

CONTENTS

PREFACE	xiii
CONTRIBUTORS	xv

SECTION I BIOLOGICAL DATA PREPROCESSING

PART A: BIOLOGICAL DATA MANAGEMENT

1 GENOME AND TRANSCRIPTOME SEQUENCE DATABASES FOR DISCOVERY, STORAGE, AND REPRESENTATION OF ALTERNATIVE SPLICING EVENTS	5
<i>Bahar Taneri and Terry Gaasterland</i>	
2 CLEANING, INTEGRATING, AND WAREHOUSING GENOMIC DATA FROM BIOMEDICAL RESOURCES	35
<i>Fouzia Moussouni and Laure Berti-Équille</i>	
3 CLEANSING OF MASS SPECTROMETRY DATA FOR PROTEIN IDENTIFICATION AND QUANTIFICATION	59
<i>Penghao Wang and Albert Y. Zomaya</i>	
4 FILTERING PROTEIN-PROTEIN INTERACTIONS BY INTEGRATION OF ONTOLOGY DATA	77
<i>Young-Rae Cho</i>	

PART B: BIOLOGICAL DATA MODELING

- 5 COMPLEXITY AND SYMMETRIES IN DNA SEQUENCES** 95
Carlo Cattani
- 6 ONTOLOGY-DRIVEN FORMAL CONCEPTUAL DATA MODELING FOR BIOLOGICAL DATA ANALYSIS** 129
Catharina Maria Keet
- 7 BIOLOGICAL DATA INTEGRATION USING NETWORK MODELS** 155
Gaurav Kumar and Shoba Ranganathan
- 8 NETWORK MODELING OF STATISTICAL EPISTASIS** 175
Ting Hu and Jason H. Moore
- 9 GRAPHICAL MODELS FOR PROTEIN FUNCTION AND STRUCTURE PREDICTION** 191
Mingjie Tang, Kean Ming Tan, Xin Lu Tan, Lee Sael, Meghana Chitale, Juan Esquivel-Rodriguez, and Daisuke Kihara

PART C: BIOLOGICAL FEATURE EXTRACTION

- 10 ALGORITHMS AND DATA STRUCTURES FOR NEXT-GENERATION SEQUENCES** 225
Francesco Vezzi, Giuseppe Lancia, and Alberto Policriti
- 11 ALGORITHMS FOR NEXT-GENERATION SEQUENCING DATA** 251
Costas S. Iliopoulos and Solon P. Pissis
- 12 GENE REGULATORY NETWORK IDENTIFICATION WITH QUALITATIVE PROBABILISTIC NETWORKS** 281
Zina M. Ibrahim, Alioune Ngom, and Ahmed Y. Tawfik

PART D: BIOLOGICAL FEATURE SELECTION

- 13 COMPARING, RANKING, AND FILTERING MOTIFS WITH CHARACTER CLASSES: APPLICATION TO BIOLOGICAL SEQUENCES ANALYSIS** 309
Matteo Comin and Davide Verzotto
- 14 STABILITY OF FEATURE SELECTION ALGORITHMS AND ENSEMBLE FEATURE SELECTION METHODS IN BIOINFORMATICS** 333
Pengyi Yang, Bing B. Zhou, Jean Yee-Hwa Yang, and Albert Y. Zomaya
- 15 STATISTICAL SIGNIFICANCE ASSESSMENT FOR BIOLOGICAL FEATURE SELECTION: METHODS AND ISSUES** 353
Juntao Li, Kwok Pui Choi, Yudi Pawitan, and Radha Krishna Murthy Karuturi

16 SURVEY OF NOVEL FEATURE SELECTION METHODS FOR CANCER CLASSIFICATION	379
<i>Oleg Okun</i>	
17 INFORMATION-THEORETIC GENE SELECTION IN EXPRESSION DATA	399
<i>Patrick E. Meyer and Gianluca Bontempi</i>	
18 FEATURE SELECTION AND CLASSIFICATION FOR GENE EXPRESSION DATA USING EVOLUTIONARY COMPUTATION	421
<i>Haider Banka, Suresh Dara, and Mourad Elloumi</i>	

SECTION II BIOLOGICAL DATA MINING

PART E: REGRESSION ANALYSIS OF BIOLOGICAL DATA

19 BUILDING VALID REGRESSION MODELS FOR BIOLOGICAL DATA USING STATA AND R	445
<i>Charles Lindsey and Simon J. Sheather</i>	
20 LOGISTIC REGRESSION IN GENOMEWIDE ASSOCIATION ANALYSIS	477
<i>Wentian Li and Yaning Yang</i>	
21 SEMIPARAMETRIC REGRESSION METHODS IN LONGITUDINAL DATA: APPLICATIONS TO AIDS CLINICAL TRIAL DATA	501
<i>Yehua Li</i>	

PART F: BIOLOGICAL DATA CLUSTERING

22 THE THREE STEPS OF CLUSTERING IN THE POST-GENOMIC ERA	521
<i>Raffaele Giancarlo, Giosué Lo Bosco, Luca Pinello, and Filippo Utró</i>	
23 CLUSTERING ALGORITHMS OF MICROARRAY DATA	557
<i>Haifa Ben Saber, Mourad Elloumi, and Mohamed Nadif</i>	
24 SPREAD OF EVALUATION MEASURES FOR MICROARRAY CLUSTERING	569
<i>Giulia Bruno and Alessandro Fiori</i>	
25 SURVEY ON BICLUSTERING OF GENE EXPRESSION DATA	591
<i>Adelaide Valente Freitas, Wassim Ayadi, Mourad Elloumi, José Luis Oliveira, and Jin-Kao Hao</i>	

- 26 MULTIOBJECTIVE BICLUSTERING OF GENE EXPRESSION DATA WITH BIOINSPIRED ALGORITHMS** **609**
Khedidja Seridi, Laetitia Jourdan, and El-Ghazali Talbi
- 27 COCLUSTERING UNDER GENE ONTOLOGY DERIVED CONSTRAINTS FOR PATHWAY IDENTIFICATION** **625**
Alessia Visconti, Francesca Cordero, Dino Ienco, and Ruggero G. Pensa

PART G: BIOLOGICAL DATA CLASSIFICATION

- 28 SURVEY ON FINGERPRINT CLASSIFICATION METHODS FOR BIOLOGICAL SEQUENCES** **645**
Bhaskar DasGupta and Lakshmi Kaligounder
- 29 MICROARRAY DATA ANALYSIS: FROM PREPARATION TO CLASSIFICATION** **657**
Luciano Cascione, Alfredo Ferro, Rosalba Giugno, Giuseppe Pigola, and Alfredo Pulvirenti
- 30 DIVERSIFIED CLASSIFIER FUSION TECHNIQUE FOR GENE EXPRESSION DATA** **675**
Sashikala Mishra, Kailash Shaw, and Debahuti Mishra
- 31 RNA CLASSIFICATION AND STRUCTURE PREDICTION: ALGORITHMS AND CASE STUDIES** **685**
Ling Zhong, Junilda Spirollari, Jason T. L. Wang, and Dongrong Wen
- 32 AB INITIO PROTEIN STRUCTURE PREDICTION: METHODS AND CHALLENGES** **703**
Jad Abbass, Jean-Christophe Nebel, and Nashat Mansour
- 33 OVERVIEW OF CLASSIFICATION METHODS TO SUPPORT HIV/AIDS CLINICAL DECISION MAKING** **725**
Khairul A. Kasmiran, Ali Al Mazari, Albert Y. Zomaya, and Roger J. Garsia

PART H: ASSOCIATION RULES LEARNING FROM BIOLOGICAL DATA

- 34 MINING FREQUENT PATTERNS AND ASSOCIATION RULES FROM BIOLOGICAL DATA** **737**
Ioannis Kavakiotis, George Tzanis, and Ioannis Vlahavas
- 35 GALOIS CLOSURE BASED ASSOCIATION RULE MINING FROM BIOLOGICAL DATA** **761**
Kartick Chandra Mondal and Nicolas Pasquier

36 INFERENCE OF GENE REGULATORY NETWORKS BASED ON ASSOCIATION RULES	803
<i>Cristian Andrés Gallo, Jessica Andrea Carballido, and Ignacio Ponzoni</i>	

PART I: TEXT MINING AND APPLICATION TO BIOLOGICAL DATA

37 CURRENT METHODOLOGIES FOR BIOMEDICAL NAMED ENTITY RECOGNITION	841
<i>David Campos, Sérgio Matos, and José Luís Oliveira</i>	

38 AUTOMATED ANNOTATION OF SCIENTIFIC DOCUMENTS: INCREASING ACCESS TO BIOLOGICAL KNOWLEDGE	869
<i>Evangelos Pafilis, Heiko Horn, and Nigel P. Brown</i>	

39 AUGMENTING BIOLOGICAL TEXT MINING WITH SYMBOLIC INFERENCE	901
<i>Jong C. Park and Hee-Jin Lee</i>	

40 WEB CONTENT MINING FOR LEARNING GENERIC RELATIONS AND THEIR ASSOCIATIONS FROM TEXTUAL BIOLOGICAL DATA	919
<i>Muhammad Abulaish and Jahiruddin</i>	

41 PROTEIN-PROTEIN RELATION EXTRACTION FROM BIOMEDICAL ABSTRACTS	943
<i>Syed Toufeeq Ahmed, Hasan Davulcu, Sukru Tikves, Radhika Nair, and Chintan Patel</i>	

PART J: HIGH-PERFORMANCE COMPUTING FOR BIOLOGICAL DATA MINING

42 ACCELERATING PAIRWISE ALIGNMENT ALGORITHMS BY USING GRAPHICS PROCESSOR UNITS	971
<i>Mourad Elloumi, Mohamed Al Sayed Issa, and Ahmed Mokaddem</i>	

43 HIGH-PERFORMANCE COMPUTING IN HIGH-THROUGHPUT SEQUENCING	981
<i>Kamer Kaya, Ayat Hatem, Hatice Gülçin Özer, Kun Huang, and Ümit V. Çatalyürek</i>	

44 LARGE-SCALE CLUSTERING OF SHORT READS FOR METAGENOMICS ON GPUs	1003
<i>Thuy Diem Nguyen, Bertil Schmidt, Zejun Zheng, and Chee Keong Kwoh</i>	

SECTION III BIOLOGICAL DATA POSTPROCESSING

PART K: BIOLOGICAL KNOWLEDGE INTEGRATION AND VISUALIZATION

45	INTEGRATION OF METABOLIC KNOWLEDGE FOR GENOME-SCALE METABOLIC RECONSTRUCTION	1027
	<i>Ali Masoudi-Nejad, Ali Salehzadeh-Yazdi, Shiva Akbari-Birgani, and Yazdan Asgari</i>	
46	INFERRING AND POSTPROCESSING HUGE PHYLOGENIES	1049
	<i>Stephen A. Smith and Alexandros Stamatakis</i>	
47	BIOLOGICAL KNOWLEDGE VISUALIZATION	1073
	<i>Rodrigo Santamaría</i>	
48	VISUALIZATION OF BIOLOGICAL KNOWLEDGE BASED ON MULTIMODAL BIOLOGICAL DATA	1109
	<i>Hendrik Rohn and Falk Schreiber</i>	
	INDEX	1127