

Contents

Preface

v

1. Bioinformatics and its Applications	1.1
1.1 Definition and Applications	1.1
2. Databases	2.1
2.1 Definition and Types	2.1
2.2 Nucleic Acid Databases	2.2
2.3 Nucleic Acid Secondary Databases	2.5
2.4 Protein Databases	2.6
2.4.1 Primary Database	2.6
2.4.2 Composite	2.8
2.4.3 Secondary Database	2.9
2.4.4 Tertiary Database	2.11
2.5 Relational Database	2.11
2.6 Object-oriented Database	2.12
2.7 Public/Private Sequence Database	2.12
2.8 Non-redundant Databases	2.12
2.9 Corrupt Databases	2.14
2.10 Improving the Quality of Database	2.15
2.10.1 Methods of Improving Quality	2.15
2.11 Errors in Protein Sequence Databases	2.15
2.12 Cleaning the GenBank	2.15
2.13 Gene Structure Sanity Checks	2.16
2.14 Types of Error in the Dataset	2.17
2.15 Cleaning the Database	2.17

2.16	Utility of Database	2.18
2.17	Storing Images—Tracking Protein Location	2.19
2.18	Formatting the Database	2.19
2.19	Species Specific Databases	2.20
3.	Data Mining	3.1
3.1	Definition	3.1
3.2	DM Methodology (DMM)	3.2
3.3	Data Warehousing	3.3
3.4	Mining Models	3.3
3.5	Data Mining Algorithm (DMA)	3.3
3.6	Training of Models	3.4
3.7	Maintaining the Validity of DMM	3.4
3.8	Pattern Recognition and Discovery	3.5
3.8.1	Steps Involved in Pattern Recognition and Discovery	3.5
3.9	Genome Database Mining	3.5
3.10	Simulation-based Analysis	3.6
4.	Analysis of Nucleotide Sequences	4.1
4.1	Sequence Comparison	4.1
4.2	Analysis of DNA Sequences	4.3
4.3	Sequence Alignment	4.3
4.4	Sequence Alignment Algorithms	4.6
4.5	Features of Eukaryotic Genes	4.6
4.6	Evidenced-based Gene Prediction	4.7
4.7	Detection of ORF (Protein Coding Sequence)/ <i>in silico</i> Gene	4.7
4.7.1	Intrinsic Approach	4.8
4.7.2	Extrinsic Approach	4.9
4.8	Application of EST in Gene Identification	4.10
4.9	Problem in Identification of a New Gene	4.11
4.10	Strategies for Identification of Genes/Protein Coding Regions	4.12
4.11	Computer Methodologies for Identifying Coding Regions	4.13
4.11.1	Gene Search by Signals	4.14
4.11.2	Gene Search by Contents	4.15
4.12	Classification Profile Analysis	4.18

4.13 Dynamic Programming (DP) Algorithm	4.21
4.14 Gene Parser Program	4.24
4.15 Computer Programs for Gene Prediction	4.26
4.16 Gene Recognition Algorithms	4.27
4.17 Methods Related to the Encoded Amino Acid Sequence	4.31
4.18 Other Global Patterns	4.33
4.19 Evaluation of the Measure	4.34
4.20 Evolution of Genes	4.35
4.21 Networks and Genetic Algorithms	4.36
4.21.1 Artificial Neural Network	4.36
4.21.2 Approach to Intron/exon Discrimination	4.40
4.21.3 In-frame Dicodon Statistics	4.41
4.22 Sensor Algorithms	4.42
4.23 Identification of New Genes	4.44
4.24 Prediction of Expressed ORFs using GeneMark	4.44
4.25 Failures to Identify Genes	4.47
4.26 Finding Noncoding RNA Genes	4.48
4.27 Prediction of miRNA	4.48
4.28 Identification of Binding Site based on Sequence Alignment	4.50
4.29 Methods of Detection of Motif	4.55
4.29.1 Motif Discovery Problems	4.55
4.29.2 Multi-step Approach to Finding Motifs	4.56
4.30 Pattern Analysis	4.57
4.31 Prediction of Promoters and Regulatory Elements	4.61
4.32 Chromatin Domain and Prediction of MAR Sequences	4.63
4.33 Identification of tRNA Genes	4.63
4.33.1 Detection and Prediction of Potential Nuclear tRNA-coding Genes on Genomic DNA	4.64
4.33.2 Defining a Pattern Containing Several Motifs	4.64
4.33.3 Detection of Identities in tRNA Genes	4.64
4.34 Prediction of RNA Structure	4.64
4.34.1 Methods of Prediction	4.65
4.34.2 Prediction of the Structure in mRNA	4.65
4.34.3 Detecting 'Pseudoknots' and Other Base Pairing Similarities in RNA Sequences	4.66

4.35 Catalytic RNA	4.66
4.36 <i>Ab initio</i> Prediction	4.67
4.37 Linguistic Gene Recognition—The Use of Gene Grammar	4.70
4.38 Determining Gene Function	4.72
4.39 Genome Annotation	4.72
4.40 Gene Hunting (Mining)	4.73
5. Analysis of Protein Sequences	5.1
5.1 Prediction of Protein Sequence	5.1
5.2 Protein Sequence Comparisons	5.2
5.3 Aligning of Sequences	5.4
5.4 Protein Sequence Analysis	5.6
5.5 Alignment Algorithms	5.11
5.5.1 Needleman-Wunsch Algorithm	5.12
5.6 Pair-wise Sequence Alignment Algorithm	5.16
5.6.1 Dayhoff Mutation Data Matrix	5.18
5.6.2 BLOSUM (Block Substitution Matrix) Series Matrices	5.25
5.6.3 Testing of Reliability of Sequence Alignments	5.27
5.7 Smith-Waterman Local Alignment Algorithm	5.29
5.8 Multiple Sequence Alignment	5.31
5.8.1 Multiple-alignment Algorithm	5.33
5.8.2 Progressive Clustered Alignment	5.33
5.8.3 Alternatives to Progressive Alignment	5.34
5.9 Evaluation of Alignment	5.35
5.10 Profile Alignment	5.36
5.10.1 Hidden Markov Model (HMM) Alignment	5.38
5.11 Prediction of a Protein's Structure or Function from its Sequence	5.42
5.11.1 3-Dimensional Predictive Scheme	5.42
5.11.2 <i>AB Initio</i> Predictors	5.43
5.12 Prediction of Protein Function	5.44
5.13 Prediction of Protein-protein Interaction	5.45
5.14 Errors in Sequence Alignment	5.48
5.15 Sequential Structure Alignment Program (SSAP)	5.49
5.16 Dot Plot	5.50
5.17 Graphic Matrix Display	5.52

5.18 Motifs	5.54
5.18.1 Description of Classes of Motifs	5.56
5.18.2 Search for Patterns of Motifs in Protein or Nucleic Acid	5.56
5.18.3 Recognition of Patterns in Proteins	5.58
5.19 Prediction of Subcellular Localization of Eukaryotic Proteins	5.59
6. Polymorphism and Phylogenetic Inference	6.1
6.1 Polymorphisms and their Applications	6.1
6.2 Phylogenetic Inference	6.2
6.3 Approaches to Deriving Phylogenetic Trees	6.5
6.4 Clustering Methods	6.5
6.5 Characters for Constructing Tree	6.12
6.6 Measurement of Genetic Distance at Nucleic Acid Levels Instead of Proteins	6.13
6.7 Evolutionary Distance	6.13
6.8 Phylogenetic Networks	6.15
6.8.1 Phylogenetic Network Reconstruction and Evaluation	6.16
7. Tools and Software Packages	7.1
7.1 Basic Local Alignment Search Tool (BLAST)	7.4
7.2 BLASTing	7.4
7.3 BLAST Algorithm	7.6
7.4 Variants of BLAST	7.9
7.4.1 PSI-BLAST	7.10
7.4.2 Power BLAST	7.14
7.4.3 Pattern-hit initiated BLAST (PHI-BLAST)	7.14
7.4.4 BLAST2	7.15
7.4.5 Blastall	7.15
7.4.6 BLAT	7.16
7.5 FASTA	7.16
7.5.1 TFASTA	7.19
7.5.2 LFASTA	7.20
7.5.3 FASTA Format	7.21
7.6 HMM- based Searches	7.21
7.7 Definition of Gene Structure	7.22
7.7.1 Output of GENSCAN	7.23
7.7.2 HAMMER	7.23

7.8	Annotation	7.25
7.9	CLUSTAL W	7.29
7.10	CLUSTAL X	7.29
7.11	ALIGN Program	7.30
7.12	Fuzzy Algorithms	7.31
7.13	Visual Mining Tool	7.31
7.13.1	Viewing the Full Genomic Context of the Locus	7.32
7.13.2	Viewing of Interaction Paths	7.32
7.13.3	Viewing Various Maps	7.32
7.14	Utility of Graphs	7.33
7.15	Computer Programs for use in SNP Detection	7.33
<i>References</i>		R.1-R.12
<i>Index</i>		I.1-I.7