

Contents

Preface ■ ix

SECTION I INTRODUCTION AND BIOLOGICAL DATABASES

- 1 Introduction ■ 3**
 - What Is Bioinformatics? ■ 4
 - Goal ■ 5
 - Scope ■ 5
 - Applications ■ 6
 - Limitations ■ 7
 - New Themes ■ 8
 - Further Reading ■ 8
- 2 Introduction to Biological Databases ■ 10**
 - What Is a Database? ■ 10
 - Types of Databases ■ 10
 - Biological Databases ■ 13
 - Pitfalls of Biological Databases ■ 17
 - Information Retrieval from Biological Databases ■ 18
 - Summary ■ 27
 - Further Reading ■ 27

SECTION II SEQUENCE ALIGNMENT

- 3 Pairwise Sequence Alignment ■ 31**
 - Evolutionary Basis ■ 31
 - Sequence Homology versus Sequence Similarity ■ 32
 - Sequence Similarity versus Sequence Identity ■ 33
 - Methods ■ 34
 - Scoring Matrices ■ 41
 - Statistical Significance of Sequence Alignment ■ 47
 - Summary ■ 48
 - Further Reading ■ 49
- 4 Database Similarity Searching ■ 51**
 - Unique Requirements of Database Searching ■ 51
 - Heuristic Database Searching ■ 52
 - Basic Local Alignment Search Tool (BLAST) ■ 52
 - FASTA ■ 57
 - Comparison of FASTA and BLAST ■ 60
 - Database Searching with the Smith-Waterman Method ■ 61

- Summary ■ 61
- Further Reading ■ 62
- 5 Multiple Sequence Alignment ■ 63**
 - Scoring Function ■ 63
 - Exhaustive Algorithms ■ 64
 - Heuristic Algorithms ■ 65
 - Practical Issues ■ 71
 - Summary ■ 73
 - Further Reading ■ 74
- 6 Profiles and Hidden Markov Models ■ 75**
 - Position-Specific Scoring Matrices ■ 75
 - Profiles ■ 77
 - Markov Model and Hidden Markov Model ■ 79
 - Summary ■ 84
 - Further Reading ■ 84
- 7 Protein Motifs and Domain Prediction ■ 85**
 - Identification of Motifs and Domains in Multiple Sequence Alignment ■ 86
 - Motif and Domain Databases Using Regular Expressions ■ 86
 - Motif and Domain Databases Using Statistical Models ■ 87
 - Protein Family Databases ■ 90
 - Motif Discovery in Unaligned Sequences ■ 91
 - Sequence Logos ■ 92
 - Summary ■ 93
 - Further Reading ■ 94

SECTION III GENE AND PROMOTER PREDICTION

- 8 Gene Prediction ■ 97**
 - Categories of Gene Prediction Programs ■ 97
 - Gene Prediction in Prokaryotes ■ 98
 - Gene Prediction in Eukaryotes ■ 103
 - Summary ■ 111
 - Further Reading ■ 111
- 9 Promoter and Regulatory Element Prediction ■ 113**
 - Promoter and Regulatory Elements in Prokaryotes ■ 113
 - Promoter and Regulatory Elements in Eukaryotes ■ 114
 - Prediction Algorithms ■ 115
 - Summary ■ 123
 - Further Reading ■ 124

SECTION IV MOLECULAR PHYLOGENETICS

- 10 Phylogenetics Basics ■ 127**
 - Molecular Evolution and Molecular Phylogenetics ■ 127
 - Terminology ■ 128
 - Gene Phylogeny versus Species Phylogeny ■ 130

- Forms of Tree Representation ■ 131
- Why Finding a True Tree Is Difficult ■ 132
- Procedure ■ 133
- Summary ■ 140
- Further Reading ■ 141

11 Phylogenetic Tree Construction Methods and Programs ■ 142

- Distance-Based Methods ■ 142
- Character-Based Methods ■ 150
- Phylogenetic Tree Evaluation ■ 163
- Phylogenetic Programs ■ 167
- Summary ■ 168
- Further Reading ■ 169

SECTION V STRUCTURAL BIOINFORMATICS

12 Protein Structure Basics ■ 173

- Amino Acids ■ 173
- Peptide Formation ■ 174
- Dihedral Angles ■ 175
- Hierarchy ■ 176
- Secondary Structures ■ 178
- Tertiary Structures ■ 180
- Determination of Protein Three-Dimensional Structure ■ 181
- Protein Structure Database ■ 182
- Summary ■ 185
- Further Reading ■ 186

13 Protein Structure Visualization, Comparison, and Classification ■ 187

- Protein Structural Visualization ■ 187
- Protein Structure Comparison ■ 190
- Protein Structure Classification ■ 195
- Summary ■ 199
- Further Reading ■ 199

14 Protein Secondary Structure Prediction ■ 200

- Secondary Structure Prediction for Globular Proteins ■ 201
- Secondary Structure Prediction for Transmembrane Proteins ■ 208
- Coiled Coil Prediction ■ 211
- Summary ■ 212
- Further Reading ■ 213

15 Protein Tertiary Structure Prediction ■ 214

- Methods ■ 215
- Homology Modeling ■ 215
- Threading and Fold Recognition ■ 223
- Ab Initio Protein Structural Prediction ■ 227
- CASP ■ 228
- Summary ■ 229
- Further Reading ■ 230

16 RNA Structure Prediction ■ 231

Introduction ■ 231

Types of RNA Structures ■ 233

RNA Secondary Structure Prediction Methods ■ 234

Ab Initio Approach ■ 234

Comparative Approach ■ 237

Performance Evaluation ■ 239

Summary ■ 239

Further Reading ■ 240

SECTION VI GENOMICS AND PROTEOMICS**17 Genome Mapping, Assembly, and Comparison ■ 243**

Genome Mapping ■ 243

Genome Sequencing ■ 245

Genome Sequence Assembly ■ 246

Genome Annotation ■ 250

Comparative Genomics ■ 255

Summary ■ 259

Further Reading ■ 259

18 Functional Genomics ■ 261

Sequence-Based Approaches ■ 261

Microarray-Based Approaches ■ 267

Comparison of SAGE and DNA Microarrays ■ 278

Summary ■ 279

Further Reading ■ 280

19 Proteomics ■ 281

Technology of Protein Expression Analysis ■ 281

Posttranslational Modification ■ 287

Protein Sorting ■ 289

Protein-Protein Interactions ■ 291

Summary ■ 296

Further Reading ■ 296

APPENDIX

Appendix 1. Practical Exercises ■ 301

Appendix 2. Glossary ■ 318

Index ■ 331