

---

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	The Genesis of Bioinformatics	1
1.2	Bioinformatics Versus Other Disciplines	2
1.3	Further Developments: from Linear Information to Multidimensional Structure Organization.	4
1.4	Mathematical and Computational Methods	5
1.4.1	Why Mathematical Modeling?	6
1.4.2	Fitting Models to Data	7
1.4.3	Computer Software	7
1.5	Applications	8

---

## Part I Mathematical and Computational Methods

---

<b>2</b>	<b>Probability and Statistics</b>	<b>13</b>
2.1	The Rules of Probability Calculus	13
2.1.1	Independence, Conditional Probabilities and Bayes' Rules	14
2.2	Random Variables	15
2.2.1	Vector Random Variables	16
2.2.2	Marginal Distributions	17
2.2.3	Operations on Random Variables	17
2.2.4	Notation	19
2.2.5	Expectation and Moments of Random Variables	19
2.2.6	Probability-Generating Functions and Characteristic Functions	20
2.3	A Collection of Discrete and Continuous Distributions	22
2.3.1	Bernoulli Trials and the Binomial Distribution	22
2.3.2	The Geometric Distribution	23
2.3.3	The Negative Binomial Distribution	23
2.3.4	The Poisson Distribution	24

2.3.5	The Multinomial Distribution .....	25
2.3.6	The Hypergeometric Distribution .....	25
2.3.7	The Normal (Gaussian) Distribution .....	26
2.3.8	The Exponential Distribution .....	26
2.3.9	The Gamma Distribution .....	27
2.3.10	The Beta Distribution .....	27
2.4	Likelihood maximization .....	28
2.4.1	Binomial Distribution .....	29
2.4.2	Multinomial distribution .....	29
2.4.3	Poisson Distribution .....	29
2.4.4	Geometric Distribution .....	30
2.4.5	Normal Distribution .....	30
2.4.6	Exponential Distribution .....	31
2.5	Other Methods of Estimating Parameters: a Comparison .....	31
2.5.1	Example 1. Uniform Distribution .....	31
2.5.2	Example 2. Cauchy Distribution .....	33
2.5.3	Minimum Variance Parameter Estimation .....	35
2.6	The Expectation Maximization Method .....	37
2.6.1	The Derivations of the Algorithm .....	38
2.6.2	Examples of Recursive Estimation of Parameters by Using the EM Algorithm .....	41
2.7	Statistical Tests .....	45
2.7.1	The Idea .....	45
2.7.2	Parametric Tests .....	47
2.7.3	Nonparametric Tests .....	48
2.7.4	Type I and II statistical errors .....	49
2.8	Markov Chains .....	49
2.8.1	Transition Probability Matrix and State Transition Graph .....	50
2.8.2	Time Evolution of Probability Distributions of States ..	51
2.8.3	Classification of States .....	52
2.8.4	Ergodicity .....	54
2.8.5	Stationary Distribution .....	54
2.8.6	Reversible Markov Chains .....	55
2.8.7	Time-Continuous Markov Chains .....	56
2.9	Markov Chain Monte Carlo (MCMC) Methods .....	57
2.9.1	Acceptance–Rejection Rule .....	59
2.9.2	Applications of the Metropolis–Hastings Algorithm ....	59
2.9.3	Simulated Annealing and MC3 .....	59
2.10	Hidden Markov Models .....	60
2.10.1	Probability of Occurrence of a Sequence of Symbols ....	60
2.10.2	Backward Algorithm. ....	61
2.10.3	Forward Algorithm .....	61
2.10.4	Viterbi Algorithm .....	62
2.10.5	The Baum–Welch algorithm .....	63

2.11 Exercises .....	63
<b>3 Computer Science Algorithms .....</b>	<b>67</b>
3.1 Algorithms .....	67
3.2 Sorting and Quicksort .....	68
3.2.1 Simple Sort .....	69
3.2.2 Quicksort .....	69
3.3 String Searches. Fast Search .....	70
3.3.1 Easy Search .....	71
3.3.2 Fast Search .....	71
3.4 Index Structures for Strings. Search Tries. Suffix Trees .....	73
3.4.1 A Treelike Structure in Computer Memory .....	74
3.4.2 Search Tries .....	75
3.4.3 Compact Search Tries .....	76
3.4.4 Suffix Tries and Suffix Trees .....	77
3.4.5 Suffix Arrays .....	80
3.4.6 Algorithms for Searching Tries .....	80
3.4.7 Building Tries .....	83
3.4.8 Remarks on the Efficiency of the Algorithms .....	85
3.5 The Burrows–Wheeler Transform .....	85
3.5.1 Inverse transform. ....	86
3.5.2 BW Transform as a Compression Tool .....	88
3.5.3 BW Transform as a Search Tool for Patterns .....	89
3.5.4 BW Transform as an Associative, Compressed Memory .....	90
3.5.5 Computational Complexity of BW Transform .....	91
3.6 Hashing .....	91
3.6.1 Hashing functions for addressing variables .....	91
3.6.2 Collisions .....	92
3.6.3 Statistics of Memory Access Time with Hashing .....	93
3.6.4 Inquiring About Repetitive Structure of Sequences, Comparing Sequences and Detecting Sequence Overlap by Hashing .....	94
3.7 Exercises .....	95
<b>4 Pattern Analysis .....</b>	<b>97</b>
4.1 Feature Extraction .....	97
4.2 Classification .....	98
4.2.1 Linear Classifiers .....	98
4.2.2 Linear Classifier Functions and Artificial Neurons .....	100
4.2.3 Artificial Neural Networks .....	100
4.2.4 Support Vector Machines .....	102
4.3 Clustering .....	103
4.3.1 K-means Clustering .....	104
4.3.2 Hierarchical Clustering .....	105
4.4 Dimensionality Reduction, Principal Component Analysis .....	107

4.4.1	Singular-Value Decomposition (SVD) .....	108
4.4.2	Geometric Interpretation of SVD .....	109
4.4.3	Partial-Least-Squares (PLS) Method .....	115
4.5	Parametric Transformations .....	116
4.5.1	Hough Transform .....	117
4.5.2	Generalized Hough Transforms .....	118
4.5.3	Geometric Hashing .....	119
4.6	Exercises .....	119
<b>5</b>	<b>Optimization</b> .....	<b>123</b>
5.1	Static Optimization .....	124
5.1.1	Convexity and Concavity .....	126
5.1.2	Constrained Optimization with Equality Constraints ..	128
5.1.3	Constrained Optimization with Inequality Constraints .	131
5.1.4	Sufficiency of Optimality Conditions for Constrained Problems .....	133
5.1.5	Computing Solutions to Optimization Problems .....	133
5.1.6	Linear Programming .....	136
5.1.7	Quadratic Programming.....	137
5.1.8	Recursive Optimization Algorithms .....	137
5.2	Dynamic Programming.....	140
5.2.1	Dynamic Programming Algorithm for a Discrete-Time System.....	141
5.2.2	Tracing a Path in a Plane .....	143
5.2.3	Shortest Paths in Arrays and Graphs .....	145
5.3	Combinatorial Optimization .....	147
5.3.1	Examples of Combinatorial Optimization Problems ...	148
5.3.2	Time Complexity .....	148
5.3.3	Decision and Optimization Problems .....	149
5.3.4	Classes of Problems and Algorithms .....	149
5.3.5	Suboptimal Algorithms .....	150
5.3.6	Unsolved Problems .....	150
5.4	Exercises .....	151

---

## Part II Applications

---

<b>6</b>	<b>Sequence Alignment</b> .....	<b>155</b>
6.1	Number of Possible Alignments .....	157
6.2	Dot Matrices.....	159
6.3	Scoring Correspondences and Mismatches .....	160
6.4	Developing Scoring Functions .....	162
6.4.1	Estimating Probabilities of Nucleotide Substitution ...	162
6.4.2	Parametric Models of Nucleotide Substitution .....	163
6.4.3	Computing Transition Probabilities .....	165

6.4.4	Fitting Nucleotide Substitution Models to Data . . . . .	168
6.4.5	Breaking the Loop of Dependencies . . . . .	173
6.4.6	Scaling Substitution Probabilities . . . . .	173
6.4.7	Amino Acid Substitution Matrices . . . . .	173
6.4.8	Gaps . . . . .	177
6.5	Sequence Alignment by Dynamic Programming . . . . .	178
6.5.1	The Needleman–Wunsch Alignment Algorithm . . . . .	178
6.5.2	The Smith–Waterman Algorithm . . . . .	181
6.6	Aligning Sequences Against Databases . . . . .	182
6.7	Methods of Multiple Alignment . . . . .	183
6.8	Exercises . . . . .	184
<b>7</b>	<b>Molecular Phylogenetics . . . . .</b>	<b>187</b>
7.1	Trees: Vocabulary and Methods . . . . .	187
7.1.1	The Vocabulary of Trees . . . . .	188
7.2	Overview of Tree-Building Methodologies . . . . .	189
7.3	Distance-Based Trees . . . . .	190
7.3.1	Tree-Derived Distance . . . . .	191
7.3.2	Ultrametric Distances and Molecular-Clock Trees . . . . .	191
7.3.3	Unweighted Pair Group Method with Arithmetic Mean (UPGMA) Algorithm . . . . .	193
7.3.4	Neighbor-Joining Trees . . . . .	193
7.4	Maximum Likelihood (Felsenstein) Trees . . . . .	194
7.4.1	Hypotheses and Steps: . . . . .	196
7.4.2	The Pulley Principle . . . . .	197
7.4.3	Estimating Branch Lengths . . . . .	197
7.4.4	Estimating the Tree Topology . . . . .	198
7.5	Maximum-Parsimony Trees . . . . .	198
7.5.1	Minimal Number of Evolutionary Events for a Given Tree . . . . .	199
7.5.2	Searching for the Optimal Tree Topology . . . . .	199
7.6	Miscellaneous Topics in Phylogenetic Tree Models . . . . .	200
7.6.1	The Nonparametric Bootstrap Method . . . . .	200
7.6.2	Variable Substitution Rates, the Felsenstein-Churchill Algorithm and Related Methods . . . . .	201
7.6.3	The Evolutionary Trace Method and Functional Sites in Proteins . . . . .	201
7.7	Coalescence Theory . . . . .	202
7.7.1	Neutral Evolution: Interaction of Genetic Drift and Mutation . . . . .	202
7.7.2	Modeling Genetic Drift . . . . .	203
7.7.3	Modeling Mutation . . . . .	204
7.7.4	Coalescence Under Different Demographic Scenarios . . . . .	204
7.7.5	Statistical Inference on Demographic Hypotheses and Parameters . . . . .	207

7.7.6	Markov Chain Monte Carlo (MCMC) Methods . . . . .	207
7.7.7	Approximate Approaches . . . . .	208
7.8	Exercises . . . . .	212
<b>8</b>	<b>Genomics</b> . . . . .	<b>213</b>
8.1	The DNA Molecule and the Central Dogma of Molecular Biology . . . . .	214
8.2	Genome Structure . . . . .	220
8.3	Genome Sequencing . . . . .	223
8.3.1	Restriction Enzymes . . . . .	224
8.3.2	Electrophoresis . . . . .	224
8.3.3	Southern Blot . . . . .	224
8.3.4	The Polymerase Chain Reaction . . . . .	225
8.3.5	DNA Cloning . . . . .	226
8.3.6	Chain Termination DNA Sequencing . . . . .	226
8.3.7	Genome Shotgun Sequencing . . . . .	228
8.4	Genome Assembly Algorithms . . . . .	230
8.4.1	Growing Contigs from Fragments . . . . .	230
8.4.2	Detection of Overlaps Between Reads . . . . .	230
8.4.3	Repetitive Structure of DNA . . . . .	232
8.4.4	The Shortest Superstring Problem . . . . .	233
8.4.5	Overlap Graphs and the Hamiltonian Path Problem . . . . .	234
8.4.6	Sequencing by Hybridization . . . . .	235
8.4.7	De Bruijn Graphs . . . . .	238
8.4.8	All $l$ -mers in the Reads . . . . .	238
8.4.9	The Euler Superpath Problem . . . . .	239
8.4.10	Further Aspects of DNA Assembly Algorithms . . . . .	240
8.5	Statistics of the Genome Coverage . . . . .	243
8.5.1	Contigs, Gaps and Anchored Contigs . . . . .	244
8.5.2	Statistics with Minimum Overlaps Between Fragments, Anchored Contigs . . . . .	246
8.5.3	Genome Length and Structure Estimation by Sampling $l$ -mers . . . . .	247
8.5.4	Polymorphisms . . . . .	252
8.6	Genome Annotation . . . . .	252
8.6.1	Research Tools for Genome Annotation . . . . .	254
8.6.2	Gene Identification . . . . .	254
8.6.3	DNA Motifs . . . . .	257
8.6.4	Annotation by Words and Comparisons of Genome Assemblies . . . . .	258
8.6.5	Human Chromosome 14 . . . . .	258
8.7	Exercises . . . . .	259

<b>9</b>	<b>Proteomics</b>	261
9.1	Protein Structure	262
9.1.1	Amino Acids	262
9.1.2	Peptide Bonds	265
9.1.3	Primary Structure	266
9.1.4	Secondary Structure	266
9.1.5	Tertiary Structure	268
9.1.6	Quaternary Structure	271
9.2	Experimental Determination of Amino Acid Sequences and Protein Structures	271
9.2.1	Electrophoresis	272
9.2.2	Protein 2D Gels	272
9.2.3	Protein Western Blots	273
9.2.4	Mass Spectrometry	273
9.2.5	Chemical Identification of Amino Acids in Peptides	274
9.2.6	Analysis of Protein 3D Structure by X Ray Diffraction and NMR	275
9.2.7	Other Assays for Protein Compositions and Interactions	275
9.3	Computational Methods for Modeling Molecular Structures	275
9.3.1	Molecular-Force-Field Model	276
9.3.2	Molecular Dynamics	281
9.3.3	Hydrogen Bonds	281
9.3.4	Computation and Minimization of RMSD	282
9.3.5	Solutions to the Problem of Minimization of RMSD over Rotations	284
9.3.6	Solutions to the Problem of Minimization of RMSD over Rotations and Translations	290
9.3.7	Solvent-Accessible Surface of a Protein	290
9.4	Computational Prediction of Protein Structure and Function	290
9.4.1	Inferring Structures of Proteins	291
9.4.2	Protein Annotation	292
9.4.3	De Novo Methods	292
9.4.4	Comparative Modeling	293
9.4.5	Protein–Ligand Binding Analysis	295
9.4.6	Classification Based on Proteomic Assays	295
9.5	Exercises	296
<b>10</b>	<b>RNA</b>	299
10.1	The RNA World Hypothesis	300
10.2	The Functions of RNA	300
10.3	Reverse Transcription, Sequencing RNA Chains	301
10.4	The Northern Blot	302
10.5	RNA Primary Structure	302
10.6	RNA Secondary Structure	302
10.7	RNA Tertiary Structure	302

10.8	Computational Prediction of RNA Secondary Structure . . . . .	303
10.8.1	Nested Structure . . . . .	304
10.8.2	Maximizing the Number of Pairings Between Bases . . . . .	304
10.8.3	Minimizing the Energy of RNA Secondary Structure . . . . .	306
10.8.4	Pseudoknots . . . . .	310
10.9	Prediction of RNA Structure by Comparative Sequence Analysis . . . . .	311
10.10	Exercises . . . . .	311
<b>11</b>	<b>DNA Microarrays . . . . .</b>	<b>313</b>
11.1	Design of DNA Microarrays . . . . .	315
11.2	Kinetics of the Binding Process . . . . .	318
11.3	Data Preprocessing and Normalization . . . . .	320
11.3.1	Normalization Procedures for Single Microarrays . . . . .	321
11.3.2	Normalization Based on Spiked-in Control RNA . . . . .	323
11.3.3	RMA Normalization Procedure . . . . .	326
11.3.4	Correction of Ratio-Intensity Plots for cDNA . . . . .	328
11.4	Statistics of Gene Expression Profiles . . . . .	328
11.4.1	Modeling Probability Distributions of Gene Expressions . . . . .	331
11.5	Class Prediction and Class Discovery . . . . .	336
11.6	Dimensionality Reduction . . . . .	337
11.6.1	Example of Application of PCA to Microarray Data . . . . .	338
11.7	Class Discovery . . . . .	338
11.7.1	Hierarchical Clustering . . . . .	339
11.8	Class Prediction. Differentially Expressed Genes . . . . .	340
11.9	Multiple Testing, and Analysis of False Discovery Rate (FDR) . . . . .	341
11.9.1	FDR analysis in ALL versus AML gene expression data . . . . .	344
11.10	The Gene Ontology Database . . . . .	344
11.10.1	Structure of GO . . . . .	345
11.10.2	Other Vocabularies of Terms . . . . .	346
11.10.3	Supporting Results of DNA Microarray Analyses with GO and other Vocabulary Terms . . . . .	347
11.11	Exercises . . . . .	347
<b>12</b>	<b>Bioinformatic Databases and Bioinformatic Internet Resources . . . . .</b>	<b>349</b>
12.1	Genomic Databases . . . . .	350
12.2	Proteomic Databases . . . . .	350
12.3	RNA Databases . . . . .	350
12.4	Gene Expression Databases . . . . .	351
12.5	Ontology Databases . . . . .	351
12.6	Databases of Genetic and Proteomic Pathways . . . . .	351
12.7	Programs and Services . . . . .	352
12.8	Clinical Databases . . . . .	352



**References** ..... 355

**Index** ..... 371



<http://www.springer.com/978-3-540-24166-9>

Bioinformatics

Polanski, A.; Kimmel, M.

2007, XVII, 376 p., Hardcover

ISBN: 978-3-540-24166-9