

Contents

Preface	ix
Book Overview	xi
Exercises	xii
Errors	xii
Acknowledgments	xiii
1 Basic Concepts of Molecular Biology	1
1.1 Life	1
1.2 Proteins	2
1.3 Nucleic Acids	5
1.3.1 DNA	5
1.3.2 RNA	8
1.4 The Mechanisms of Molecular Genetics	9
1.4.1 Genes and the Genetic Code	9
1.4.2 Transcription, Translation, and Protein Synthesis	10
1.4.3 Junk DNA and Reading Frames	12
1.4.4 Chromosomes	13
1.4.5 Is the Genome like a Computer Program?	15
1.5 How the Genome Is Studied	15
1.5.1 Maps and Sequences	16
1.5.2 Specific Techniques	17
1.6 The Human Genome Project	21
1.7 Sequence Databases	23
Exercises	30
Bibliographic Notes	30
2 Strings, Graphs, and Algorithms	33
2.1 Strings	33
2.2 Graphs	35
2.3 Algorithms	38
Exercises	43
Bibliographic Notes	45

3	Sequence Comparison and Database Search	47
3.1	Biological Background	47
3.2	Comparing Two Sequences	49
3.2.1	Global Comparison — The Basic Algorithm	49
3.2.2	Local Comparison	55
3.2.3	Semiglobal Comparison	56
3.3	Extensions to the Basic Algorithms	58
3.3.1	Saving Space	58
3.3.2	General Gap Penalty Functions	60
3.3.3	Affine Gap Penalty Functions	64
3.3.4	Comparing Similar Sequences	66
3.4	Comparing Multiple Sequences	69
3.4.1	The SP Measure	70
3.4.2	Star Alignments	76
3.4.3	Tree Alignments	79
3.5	Database Search	80
3.5.1	PAM Matrices	80
3.5.2	BLAST	84
3.5.3	FAST	87
3.6	Other Issues	89
* 3.6.1	Similarity and Distance	89
3.6.2	Parameter Choice in Sequence Comparison	96
3.6.3	String Matching and Exact Sequence Comparison	98
	Summary	100
	Exercises	101
	Bibliographic Notes	103
4	Fragment Assembly of DNA	105
4.1	Biological Background	105
4.1.1	The Ideal Case	106
4.1.2	Complications	107
4.1.3	Alternative Methods for DNA Sequencing	113
4.2	Models	114
4.2.1	Shortest Common Superstring	114
4.2.2	Reconstruction	116
4.2.3	Multicontig	117
* 4.3	Algorithms	119
4.3.1	Representing Overlaps	119
4.3.2	Paths Originating Superstrings	120
4.3.3	Shortest Superstrings as Paths	122
4.3.4	The Greedy Algorithm	124
4.3.5	Acyclic Subgraphs	126
4.4	Heuristics	132
4.4.1	Finding Overlaps	134

Exercises	139
Bibliographic Notes	141
5 Physical Mapping of DNA	143
5.1 Biological Background	143
5.1.1 Restriction Site Mapping	145
5.1.2 Hybridization Mapping	146
5.2 Models	147
5.2.1 Restriction Site Models	147
5.2.2 Interval Graph Models	149
5.2.3 The Consecutive Ones Property	150
5.2.4 Algorithmic Implications	152
5.3 An Algorithm for the CIP Problem	153
5.4 An Approximation for Hybridization Mapping with Errors	160
5.4.1 A Graph Model	160
5.4.2 A Guarantee	162
5.4.3 Computational Practice	164
5.5 Heuristics for Hybridization Mapping	167
5.5.1 Screening Chimeric Clones	167
5.5.2 Obtaining a Good Probe Ordering	168
Summary	169
Exercises	170
Bibliographic Notes	172
6 Phylogenetic Trees	175
6.1 Character States and the Perfect Phylogeny Problem	177
6.2 Binary Character States	182
6.3 Two Characters	186
6.4 Parsimony and Compatibility in Phylogenies	190
6.5 Algorithms for Distance Matrices	192
6.5.1 Reconstructing Additive Trees	193
* 6.5.2 Reconstructing Ultrametric Trees	196
6.6 Agreement Between Phylogenies	204
Summary	209
Exercises	209
Bibliographic Notes	211
7 Genome Rearrangements	215
7.1 Biological Background	215
7.2 Oriented Blocks	217
7.2.1 Definitions	219
7.2.2 Breakpoints	221
7.2.3 The Diagram of Reality and Desire	222
7.2.4 Interleaving Graph	228
7.2.5 Bad Components	231
7.2.6 Algorithm	234
7.3 Unoriented Blocks	236

7.3.1 Strips	238
7.3.2 Algorithm	241
Summary	242
Exercises	243
Bibliographic Notes	244
8 Molecular Structure Prediction	245
8.1 RNA Secondary Structure Prediction	246
8.2 The Protein Folding Problem	252
8.3 Protein Threading	254
Summary	259
Exercises	259
Bibliographic Notes	260
9 Epilogue: Computing with DNA	261
9.1 The Hamiltonian Path Problem	261
9.2 Satisfiability	264
9.3 Problems and Promises	267
Exercises	268
Bibliographic Notes and Further Sources	268
Answers to Selected Exercises	271
References	277
Index	289