Contents

Preface ..... **V**

1 Introduction to Bioinformatics **1**
1.1 Introduction 1
1.2 Needs of Bioinformatics Technologies 2
1.3 An Overview of Bioinformatics Technologies 5
1.4 A Brief Discussion on the Chapters 8
References 12

2 Overview of Structural Bioinformatics **15**
2.1 Introduction 15
2.2 Organization of Structural Bioinformatics 17
2.3 Primary Resource: Protein Data Bank 18
2.3.1 Data Format 18
2.3.2 Growth of Data 18
2.3.3 Data Processing and Quality Control 20
2.3.4 The Future of the PDB 21
2.3.5 Visualization 21
2.4 Secondary Resources and Applications 22
2.4.1 Structural Classification 22
2.4.2 Structure Prediction 28
2.4.3 Functional Assignments in Structural Genomics 30
2.4.4 Protein-Protein Interactions 32
2.4.5 Protein-Ligand Interactions 34
2.5 Using Structural Bioinformatics Approaches in Drug Design 37
2.6 The Future 39
2.6.1 Integration over Multiple Resources 39
2.6.2 The Impact of Structural Genomics 39
2.6.3 The Role of Structural Bioinformatics in Systems Biology 39
References 40

3 Database Warehousing in Bioinformatics **45**
3.1 Introduction 45
3.2 Bioinformatics Data 48
3.3 Transforming Data to Knowledge 51
3.4 Data Warehousing 54
3.5 Data Warehouse Architecture 56
3.6 Data Quality 58
3.7 Concluding Remarks 60

References 61

4 Data Mining for Bioinformatics 63
4.1 Introduction 63
4.2 Biomedical Data Analysis 64
4.2.1 Major Nucleotide Sequence Database, Protein Sequence Database, and Gene Expression Database 65
4.2.2 Software Tools for Bioinformatics Research 68
4.3 DNA Data Analysis 71
4.3.1 DNA Sequence 71
4.3.2 DNA Data Analysis 76
4.4 Protein Data Analysis 92
4.4.1 Protein and Amino Acid Sequence 92
4.4.2 Protein Data Analysis 99
References 109

5 Machine Learning in Bioinformatics 117
5.1 Introduction 117
5.2 Artificial Neural Network 120
5.3 Neural Network Architectures and Applications 128
5.3.1 Neural Network Architecture 128
5.3.2 Neural Network Learning Algorithms 131
5.3.3 Neural Network Applications in Bioinformatics 134
5.4 Genetic Algorithm 135
5.5 Fuzzy System 141
References 147

6 Systems Biotechnology: a New Paradigm in Biotechnology Development 155
6.1 Introduction 155
6.2 Why Systems Biotechnology? 156
6.3 Tools for Systems Biotechnology 158
6.3.1 Genome Analyses 158
6.3.2 Transcriptome Analyses 159
6.3.3 Proteome Analyses 161
6.3.4 Metabolome/Fluxome Analyses 163
6.4 Integrative Approaches 164
6.5 In Silico Modeling and Simulation of Cellular Processes 166
6.5.1 Statistical Modeling 167
6.5.2 Dynamic Modeling 169
6.6 Conclusion 170
References 171

7 Computational Modeling of Biological Processes with Petri Net-Based Architecture 179
7.1 Introduction 179
7.2 Hybrid Petri Net and Hybrid Dynamic Net 183
7.3 Hybrid Functional Petri Net 190
7.4 Hybrid Functional Petri Net with Extension 191
7.4.1 Definitions 191
7.4.2 Relationships with Other Petri Nets 197
7.4.3 Implementation of HFPNe in Genomic Object Net 198
7.5 Modeling of Biological Processes with HFPNe 198
7.5.1 From DNA to mRNA in Eucaryotes – Alternative Splicing . 199
7.5.2 Translation of mRNA – Frameshift 203
7.5.3 Huntington's Disease 203
7.5.4 Protein Modification – p53 207
7.6 Related Works with HFPNe 211
7.7 Genomic Object Net: GON 212
7.7.1 GON Features That Derived from HFPNe Features 214
7.7.2 GON GUI and Other Features 214
7.7.3 GONML and Related Works with GONML 220
7.7.4 Related Works with GON 222
7.8 Visualizer 224
7.8.1 Bio-processes on Visualizer 226
7.8.2 Related Works with Visualizer 231
7.9 BPE 233
7.10 Conclusion 236
References 236

8 Biological Sequence Assembly and Alignment 243
8.1 Introduction 243
8.2 Large-Scale Sequence Assembly 245
8.2.1 Related Research 245
8.2.2 Euler Sequence Assembly 249
8.2.3 PESA Sequence Assembly Algorithm 249
8.3 Large-Scale Pairwise Sequence Alignment 254
8.3.1 Pairwise Sequence Alignment 254
8.3.2 Large Smith-Waterman Pairwise Sequence Alignment 256
8.4 Large-Scale Multiple Sequence Alignment 257
8.4.1 Multiple Sequence Alignment 257
8.4.2 Large-Scale Clustal W Multiple Sequence Alignment 258
8.5 Load Balancing and Communication Overhead 259
8.6 Conclusion 259
References 260

9 Modeling for Bioinformatics 263
9.1 Introduction 263
9.2 Hidden Markov Modeling for Biological Data Analysis 264
9.2.1 Hidden Markov Modeling for Sequence Identification 264
9.2.2 Hidden Markov Modeling for Sequence Classification 273
9.2.3 Hidden Markov Modeling for Multiple Alignment Generation 278
9.2.4 Conclusion 280
9.3 Comparative Modeling 281
9.3.1 Protein Comparative Modeling 281
9.3.2 Comparative Genomic Modeling 284
9.4 Probabilistic Modeling 287
9.4.1 Bayesian Networks 287
9.4.2 Stochastic Context-Free Grammars 288
9.4.3 Probabilistic Boolean Networks 288
9.5 Molecular Modeling 290
9.5.1 Molecular and Related Visualization Applications 290
9.5.2 Molecular Mechanics 294
9.5.3 Modern Computer Programs for Molecular Modeling 295
References 297

10 Pattern Matching for Motifs 299
10.1 Introduction 299
10.2 Gene Regulation 301
10.2.1 Promoter Organization 302
10.3 Motif Recognition 303
10.4 Motif Detection Strategies 305
10.4.1 Multi-genes, Single Species Approach 306
10.5 Single Gene, Multi-species Approach 307
10.6 Multi-genes, Multi-species Approach 309
10.7 Summary 309
References 310

11 Visualization and Fractal Analysis of Biological Sequences 313
11.1 Introduction 313
11.2 Fractal Analysis 317
11.2.1 What Is a Fractal? 317
11.2.2 Recurrent Iterated Function System Model 319
11.2.3 Moment Method to Estimate the Parameters of the IFS (RIFS) Model 320
11.2.4 Multifractal Analysis 321
11.3 DNA Walk Models 323

11.3.1 One-Dimensional DNA Walk 323
11.3.2 Two-Dimensional DNA Walk 324
11.3.3 Higher-Dimensional DNA Walk 325
11.4 Chaos Game Representation of Biological Sequences 325
11.4.1 Chaos Game Representation of DNA Sequences 325
11.4.2 Chaos Game Representation of Protein Sequences 326
11.4.3 Chaos Game Representation of Protein Structures 326
11.4.4 Chaos Game Representation of Amino Acid Sequences Based on the Detailed HP Model 327
11.5 Two-Dimensional Portrait Representation of DNA Sequences . 330
11.5.1 Graphical Representation of Counters 330
11.5.2 Fractal Dimension of the Fractal Set for a Given Tag 332
11.6 One-Dimensional Measure Representation of Biological Sequences 335
11.6.1 Measure Representation of Complete Genomes 335
11.6.2 Measure Representation of Linked Protein Sequences 340
11.6.3 Measure Representation of Protein Sequences Based on Detailed HP Model 344
References 348

12 Microarray Data Analysis 353
12.1 Introduction 353
12.2 Microarray Technology for Genome Expression Study 354
12.3 Image Analysis for Data Extraction 356
12.3.1 Image Preprocessing 357
12.3.2 Block Segmentation 359
12.3.3 Automatic Gridding 360
12.3.4 Spot Extraction 360
12.3.5 Background Correction, Data Normalization and Filtering, and Missing Value Estimation 361
12.4 Data Analysis for Pattern Discovery 363
12.4.1 Cluster Analysis 363
12.4.2 Temporal Expression Profile Analysis and Gene Regulation 371
12.4.3 Gene Regulatory Network Analysis 382
References 384

Index 389