
Contents

Part I Sequence Analysis

1	Introduction: Biological Sequences	3
2	Sequence Alignment	7
2.1	Sequence Similarity	7
2.2	Dynamic Programming: Global Alignment	9
2.3	Dynamic Programming: Local Alignment	10
2.4	Alignment with Affine Gap Model	12
2.5	Heuristic Alignment Algorithms	14
2.5.1	FASTA	14
2.5.2	BLAST	16
2.6	Significance of Scores	16
2.7	Multiple Alignment	16
2.7.1	MSA	19
2.7.2	Progressive Alignment	20
	Exercises	23
3	Markov Chains and Hidden Markov Models	25
3.1	Markov Chains	25
3.2	Hidden Markov Models	33
3.3	The Viterbi Algorithm	37
3.4	The Forward Algorithm	40
3.5	The Backward Algorithm and Posterior Decoding	41
3.6	Parameter Estimation for HMMs	45
3.6.1	Estimation when Paths are Known	46
3.6.2	Estimation when Paths are Unknown	47
3.7	HMMs with Silent States	56
3.8	Profile HMMs	63
3.9	Multiple Sequence Alignment by Profile HMMs	66
	Exercises	68

4 Protein Folding	75
4.1 Levels of Protein Structure	75
4.2 Prediction by Profile HMMs	79
4.3 Threading	79
4.4 Molecular Modeling	82
4.5 Lattice <i>HP</i> -Model	84
Exercises	88
5 Phylogenetic Reconstruction	89
5.1 Phylogenetic Trees	89
5.2 Parsimony Methods	94
5.3 Distance Methods	96
5.4 Evolutionary Models	122
5.4.1 The Jukes-Cantor Model	127
5.4.2 The Kimura Model	128
5.4.3 The Felsenstein Model	129
5.4.4 The Hasegawa-Kishino-Yano (HKY) Model	130
5.5 Maximum Likelihood Method	130
5.6 Model Comparison	137
Exercises	139

Part II Mathematical Background for Sequence Analysis

6 Elements of Probability Theory	145
6.1 Sample Spaces and Events	145
6.2 Probability Measure	151
6.3 Conditional Probability	159
6.4 Random Variables	162
6.5 Integration of Random Variables	163
6.6 Monotone Functions on the Real Line	172
6.7 Distribution Functions	176
6.8 Common Types of Random Variables	181
6.8.1 The Discrete Type	181
6.8.2 The Continuous Type	182
6.9 Common Discrete and Continuous Distributions	184
6.9.1 The Discrete Case	184
6.9.2 The Continuous Case	188
6.10 Vector-Valued Random Variables	191
6.11 Sequences of Random Variables	196
Exercises	205

7	Significance of Sequence Alignment Scores	209
	7.1 The Problem	209
	7.2 Random Walks	211
	7.3 Significance of Scores	220
	Exercises	228
8	Elements of Statistics	231
	8.1 Statistical Modeling	231
	8.2 Parameter Estimation	235
	8.3 Hypothesis Testing	256
	8.4 Significance of Scores for Global Alignments	266
	Exercises	269
9	Substitution Matrices	271
	9.1 The General Form of a Substitution Matrix	271
	9.2 PAM Substitution Matrices	273
	9.3 BLOSUM Substitution Matrices	279
	Exercises	283
	References	285
	Index	289