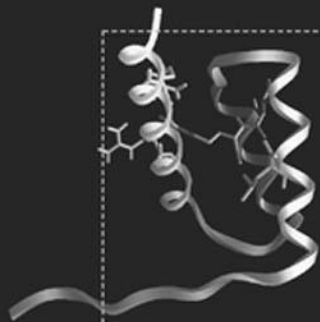


SECOND EDITION



# BIOINFORMATICS

A Practical Guide to the Analysis of Genes and Proteins

EDITED BY  
ANDREAS D. BAXEVANIS  
B. F. FRANCIS OUELLETTE



WWW.  
SUPPLEMENT

# BIOINFORMATICS

---

SECOND EDITION





---

# BIOINFORMATICS

## A Practical Guide to the Analysis of Genes and Proteins

---

SECOND EDITION

---

Andreas D. Baxevanis  
Genome Technology Branch  
National Human Genome Research Institute  
National Institutes of Health  
Bethesda, Maryland  
USA

B. F. Francis Ouellette  
Centre for Molecular Medicine and Therapeutics  
Children's and Women's Health Centre of British Columbia  
University of British Columbia  
Vancouver, British Columbia  
Canada



A JOHN WILEY & SONS, INC., PUBLICATION

New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

Designations used by companies to distinguish their products are often claimed as trademarks. In all instances where John Wiley & Sons, Inc., is aware of a claim, the product names appear in initial capital or ALL CAPITAL LETTERS. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

Copyright © 2001 by John Wiley & Sons, Inc. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic or mechanical, including uploading, downloading, printing, decompiling, recording or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6011, fax (212) 850-6008, E-Mail: PERMREQ@WILEY.COM.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold with the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional person should be sought.

This title is also available in print as ISBN 0-471-38390-2 (cloth) and ISBN 0-471-38391-0 (paper).

For more information about Wiley products, visit our website at [www.Wiley.com](http://www.Wiley.com).

*ADB dedicates this book to his Goddaughter, Anne Terzian, for her constant kindness, good humor, and love—and for always making me smile.*

*BFFO dedicates this book to his daughter, Maya. Her sheer joy and delight in the simplest of things lights up my world everyday.*



---

# CONTENTS

---

Foreword .....	xiii
Preface .....	xv
Contributors .....	xvii

## **1 BIOINFORMATICS AND THE INTERNET 1**

*Andreas D. Baxevanis*

Internet Basics .....	2
Connecting to the Internet .....	4
Electronic Mail .....	7
File Transfer Protocol .....	10
The World Wide Web .....	13
Internet Resources for Topics Presented in Chapter 1 .....	16
References .....	17

## **2 THE NCBI DATA MODEL 19**

*James M. Ostell, Sarah J. Wheelan, and Jonathan A. Kans*

Introduction .....	19
PUBs: Publications or Perish .....	24
SEQ-Ids: What's in a Name? .....	28
BIOSEQs: Sequences .....	31
BIOSEQ-SETs: Collections of Sequences .....	34
SEQ-ANNOT: Annotating the Sequence .....	35
SEQ-DESCR: Describing the Sequence .....	40
Using the Model .....	41
Conclusions .....	43
References .....	43

## **3 THE GENBANK SEQUENCE DATABASE 45**

*Ilene Karsch-Mizrachi and B. F. Francis Ouellette*

Introduction .....	45
Primary and Secondary Databases .....	47
Format vs. Content: Computers vs. Humans .....	47
The Database .....	49



The GenBank Flatfile: A Dissection .....	49
Concluding Remarks .....	58
Internet Resources for Topics Presented in Chapter 3 .....	58
References .....	59
Appendices .....	59
Appendix 3.1 Example of GenBank Flatfile Format .....	59
Appendix 3.2 Example of EMBL Flatfile Format .....	61
Appendix 3.3 Example of a Record in CON Division .....	63
<b>4 SUBMITTING DNA SEQUENCES TO THE DATABASES</b> .....	<b>65</b>
<i>Jonathan A. Kans and B. F. Francis Ouellette</i>	
Introduction .....	65
Why, Where, and What to Submit? .....	66
DNA/RNA .....	67
Population, Phylogenetic, and Mutation Studies .....	69
Protein-Only Submissions .....	69
How to Submit on the World Wide Web .....	70
How to Submit with Sequin .....	70
Updates .....	77
Consequences of the Data Model .....	77
EST/STS/GSS/HTG/SNP and Genome Centers .....	79
Concluding Remarks .....	79
Contact Points for Submission of Sequence Data to	
DDBJ/EMBL/GenBank .....	80
Internet Resources for Topics Presented in Chapter 4 .....	80
References .....	81
<b>5 STRUCTURE DATABASES</b> .....	<b>83</b>
<i>Christopher W. V. Hogue</i>	
Introduction to Structures .....	83
PDB: Protein Data Bank at the Research Collaboratory for	
Structural Bioinformatics (RCSB) .....	87
MMDB: Molecular Modeling Database at NCBI .....	91
Structure File Formats .....	94
Visualizing Structural Information .....	95
Database Structure Viewers .....	100
Advanced Structure Modeling .....	103
Structure Similarity Searching .....	103
Internet Resources for Topics Presented in Chapter 5 .....	106
Problem Set .....	107
References .....	107
<b>6 GENOMIC MAPPING AND MAPPING DATABASES</b> .....	<b>111</b>
<i>Peter S. White and Tara C. Matise</i>	
Interplay of Mapping and Sequencing .....	112
Genomic Map Elements .....	113

Types of Maps .....	115
Complexities and Pitfalls of Mapping .....	120
Data Repositories .....	122
Mapping Projects and Associated Resources .....	127
Practical Uses of Mapping Resources .....	142
Internet Resources for Topics Presented in Chapter 6 .....	146
Problem Set .....	148
References .....	149

**7 INFORMATION RETRIEVAL FROM BIOLOGICAL DATABASES 155**

*Andreas D. Baxevanis*

Integrated Information Retrieval: The Entrez System .....	156
LocusLink .....	172
Sequence Databases Beyond NCBI .....	178
Medical Databases .....	181
Internet Resources for Topics Presented in Chapter 7 .....	183
Problem Set .....	184
References .....	185

**8 SEQUENCE ALIGNMENT AND DATABASE SEARCHING 187**

*Gregory D. Schuler*

Introduction .....	187
The Evolutionary Basis of Sequence Alignment .....	188
The Modular Nature of Proteins .....	190
Optimal Alignment Methods .....	193
Substitution Scores and Gap Penalties .....	195
Statistical Significance of Alignments .....	198
Database Similarity Searching .....	198
FASTA .....	200
BLAST .....	202
Database Searching Artifacts .....	204
Position-Specific Scoring Matrices .....	208
Spliced Alignments .....	209
Conclusions .....	210
Internet Resources for Topics Presented in Chapter 8 .....	212
References .....	212

**9 CREATION AND ANALYSIS OF PROTEIN MULTIPLE SEQUENCE ALIGNMENTS 215**

*Geoffrey J. Barton*

Introduction .....	215
What is a Multiple Alignment, and Why Do It? .....	216
Structural Alignment or Evolutionary Alignment? .....	216
How to Multiply Align Sequences .....	217

Tools to Assist the Analysis of Multiple Alignments .....	222
Collections of Multiple Alignments .....	227
Internet Resources for Topics Presented in Chapter 9 .....	228
Problem Set .....	229
References .....	230
<b>10 PREDICTIVE METHODS USING DNA SEQUENCES</b> .....	<b>233</b>
<i>Andreas D. Baxevanis</i>	
GRAIL .....	235
FGENEH/FGENES .....	236
MZEF .....	238
GENSCAN .....	240
PROCRUSTES .....	241
How Well Do the Methods Work? .....	246
Strategies and Considerations .....	248
Internet Resources for Topics Presented in Chapter 10 .....	250
Problem Set .....	251
References .....	251
<b>11 PREDICTIVE METHODS USING PROTEIN SEQUENCES</b> .....	<b>253</b>
<i>Sharmila Banerjee-Basu and Andreas D. Baxevanis</i>	
Protein Identity Based on Composition .....	254
Physical Properties Based on Sequence .....	257
Motifs and Patterns .....	259
Secondary Structure and Folding Classes .....	263
Specialized Structures or Features .....	269
Tertiary Structure .....	274
Internet Resources for Topics Presented in Chapter 11 .....	277
Problem Set .....	278
References .....	279
<b>12 EXPRESSED SEQUENCE TAGS (ESTs)</b> .....	<b>283</b>
<i>Tyra G. Wolfsberg and David Landsman</i>	
What is an EST? .....	284
EST Clustering .....	288
TIGR Gene Indices .....	293
STACK .....	293
ESTs and Gene Discovery .....	294
The Human Gene Map .....	294
Gene Prediction in Genomic DNA .....	295
ESTs and Sequence Polymorphisms .....	296
Assessing Levels of Gene Expression Using ESTs .....	296
Internet Resources for Topics Presented in Chapter 12 .....	298
Problem Set .....	298
References .....	299

**13 SEQUENCE ASSEMBLY AND FINISHING METHODS 303***Rodger Staden, David P. Judge, and James K. Bonfield*

The Use of Base Cell Accuracy Estimates or Confidence Values	305
The Requirements for Assembly Software	306
Global Assembly	306
File Formats	307
Preparing Readings for Assembly	308
Introduction to Gap4	311
The Contig Selector	311
The Contig Comparator	312
The Template Display	313
The Consistency Display	316
The Contig Editor	316
The Contig Joining Editor	319
Disassembling Readings	319
Experiment Suggestion and Automation	319
Concluding Remarks	321
Internet Resources for Topics Presented in Chapter 13	321
Problem Set	322
References	322

**14 PHYLOGENETIC ANALYSIS 323***Fiona S. L. Brinkman and Detlef D. Leipe*

Fundamental Elements of Phylogenetic Models	325
Tree Interpretation—The Importance of Identifying Paralogs and Orthologs	327
Phylogenetic Data Analysis: The Four Steps	327
Alignment: Building the Data Model	329
Alignment: Extraction of a Phylogenetic Data Set	333
Determining the Substitution Model	335
Tree-Building Methods	340
Distance, Parsimony, and Maximum Likelihood: What's the Difference?	345
Tree Evaluation	346
Phylogenetics Software	348
Internet-Accessible Phylogenetic Analysis Software	354
Some Simple Practical Considerations	356
Internet Resources for Topics Presented in Chapter 14	356
References	357

**15 COMPARATIVE GENOME ANALYSIS 359***Michael Y. Galperin and Eugene V. Koonin*

Progress in Genome Sequencing	360
Genome Analysis and Annotation	366
Application of Comparative Genomics—Reconstruction of Metabolic Pathways	382
Avoiding Common Problems in Genome Annotation	385

Conclusions .....	387
Internet Resources for Topics Presented in Chapter 15 .....	387
Problems for Additional Study .....	389
References .....	390
<b>16 LARGE-SCALE GENOME ANALYSIS</b> .....	<b>393</b>
<i>Paul S. Meltzer</i>	
Introduction .....	393
Technologies for Large-Scale Gene Expression .....	394
Computational Tools for Expression Analysis .....	399
Hierarchical Clustering .....	407
Prospects for the Future .....	409
Internet Resources for Topics Presented in Chapter 16 .....	410
References .....	410
<b>17 USING PERL TO FACILITATE BIOLOGICAL ANALYSIS</b> .....	<b>413</b>
<i>Lincoln D. Stein</i>	
Getting Started .....	414
How Scripts Work .....	416
Strings, Numbers, and Variables .....	417
Arithmetic .....	418
Variable Interpolation .....	419
Basic Input and Output .....	420
Filehandles .....	422
Making Decisions .....	424
Conditional Blocks .....	427
What is Truth? .....	430
Loops .....	430
Combining Loops with Input .....	432
Standard Input and Output .....	433
Finding the Length of a Sequence File .....	435
Pattern Matching .....	436
Extracting Patterns .....	440
Arrays .....	441
Arrays and Lists .....	444
Split and Join .....	444
Hashes .....	445
A Real-World Example .....	446
Where to Go From Here .....	449
Internet Resources for Topics Presented in Chapter 17 .....	449
Suggested Reading .....	449
Glossary .....	451
Index .....	457

---

# FOREWORD

---



I am writing these words on a watershed day in molecular biology. This morning, a paper was officially published in the journal *Nature* reporting an initial sequence and analysis of the human genome. One of the fruits of the Human Genome Project, the paper describes the broad landscape of the nearly 3 billion bases of the euchromatic portion of the human chromosomes.

In the most narrow sense, the paper was the product of a remarkable international collaboration involving six countries, twenty genome centers, and more than a thousand scientists (myself included) to produce the information and to make it available to the world freely and without restriction.

In a broader sense, though, the paper is the product of a century-long scientific program to understand genetic information. The program began with the rediscovery of Mendel's laws at the beginning of the 20th century, showing that information was somehow transmitted from generation to generation in discrete form. During the first quarter-century, biologists found that the cellular basis of the information was the chromosomes. During the second quarter-century, they discovered that the molecular basis of the information was DNA. During the third quarter-century, they unraveled the mechanisms by which cells read this information and developed the recombinant DNA tools by which scientists can do the same. During the last quarter-century, biologists have been trying voraciously to gather genetic information—first from genes, then entire genomes.

The result is that biology in the 21st century is being transformed from a purely laboratory-based science to an information science as well. The information includes comprehensive global views of DNA sequence, RNA expression, protein interactions or molecular conformations. Increasingly, biological studies begin with the study of huge databases to help formulate specific hypotheses or design large-scale experiments. In turn, laboratory work ends with the accumulation of massive collections of data that must be sifted. These changes represent a dramatic shift in the biological sciences.

One of the crucial steps in this transformation will be training a new generation of biologists who are both computational scientists and laboratory scientists. This major challenge requires both vision and hard work: vision to set an appropriate agenda for the computational biologist of the future and hard work to develop a curriculum and textbook.

James Watson changed the world with his co-discovery of the double-helical structure of DNA in 1953. But, he also helped train a new generation to inhabit that new world in the 1960s and beyond through his textbook, *The Molecular Biology of the Gene*. Discovery and teaching go hand-in-hand in changing the world.

In this book, Andy Baxevanis and Francis Ouellette have taken on the tremendously important challenge of training the 21st century computational biologist. Toward this end, they have undertaken the difficult task of organizing the knowledge in this field in a logical progression and presenting it in a digestible form. And, they have done an excellent job. This fine text will make a major impact on biological research and, in turn, on progress in biomedicine. We are all in their debt.

Eric S. Lander

*February 15, 2001  
Cambridge, Massachusetts*

---

# PREFACE

---



With the advent of the new millenium, the scientific community marked a significant milestone in the study of biology—the completion of the “working draft” of the human genome. This work, which was chronicled in special editions of *Nature* and *Science* in early 2001, signals a new beginning for modern biology, one in which the majority of biological and biomedical research would be conducted in a “sequence-based” fashion. This new approach, long-awaited and much-debated, promises to quickly lead to advances not only in the understanding of basic biological processes, but in the prevention, diagnosis, and treatment of many genetic and genomic disorders. While the fruits of sequencing the human genome may not be known or appreciated for another hundred years or more, the implications to the basic way in which science and medicine will be practiced in the future are staggering. The availability of this flood of raw information has had a significant effect on the field of bioinformatics as well, with a significant amount of effort being spent on how to effectively and efficiently warehouse and access these data, as well as on new methods aimed at mining this warehoused data in order to make novel biological discoveries.

This new edition of *Bioinformatics* attempts to keep up with the quick pace of change in this field, reinforcing concepts that have stood the test of time while making the reader aware of new approaches and algorithms that have emerged since the publication of the first edition. Based on our experience both as scientists and as teachers, we have tried to improve upon the first edition by introducing a number of new features in the current version. Five chapters have been added on topics that have emerged as being important enough in their own right to warrant distinct and separate discussion: expressed sequence tags, sequence assembly, comparative genomics, large-scale genome analysis, and BioPerl. We have also included problem sets at the end of most of the chapters with the hopes that the readers will work through these examples, thereby reinforcing their command of the concepts presented therein. The solutions to these problems are available through the book’s Web site, at [www.wiley.com/bioinformatics](http://www.wiley.com/bioinformatics). We have been heartened by the large number of instructors who have adopted the first edition as their book of choice, and hope that these new features will continue to make the book useful both in the classroom and at the bench.

There are many individuals we both thank, without whose efforts this volume would not have become a reality. First and foremost, our thanks go to all of the authors whose individual contributions make up this book. The expertise and professional viewpoints that these individuals bring to bear go a long way in making this book’s contents as strong as it is. That, coupled with their general good-



naturedness under tight time constraints, has made working with these men and women an absolute pleasure.

Since the databases and tools discussed in this book are unique in that they are freely shared amongst fellow academics, we would be remiss if we did not thank all of the people who, on a daily basis, devote their efforts to curating and maintaining the public databases, as well as those who have developed the now-indispensible tools for mining the data contained in those databases. As we pointed out in the preface to the first edition, the bioinformatics community is truly unique in that the *esprit de corps* characterizing this group is one of openness, and this underlying philosophy is one that has enabled the field of bioinformatics to make the substantial strides that it has in such a short period of time.

We also thank our editor, Luna Han, for her steadfast patience and support throughout the entire process of making this new edition a reality. Through our extended discussions both on the phone and in person, and in going from deadline to deadline, we've developed a wonderful relationship with Luna, and look forward to working with her again on related projects in the future. We also would like to thank Camille Carter and Danielle Lacourciere at Wiley for making the entire copy-editing process a quick and (relatively) painless one, as well as Eloise Nelson for all of her hard work in making sure all of the loose ends came together on schedule.

BFFO would like to acknowledge the continued support of Nancy Ryder. Nancy is not only a friend, spouse, and mother to our daughter Maya, but a continuous source of inspiration to do better, and to challenge; this is something that I try to do every day, and her love and support enables this. BFFO also wants to acknowledge the continued friendship and support from ADB throughout both of these editions. It has been an honor and a privilege to be a co-editor with him. Little did we know seven years ago, in the *second* basement of the Lister Hill Building at NIH where we shared an office, that so many words would be shared between our respective computers.

ADB would also like to specifically thank Debbie Wilson for all of her help throughout the editing process, whose help and moral support went a long way in making sure that this project got done the right way the first time around. I would also like to extend special thanks to Jeff Trent, who I have had the pleasure of working with for the past several years and with whom I've developed a special bond, both professionally and personally. Jeff has enthusiastically provided me the latitude to work on projects like these and has been a wonderful colleague and friend, and I look forward to our continued associations in the future.

*Andreas D. Baxevanis  
B. F. Francis Ouellette*

---

# CONTRIBUTORS

---

**Sharmila Banerjee-Basu**, Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland

**Geoffrey J. Barton**, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

**Andreas D. Baxevanis**, Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland

**James K. Bonfield**, Medical Research Council, Laboratory of Molecular Biology, Cambridge, United Kingdom

**Fiona S. L. Brinkman**, Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada

**Michael Y. Galperin**, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

**Christopher W. V. Hogue**, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada

**David P. Judge**, Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

**Jonathan A. Kans**, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

**Ilene Karsch-Mizrachi**, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

**Eugene V. Koonin**, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

**David Landsman**, Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

**Detlef D. Leipe**, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

**Tara C. Matise**, Department of Genetics, Rutgers University, New Brunswick, New Jersey

**Paul S. Meltzer**, Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland

**James M. Ostell**, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

**B. F. Francis Ouellette**, Centre for Molecular Medicine and Therapeutics, Children's and Women's Health Centre of British Columbia, The University of British Columbia, Vancouver, British Columbia, Canada

**Gregory D. Schuler**, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

**Rodger Staden**, Medical Research Council, Laboratory of Molecular Biology, Cambridge, United Kingdom

**Lincoln D. Stein**, The Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

**Sarah J. Wheelan**, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland and Department of Molecular Biology and Genetics, The Johns Hopkins School of Medicine, Baltimore, Maryland

**Peter S. White**, Department of Pediatrics, University of Pennsylvania, Philadelphia, Pennsylvania

**Tyra G. Wolfsberg**, Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland