

Contents

Foreword	vii
Preface	ix
About the Authors	xv
1 Introduction	1
1.1 Biological Motivations	1
1.2 Alignment: A Model for Sequence Comparison	2
1.2.1 Definition	2
1.2.2 Alignment Graph	3
1.3 Scoring Alignment	7
1.4 Computing Sequence Alignment	8
1.4.1 Global Alignment Problem	9
1.4.2 Local Alignment Problem	10
1.5 Multiple Alignment	11
1.6 What Alignments are Meaningful?	12
1.7 Overview of the Book	12
1.8 Bibliographic Notes and Further Reading	13
Part I. Algorithms and Techniques	15
2 Basic Algorithmic Techniques	17
2.1 Algorithms and their Complexity	18
2.2 Greedy Algorithms	19
2.2.1 Huffman Codes	19
2.3 Divide-and-Conquer Strategies	21
2.3.1 Mergesort	22
2.4 Dynamic Programming	23
2.4.1 Fibonacci Numbers	25
2.4.2 The Maximum-Sum Segment Problem	26
2.4.3 Longest Increasing Subsequences	27

2.4.4	Longest Common Subsequences	29
2.5	Bibliographic Notes and Further Reading	32
3	Pairwise Sequence Alignment	35
3.1	Introduction	36
3.2	Dot Matrix	37
3.3	Global Alignment	37
3.4	Local Alignment	42
3.5	Various Scoring Schemes	46
3.5.1	Affine Gap Penalties	46
3.5.2	Constant Gap Penalties	48
3.5.3	Restricted Affine Gap Penalties	48
3.6	Space-Saving Strategies	50
3.7	Other Advanced Topics	54
3.7.1	Constrained Sequence Alignment	54
3.7.2	Similar Sequence Alignment	57
3.7.3	Suboptimal Alignment	58
3.7.4	Robustness Measurement	59
3.8	Bibliographic Notes and Further Reading	60
4	Homology Search Tools	63
4.1	Finding Exact Word Matches	64
4.1.1	Hash Tables	64
4.1.2	Suffix Trees	66
4.1.3	Suffix Arrays	67
4.2	FASTA	68
4.3	BLAST	69
4.3.1	Ungapped BLAST	70
4.3.2	Gapped BLAST	72
4.3.3	PSI-BLAST	73
4.4	BLAT	74
4.5	PatternHunter	75
4.6	Bibliographic Notes and Further Reading	77
5	Multiple Sequence Alignment	81
5.1	Aligning Multiple Sequences	81
5.2	Scoring Multiple Sequence Alignment	82
5.3	An Exact Method for Aligning Three Sequences	84
5.4	Progressive Alignment	85
5.5	Bibliographic Notes and Further Reading	86
Part II.	Theory	89

6	Anatomy of Spaced Seeds	91
6.1	Filtration Technique in Homology Search	92
6.1.1	Spaced Seed	92
6.1.2	Sensitivity and Specificity	92
6.2	Basic Formulas on Hit Probability	93
6.2.1	A Recurrence System for Hit Probability	95
6.2.2	Computing Non-Hit Probability	97
6.2.3	Two Inequalities	98
6.3	Distance between Non-Overlapping Hits	99
6.3.1	A Formula for μ_π	100
6.3.2	An Upper Bound for μ_π	101
6.3.3	Why Do Spaced Seeds Have More Hits?	103
6.4	Asymptotic Analysis of Hit Probability	103
6.4.1	Consecutive Seeds	104
6.4.2	Spaced Seeds	108
6.5	Spaced Seed Selection	110
6.5.1	Selection Methods	110
6.5.2	Good Spaced Seeds	111
6.6	Generalizations of Spaced Seeds	112
6.6.1	Transition Seeds	112
6.6.2	Multiple Spaced Seeds	114
6.6.3	Vector Seed	115
6.7	Bibliographic Notes and Further Reading	115
7	Local Alignment Statistics	119
7.1	Introduction	120
7.2	Ungapped Local Alignment Scores	122
7.2.1	Maximum Segment Scores	123
7.2.2	E-value and P-value Estimation	128
7.2.3	The Number of High-Scoring Segments	130
7.2.4	Karlin-Altschul Sum Statistic	131
7.2.5	Local Ungapped Alignment	132
7.2.6	Edge Effects	133
7.3	Gapped Local Alignment Scores	134
7.3.1	Effects of Gap Penalty	134
7.3.2	Estimation of Statistical Parameters	135
7.3.3	Statistical Parameters for BLOSUM and PAM Matrices	138
7.4	BLAST Database Search	139
7.4.1	Calculation of <i>P</i> -values and Expect Values	140
7.4.2	BLAST Printouts	142
7.5	Bibliographic Notes and Further Reading	146

8	Scoring Matrices	149
8.1	The PAM Scoring Matrices	150
8.2	The BLOSUM Scoring Matrices	158
8.3	General Form of the Scoring Matrices	163
8.4	How to Select A Scoring Matrix?	164
8.5	Compositional Adjustment of Scoring Matrices	166
8.6	DNA Scoring Matrices	168
8.7	Gap Cost in Gapped Alignments	171
8.8	Bibliographic Notes and Further Reading	171
A	Basic Concepts in Molecular Biology	175
A.1	The Nucleic Acids: DNA and RNA	175
A.2	Proteins	176
A.3	Genes	177
A.4	The Genomes	177
B	Elementary Probability Theory	179
B.1	Events and Probabilities	179
B.2	Random Variables	180
B.3	Major Discrete Distributions	181
B.3.1	Bernoulli Distribution	181
B.3.2	Binomial Distribution	182
B.3.3	Geometric and Geometric-Like Distributions	182
B.3.4	The Poisson Distribution	182
B.3.5	Probability Generating Function	183
B.4	Major Continuous Distributions	183
B.4.1	Uniform Distribution	184
B.4.2	Exponential Distribution	184
B.4.3	Normal Distribution	185
B.5	Mean, Variance and Moments	185
B.5.1	The Mean of a Random Variable	185
B.5.2	The Variance of a Random Variable	186
B.5.3	The Moment-Generating Function	187
B.6	Relative Entropy of Probability Distributions	189
B.7	Discrete-time Finite Markov Chains	189
B.7.1	Basic Definitions	189
B.7.2	Markov Chains with No Absorbing States	191
B.7.3	Markov Chains with Absorbing States	192
B.7.4	Random Walks	192
B.7.5	High-Order Markov Chains	193
B.8	Recurrent Events and the Renewal Theorem	193
C	Software Packages for Sequence Alignment	195
	References	197

Contents	xxi
Index	207