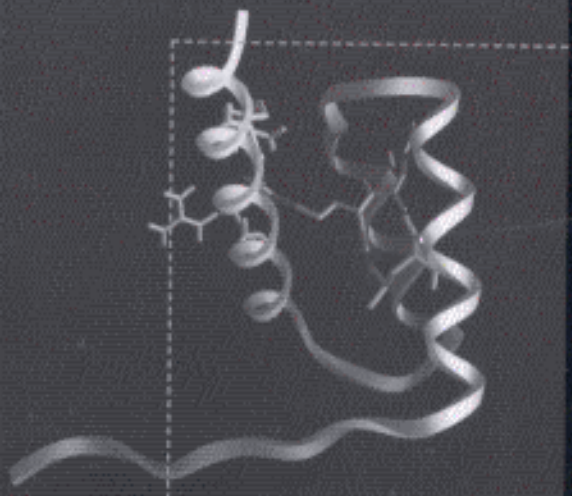
SECOND EDITION



BIOINFORMATICS

A Practical Guide to the Analysis of Genes and Proteins

EDITED BY

ANDREAS D. BAXEVANIS

B. F. FRANCIS OUELLETTE

WWW.

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

4 SUBMITTING DNA SEQUENCES TO THE DATABASES 65
Jonathan A. Kans and B. F. Francis Ouellette

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

5 STRUCTURE DATABASES 83
Christopher W. V. Hogue

Introduction to Structures 83
PDB: Protein Data Bank at the Research Collaboratory for Structural Bioinformatics (RCSB) 87
MMDB: Molecular Modeling Database at NCBI 91
Stucture 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

6 GENOMIC MAPPING AND MAPPING DATABASES 111
Peter S. White and Tara C. Matise

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

10 **PREDICTIVE METHODS USING DNA SEQUENCES** **233**
Andreas D. Baxevanis

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

11 **PREDICTIVE METHODS USING PROTEIN SEQUENCES** **253**
Sharmila Banerjee-Basu and Andreas D. Baxevanis

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

12 **EXPRESSED SEQUENCE TAGS (ESTs)** **283**
Tyra G. Wolfsberg and David Landsman

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

16 LARGE-SCALE GENOME ANALYSIS 393
Paul S. Meltzer

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

17 USING PERL TO FACILITATE BIOLOGICAL ANALYSIS 413
Lincoln D. Stein

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