
Contents

1	Introduction	1
1.1	Reading and Writing	1
1.2	Design and Scope of This Book	3
1.2.1	Sequences in Space	4
1.2.2	Sequences in Time	7

Part I Sequences in Space

2	Optimal Pairwise Alignment	11
2.1	What Is an Alignment?	14
2.2	Biological Interpretation of the Alignment Problem	15
2.3	Scoring Alignments	15
2.4	Amino Acid Substitution Matrices	16
2.4.1	PAM Matrices	18
2.4.2	BLOSUM Matrices	22
2.4.3	Comparison between PAM and BLOSUM	25
2.4.4	Application of Substitution Matrices	27
2.5	The Number of Possible Alignments	27
2.6	Global Alignment	30
2.7	Shotgun Sequencing and Overlap Alignment	33
2.8	Local Alignment	35
2.9	Accommodating Affine Gap Costs	36
2.10	Maximizing vs. Minimizing Scores	38
2.11	Example Application of Global, Local, and Overlap Alignment	39
2.12	Summary	39
2.13	Further Reading	40
2.14	Exercises and Software Demonstrations	40

3	Biological Sequences and the Exact String Matching Problem	43
3.1	Exact vs. Inexact String Matching	43
3.2	Naïve Pattern Matching	44
3.3	String Searching in Linear Time	45
3.4	Trees	46
3.5	Set Matching Using Keyword Trees	48
3.6	Suffix Trees	51
3.7	Suffix Tree Construction	54
3.8	Suffix Arrays	55
3.9	Repetitive Sequences in Genomics—the <i>C</i> -value Paradox	56
3.10	Detection of Repeated and Unique Substrings Using Suffix Trees	57
3.11	Maximal Repeats	59
3.12	Generalized Suffix Tree	59
3.13	Longest Common Substring Problem	60
3.14	<i>k</i> -Mismatches	60
3.15	Summary	62
3.16	Further Reading	62
3.17	Exercises and Software Demonstrations	63
4	Fast Alignment: Genome Comparison and Database Searching	65
4.1	Global Alignment	67
4.2	Local Alignment	69
4.2.1	Global/Local Alignment: <i>k</i> -Error Matching	71
4.2.2	Examples of Database Search Programs	73
4.3	Database Composition	79
4.4	Heuristic vs. Optimal Alignment Methods	79
4.5	Application: Determining Gene Families	79
4.6	Statistics of Local Alignments	81
4.6.1	Maximum Local Alignment Scores	81
4.6.2	Choosing a Substitution Matrix	84
4.7	Bit Scores	85
4.8	Summary	85
4.9	Further Reading	86
4.10	Exercises and Software Demonstrations	86
5	Multiple Sequence Alignment	91
5.1	Scoring Multiple Alignments	94
5.2	Multiple Alignment by Dynamic Programming	94
5.3	Heuristic Multiple Alignment	97
5.4	Summary	98
5.5	Further Reading	99
5.6	Exercises and Study Questions	99

6	Sequence Profiles and Hidden Markov Models	101
6.1	Profile Analysis	101
6.2	Hidden Markov Models	106
6.3	Profile Hidden Markov Models	111
6.4	Summary	113
6.5	Further Reading	114
6.6	Exercises and Software Demonstration	114
7	Gene Prediction	117
7.1	What is a Gene?	117
7.2	Computational Gene Finding	118
7.3	Measuring the Accuracy of Gene Predictions	121
7.4	<i>Ab initio</i> Methods: Searching for Signals and Content	124
7.4.1	Codon Usage	126
7.4.2	Finding Splice Sites with a Sequence Profile	126
7.4.3	Exon Chaining	131
7.5	Comparative Methods	134
7.5.1	General Remarks	134
7.5.2	Comparative Gene Prediction at the <i>Adh</i> Locus	135
7.6	Problems and Perspectives	138
7.7	Summary	139
7.8	Further Reading	139
7.9	Exercises	140

Part II Sequences in Time

8	Phylogeny	143
8.1	Is There a Tree?—Statistical Geometry	145
8.2	Likelihood-Mapping	146
8.3	The Number of Possible Phylogenies	148
8.4	Distance Methods	150
8.4.1	Average Linkage Clustering	152
8.4.2	Neighbor-Joining	155
8.5	Maximum Parsimony	157
8.6	Maximum Likelihood	159
8.7	Searching Through Tree Space	161
8.7.1	Nearest Neighbor Interchange	162
8.7.2	Subtree Pruning and Regrafting	163
8.7.3	Branch and Bound	163
8.8	Bootstrapping Phylogenies	164
8.9	Summary	166
8.10	Further Reading	167
8.11	Exercises and Study Questions	167

9	Sequence Variation and Molecular Evolution	169
9.1	The Record of Past Events	170
9.2	Mutations and Substitutions	171
9.3	The Molecular Clock	172
9.4	Explicit Models of Molecular Evolution	173
9.5	Estimating Evolutionary Rates	175
9.6	Coding Sequences: Synonymous and Non-Synonymous Substitutions	177
9.7	Substitutions in Globin Sequences	180
9.8	Applications of K_a/K_s	182
9.8.1	A Language Gene?	182
9.8.2	Selection in the Human Genome	183
9.9	Summary	184
9.10	Further Reading	185
9.11	Exercises	185
10	Genes in Populations: Forward in Time	187
10.1	Polymorphism and Genetic Diversity	187
10.2	The Neutral Theory	191
10.3	Modeling Evolution Forward in Time	193
10.4	The Neutral Wright-Fisher Model	194
10.4.1	Fixation and Loss of Alleles	195
10.4.2	The Hardy-Weinberg Law	197
10.4.3	Fixation Probability and Time to Fixation	197
10.4.4	Loss of Genetic Diversity	200
10.5	Adding Mutation to the Model	201
10.5.1	Finite Alleles Model	202
10.5.2	Infinite Alleles Model	203
10.5.3	Infinite Sites Model	203
10.6	Mutation Drift Balance	203
10.6.1	The Rate of Fixation	203
10.6.2	Number of Alleles	205
10.6.3	Genetic Diversity	207
10.7	Sampling Alleles from Populations	209
10.7.1	Ewens' Sampling Formula	209
10.7.2	Application	212
10.8	Selection	212
10.9	Summary	214
10.10	Further Reading	215
10.11	Exercises and Software Demonstration	215

11	Genes in Populations: Backward in Time	217
11.1	Individuals' Genealogies vs. Gene Genealogies	217
11.2	Forward vs. Backward in Time	218
11.3	The Coalescent	222
11.4	Coalescent vs. Phylogenetic Trees	224
11.5	The Infinite Sites Model and the Number of SNPs	224
11.6	Mathematical Properties of the Neutral Coalescent	225
11.6.1	Tree Depth, Tree Size and the Number of Segregating Sites	225
11.6.2	Heterozygosity	232
11.6.3	The Distribution of Segregating Sites	233
11.7	Simulation Example	233
11.8	Recombination	233
11.9	Selection	237
11.10	Combining Recombination and Selection	238
11.11	Summary	242
11.12	Further Reading	242
11.13	Software Demonstrations and Exercises	242
12	Testing Evolutionary Hypotheses	245
12.1	Hudson-Kreitman-Aguadé (HKA) Test	245
12.2	Tajima's Test	248
12.3	Fu and Li's Test	251
12.4	McDonald-Kreitman Test	253
12.5	Minimum Number of Recombination Events	253
12.6	Detecting Linkage Disequilibrium	255
12.7	Implementations	257
12.8	Summary	257
12.9	Exercises and Software Demonstration	257
A	bioinform	259
A.1	Alignment	259
A.1.1	Protein Substitution Matrices	259
A.1.2	Number of Alignments	261
A.1.3	Pairwise Alignment	262
A.2	Match	263
A.2.1	String Matching	263
A.2.2	Suffix Tree	264
A.2.3	Repeat Searching	265
A.2.4	Hash Table	265
A.2.5	Dotplot	266
A.3	Probability	266
A.3.1	Hidden Markov Model	267
A.4	Evolution	268
A.4.1	Phylogeny	268
A.4.2	Drift	270

A.4.3	Wright-Fisher	271
A.4.4	Coalescent	273
B	Probability	275
C	Molecular Biology Figures and Tables	279
D	Resources	285
	Answers to Exercises	287
	References	299
	Glossary	313
	Author Index	321
	Subject Index	323