
Contents

1	Introduction	1
1.1	Introduction	1
1.2	Biological Sequences	2
1.3	Biological Networks	3
1.4	The Need for Distributed Algorithms	6
1.5	Outline of the Book	7
	Reference	8

Part I Background

2	Introduction to Molecular Biology	11
2.1	Introduction	11
2.2	The Cell	12
2.2.1	DNA	13
2.2.2	RNA	14
2.2.3	Genes	15
2.2.4	Proteins	15
2.3	Central Dogma of Life	16
2.3.1	Transcription	17
2.3.2	The Genetic Code	18
2.3.3	Translation	18
2.3.4	Mutations	19
2.4	Biotechnological Methods	20
2.4.1	Cloning	20
2.4.2	Polymerase Chain Reaction	20
2.4.3	DNA Sequencing	21
2.5	Databases	22
2.5.1	Nucleotide Databases	22
2.5.2	Protein Sequence Databases	22
2.6	Human Genome Project	23
2.7	Chapter Notes	23
	References	24

3	Graphs, Algorithms, and Complexity	27
3.1	Introduction	27
3.2	Graphs	27
3.2.1	Types of Graphs	29
3.2.2	Graph Representations	29
3.2.3	Paths, Cycles, and Connectivity	30
3.2.4	Trees	32
3.2.5	Spectral Properties of Graphs	32
3.3	Algorithms	33
3.3.1	Time and Space Complexities	33
3.3.2	Recurrences	34
3.3.3	Fundamental Approaches	35
3.3.4	Dynamic Programming	35
3.3.5	Graph Algorithms	36
3.3.6	Special Subgraphs	41
3.4	NP-Completeness	43
3.4.1	Reductions	44
3.4.2	Coping with NP-Completeness	45
3.5	Chapter Notes	47
	References	49
4	Parallel and Distributed Computing	51
4.1	Introduction	51
4.2	Architectures for Parallel and Distributed Computing	52
4.2.1	Interconnection Networks	52
4.2.2	Multiprocessors and Multicomputers	53
4.2.3	Flynn's Taxonomy	54
4.3	Parallel Computing	54
4.3.1	Complexity of Parallel Algorithms	55
4.3.2	Parallel Random Access Memory Model	55
4.3.3	Parallel Algorithm Design Methods	57
4.3.4	Shared Memory Programming	59
4.3.5	Multi-threaded Programming	63
4.3.6	Parallel Processing in UNIX	66
4.4	Distributed Computing	68
4.4.1	Distributed Algorithm Design	69
4.4.2	Threads Re-visited	69
4.4.3	Message Passing Interface	70
4.4.4	Distributed Processing in UNIX	73
4.5	Chapter Notes	74
	References	76

Part II Biological Sequences

5	String Algorithms	81
5.1	Introduction	81
5.2	Exact String Matching	82
5.2.1	Sequential Algorithms	82
5.2.2	Distributed String Matching	90
5.3	Approximate String Matching	91
5.4	Longest Subsequence Problems	92
5.4.1	Longest Common Subsequence	92
5.4.2	Longest Increasing Subsequence	95
5.5	Suffix Trees	96
5.5.1	Construction of Suffix Trees	97
5.5.2	Applications of Suffix Trees	102
5.5.3	Suffix Arrays	104
5.6	Chapter Notes	107
	References	109
6	Sequence Alignment	111
6.1	Introduction	111
6.2	Problem Statement	112
6.2.1	The Objective Function	112
6.2.2	Scoring Matrices for Proteins	114
6.3	Pairwise Alignment	115
6.3.1	Global Alignment	115
6.3.2	Local Alignment	118
6.4	Multiple Sequence Alignment	120
6.4.1	Center Star Method	121
6.4.2	Progressive Alignment	122
6.5	Alignment with Suffix Trees	123
6.6	Database Search	124
6.6.1	FASTA	124
6.6.2	BLAST	125
6.7	Parallel and Distributed Sequence Alignment	126
6.7.1	Parallel and Distributed SW Algorithm	126
6.7.2	Distributed BLAST	127
6.7.3	Parallel/Distributed CLUSTALW	129
6.8	Chapter Notes	130
	References	132
7	Clustering of Biological Sequences	135
7.1	Introduction	135
7.2	Analysis	136
7.2.1	Distance and Similarity Measures	136
7.2.2	Validation of Cluster Quality	137

7.3	Classical Methods	138
7.3.1	Hierarchical Algorithms	138
7.3.2	Partitional Algorithms	140
7.3.3	Other Methods	143
7.4	Clustering Algorithms Targeting Biological Sequences.	144
7.4.1	Alignment-Based Clustering	144
7.4.2	Other Similarity-Based Methods	144
7.4.3	Graph-Based Clustering	145
7.5	Distributed Clustering	146
7.5.1	Hierarchical Clustering	146
7.5.2	k -means Clustering	152
7.5.3	Graph-Based Clustering	154
7.5.4	Review of Existing Algorithms	155
7.6	Chapter Notes.	156
	References.	159
8	Sequence Repeats	161
8.1	Introduction	161
8.2	Tandem Repeats	163
8.2.1	Stoye and Gusfield Algorithm.	164
8.2.2	Distributed Tandem Repeat Search	166
8.3	Sequence Motifs	166
8.3.1	Probabilistic Approaches	169
8.3.2	Combinatorial Methods	171
8.3.3	Parallel and Distributed Motif Search	174
8.3.4	A Survey of Recent Distributed Algorithms	178
8.4	Chapter Notes.	179
	References.	181
9	Genome Analysis	183
9.1	Introduction	183
9.2	Gene Finding	184
9.2.1	Fundamental Methods	185
9.2.2	Hidden Markov Models	186
9.2.3	Nature Inspired Methods	187
9.2.4	Distributed Gene Finding.	189
9.3	Genome Rearrangement.	190
9.3.1	Sorting by Reversals	191
9.3.2	Unsigned Reversals.	193
9.3.3	Signed Reversals.	196
9.3.4	Distributed Genome Rearrangement Algorithms	199
9.4	Haplotype Inference	200
9.4.1	Problem Statement	202
9.4.2	Clark's Algorithm	202

9.4.3	EM Algorithm	203
9.4.4	Distributed Haplotype Inference Algorithms	204
9.5	Chapter Notes.	206
	References.	208

Part III Biological Networks

10	Analysis of Biological Networks.	213
10.1	Introduction	213
10.2	Networks in the Cell	214
10.2.1	Metabolic Networks	214
10.2.2	Gene Regulation Networks.	215
10.2.3	Protein Interaction Networks.	216
10.3	Networks Outside the Cell	217
10.3.1	Networks of the Brain	217
10.3.2	Phylogenetic Networks	219
10.3.3	The Food Web	220
10.4	Properties of Biological Networks.	221
10.4.1	Distance.	221
10.4.2	Vertex Degrees.	222
10.4.3	Clustering Coefficient	223
10.4.4	Matching Index.	223
10.5	Centrality.	224
10.5.1	Degree Centrality	224
10.5.2	Closeness Centrality	225
10.5.3	Betweenness Centrality	225
10.5.4	Eigenvalue Centrality	229
10.6	Network Models	230
10.6.1	Random Networks.	230
10.6.2	Small World Networks	231
10.6.3	Scale-Free Networks	232
10.6.4	Hierarchical Networks	233
10.7	Module Detection	234
10.8	Network Motifs	235
10.9	Network Alignment.	235
10.10	Chapter Notes.	236
	References.	239
11	Cluster Discovery in Biological Networks.	241
11.1	Introduction	241
11.2	Analysis.	242
11.2.1	Quality Metrics.	242
11.2.2	Classification of Clustering Algorithms	245

11.3	Hierarchical Clustering	246
11.3.1	MST-Based Clustering.	247
11.3.2	Edge-Betweenness-Based Clustering	250
11.4	Density-Based Clustering.	252
11.4.1	Clique Algorithms.	252
11.4.2	k -core Decomposition	254
11.4.3	Highly Connected Subgraphs Algorithm	258
11.4.4	Modularity-Based Clustering	259
11.5	Flow Simulation-Based Approaches.	263
11.5.1	Markov Clustering Algorithm.	263
11.5.2	Distributed Markov Clustering Algorithm Proposal	265
11.6	Spectral Clustering	267
11.7	Chapter Notes.	269
	References.	272
12	Network Motif Search.	275
12.1	Introduction	275
12.2	Problem Statement	276
12.2.1	Methods of Motif Discovery.	277
12.2.2	Relation to Graph Isomorphism	278
12.2.3	Frequency Concepts	279
12.2.4	Random Graph Generation.	280
12.2.5	Statistical Significance.	280
12.3	A Review of Sequential Motif Searching Algorithms.	281
12.3.1	Network Centric Algorithms.	282
12.3.2	Motif Centric Algorithms.	286
12.4	Distributed Motif Discovery	291
12.4.1	A General Framework	291
12.4.2	Review of Distributed Motif Searching Algorithms	292
12.4.3	Wang et al.'s Algorithm.	292
12.4.4	Schatz et al.'s Algorithm	294
12.4.5	Riberio et al.'s Algorithms	294
12.5	Chapter Notes.	299
	References.	301
13	Network Alignment.	303
13.1	Introduction	303
13.2	Problem Statement	304
13.2.1	Relation to Graph Isomorphism	304
13.2.2	Relation to Bipartite Graph Matching	305
13.2.3	Evaluation of Alignment Quality.	305
13.2.4	Network Alignment Methods	307

13.3	Review of Sequential Network Alignment Algorithms	308
13.3.1	PathBlast	308
13.3.2	IsoRank	309
13.3.3	MaWish	309
13.3.4	GRAAL	310
13.3.5	Recent Algorithms	310
13.4	Distributed Network Alignment	311
13.4.1	A Distributed Greedy Approximation Algorithm Proposal	311
13.4.2	Distributed Hoepman's Algorithm	314
13.4.3	Distributed Auction Algorithms	316
13.5	Chapter Notes	318
	References	320
14	Phylogenetics	323
14.1	Introduction	323
14.2	Terminology	324
14.3	Phylogenetic Trees	325
14.3.1	Distance-Based Algorithms	326
14.3.2	Maximum Parsimony	335
14.3.3	Maximum Likelihood	342
14.4	Phylogenetic Networks	343
14.4.1	Reconstruction Methods	344
14.5	Chapter Notes	345
	References	348
15	Epilogue	351
15.1	Introduction	351
15.2	Current Challenges	352
15.2.1	Big Data Analysis	352
15.2.2	Disease Analysis	353
15.2.3	Bioinformatics Education	355
15.3	Specific Challenges	356
15.3.1	Sequence Analysis	356
15.3.2	Network Analysis	357
15.4	Future Directions	360
15.4.1	Big Data Gets Bigger	360
15.4.2	New Paradigms on Disease Analysis	360
15.4.3	Personalized Medicine	361
	References	362
	Index	363

Distributed and Sequential Algorithms for
Bioinformatics

Erciyes, K.

2015, XVII, 367 p. 157 illus. in color., Hardcover

ISBN: 978-3-319-24964-3