

Chapman & Hall/CRC Mathematical and Computational Biology Series

ALGORITHMS IN BIOINFORMATICS

A PRACTICAL INTRODUCTION

WING-KIN SUNG



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an *informa* business

A CHAPMAN & HALL BOOK

Contents

Preface	xv
1 Introduction to Molecular Biology	1
1.1 DNA, RNA, and Protein	1
1.1.1 Proteins	1
1.1.2 DNA	4
1.1.3 RNA	9
1.2 Genome, Chromosome, and Gene	10
1.2.1 Genome	10
1.2.2 Chromosome	10
1.2.3 Gene	11
1.2.4 Complexity of the Organism versus Genome Size	11
1.2.5 Number of Genes versus Genome Size	11
1.3 Replication and Mutation of DNA	12
1.4 Central Dogma (from DNA to Protein)	13
1.4.1 Transcription (Prokaryotes)	13
1.4.2 Transcription (Eukaryotes)	14
1.4.3 Translation	15
1.5 Post-Translation Modification (PTM)	17
1.6 Population Genetics	18
1.7 Basic Biotechnological Tools	18
1.7.1 Restriction Enzymes	19
1.7.2 Sonication	19
1.7.3 Cloning	19
1.7.4 PCR	20
1.7.5 Gel Electrophoresis	22
1.7.6 Hybridization	23
1.7.7 Next Generation DNA Sequencing	24
1.8 Brief History of Bioinformatics	26
1.9 Exercises	27
2 Sequence Similarity	29
2.1 Introduction	29
2.2 Global Alignment Problem	30
2.2.1 Needleman-Wunsch Algorithm	32
2.2.2 Running Time Issue	34
2.2.3 Space Efficiency Issue	35

2.2.4	More on Global Alignment	38
2.3	Local Alignment	39
2.4	Semi-Global Alignment	41
2.5	Gap Penalty	42
2.5.1	General Gap Penalty Model	43
2.5.2	Affine Gap Penalty Model	43
2.5.3	Convex Gap Model	45
2.6	Scoring Function	50
2.6.1	Scoring Function for DNA	50
2.6.2	Scoring Function for Protein	51
2.7	Exercises	53
3	Suffix Tree	57
3.1	Introduction	57
3.2	Suffix Tree	57
3.3	Simple Applications of a Suffix Tree	59
3.3.1	Exact String Matching Problem	59
3.3.2	Longest Repeated Substring Problem	60
3.3.3	Longest Common Substring Problem	60
3.3.4	Longest Common Prefix (LCP)	61
3.3.5	Finding a Palindrome	62
3.3.6	Extracting the Embedded Suffix Tree of a String from the Generalized Suffix Tree	63
3.3.7	Common Substring of 2 or More Strings	64
3.4	Construction of a Suffix Tree	65
3.4.1	Step 1: Construct the Odd Suffix Tree	68
3.4.2	Step 2: Construct the Even Suffix Tree	69
3.4.3	Step 3: Merge the Odd and the Even Suffix Trees	70
3.5	Suffix Array	72
3.5.1	Construction of a Suffix Array	73
3.5.2	Exact String Matching Using a Suffix Array	73
3.6	FM-Index	76
3.6.1	Definition	77
3.6.2	The <i>occ</i> Data Structure	78
3.6.3	Exact String Matching Using the FM-Index	79
3.7	Approximate Searching Problem	81
3.8	Exercises	82
4	Genome Alignment	87
4.1	Introduction	87
4.2	Maximum Unique Match (MUM)	88
4.2.1	How to Find MUMs	89
4.3	MUMmer1: LCS	92
4.3.1	Dynamic Programming Algorithm in $O(n^2)$ Time	93
4.3.2	An $O(n \log n)$ -Time Algorithm	93

4.4	MUMmer2 and MUMmer3	96
4.4.1	Reducing Memory Usage	97
4.4.2	Employing a New Alternative Algorithm for Finding MUMs	97
4.4.3	Clustering Matches	97
4.4.4	Extension of the Definition of MUM	98
4.5	Mutation Sensitive Alignment	99
4.5.1	Concepts and Definitions	99
4.5.2	The Idea of the Heuristic Algorithm	100
4.5.3	Experimental Results	102
4.6	Dot Plot for Visualizing the Alignment	103
4.7	Further Reading	105
4.8	Exercises	105
5	Database Search	109
5.1	Introduction	109
5.1.1	Biological Database	109
5.1.2	Database Searching	109
5.1.3	Types of Algorithms	110
5.2	Smith-Waterman Algorithm	111
5.3	FastA	111
5.3.1	FastP Algorithm	112
5.3.2	FastA Algorithm	113
5.4	BLAST	114
5.4.1	BLAST1	115
5.4.2	BLAST2	116
5.4.3	BLAST1 versus BLAST2	118
5.4.4	BLAST versus FastA	118
5.4.5	Statistics for Local Alignment	119
5.5	Variations of the BLAST Algorithm	120
5.5.1	MegaBLAST	120
5.5.2	BLAT	121
5.5.3	PatternHunter	121
5.5.4	PSI-BLAST (Position-Specific Iterated BLAST)	123
5.6	Q-gram Alignment based on Suffix ARrays (QUASAR)	124
5.6.1	Algorithm	124
5.6.2	Speeding Up and Reducing the Space for QUASAR	127
5.6.3	Time Analysis	127
5.7	Locality-Sensitive Hashing	128
5.8	BWT-SW	130
5.8.1	Aligning Query Sequence to Suffix Tree	130
5.8.2	Meaningful Alignment	133
5.9	Are Existing Database Searching Methods Sensitive Enough?	136
5.10	Exercises	136

6	Multiple Sequence Alignment	139
6.1	Introduction	139
6.2	Formal Definition of the Multiple Sequence Alignment Problem	139
6.3	Methods for Solving the MSA Problem	141
6.4	Dynamic Programming Method	142
6.5	Center Star Method	143
6.6	Progressive Alignment Method	146
6.6.1	ClustalW	147
6.6.2	Profile-Profile Alignment	147
6.6.3	Limitation of Progressive Alignment Construction	149
6.7	Iterative Method	149
6.7.1	MUSCLE	150
6.7.2	Log-Expectation (LE) Score	151
6.8	Further Reading	151
6.9	Exercises	152
7	Phylogeny Reconstruction	155
7.1	Introduction	155
7.1.1	Mitochondrial DNA and Inheritance	155
7.1.2	The Constant Molecular Clock	155
7.1.3	Phylogeny	156
7.1.4	Applications of Phylogeny	157
7.1.5	Phylogenetic Tree Reconstruction	158
7.2	Character-Based Phylogeny Reconstruction Algorithm	159
7.2.1	Maximum Parsimony	159
7.2.2	Compatibility	165
7.2.3	Maximum Likelihood Problem	172
7.3	Distance-Based Phylogeny Reconstruction Algorithm	178
7.3.1	Additive Metric and Ultrametric	179
7.3.2	Unweighted Pair Group Method with Arithmetic Mean (UPGMA)	184
7.3.3	Additive Tree Reconstruction	187
7.3.4	Nearly Additive Tree Reconstruction	189
7.3.5	Can We Apply Distance-Based Methods Given a Character-State Matrix?	190
7.4	Bootstrapping	191
7.5	Can Tree Reconstruction Methods Infer the Correct Tree?	192
7.6	Exercises	193
8	Phylogeny Comparison	199
8.1	Introduction	199
8.2	Similarity Measurement	200
8.2.1	Computing MAST by Dynamic Programming	201
8.2.2	MAST for Unrooted Trees	202

8.3	Dissimilarity Measurements	203
8.3.1	Robinson-Foulds Distance	204
8.3.2	Nearest Neighbor Interchange Distance (NNI)	209
8.3.3	Subtree Transfer Distance (STT)	210
8.3.4	Quartet Distance	211
8.4	Consensus Tree Problem	214
8.4.1	Strict Consensus Tree	215
8.4.2	Majority Rule Consensus Tree	216
8.4.3	Median Consensus Tree	218
8.4.4	Greedy Consensus Tree	218
8.4.5	R^* Tree	219
8.5	Further Reading	220
8.6	Exercises	222
9	Genome Rearrangement	225
9.1	Introduction	225
9.2	Types of Genome Rearrangements	225
9.3	Computational Problems	227
9.4	Sorting an Unsigned Permutation by Reversals	227
9.4.1	Upper and Lower Bound on an Unsigned Reversal Distance	228
9.4.2	4-Approximation Algorithm for Sorting an Unsigned Permutation	229
9.4.3	2-Approximation Algorithm for Sorting an Unsigned Permutation	230
9.5	Sorting a Signed Permutation by Reversals	232
9.5.1	Upper Bound on Signed Reversal Distance	232
9.5.2	Elementary Intervals, Cycles, and Components	233
9.5.3	The Hannenhalli-Pevzner Theorem	238
9.6	Further Reading	243
9.7	Exercises	244
10	Motif Finding	247
10.1	Introduction	247
10.2	Identifying Binding Regions of TFs	248
10.3	Motif Model	250
10.4	The Motif Finding Problem	252
10.5	Scanning for Known Motifs	253
10.6	Statistical Approaches	254
10.6.1	Gibbs Motif Sampler	255
10.6.2	MEME	257
10.7	Combinatorial Approaches	260
10.7.1	Exhaustive Pattern-Driven Algorithm	261
10.7.2	Sample-Driven Approach	262
10.7.3	Suffix Tree-Based Algorithm	263

10.7.4	Graph-Based Method	265
10.8	Scoring Function	266
10.9	Motif Ensemble Methods	267
10.9.1	Approach of MotifVoter	268
10.9.2	Motif Filtering by the Discriminative and Consensus Criteria	268
10.9.3	Sites Extraction and Motif Generation	270
10.10	Can Motif Finders Discover the Correct Motifs?	271
10.11	Motif Finding Utilizing Additional Information	274
10.11.1	Regulatory Element Detection Using Correlation with Expression	274
10.11.2	Discovery of Regulatory Elements by Phylogenetic Footprinting	277
10.12	Exercises	279
11	RNA Secondary Structure Prediction	281
11.1	Introduction	281
11.1.1	Base Interactions in RNA	282
11.1.2	RNA Structures	282
11.2	Obtaining RNA Secondary Structure Experimentally	285
11.3	RNA Structure Prediction Based on Sequence Only	286
11.4	Structure Prediction with the Assumption That There is No Pseudoknot	286
11.5	Nussinov Folding Algorithm	288
11.6	ZUKER Algorithm	290
11.6.1	Time Analysis	292
11.6.2	Speeding up Multi-Loops	292
11.6.3	Speeding up Internal Loops	294
11.7	Structure Prediction with Pseudoknots	296
11.7.1	Definition of a Simple Pseudoknot	296
11.7.2	Akutsu's Algorithm for Predicting an RNA Secondary Structure with Simple Pseudoknots	297
11.8	Exercises	300
12	Peptide Sequencing	305
12.1	Introduction	305
12.2	Obtaining the Mass Spectrum of a Peptide	306
12.3	Modeling the Mass Spectrum of a Fragmented Peptide	310
12.3.1	Amino Acid Residue Mass	310
12.3.2	Fragment Ion Mass	310
12.4	De Novo Peptide Sequencing Using Dynamic Programming	312
12.4.1	Scoring by Considering y-Ions	313
12.4.2	Scoring by Considering y-Ions and b-Ions	314
12.5	De Novo Sequencing Using Graph-Based Approach	317
12.6	Peptide Sequencing via Database Search	319

12.7	Further Reading	320
12.8	Exercises	321
13	Population Genetics	323
13.1	Introduction	323
13.1.1	Locus, Genotype, Allele, and SNP	323
13.1.2	Genotype Frequency and Allele Frequency	324
13.1.3	Haplotype and Phenotype	325
13.1.4	Technologies for Studying the Human Population	325
13.1.5	Bioinformatics Problems	325
13.2	Hardy-Weinberg Equilibrium	326
13.3	Linkage Disequilibrium	327
13.3.1	D and D'	328
13.3.2	r^2	328
13.4	Genotype Phasing	328
13.4.1	Clark's Algorithm	329
13.4.2	Perfect Phylogeny Haplotyping Problem	330
13.4.3	Maximum Likelihood Approach	334
13.4.4	Phase Algorithm	336
13.5	Tag SNP Selection	337
13.5.1	Zhang et al.'s Algorithm	338
13.5.2	IdSelect	339
13.6	Association Study	339
13.6.1	Categorical Data Analysis	340
13.6.2	Relative Risk and Odds Ratio	341
13.6.3	Linear Regression	342
13.6.4	Logistic Regression	343
13.7	Exercises	344
	References	349
	Index	375