NIH).

Peter S. White, Ph.D. is a Research Assistant Professor in the Department of Pediatrics at the University of Pennsylvania, and in the Division of Oncology, Childrens Hospital of Philadelphia (CHOP). Dr. White holds the David Lawrence Altschuler Endowed Chair in Genomics and Computational Biology at CHOP. He is a member of Penn's Center for Bioinformatics and Penns Genomics Institute, and he is the Faculty Director of CHOP's bioinformatics core facility.

David S. Wishart, Ph.D. is a Professor of Biological Sciences and Computing Science at the University of Alberta. He also holds the Bristol Myers Squibb Chair in protein chemistry and is a Senior Research Officer with the National Institute for Nanotechnology (NINT) in Edmonton. In addition to starting two bioinformatics companies (BioTools and Chenomx) in the 1990's, Dr. Wishart has been actively involved in teaching bioinformatics for nearly a decade, including several undergraduate and graduate courses at the University of Alberta. He has been a principal instructor for numerous weeklong training workshops offered through the Canadian Bioinformatics Workshop series (CBW), the Canadian Proteomics Initiative (CPI) and Genome Canada (ACGC).

Tyra G. Wolfsberg, Ph.D. is an Associate Investigator and the Associate Director of the Bioinformatics and Scientific Programming Core at the National Human Genome Research Institute, National Institutes of Health. She lectures and publishes extensively on using bioinformatics tools, especially online genome browsers, to mine genomic sequence information.

Michael Zuker, Ph.D. is a Professor of Mathematical Sciences and Biology at Rensselaer Polytechnic Institute. He works on the development of algorithms to predict folding, hybridization and melting profiles in nucleic acids. His educational activities include developing and teaching his own bioinformatics course at Rensselaer, and participating in both a Chautauqua short course in bioinformatics for college teachers and an intensive bioinformatics course at the University of Michigan.

Contents

PART ONE

BIOLOGICAL DATABASES