

- 4.4 Repetitive Structures in Degenerate Strings / 79
 - 4.4.1 Using the Masking Technique / 79
 - 4.4.2 Computing the Smallest Cover of the Degenerate String x / 79
 - 4.4.3 Computing Maximal Local Covers of x / 81
 - 4.4.4 Computing All Covers of x / 84
 - 4.4.5 Computing the Seeds of x / 84
- 4.5 Conservative String Covering in Degenerate Strings / 84
 - 4.5.1 Finding Constrained Pattern p in Degenerate String T / 85
 - 4.5.2 Computing λ -Conservative Covers of Degenerate Strings / 86
 - 4.5.3 Computing λ -Conservative Seeds of Degenerate Strings / 87
- 4.6 Conclusion / 88
- References / 89

5 EXACT SEARCH ALGORITHMS FOR BIOLOGICAL SEQUENCES

91

Eric Rivals, Leena Salmela, and Jorma Tarhio

- 5.1 Introduction / 91
- 5.2 Single Pattern Matching Algorithms / 93
 - 5.2.1 Algorithms for DNA Sequences / 94
 - 5.2.2 Algorithms for Amino Acids / 96
- 5.3 Algorithms for Multiple Patterns / 97
 - 5.3.1 Trie-Based Algorithms / 97
 - 5.3.2 Filtering Algorithms / 100
 - 5.3.3 Other Algorithms / 103
- 5.4 Application of Exact Set Pattern Matching for Read Mapping / 103
 - 5.4.1 MPSCAN: An Efficient Exact Set Pattern Matching Tool for DNA/RNA Sequences / 103
 - 5.4.2 Other Solutions for Mapping Reads / 104
 - 5.4.3 Comparison of Mapping Solutions / 105
- 5.5 Conclusions / 107
- References / 108

6 ALGORITHMIC ASPECTS OF ARC-ANNOTATED SEQUENCES 113

Guillaume Blin, Maxime Crochemore, and Stéphane Vialette

- 6.1 Introduction / 113
- 6.2 Preliminaries / 114
 - 6.2.1 Arc-Annotated Sequences / 114
 - 6.2.2 Hierarchy / 114
 - 6.2.3 Refined Hierarchy / 115

ALGORITHMS IN COMPUTATIONAL MOLECULAR BIOLOGY

Techniques, Approaches
and Applications

Edited by

Mourad Elloumi

Unit of Technologies of Information and Communication
and University of Tunis-El Manar, Tunisia

Albert Y. Zomaya

The University of Sydney, Australia



WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

CONTENTS

PREFACE	xxxi
CONTRIBUTORS	xxxiii
I STRINGS PROCESSING AND APPLICATION TO BIOLOGICAL SEQUENCES	1
1 STRING DATA STRUCTURES FOR COMPUTATIONAL MOLECULAR BIOLOGY	3
<i>Christos Makris and Evangelos Theodoridis</i>	
1.1 Introduction / 3	
1.2 Main String Indexing Data Structures / 6	
1.2.1 Suffix Trees / 6	
1.2.2 Suffix Arrays / 8	
1.3 Index Structures for Weighted Strings / 12	
1.4 Index Structures for Indeterminate Strings / 14	
1.5 String Data Structures in Memory Hierarchies / 17	
1.6 Conclusions / 20	
References / 20	
2 EFFICIENT RESTRICTED-CASE ALGORITHMS FOR PROBLEMS IN COMPUTATIONAL BIOLOGY	27
<i>Patricia A. Evans and H. Todd Wareham</i>	
2.1 The Need for Special Cases / 27	
2.2 Assessing Efficient Solvability Options for General Problems and Special Cases / 28	
2.3 String and Sequence Problems / 30	
2.4 Shortest Common Superstring / 31	
2.4.1 Solving the General Problem / 32	
2.4.2 Special Case: SCSt for Short Strings Over Small Alphabets / 34	
2.4.3 Discussion / 35	

2.5	Longest Common Subsequence / 36
2.5.1	Solving the General Problem / 37
2.5.2	Special Case: LCS of Similar Sequences / 39
2.5.3	Special Case: LCS Under Symbol-Occurrence Restrictions / 39
2.5.4	Discussion / 40
2.6	Common Approximate Substring / 41
2.6.1	Solving the General Problem / 42
2.6.2	Special Case: Common Approximate String / 44
2.6.3	Discussion / 45
2.7	Conclusion / 46
	References / 47

3 FINITE AUTOMATA IN PATTERN MATCHING

51

Jan Holub

3.1	Introduction / 51
3.1.1	Preliminaries / 52
3.2	Direct Use of DFA in Stringology / 53
3.2.1	Forward Automata / 53
3.2.2	Degenerate Strings / 56
3.2.3	Indexing Automata / 57
3.2.4	Filtering Automata / 59
3.2.5	Backward Automata / 59
3.2.6	Automata with Fail Function / 60
3.3	NFA Simulation / 60
3.3.1	Basic Simulation Method / 61
3.3.2	Bit Parallelism / 61
3.3.3	Dynamic Programming / 63
3.3.4	Basic Simulation Method with Deterministic State Cache / 66
3.4	Finite Automaton as Model of Computation / 66
3.5	Finite Automata Composition / 67
3.6	Summary / 67
	References / 69

4 NEW DEVELOPMENTS IN PROCESSING OF DEGENERATE SEQUENCES

73

Pavlos Antoniou and Costas S. Iliopoulos

4.1	Introduction / 73
4.1.1	Degenerate Primer Design Problem / 74
4.2	Background / 74
4.3	Basic Definitions / 76

- 4.4 Repetitive Structures in Degenerate Strings / 79
 - 4.4.1 Using the Masking Technique / 79
 - 4.4.2 Computing the Smallest Cover of the Degenerate String x / 79
 - 4.4.3 Computing Maximal Local Covers of x / 81
 - 4.4.4 Computing All Covers of x / 84
 - 4.4.5 Computing the Seeds of x / 84
- 4.5 Conservative String Covering in Degenerate Strings / 84
 - 4.5.1 Finding Constrained Pattern p in Degenerate String T / 85
 - 4.5.2 Computing λ -Conservative Covers of Degenerate Strings / 86
 - 4.5.3 Computing λ -Conservative Seeds of Degenerate Strings / 87
- 4.6 Conclusion / 88
- References / 89

5 EXACT SEARCH ALGORITHMS FOR BIOLOGICAL SEQUENCES

91

Eric Rivals, Leena Salmela, and Jorma Tarhio

- 5.1 Introduction / 91
- 5.2 Single Pattern Matching Algorithms / 93
 - 5.2.1 Algorithms for DNA Sequences / 94
 - 5.2.2 Algorithms for Amino Acids / 96
- 5.3 Algorithms for Multiple Patterns / 97
 - 5.3.1 Trie-Based Algorithms / 97
 - 5.3.2 Filtering Algorithms / 100
 - 5.3.3 Other Algorithms / 103
- 5.4 Application of Exact Set Pattern Matching for Read Mapping / 103
 - 5.4.1 MPSCAN: An Efficient Exact Set Pattern Matching Tool for DNA/RNA Sequences / 103
 - 5.4.2 Other Solutions for Mapping Reads / 104
 - 5.4.3 Comparison of Mapping Solutions / 105
- 5.5 Conclusions / 107
- References / 108

6 ALGORITHMIC ASPECTS OF ARC-ANNOTATED SEQUENCES 113

Guillaume Blin, Maxime Crochemore, and Stéphane Vialette

- 6.1 Introduction / 113
- 6.2 Preliminaries / 114
 - 6.2.1 Arc-Annotated Sequences / 114
 - 6.2.2 Hierarchy / 114
 - 6.2.3 Refined Hierarchy / 115

- 6.2.4 Alignment / 115
- 6.2.5 Edit Operations / 116
- 6.3 Longest Arc-Preserving Common Subsequence / 117
 - 6.3.1 Definition / 117
 - 6.3.2 Classical Complexity / 118
 - 6.3.3 Parameterized Complexity / 119
 - 6.3.4 Approximability / 120
- 6.4 Arc-Preserving Subsequence / 120
 - 6.4.1 Definition / 120
 - 6.4.2 Classical Complexity / 121
 - 6.4.3 Classical Complexity for the Refined Hierarchy / 121
 - 6.4.4 Open Problems / 122
- 6.5 Maximum Arc-Preserving Common Subsequence / 122
 - 6.5.1 Definition / 122
 - 6.5.2 Classical Complexity / 123
 - 6.5.3 Open Problems / 123
- 6.6 Edit Distance / 123
 - 6.6.1 Definition / 123
 - 6.6.2 Classical Complexity / 123
 - 6.6.3 Approximability / 125
 - 6.6.4 Open Problems / 125
- References / 125

7 ALGORITHMIC ISSUES IN DNA BARCODING PROBLEMS **129**

Bhaskar DasGupta, Ming-Yang Kao, and Ion Măndoiu

- 7.1 Introduction / 129
- 7.2 Test Set Problems: A General Framework for Several Barcoding Problems / 130
- 7.3 A Synopsis of Biological Applications of Barcoding / 132
- 7.4 Survey of Algorithmic Techniques on Barcoding / 133
 - 7.4.1 Integer Programming / 134
 - 7.4.2 Lagrangian Relaxation and Simulated Annealing / 134
 - 7.4.3 Provably Asymptotically Optimal Results / 134
- 7.5 Information Content Approach / 135
- 7.6 Set-Covering Approach / 136
 - 7.6.1 Set-Covering Implementation in More Detail / 137
- 7.7 Experimental Results and Software Availability / 139
 - 7.7.1 Randomly Generated Instances / 139
 - 7.7.2 Real Data / 140
 - 7.7.3 Software Availability / 140
- 7.8 Concluding Remarks / 140
- References / 141

8 RECENT ADVANCES IN WEIGHTED DNA SEQUENCES 143*Manolis Christodoulakis and Costas S. Iliopoulos*

- 8.1 Introduction / 143
- 8.2 Preliminaries / 146
 - 8.2.1 Strings / 146
 - 8.2.2 Weighted Sequences / 147
- 8.3 Indexing / 148
 - 8.3.1 Weighted Suffix Tree / 148
 - 8.3.2 Property Suffix Tree / 151
- 8.4 Pattern Matching / 152
 - 8.4.1 Pattern Matching Using the Weighted Suffix Tree / 152
 - 8.4.2 Pattern Matching Using Match Counts / 153
 - 8.4.3 Pattern Matching with Gaps / 154
 - 8.4.4 Pattern Matching with Swaps / 156
- 8.5 Approximate Pattern Matching / 157
 - 8.5.1 Hamming Distance / 157
- 8.6 Repetitions, Covers, and Tandem Repeats / 160
 - 8.6.1 Finding Simple Repetitions with the Weighted Suffix Tree / 161
 - 8.6.2 Fixed-Length Simple Repetitions / 161
 - 8.6.3 Fixed-Length Strict Repetitions / 163
 - 8.6.4 Fixed-Length Tandem Repeats / 163
 - 8.6.5 Identifying Covers / 164
- 8.7 Motif Discovery / 164
 - 8.7.1 Approximate Motifs in a Single Weighted Sequence / 164
 - 8.7.2 Approximate Common Motifs in a Set of Weighted Sequences / 165
- 8.8 Conclusions / 166
- References / 167

9 DNA COMPUTING FOR SUBGRAPH ISOMORPHISM PROBLEM AND RELATED PROBLEMS 171*Sun-Yuan Hsieh, Chao-Wen Huang, and Hsin-Hung Chou*

- 9.1 Introduction / 171
- 9.2 Definitions of Subgraph Isomorphism Problem and Related Problems / 172
- 9.3 DNA Computing Models / 174
 - 9.3.1 The Stickers / 174
 - 9.3.2 The Adleman–Lipton Model / 175
- 9.4 The Sticker-based Solution Space / 175
 - 9.4.1 Using Stickers for Generating the Permutation Set / 176
 - 9.4.2 Using Stickers for Generating the Solution Space / 177

- 9.5 Algorithms for Solving Problems / 179
 - 9.5.1 Solving the Subgraph Isomorphism Problem / 179
 - 9.5.2 Solving the Graph Isomorphism Problem / 183
 - 9.5.3 Solving the Maximum Common Subgraph Problem / 184
- 9.6 Experimental Data / 187
- 9.7 Conclusion / 188
- References / 188

II ANALYSIS OF BIOLOGICAL SEQUENCES

191

10 GRAPHS IN BIOINFORMATICS

193

Elsa Chacko and Shoba Ranganathan

- 10.1 Graph theory—Origin / 193
 - 10.1.1 What is a Graph? / 193
 - 10.1.2 Types of Graphs / 194
 - 10.1.3 Well-Known Graph Problems and Algorithms / 200
- 10.2 Graphs and the Biological World / 207
 - 10.2.1 Alternative Splicing and Graphs / 207
 - 10.2.2 Evolutionary Tree Construction / 208
 - 10.2.3 Tracking the Temporal Variation of Biological Systems / 209
 - 10.2.4 Identifying Protein Domains by Clustering Sequence Alignments / 210
 - 10.2.5 Clustering Gene Expression Data / 211
 - 10.2.6 Protein Structural Domain Decomposition / 212
 - 10.2.7 Optimal Design of Thermally Stable Proteins / 212
 - 10.2.8 The Sequencing by Hybridization (SBH) Problem / 214
 - 10.2.9 Predicting Interactions in Protein Networks by Completing Defective Cliques / 215
- 10.3 Conclusion / 216
- References / 216

11 A FLEXIBLE DATA STORE FOR MANAGING BIOINFORMATICS DATA

221

Bassam A. Alqaralleh, Chen Wang, Bing Bing Zhou, and Albert Y. Zomaya

- 11.1 Introduction / 221
 - 11.1.1 Background / 222
 - 11.1.2 Scalability Challenges / 222
- 11.2 Data Model and System Overview / 223

11.3	Replication and Load Balancing /	227
11.3.1	Replicating an Index Node /	228
11.3.2	Answering Range Queries with Replicas /	229
11.4	Evaluation /	230
11.4.1	Point Query Processing Performance /	230
11.4.2	Range Query Processing Performance /	233
11.4.3	Growth of the Replicas of an Indexing Node /	235
11.5	Related Work /	237
11.6	Summary /	237
	References /	238

12 ALGORITHMS FOR THE ALIGNMENT OF BIOLOGICAL SEQUENCES

241

Ahmed Mokaddem and Mourad Elloumi

12.1	Introduction /	241
12.2	Alignment Algorithms /	242
12.2.1	Pairwise Alignment Algorithms /	242
12.2.2	Multiple Alignment Algorithms /	245
12.3	Score Functions /	251
12.4	Benchmarks /	252
12.5	Conclusion /	255
	Acknowledgments /	255
	References /	255

13 ALGORITHMS FOR LOCAL STRUCTURAL ALIGNMENT AND STRUCTURAL MOTIF IDENTIFICATION

261

Sanguthevar Rajasekaran, Vamsi Kundeti, and Martin Schiller

13.1	Introduction /	261
13.2	Problem Definition of Local Structural Alignment /	262
13.3	Variable-Length Alignment Fragment Pair (VLAFP) Algorithm /	263
13.3.1	Alignment Fragment Pairs /	263
13.3.2	Finding the Optimal Local Alignments Based on the VLAFP Cost Function /	264
13.4	Structural Alignment based on Center of Gravity: SACG /	266
13.4.1	Description of Protein Structure in PDB Format /	266
13.4.2	Related Work /	267
13.4.3	Center-of-Gravity-Based Algorithm /	267
13.4.4	Extending Theorem 13.1 for Atomic Coordinates in Protein Structure /	269
13.4.5	Building VCOST(i,j,q) Function Based on Center of Gravity /	270

13.5	Searching Structural Motifs / 270
13.6	Using SACG Algorithm for Classification of New Protein Structures / 273
13.7	Experimental Results / 273
13.8	Accuracy Results / 273
13.9	Conclusion / 274
	Acknowledgments / 275
	References / 276

14 EVOLUTION OF THE CLUSTAL FAMILY OF MULTIPLE SEQUENCE ALIGNMENT PROGRAMS

277

Mohamed Radhouene Aniba and Julie Thompson

14.1	Introduction / 277
14.2	Clustal-ClustalV / 278
14.2.1	Pairwise Similarity Scores / 279
14.2.2	Guide Tree / 280
14.2.3	Progressive Multiple Alignment / 282
14.2.4	An Efficient Dynamic Programming Algorithm / 282
14.2.5	Profile Alignments / 284
14.3	ClustalW / 284
14.3.1	Optimal Pairwise Alignments / 284
14.3.2	More Accurate Guide Tree / 284
14.3.3	Improved Progressive Alignment / 285
14.4	ClustalX / 289
14.4.1	Alignment Quality Analysis / 290
14.5	ClustalW and ClustalX 2.0 / 292
14.6	DbClustal / 293
14.6.1	Anchored Global Alignment / 294
14.7	Perspectives / 295
	References / 296

15 FILTERS AND SEEDS APPROACHES FOR FAST HOMOLOGY SEARCHES IN LARGE DATASETS

299

Nadia Pisanti, Mathieu Giraud, and Pierre Peterlongo

15.1	Introduction / 299
15.1.1	Homologies and Large Datasets / 299
15.1.2	Filter Preprocessing or Heuristics / 300
15.1.3	Contents / 300
15.2	Methods Framework / 301
15.2.1	Strings and Repeats / 301
15.2.2	Filters—Fundamental Concepts / 301

- 15.3 Lossless filters / 303
 - 15.3.1 History of Lossless Filters / 303
 - 15.3.2 QUASAR and SWIFT—Filtering Repeats with Edit Distance / 304
 - 15.3.3 NIMBUS—Filtering Multiple Repeats with Hamming Distance / 305
 - 15.3.4 TUIUIU—Filtering Multiple Repeats with Edit Distance / 308
- 15.4 Lossy Seed-Based Filters / 309
 - 15.4.1 Seed-Based Heuristics / 310
 - 15.4.2 Advanced Seeds / 311
 - 15.4.3 Latencies and Neighborhood Indexing / 311
 - 15.4.4 Seed-Based Heuristics Implementations / 313
- 15.5 Conclusion / 315
- 15.6 Acknowledgments / 315
- References / 315

16 NOVEL COMBINATORIAL AND INFORMATION-THEORETIC ALIGNMENT-FREE DISTANCES FOR BIOLOGICAL DATA MINING

321

Chiara Epifanio, Alessandra Gabriele, Raffaele Giancarlo, and Marinella Sciortino

- 16.1 Introduction / 321
- 16.2 Information-Theoretic Alignment-Free Methods / 323
 - 16.2.1 Fundamental Information Measures, Statistical Dependency, and Similarity of Sequences / 324
 - 16.2.2 Methods Based on Relative Entropy and Empirical Probability Distributions / 325
 - 16.2.3 A Method Based on Statistical Dependency, via Mutual Information / 329
- 16.3 Combinatorial Alignment-Free Methods / 331
 - 16.3.1 The Average Common Substring Distance / 332
 - 16.3.2 A Method Based on the EBWT Transform / 333
 - 16.3.3 N -Local Decoding / 334
- 16.4 Alignment-Free Compositional Methods / 336
 - 16.4.1 The k -String Composition Approach / 337
 - 16.4.2 Complete Composition Vector / 338
 - 16.4.3 Fast Algorithms to Compute Composition Vectors / 339
- 16.5 Alignment-Free Exact Word Matches Methods / 340
 - 16.5.1 D_2 and its Distributional Regimes / 340
 - 16.5.2 An Extension to Mismatches and the Choice of the Optimal Word Size / 342
 - 16.5.3 The Transformation of D_2 into a Method Assessing the Statistical Significance of Sequence Similarity / 343

16.6	Domains of Biological Application /	344
16.6.1	Phylogeny: Information Theoretic and Combinatorial Methods /	345
16.6.2	Phylogeny: Compositional Methods /	346
16.6.3	CIS Regulatory Modules /	347
16.6.4	DNA Sequence Dependencies /	348
16.7	Datasets and Software for Experimental Algorithmics /	349
16.7.1	Datasets /	350
16.7.2	Software /	353
16.8	Conclusions /	354
	References /	355

17 **IN SILICO METHODS FOR THE ANALYSIS OF METABOLITES AND DRUG MOLECULES** 361

Varun Khanna and Shoba Ranganathan

17.1	Introduction /	361
17.1.1	Chemoinformatics and “Drug-Likeness” /	361
17.2	Molecular Descriptors /	363
17.2.1	One-Dimensional (1-D) Descriptors /	363
17.2.2	Two-Dimensional (2-D) Descriptors /	364
17.2.3	Three-Dimensional (3-D) Descriptors /	366
17.3	Databases /	367
17.3.1	PubChem /	367
17.3.2	Chemical Entities of Biological Interest (ChEBI) /	369
17.3.3	ChemBank /	369
17.3.4	ChemIDplus /	369
17.3.5	ChemDB /	369
17.4	Methods and Data Analysis Algorithms /	370
17.4.1	Simple Count Methods /	370
17.4.2	Enhanced Simple Count Methods, Using Structural Features /	371
17.4.3	ML Methods /	372
17.5	Conclusions /	376
	Acknowledgments /	377
	References /	377

III **MOTIF FINDING AND STRUCTURE PREDICTION** 383

18 **MOTIF FINDING ALGORITHMS IN BIOLOGICAL SEQUENCES** 385

Tarek El Falah, Mourad Elloumi, and Thierry Lecroq

18.1	Introduction /	385
------	----------------	-----

18.2	Preliminaries /	386
18.3	The Planted (l, d) -Motif Problem /	387
18.3.1	Formulation /	387
18.3.2	Algorithms /	387
18.4	The Extended (l, d) -Motif Problem /	391
18.4.1	Formulation /	391
18.4.2	Algorithms /	391
18.5	The Edited Motif Problem /	392
18.5.1	Formulation /	392
18.5.2	Algorithms /	393
18.6	The Simple Motif Problem /	393
18.6.1	Formulation /	393
18.6.2	Algorithms /	394
18.7	Conclusion /	395
	References /	396

19 COMPUTATIONAL CHARACTERIZATION OF REGULATORY REGIONS

397

Enrique Blanco

19.1	The Genome Regulatory Landscape /	397
19.2	Qualitative Models of Regulatory Signals /	400
19.3	Quantitative Models of Regulatory Signals /	401
19.4	Detection of Dependencies in Sequences /	403
19.5	Repositories of Regulatory Information /	405
19.6	Using Predictive Models to Annotate Sequences /	406
19.7	Comparative Genomics Characterization /	408
19.8	Sequence Comparisons /	410
19.9	Combining Motifs and Alignments /	412
19.10	Experimental Validation /	414
19.11	Summary /	417
	References /	417

20 ALGORITHMIC ISSUES IN THE ANALYSIS OF CHIP-SEQ DATA 425

Federico Zambelli and Giulio Pavesi

20.1	Introduction /	425
20.2	Mapping Sequences on the Genome /	429
20.3	Identifying Significantly Enriched Regions /	434
20.3.1	ChIP-Seq Approaches to the Identification of DNA Structure Modifications /	437
20.4	Deriving Actual Transcription Factor Binding Sites /	438

20.5 Conclusions / 444

References / 444

21 APPROACHES AND METHODS FOR OPERON PREDICTION BASED ON MACHINE LEARNING TECHNIQUES

449

*Yan Wang, You Zhou, Chunguang Zhou, Shuqin Wang, Wei Du, Chen Zhang,
and Yanchun Liang*

21.1 Introduction / 449

21.2 Datasets, Features, and Preprocesses for Operon Prediction / 451

21.2.1 Operon Datasets / 451

21.2.2 Features / 454

21.2.3 Preprocess Methods / 459

21.3 Machine Learning Prediction Methods for Operon Prediction / 460

21.3.1 Hidden Markov Model / 461

21.3.2 Linkage Clustering / 462

21.3.3 Bayesian Classifier / 464

21.3.4 Bayesian Network / 467

21.3.5 Support Vector Machine / 468

21.3.6 Artificial Neural Network / 470

21.3.7 Genetic Algorithms / 471

21.3.8 Several Combinations / 472

21.4 Conclusions / 474

21.5 Acknowledgments / 475

References / 475

22 PROTEIN FUNCTION PREDICTION WITH DATA-MINING TECHNIQUES

479

Xing-Ming Zhao and Luonan Chen

22.1 Introduction / 479

22.2 Protein Annotation Based on Sequence / 480

22.2.1 Protein Sequence Classification / 480

22.2.2 Protein Subcellular Localization Prediction / 483

22.3 Protein Annotation Based on Protein Structure / 484

22.4 Protein Function Prediction Based on Gene-Expression Data / 485

22.5 Protein Function Prediction Based on Protein Interactome Map / 486

22.5.1 Protein Function Prediction Based on Local Topology
Structure of Interaction Map / 486

22.5.2 Protein Function Prediction Based on Global Topology
of Interaction Map / 488

22.6	Protein Function Prediction Based on Data Integration /	489
22.7	Conclusions and Perspectives /	491
	References /	493

23 PROTEIN DOMAIN BOUNDARY PREDICTION **501**

Paul D. Yoo, Bing Bing Zhou, and Albert Y. Zomaya

23.1	Introduction /	501
23.2	Profiling Technique /	503
23.2.1	Nonlocal Interaction and Vanishing Gradient Problem /	506
23.2.2	Hierarchical Mixture of Experts /	506
23.2.3	Overall Modular Kernel Architecture /	508
23.3	Results /	510
23.4	Discussion /	512
23.4.1	Nonlocal Interactions in Amino Acids /	512
23.4.2	Secondary Structure Information /	513
23.4.3	Hydrophobicity and Profiles /	514
23.4.4	Domain Assignment Is More Accurate for Proteins with Fewer Domains /	514
23.5	Conclusions /	515
	References /	515

24 AN INTRODUCTION TO RNA STRUCTURE AND PSEUDOKNOT PREDICTION **521**

Jana Sperschneider and Amitava Datta

24.1	Introduction /	521
24.2	RNA Secondary Structure Prediction /	522
24.2.1	Minimum Free Energy Model /	524
24.2.2	Prediction of Minimum Free Energy Structure /	526
24.2.3	Partition Function Calculation /	530
24.2.4	Base Pair Probabilities /	533
24.3	RNA Pseudoknots /	534
24.3.1	Biological Relevance /	536
24.3.2	RNA Pseudoknot Prediction /	537
24.3.3	Dynamic Programming /	538
24.3.4	Heuristic Approaches /	541
24.3.5	Pseudoknot Detection /	542
24.3.6	Overview /	542
24.4	Conclusions /	543
	References /	544

IV	PHYLOGENY RECONSTRUCTION	547
25	PHYLOGENETIC SEARCH ALGORITHMS FOR MAXIMUM LIKELIHOOD	549
	<i>Alexandros Stamatakis</i>	
25.1	Introduction / 549	
25.1.1	Phylogenetic Inference / 550	
25.2	Computing the Likelihood / 552	
25.3	Accelerating the PLF by Algorithmic Means / 555	
25.3.1	Reuse of Values Across Probability Vectors / 555	
25.3.2	Gappy Alignments and Pointer Meshes / 557	
25.4	Alignment Shapes / 558	
25.5	General Search Heuristics / 559	
25.5.1	Lazy Evaluation Strategies / 563	
25.5.2	Further Heuristics / 564	
25.5.3	Rapid Bootstrapping / 565	
25.6	Computing the Robinson Foulds Distance / 566	
25.7	Convergence Criteria / 568	
25.7.1	Asymptotic Stopping / 569	
25.8	Future Directions / 572	
	References / 573	
26	HEURISTIC METHODS FOR PHYLOGENETIC RECONSTRUCTION WITH MAXIMUM PARSIMONY	579
	<i>Adrien Goëffon, Jean-Michel Richer, and Jin-Kao Hao</i>	
26.1	Introduction / 579	
26.2	Definitions and Formal Background / 580	
26.2.1	Parsimony and Maximum Parsimony / 580	
26.3	Methods / 581	
26.3.1	Combinatorial Optimization / 581	
26.3.2	Exact Approach / 582	
26.3.3	Local Search Methods / 582	
26.3.4	Evolutionary Metaheuristics and Genetic Algorithms / 588	
26.3.5	Memetic Methods / 590	
26.3.6	Problem-Specific Improvements / 592	
26.4	Conclusion / 594	
	References / 595	

27 MAXIMUM ENTROPY METHOD FOR COMPOSITION VECTOR METHOD

599*Raymond H.-F. Chan, Roger W. Wang, and Jeff C.-F. Wong*

- 27.1 Introduction / 599
- 27.2 Models and Entropy Optimization / 601
 - 27.2.1 Definitions / 601
 - 27.2.2 Denoising Formulas / 603
 - 27.2.3 Distance Measure / 611
 - 27.2.4 Phylogenetic Tree Construction / 613
- 27.3 Application and Discussion / 614
 - 27.3.1 Example 1 / 614
 - 27.3.2 Example 2 / 614
 - 27.3.3 Example 3 / 615
 - 27.3.4 Example 4 / 617
- 27.4 Concluding Remarks / 619
- References / 619

V MICROARRAY DATA ANALYSIS

623

28 MICROARRAY GENE EXPRESSION DATA ANALYSIS

625*Alan Wee-Chung Liew and Xiangchao Gan*

- 28.1 Introduction / 625
- 28.2 DNA Microarray Technology and Experiment / 626
- 28.3 Image Analysis and Expression Data Extraction / 627
 - 28.3.1 Image Preprocessing / 628
 - 28.3.2 Block Segmentation / 628
 - 28.3.3 Automatic Gridding / 628
 - 28.3.4 Spot Extraction / 628
- 28.4 Data Processing / 630
 - 28.4.1 Background Correction / 630
 - 28.4.2 Normalization / 630
 - 28.4.3 Data Filtering / 631
- 28.5 Missing Value Imputation / 631
- 28.6 Temporal Gene Expression Profile Analysis / 634
- 28.7 Cyclic Gene Expression Profiles Detection / 640
 - 28.7.1 SSA-AR Spectral Estimation / 643
 - 28.7.2 Spectral Estimation by Signal Reconstruction / 644
 - 28.7.3 Statistical Hypothesis Testing for Periodic Profile Detection / 646
- 28.8 Summary / 647
- Acknowledgments / 648
- References / 649

29	BICLUSTERING OF MICROARRAY DATA	651
	<i>Wassim Ayadi and Mourad Elloumi</i>	
29.1	Introduction / 651	
29.2	Types of Biclusters / 652	
29.3	Groups of Biclusters / 653	
29.4	Evaluation Functions / 654	
29.5	Systematic and Stochastic Biclustering Algorithms / 656	
29.6	Biological Validation / 659	
29.7	Conclusion / 661	
	References / 661	
30	COMPUTATIONAL MODELS FOR CONDITION-SPECIFIC GENE AND PATHWAY INFERENCE	665
	<i>Yu-Qing Qiu, Shihua Zhang, Xiang-Sun Zhang, and Luonan Chen</i>	
30.1	Introduction / 665	
30.2	Condition-Specific Pathway Identification / 666	
30.2.1	Gene Set Analysis / 667	
30.2.2	Condition-Specific Pathway Inference / 671	
30.3	Disease Gene Prioritization and Genetic Pathway Detection / 681	
30.4	Module Networks / 684	
30.5	Summary / 685	
	Acknowledgments / 685	
	References / 685	
31	HETEROGENEITY OF DIFFERENTIAL EXPRESSION IN CANCER STUDIES: ALGORITHMS AND METHODS	691
	<i>Radha Krishna Murthy Karuturi</i>	
31.1	Introduction / 691	
31.2	Notations / 692	
31.3	Differential Mean of Expression / 694	
31.3.1	Single Factor Differential Expression / 695	
31.3.2	Multifactor Differential Expression / 697	
31.3.3	Empirical Bayes Extension / 698	
31.4	Differential Variability of Expression / 699	
31.4.1	<i>F</i> -Test for Two-Group Differential Variability Analysis / 699	
31.4.2	Bartlett's and Levene's Tests for Multigroup Differential Variability Analysis / 700	
31.5	Differential Expression in Compendium of Tumors / 701	
31.5.1	Gaussian Mixture Model (GMM) for Finite Levels of Expression / 701	
31.5.2	Outlier Detection Strategy / 703	
31.5.3	Kurtosis Excess / 704	

31.6	Differential Expression by Chromosomal Aberrations: The Local Properties / 705
31.6.1	Wavelet Variance Scanning (WAVES) for Single-Sample Analysis / 708
31.6.2	Local Singular Value Decomposition (LSVD) for Compendium of Tumors / 709
31.6.3	Locally Adaptive Statistical Procedure (LAP) for Compendium of Tumors with Control Samples / 710
31.7	Differential Expression in Gene Interactome / 711
31.7.1	Friendly Neighbors Algorithm: A Multiplicative Interactome / 711
31.7.2	GeneRank: A Contributing Interactome / 712
31.7.3	Top Scoring Pairs (TSP): A Differential Interactome / 713
31.8	Differential Coexpression: Global MultiDimensional Interactome / 714
31.8.1	Kostka and Spang's Differential Coexpression Algorithm / 715
31.8.2	Differential Expression Linked Differential Coexpression / 718
31.8.3	Differential Friendly Neighbors (DiffFNs) / 718
	Acknowledgments / 720
	References / 720

VI ANALYSIS OF GENOMES 723

32 COMPARATIVE GENOMICS: ALGORITHMS AND APPLICATIONS 725

Xiao Yang and Srinivas Aluru

32.1	Introduction / 725
32.2	Notations / 727
32.3	Ortholog Assignment / 727
32.3.1	Sequence Similarity-Based Method / 729
32.3.2	Phylogeny-Based Method / 731
32.3.3	Rearrangement-Based Method / 732
32.4	Gene Cluster and Synteny Detection / 734
32.4.1	Synteny Detection / 736
32.4.2	Gene Cluster Detection / 739
32.5	Conclusions / 743
	References / 743

33 ADVANCES IN GENOME REARRANGEMENT ALGORITHMS 749*Masud Hasan and M. Sohel Rahman*

- 33.1 Introduction / 749
- 33.2 Preliminaries / 752
- 33.3 Sorting by Reversals / 753
 - 33.3.1 Approaches to Approximation Algorithms / 754
 - 33.3.2 Signed Permutations / 757
- 33.4 Sorting by Transpositions / 759
 - 33.4.1 Approximation Results / 760
 - 33.4.2 Improved Running Time and Simpler Algorithms / 761
- 33.5 Other Operations / 761
 - 33.5.1 Sorting by Prefix Reversals / 761
 - 33.5.2 Sorting by Prefix Transpositions / 762
 - 33.5.3 Sorting by Block Interchange / 762
 - 33.5.4 Short Swap and Fixed-Length Reversals / 763
- 33.6 Sorting by More Than One Operation / 763
 - 33.6.1 Unified Operation: Doule Cut and Join / 764
- 33.7 Future Research Directions / 765
- 33.8 Notes on Software / 766
- References / 767

34 COMPUTING GENOMIC DISTANCES: AN ALGORITHMIC VIEWPOINT 773*Guillaume Fertin and Irena Rusu*

- 34.1 Introduction / 773
 - 34.1.1 What this Chapter is About / 773
 - 34.1.2 Definitions and Notations / 774
 - 34.1.3 Organization of the Chapter / 775
- 34.2 Interval-Based Criteria / 775
 - 34.2.1 Brief Introduction / 775
 - 34.2.2 The Context and the Problems / 776
 - 34.2.3 Common Intervals in Permutations and the Commuting Generators Strategy / 778
 - 34.2.4 Conserved Intervals in Permutations and the Bound-and-Drop Strategy / 782
 - 34.2.5 Common Intervals in Strings and the Element Plotting Strategy / 783
 - 34.2.6 Variants / 785
- 34.3 Character-Based Criteria / 785
 - 34.3.1 Introduction and Definition of the Problems / 785
 - 34.3.2 An Approximation Algorithm for BAL-FMB / 787

34.3.3	An Exact Algorithm for UNBAL-FMB. /	791
34.3.4	Other Results and Open Problems /	795
34.4	Conclusion /	795
	References /	796

35 WAVELET ALGORITHMS FOR DNA ANALYSIS

799

Carlo Cattani

35.1	Introduction /	799
35.2	DNA Representation /	802
35.2.1	Preliminary Remarks on DNA /	802
35.2.2	Indicator Function /	803
35.2.3	Representation /	806
35.2.4	Representation Models /	807
35.2.5	Constraints on the Representation in $\mathbb{R}^2$ /	808
35.2.6	Complex Representation /	810
35.2.7	DNA Walks /	810
35.3	Statistical Correlations in DNA /	812
35.3.1	Long-Range Correlation /	812
35.3.2	Power Spectrum /	814
35.3.3	Complexity /	817
35.4	Wavelet Analysis /	818
35.4.1	Haar Wavelet Basis /	819
35.4.2	Haar Series /	819
35.4.3	Discrete Haar Wavelet Transform /	821
35.5	Haar Wavelet Coefficients and Statistical Parameters /	823
35.6	Algorithm of the Short Haar Discrete Wavelet Transform /	826
35.7	Clusters of Wavelet Coefficients /	828
35.7.1	Cluster Analysis of the Wavelet Coefficients of the Complex DNA Representation /	830
35.7.2	Cluster Analysis of the Wavelet Coefficients of DNA Walks /	834
35.8	Conclusion /	838
	References /	839

36 HAPLOTYPE INFERENCE MODELS AND ALGORITHMS

843

Ling-Yun Wu

36.1	Introduction /	843
36.2	Problem Statement and Notations /	844
36.3	Combinatorial Methods /	846
36.3.1	Clark's Inference Rule /	846

36.3.2	Pure Parsimony Model /	848
36.3.3	Phylogeny Methods /	849
36.4	Statistical Methods /	851
36.4.1	Maximum Likelihood Methods /	851
36.4.2	Bayesian Methods /	852
36.4.3	Markov Chain Methods /	852
36.5	Pedigree Methods /	853
36.5.1	Minimum Recombinant Haplotype Configurations /	854
36.5.2	Zero Recombinant Haplotype Configurations /	854
36.5.3	Statistical Methods /	855
36.6	Evaluation /	856
36.6.1	Evaluation Measurements /	856
36.6.2	Comparisons /	857
36.6.3	Datasets /	857
36.7	Discussion /	858
	References /	859

VII ANALYSIS OF BIOLOGICAL NETWORKS 865

37 UNTANGLING BIOLOGICAL NETWORKS USING BIOINFORMATICS 867

Gaurav Kumar, Adrian P. Cootes, and Shoba Ranganathan

37.1	Introduction /	867
37.1.1	Predicting Biological Processes: A Major Challenge to Understanding Biology /	867
37.1.2	Historical Perspective and Mathematical Preliminaries of Networks /	868
37.1.3	Structural Properties of Biological Networks /	870
37.1.4	Local Topology of Biological Networks: Functional Motifs, Modules, and Communities /	873
37.2	Types of Biological Networks /	878
37.2.1	Protein-Protein Interaction Networks /	878
37.2.2	Metabolic Networks /	879
37.2.3	Transcriptional Networks /	881
37.2.4	Other Biological Networks /	883
37.3	Network Dynamic, Evolution and Disease /	884
37.3.1	Biological Network Dynamic and Evolution /	884
37.3.2	Biological Networks and Disease /	886
37.4	Future Challenges and Scope /	887
	Acknowledgments /	887
	References /	888

38 PROBABILISTIC APPROACHES FOR INVESTIGATING BIOLOGICAL NETWORKS **893**

J  r  mie Bourdon and Damien Eveillard

- 38.1 Probabilistic Models for Biological Networks / 894
 - 38.1.1 Boolean Networks / 895
 - 38.1.2 Probabilistic Boolean Networks: A Natural Extension / 900
 - 38.1.3 Inferring Probabilistic Models from Experiments / 901
- 38.2 Interpretation and Quantitative Analysis of Probabilistic Models / 902
 - 38.2.1 Dynamical Analysis and Temporal Properties / 902
 - 38.2.2 Impact of Update Strategies for Analyzing Probabilistic Boolean Networks / 905
 - 38.2.3 Simulations of a Probabilistic Boolean Network / 906
- 38.3 Conclusion / 911
- Acknowledgments / 911
- References / 911

39 MODELING AND ANALYSIS OF BIOLOGICAL NETWORKS WITH MODEL CHECKING **915**

Dragan Bo  na  ki, Peter A.J. Hilbers, Ronny S. Mans, and Erik P. de Vink

- 39.1 Introduction / 915
- 39.2 Preliminaries / 916
 - 39.2.1 Model Checking / 916
 - 39.2.2 SPIN and Promela / 917
 - 39.2.3 LTL / 918
- 39.3 Analyzing Genetic Networks with Model Checking / 919
 - 39.3.1 Boolean Regulatory Networks / 919
 - 39.3.2 A Case Study / 919
 - 39.3.3 Translating Boolean Regulatory Graphs into Promela / 921
 - 39.3.4 Some Results / 922
 - 39.3.5 Concluding Remarks / 924
 - 39.3.6 Related Work and Bibliographic Notes / 924
- 39.4 Probabilistic Model Checking for Biological Systems / 925
 - 39.4.1 Motivation and Background / 926
 - 39.4.2 A Kinetic Model of mRNA Translation / 927
 - 39.4.3 Probabilistic Model Checking / 928
 - 39.4.4 The Prism Model / 929
 - 39.4.5 Insertion Errors / 933
 - 39.4.6 Concluding Remarks / 934
 - 39.4.7 Related Work and Bibliographic Notes / 935
- References / 936

40	REVERSE ENGINEERING OF MOLECULAR NETWORKS FROM A COMMON COMBINATORIAL APPROACH	941
	<i>Bhaskar DasGupta, Paola Vera-Licona, and Eduardo Sontag</i>	
40.1	Introduction / 941	
40.2	Reverse-Engineering of Biological Networks / 942	
40.2.1	Evaluation of the Performance of Reverse-Engineering Methods / 945	
40.3	Classical Combinatorial Algorithms: A Case Study / 946	
40.3.1	Benchmarking RE Combinatorial-Based Methods / 947	
40.3.2	Software Availability / 950	
40.4	Concluding Remarks / 951	
	Acknowledgments / 951	
	References / 951	
41	UNSUPERVISED LEARNING FOR GENE REGULATION NETWORK INFERENCE FROM EXPRESSION DATA: A REVIEW	955
	<i>Mohamed Elati and Céline Rouveirol</i>	
41.1	Introduction / 955	
41.2	Gene Networks: Definition and Properties / 956	
41.3	Gene Expression: Data and Analysis / 958	
41.4	Network Inference as an Unsupervised Learning Problem / 959	
41.5	Correlation-Based Methods / 959	
41.6	Probabilistic Graphical Models / 961	
41.7	Constraint-Based Data Mining / 963	
41.7.1	Multiple Usages of Extracted Patterns / 965	
41.7.2	Mining Gene Regulation from Transcriptome Datasets / 966	
41.8	Validation / 969	
41.8.1	Statistical Validation of Network Inference / 970	
41.8.2	Biological Validation / 972	
41.9	Conclusion and Perspectives / 973	
	References / 974	
42	APPROACHES TO CONSTRUCTION AND ANALYSIS OF MICRORNA-MEDIATED NETWORKS	979
	<i>Ilana Lichtenstein, Albert Zomaya, Jennifer Gamble, and Mathew Vadas</i>	
42.1	Introduction / 979	
42.1.1	miRNA-mediated Genetic Regulatory Networks / 979	
42.1.2	The Four Levels of Regulation in GRNs / 981	
42.1.3	Overview of Sections / 982	

42.2	Fundamental Component Interaction Research: Predicting miRNA Genes, Regulators, and Targets /	982
42.2.1	Prediction of Novel miRNA Genes /	983
42.2.2	Prediction of miRNA Targets /	984
42.2.3	Prediction of miRNA Transcript Elements and Transcriptional Regulation /	984
42.3	Identifying miRNA-mediated Networks /	988
42.3.1	Forward Engineering—Construction of Multinode Components in miRNA-mediated Networks Using Paired Interaction Information /	988
42.3.2	Reverse Engineering—Inference of MicroRNA Modules Using Top-Down Approaches /	988
42.4	Global and Local Architecture Analysis in miRNA-Containing Networks /	993
42.4.1	Global Architecture Properties of miRNA-mediated Post-transcriptional Networks /	993
42.4.2	Local Architecture Properties of miRNA-mediated Post-transcriptional Networks /	994
42.5	Conclusion /	1001
	References /	1001