

CONTENTS

Dedication	v
Preface	vii
Chapter 1. Molecular Biology for the Practical Bioinformatician	1
1 Introduction	1
2 Our Molecular Selves	3
2.1 Cells, DNAs, and Chromosomes	4
2.2 Genes and Genetic Variations	7
2.2.1 Mutations and Genetic Diseases	9
3 Our Biological Machineries	10
3.1 From Gene to Protein: The Central Dogma	12
3.2 The Genetic Code	14
3.3 Gene Regulation and Expression	15
4 Tools of the Trade	18
4.1 Basic Operations	18
4.1.1 Cutting DNA	18
4.1.2 Copying DNA	19
4.1.3 Separating DNA	22
4.1.4 Matching DNA	23
4.2 Putting It Together	24
4.2.1 Genome Sequencing	24
4.2.2 Gene Expression Profiling	26
5 Conclusion	28
Chapter 2. Strategy and Planning of Bioinformatics Experiments	31
1 Multi-Dimensional Bioinformatics	31
2 Reasoning and Strategy	33
3 Planning	34

Chapter 3. Data Mining Techniques for the Practical Bioinformatician	35
1 Feature Selection Methods	36
1.1 Curse of Dimensionality	36
1.2 Signal-to-Noise Measure	38
1.3 T-Test Statistical Measure	40
1.4 Fisher Criterion Score	40
1.5 Entropy Measure	41
1.6 χ^2 Measure	43
1.7 Information Gain	44
1.8 Information Gain Ratio	44
1.9 Wilcoxon Rank Sum Test	45
1.10 Correlation-Based Feature Selection	46
1.11 Principal Component Analysis	47
1.12 Use of Feature Selection in Bioinformatics	49
2 Classification Methods	50
2.1 Decision Trees	50
2.2 Bayesian Methods	51
2.3 Hidden Markov Models	52
2.4 Artificial Neural Networks	52
2.5 Support Vector Machines	55
2.6 Prediction by Collective Likelihood of Emerging Patterns	57
3 Association Rules Mining Algorithms	59
3.1 Association Rules	59
3.2 The Apriori Algorithm	60
3.3 The Max-Miner Algorithm	61
3.4 Use of Association Rules in Bioinformatics	62
4 Clustering Methods	64
4.1 Factors Affecting Clustering	64
4.2 Partitioning Methods	66
4.3 Hierarchical Methods	66
4.4 Use of Clustering in Bioinformatics	68
5 Remarks	69
Chapter 4. Techniques for Recognition of Translation Initiation Sites	71
1 Translation Initiation Sites	72
2 Data Set and Evaluation	73
3 Recognition by Perceptrons	74
4 Recognition by Artificial Neural Networks	75
5 Recognition by Engineering of Support Vector Machine Kernels	76

6	Recognition by Feature Generation, Feature Selection, and Feature Integration	77
6.1	Feature Generation	78
6.2	Feature Selection	80
6.3	Feature Integration for Decision Making	82
7	Improved Recognition by Feature Generation, Feature Selection, and Feature Integration	82
8	Recognition by Linear Discriminant Function	84
9	Recognition by Ribosome Scanning	86
10	Remarks	88

Chapter 5. How Neural Networks Find Promoters Using Recognition of Micro-Structural Promoter Components

1	Motivation and Background	92
1.1	Problem Framework	93
1.2	Promoter Recognition	94
1.3	ANN-Based Promoter Recognition	95
2	Characteristic Motifs of Eukaryotic Promoters	96
3	Motif-Based Search for Promoters	97
3.1	Evaluation Study by Fickett and Hatzigeorgiou	100
3.2	Enhancers May Contribute to False Recognition	100
4	ANNs and Promoter Components	101
4.1	Description of Promoter Recognition Problem	101
4.2	Representation of Nucleotides for Network Processing	103
5	Structural Decomposition	104
5.1	Parallel Composition of Feature Detectors	104
5.2	First- and Second-Level ANNs	109
5.3	Cascade Composition of Feature Detectors	112
5.4	Structures Based on Multilayer Perceptrons	112
6	Time-Delay Neural Networks	114
6.1	Multistate TDNN	118
6.2	Pruning ANN Connections	118
7	Comments on Performance of ANN-Based Programs for Eukaryotic Promoter Prediction	119

Chapter 6. Neural-Statistical Model of TATA-Box Motifs in Eukaryotes

1	Promoter Recognition via Recognition of TATA-Box	124
2	Position Weight Matrix and Statistical Analysis of the TATA-Box and Its Neighborhood	126

2.1	TATA Motifs as One of the Targets in the Search for Eukaryotic Promoters	126
2.2	Data Sources and Data Sets	127
2.3	Recognition Quality	128
2.4	Statistical Analysis of TATA Motifs	128
2.4.1	PWM	129
2.4.2	Numerical Characterization of Segments Around TATA-Box	130
2.4.3	Position Analysis of TATA Motifs	131
2.4.4	Characteristics of Segments S_1 , S_2 , and S_3	132
2.5	Concluding Remarks	134
3	LVQ ANN for TATA-Box Recognition	135
3.1	Data Preprocessing: Phase 1	135
3.2	Data Preprocessing: Phase 2 — Principal Component Analysis	138
3.3	Learning Vector Quantization ANN	140
3.4	The Structure of an LVQ Classifier	141
3.5	LVQ ANN Training	142
3.6	Initial Values for Network Parameters	143
3.6.1	Genetic Algorithm	144
3.6.2	Searching for Good Initial Weights	144
4	Final Model of TATA-Box Motif	146
4.1	Structure of Final System for TATA-Box Recognition	146
4.2	Training of the ANN Part of the Model	146
4.3	Performance of Complete System	149
4.3.1	Comparison of MLVQ and System with Single LVQ ANN	151
4.4	The Final Test	152
5	Summary	155

Chapter 7. Tuning the Dragon Promoter Finder System for Human Promoter Recognition

1	Promoter Recognition	158
2	Dragon Promoter Finder	159
3	Model	161
4	Tuning Data	161
4.1	Promoter Data	161
4.2	Non-Promoter Data	162

5	Tuning Process	162
6	Discussions and Conclusions	164
Chapter 8. RNA Secondary Structure Prediction		167
1	Introduction to RNA Secondary Structures	168
2	RNA Secondary Structure Determination Experiments	171
3	RNA Structure Prediction Based on Sequence	172
4	Structure Prediction in the Absence of Pseudoknot	173
4.1	Loop Energy	174
4.2	First RNA Secondary Structure Prediction Algorithm	175
4.2.1	$W(j)$	175
4.2.2	$V(i, j)$	176
4.2.3	$VBI(i, j)$	176
4.2.4	$VM(i, j)$	176
4.2.5	Time Analysis	177
4.3	Speeding up Multi-Loops	178
4.3.1	Assumption on Free Energy of Multi-Loop	178
4.3.2	Modified Algorithm for Speeding Up Multi-Loop Computation	178
4.3.3	Time Complexity	179
4.4	Speeding Up Internal Loops	179
4.4.1	Assumption on Free Energy for Internal Loop	179
4.4.2	Detailed Description	180
4.4.3	Time Analysis	181
5	Structure Prediction in the Presence of Pseudoknots	181
6	Akutsu's Algorithm	182
6.1	Definition of Simple Pseudoknot	182
6.2	RNA Secondary Structure Prediction with Simple Pseudoknots	183
6.2.1	$V_L(i, j, k), V_R(i, j, k), V_M(i, j, k)$	185
6.2.2	To Compute Basis	186
6.2.3	Time Analysis	186
7	Approximation Algorithm for Predicting Secondary Structure with General Pseudoknots	187
Chapter 9. Protein Localization Prediction		193
1	Motivation	194
2	Biology of Localization	196
3	Experimental Techniques for Determining Localization Sites	198
3.1	Traditional Methods	198

3.1.1	Immunofluorescence Microscopy	198
3.1.2	Gradient Centrifugation	199
3.2	Large-Scale Experiments	199
3.2.1	Immunofluorescent Microscopy	199
3.2.2	Green Fluorescent Protein	200
3.2.3	Comments on Large-Scale Experiments	200
4	Issues and Complications	201
4.1	How Many Sites are There and What are They?	201
4.2	Is One Site Per Protein an Adequate Model?	202
4.3	How Good are the Predictions?	203
4.3.1	Which Method Should I Use?	204
4.4	A Caveat Regarding Estimated Prediction Accuracies	204
4.5	Correlation and Causality	206
5	Localization and Machine Learning	209
5.1	Standard Classifiers Using (Generalized) Amino Acid Content	210
5.2	Localization Process Modeling Approach	212
5.3	Sequence Based Machine Learning Approaches with Architectures Designed to Reflect Localization Signals	212
5.4	Nearest Neighbor Algorithms	213
5.5	Feature Discovery	213
5.6	Extraction of Localization Information from the Literature and Experimental Data	214
6	Conclusion	215
Chapter 10. Homology Search Methods		217
1	Overview	218
2	Edit Distance and Alignments	219
2.1	Edit Distance	219
2.2	Optimal Alignments	220
2.3	More Complicated Objectives	221
2.3.1	Score Matrices	221
2.3.2	Gap Penalties	222
3	Sequence Alignment: Dynamic Programming	222
3.1	Dynamic Programming Algorithm for Sequence Alignment	222
3.1.1	Reducing Memory Needs	224
3.2	Local Alignment	225
3.3	Sequence Alignment with Gap Open Penalty	227
4	Probabilistic Approaches to Sequence Alignment	228
4.1	Scoring Matrices	228

4.1.1	PAM Matrices	229
4.1.2	BLOSUM Matrices	230
4.1.3	Weaknesses of this Approach	231
4.2	Probabilistic Alignment Significance	231
5	Second Generation Homology Search: Heuristics	232
5.1	FASTA and BLAST	232
5.2	Large-Scale Global Alignment	233
6	Next-Generation Homology Search Software	234
6.1	Improved Alignment Seeds	234
6.2	Optimized Spaced Seeds and Why They Are Better	236
6.3	Computing Optimal Spaced Seeds	238
6.4	Computing More Realistic Spaced Seeds	239
6.4.1	Optimal Seeds for Coding Regions	239
6.4.2	Optimal Seeds for Variable-Length Regions	240
6.5	Approaching Smith-Waterman Sensitivity Using Multiple Seed Models	240
6.6	Complexity of Computing Spaced Seeds	241
7	Experiments	242

Chapter 11. Analysis of Phylogeny: A Case Study on Saururaceae 245

1	The What, Why, and How of Phylogeny	246
1.1	What is Phylogeny?	246
1.2	Why Study Phylogeny?	246
1.3	How to Study Phylogeny	247
2	Case Study on Phylogeny of Saururaceae	248
3	Materials and Methods	249
3.1	Plant Materials	249
3.2	DNA Extraction, PCR, and Sequencing	249
3.3	Alignment of Sequences	250
3.4	Parsimony Analysis of Separate DNA Sequences	250
3.5	Parsimony Analysis of Combined DNA Sequences	250
3.6	Parsimony Analysis of Morphological Data	250
3.7	Analysis of Each Morphological Characters	251
4	Results	251
4.1	Phylogeny of Saururaceae from 18S Nuclear Genes	251
4.2	Phylogeny of Saururaceae from <i>trnL-F</i> Chloroplast DNA Sequences	253
4.3	Phylogeny of Saururaceae from <i>matR</i> Mitochondrial Genes	253
4.4	Phylogeny of Saururaceae from Combined DNA Sequences	254
4.5	Phylogeny of Saururaceae from Morphological Data	254

5	Discussion	256
5.1	Phylogeny of Saururaceae	256
5.2	The Differences Among Topologies from 18S, <i>trnL-F</i> , and <i>matR</i>	258
5.3	Analysis of Important Morphological Characters	259
6	Suggestions	260
6.1	Sampling	260
6.2	Selecting Out-Group	261
6.3	Gaining Sequences	261
6.4	Aligning	261
6.5	Analyzing	262
6.6	Dealing with Morphological Data	263
6.7	Comparing Phylogenies Separately from Molecular Data and Morphological Data	263
6.8	Doing Experiments	264
 Chapter 12. Functional Annotation and Protein Families: From Theory to Practice		
1	Introduction	270
2	ProtoNet — Tracing Protein Families	272
2.1	The Concept	272
2.2	The Method and Principle	273
2.3	In Practice	274
3	ProtoNet-Based Tools for Structural Genomics	282
4	PANDORA — Integration of Annotations	282
4.1	The Concept	282
4.2	The Method and Principle	283
4.3	In Practice	286
5	PANDORA-Based Tools for Functional Genomics	290
 Chapter 13. Discovering Protein-Protein Interactions		
1	Introduction	294
2	Experimental Detection of Protein Interactions	294
2.1	Traditional Experimental Methods	295
2.1.1	Co-Immunoprecipitation	295
2.1.2	Synthetic Lethal Screening	296
2.2	High Throughput Experimental Methods	296
2.2.1	Yeast Two-Hybrid	297
2.2.2	Phage Display	299
2.2.3	Affinity Purification and Mass Spectrometry	301

2.2.4	Protein Microarrays	303
3	Computational Prediction Protein Interaction	304
3.1	Structure-Based Predictions	305
3.1.1	Structural Homology	306
3.2	Sequence-Based Predictions	307
3.2.1	Interacting Orthologs	307
3.2.2	Interacting Domain Pairs	308
3.3	Genome-Based Predictions	308
3.3.1	Gene Locality Context: Gene Neighborhood	309
3.3.2	Gene Locality Context: Gene Fusion	310
3.3.3	Phylogenetic Context: Phylogenetic Profiles	312
3.3.4	Phylogenetic Context: Phylogenetic Tree Similarity	313
3.3.5	Gene Expression: Correlated mRNA Expression	316
4	Conclusion	317

Chapter 14. Techniques for Analysis of Gene Expression 319

1	Microarray and Gene Expression	320
2	Diagnosis by Gene Expression	322
2.1	The Two-Step Approach	323
2.2	Shrunk Centroid Approach	325
3	Co-Regulation of Gene Expression	328
3.1	Hierarchical Clustering Approach	328
3.2	Fuzzy K-Means Approach	330
4	Inference of Gene Networks	332
4.1	Classification-Based Approach	333
4.2	Association Rules Approach	335
4.3	Interaction Generality Approach	336
5	Derivation of Treatment Plan	339

Chapter 15. Genome-Wide cDNA Oligo Probe Design and its

	Applications in <i>Schizosaccharomyces Pombe</i>	347
1	Biological Background	348
2	Problem Formulation	349
3	Algorithm Overview	350
4	Implementation Details	351
4.1	Data Schema	351
4.2	Data Objects Creation	352
4.3	Criteria for Probe Production	354
4.4	Probe Production	355