

---

# BIOINFORMATICS

## A PRACTICAL GUIDE TO THE ANALYSIS OF GENES AND PROTEINS

---

EDITED BY

**Andreas D. Baxevanis**

Genome Technology Branch  
National Human Genome Research Institute  
National Institutes of Health  
Bethesda, Maryland

**B. F. Francis Ouellette**

National Center for Biotechnology Information  
National Institutes of Health  
Bethesda, Maryland



**WILEY-INTERSCIENCE**

**A JOHN WILEY & SONS, INC., PUBLICATION**

New York / Chichester / Weinheim / Brisbane / Singapore / Toronto

---

# CONTENTS

---

<b>Preface</b>	<b>vii</b>
<b>Contributors</b>	<b>xiv</b>
<b>1 The Internet and the Biologist</b>	<b>1</b>
<i>Andreas D. Baxevanis</i>	
Internet Basics / 1	
Connecting to the Internet / 3	
Electronic Mail / 4	
File Transfer Protocol / 7	
Gopher / 8	
The World Wide Web / 8	
References / 15	
<b>2 The GenBank Sequence Database</b>	<b>16</b>
<i>B. F. Francis Ouellette</i>	
Introduction / 16	
Primary and Secondary Databases / 19	
Format vs. Content: Computers vs. Humans / 20	
The Database / 21	
The GenBank Flatfile: A Dissection / 22	
Conclusions / 33	
References / 33	
Appendices: Database File Formats / 35	
Appendix 2.1 Example of an GenBank (or DDBS) Record / 35	
Appendix 2.2 Example of an ASN.1 Record / 36	
Appendix 2.3 Example of an EMBL Record / 42	
Appendix 2.4 Example of a GenBank Summary File / 44	
<b>3 Structure Databases</b>	<b>46</b>
<i>Christopher W. V. Hogue and Stephen H. Bryant</i>	
Introduction to Structures / 46	
PDB: Protein Data Bank at Brookhaven National Laboratories / 49	
MMDB: Molecular Modeling Database at NCBI / 56	

Structure File Formats / 58  
Visualizing Structural Information / 60  
Database Structure Viewers / 68  
Can't Find a Published Structure? / 70  
References / 71  
Monographs / 72

**4 Sequence Analysis Using GCG**

**74**

*Barbara A. Butler*

Introduction / 74  
The Wisconsin Package / 75  
Databases That Accompany the Wisconsin Package / 75  
The SeqLab Environment / 76  
Analyzing Sequences with Operations and Wisconsin Package Programs / 80  
Viewing Output / 81  
Monitoring Program Progress and Troubleshooting Problems / 83  
Annotating Sequences and Graphically Displaying Annotations in the SeqLab Editor / 83  
Saving Sequences in the SeqLab Editor / 85  
Examples of Analyses That Can Be Undertaken in SeqLab / 85  
Extending SeqLab by Including Programs That Are Not Part of the Wisconsin Package / 91  
References / 91  
Appendix / 93

**5 Information Retrieval from Biological Databases**

**98**

*Andreas D. Baxevanis*

Retrieving Database Entries: The Retrieve Server / 99  
Integrated Information Retrieval: The Entrez System / 101  
Integrated Information Access: The Query Server / 111  
Sequence Databases Beyond NCBI / 115  
Medical Databases / 118  
References / 120

**6 The NCBI Data Model**

**121**

*James M. Ostell and Jonathan A. Kans*

Introduction / 121  
Pubs: Publications or Perish / 125  
Seqids: What's in a Name? / 129  
Bioseq: Sequences / 132  
BioseqSets: Collections of Sequences / 135  
Seq-annot: Annotating the Sequence / 136  
Seq-descr: Describing the Sequence / 140  
Using the Model / 141  
Conclusions / 143  
References / 144

<b>7</b>	<b>Sequence Alignment and Database Searching</b>	<b>145</b>
	<i>Gregory D. Schuler</i>	
	Introduction / 145	
	The Evolutionary Basis of Sequence Alignment / 146	
	The Modular Nature of Proteins / 148	
	Optimal Alignment Methods / 150	
	Substitution Scores and Gap Penalties / 151	
	Statistical Significance of Alignments / 155	
	Database Similarity Searching / 156	
	FASTA / 159	
	BLAST / 160	
	Low-Complexity Regions / 166	
	Repetitive Elements / 166	
	Conclusions / 169	
	References / 170	
<b>8</b>	<b>Practical Aspects of Multiple Sequence Alignment</b>	<b>172</b>
	<i>Andreas D. Baxevanis</i>	
	Progressive Alignment Methods / 173	
	Motifs and Patterns / 176	
	Presentation Methods / 184	
	References / 188	
<b>9</b>	<b>Phylogenetic Analysis</b>	<b>189</b>
	<i>Mark A. HersHKovitz and Detlef D. Leipe</i>	
	Elements of Phylogenetic Models / 190	
	Phylogenetic Data Analysis: Alignment, Substitution Model Building, Tree Building, and Tree Evaluation / 191	
	Building the Data Model (Alignment) / 191	
	Determining the Substitution Model / 197	
	Tree-Building Methods / 206	
	Searching for Trees / 211	
	Rooting Trees / 212	
	Evaluating Trees and Data / 213	
	Phylogenetics Software / 217	
	Some Simple Practical Considerations / 225	
	Acknowledgments / 227	
	References / 227	
<b>10</b>	<b>Predictive Methods Using Nucleotide Sequences</b>	<b>231</b>
	<i>James W. Fickett</i>	
	Framework / 232	
	Masking Repetitive DNA / 232	
	Database Searches / 234	
	Codon Bias Detection / 234	

Detecting Functional Sites in the DNA / 236  
Integrated Gene Parsing / 238  
Finding tRNA Genes / 238  
Future Prospects / 241  
Acknowledgments / 243  
References / 243

**11 Predictive Methods Using Protein Sequences 246**

*Andreas D. Baxeavanis and David Landsman*

Protein Identity Based on Composition / 247  
Physical Properties Based on Sequence / 250  
Secondary Structure and Folding Classes / 252  
Specialized Structures or Features / 257  
Tertiary Structure / 262  
References / 265

**12 Of Mice and Men: Navigating Public Physical Mapping Databases 268**

*Lincoln D. Stein*

Types of Physical Map / 269  
Genome-Wide Maps from Large Community Databases / 271  
Genome-Wide Maps from Individual Sources / 278  
Chromosome-Specific Human Maps / 291  
Mouse Mapping Resources / 294  
References / 297

**13 ACEDB: A Database for Genome Information 299**

*Sean Walsh, Mary Anderson, and Samuel W. Cartinhour*

General Features of ACEDB / 299  
Sequence Analysis in ACEDB / 305  
Miscellaneous Analysis Functions / 315  
Acknowledgments / 316

**14 Submitting DNA Sequence to the Databases 319**

*Jonathan A. Kans and B. F. Francis Ouellette*

Introduction / 319  
Where to Submit? / 320  
What to Submit? / 321  
How to Submit on the World Wide Web / 324  
How to Submit with Sequin / 326  
EST/STS/GSS / 348  
Genome Centers / 348  
Updates / 350

Concluding Remarks / 351  
Acknowledgments / 351  
References / 353

<b>Appendix 1: Glossary</b>	<b>355</b>
<b>Appendix 2: Sample Sequence File Formats</b>	<b>359</b>
<b>Index</b>	<b>363</b>