

**Wiley Series on Bioinformatics:
Computational Techniques and Engineering**

Yi Pan and Albert Y. Zomaya, Series Editors

Pattern Recognition in Computational Molecular Biology

TECHNIQUES AND APPROACHES



MOURAD ELLOUMI • COSTAS S. ILIOPOULOS
JASON T. L. WANG AND ALBERT Y. ZOMAYA

WILEY

PATTERN RECOGNITION IN COMPUTATIONAL MOLECULAR BIOLOGY

Techniques and Approaches

Edited by

Mourad Elloumi

Laboratory of Technologies of Information and
Communication and Electrical Engineering (LaTICE), and
University of Tunis-El Manar, Tunisia

Costas S. Iliopoulos

King's College London, UK

Jason T. L. Wang

New Jersey Institute of Technology, USA

Albert Y. Zomaya

The University of Sydney, Australia

WILEY

Copyright © 2016 by John Wiley & Sons, Inc. All rights reserved

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Pattern recognition in computational molecular biology : techniques and approaches / edited by Mourad Elloumi, Costas S. Iliopoulos, Jason T. L. Wang, Albert Y. Zomaya.

pages cm

Includes bibliographical references and index.

ISBN 978-1-118-89368-5 (cloth)

1. Molecular biology--Data processing. 2. Computational biology. 3. Pattern recognition systems. I. Elloumi, Mourad.

QH506.P38 2016

572'.8--dc23

2015029791

Cover image courtesy of iStock/Henrik5000

Typeset in 10/12pt, TimesLTStd by SPi Global, Chennai, India.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

1 2016

LIST OF CONTRIBUTORS

Andrej Aderhold, School of Biology, University of St Andrews, St Andrews, UK
Lefteris Angelis, Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

Miguel Arenas, Bioinformatics Unit, Centre for Molecular Biology “Severo Ochoa” (CSIC), Madrid, Spain; Institute of Molecular Pathology and Immunology, University of Porto (IPATIMUP), Porto, Portugal

Dunarel Badescu, Département d’informatique, Université du Québec à Montréal, Succ. Centre-Ville, Montréal, Québec, Canada

Carl Barton, The Blizzard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK

Sima Behpour, Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA

Paola Bertolazzi, Institute of Systems Analysis and Computer Science “A. Ruberti”, National Research Council, Rome, Italy

Chengpeng Bi, Bioinformatics and Intelligent Computing Lab, Division of Clinical Pharmacology, Children’s Mercy Hospitals, Kansas City, MO, USA

Kevin Byron, Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA

Weiwen Cai, Institute of Applied Genomics, College of Biological Science and Engineering, Fuzhou University, Fuzhou, China

Virginio Cantoni, Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata, Pavia, Italy

Kuo-Chen Chou, Gordon Life Science Institute, Belmont, MA, USA; King Abdulaziz University, Jeddah, Saudi Arabia

Matteo Comin, Department of Information Engineering, University of Padova, Padova, Italy

David Dao, Karlsruhe Institute of Technology, Institute for Theoretical Informatics, Postfach, Karlsruhe, Germany

- Bhaskar DasGupta**, Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA
- Wajdi Dhifli**, Clermont Université, Université Blaise Pascal, LIMOS, BP, Clermont-Ferrand, France; CNRS, UMR, LIMOS, Aubière, France; Department of Computer Science, University of Quebec At Montreal, Downtown station, Montreal (Quebec) Canada
- Carlotta Domeniconi**, Department of Computer Science, George Mason University, Fairfax, VA, USA
- Maryam Faridounnia**, Bijvoet Center for Biomolecular Research, Utrecht University, Utrecht, The Netherlands
- Giovanni Felici**, Institute of Systems Analysis and Computer Science “A. Ruberti”, National Research Council, Rome, Italy
- Marco Ferretti**, Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata, Pavia, Italy
- Giulia Fiscon**, Institute of Systems Analysis and Computer Science “A. Ruberti”, National Research Council, Rome, Italy; Department of Computer, Control and Management Engineering, Sapienza University, Rome, Italy
- Tomáš Flouri**, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
- Adelaide Freitas**, Department of Mathematics, and Center for Research & Development in Mathematics and Applications (CIDMA), University of Aveiro, Aveiro, Portugal
- Valentina Fustaino**, Cellular Biology and Neurobiology Institute, National Research Council, Rome, Italy; Institute of Systems Analysis and Computer Science “A. Ruberti”, National Research Council, Rome, Italy
- Yann Guermeur**, LORIA, Université de Lorraine-CNRS, Nancy, France
- Michael Hall**, Department of Computing Science, University of Alberta, Edmonton, Canada
- Robert Harrison**, Department of Computer Science, Georgia State University, Atlanta, GA, USA
- Dirk Husmeier**, School of Mathematics and Statistics, University of Glasgow, Glasgow, UK
- Costas S. Iliopoulos**, Department of Informatics, King’s College London, London, United Kingdom
- Samad Jahandideh**, Bioinformatics and Systems Biology Program, Sanford-Burnham Medical Research Institute, La Jolla, CA, USA
- Giuseppe Lancia**, Dipartimento di Matematica e Informatica, University of Udine, Udine, Italy
- Fabien Lauer**, LORIA, Université de Lorraine-CNRS, Nancy, France
- Vladimir Makarenkov**, Département d’informatique, Université du Québec à Montréal, Succ. Centre-Ville, Montréal, Québec, Canada

- Christos Makris**, Department of Computer Engineering and Informatics, University of Patras, Patras, Greece
- Mina Maleki**, School of Computer Science, University of Windsor, Windsor, Canada
- Diego Mallo**, Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain
- David A. Morrison**, Systematic Biology, Uppsala University, Norbyvägen, Uppsala, Sweden
- Ahmed Moussa**, LabTIC Laboratory, ENSA, Abdelmalek Essaadi University, Tangier, Morocco
- Sami Muhaidat**, ECE Department, Khalifa University, Abu Dhabi, UAE; EE Department, University of Surrey, Guildford, UK
- Mirto Musci**, Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata, Pavia, Italy
- Engelbert Mephu Nguifo**, LIMOS–Blaise Pascal University–Clermont University, Clermont-Ferrand, France; LIMOS–CNRS UMR, Aubière, France
- Stavroula Ntoufa**, Hematology Department and HCT Unit, G. Papanikolaou Hospital, Thessaloniki, Greece; Institute of Applied Biosciences, C.E.R.TH, Thessaloniki, Greece
- Nahumi Nugrahaningsih**, Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata 3, 27100 Pavia, Italy
- Hasan Oğul**, Department of Computer Engineering, Başkent University, Ankara, Turkey
- Nikos Papakonstantinou**, Hematology Department and HCT Unit, G. Papanikolaou Hospital, Thessaloniki, Greece; Institute of Applied Biosciences, C.E.R.TH, Thessaloniki, Greece
- Sheng-Lung Peng**, Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan
- Solon P. Pissis**, Department of Informatics, King's College London, London, United Kingdom
- Huzefa Rangwala**, Department of Computer Science, George Mason University, Fairfax, VA, USA
- Sara Roque**, Department of Mathematics, University of Aveiro, Aveiro, Portugal
- Luis Rueda**, School of Computer Science, University of Windsor, Windsor, Canada
- Agustín Sánchez-Cobos**, Bioinformatics Unit, Centre for Molecular Biology “Severo Ochoa” (CSIC), Madrid, Spain
- Maad Shatnawi**, Higher Colleges of Technology, Abu Dhabi, UAE
- Soheila Shokrollahzade**, Department of Medicinal Biotechnology, Iran University of Medical Sciences, Tehran, Iran

- Carina Silva**, Lisbon School of Health Technology, Lisbon, Portugal; Center of Statistics and Applications of Lisbon University (CEAUL), Lisbon, Portugal
- Tiratha Raj Singh**, Biotechnolgy and Bioinformatics Department, Jaypee University of Information and Technology (JUIT), Solan, Himachal Pradesh, India
- V Anne Smith**, School of Biology, University of St Andrews, St Andrews, UK
- Jiangning Song**, National Engineering Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China; Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, VIC, Australia
- Yang Song**, Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA
- Lisete Sousa**, Department of Statistics and Operations Research, Lisbon University and Center of Statistics and Applications of Lisbon University (CEAUL), Lisbon, Portugal, Lisbon, Portugal
- Alexandros Stamatakis**, Karlsruhe Institute of Technology, Institute for Theoretical Informatics, Postfach, Karlsruhe, Germany; Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
- Kostas Stamatopoulos**, Hematology Department and HCT Unit, G. Papanikolaou Hospital, Thessaloniki, Greece; Institute of Applied Biosciences, C.E.R.TH, Thessaloniki, Greece
- Kamal Taha**, ECE Department, Khalifa University, Abu Dhabi, UAE
- Nadia Tahiri**, Département d'informatique, Université du Québec à Montréal, Succ. Centre-Ville, Montréal, Québec, Canada
- Li Teng**, Department of Internal Medicine, University of Iowa, Iowa City, IA, USA
- Evangelos Theodoridis**, Computer Technology Institute and Press "Diophantus," Patras, Greece
- Athina Tsanoua**, Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
- Yu-Wei Tsay**, Institute of Information Science, Academia Sinica, Taipei, Taiwan
- Chinua Umoja**, Department of Computer Science, Georgia State University, Atlanta, GA, USA
- Brigitte Vannier**, Receptors, Regulation and Tumor Cells (2RTC) Laboratory, University of Poitiers, Poitiers, France
- Meghana Vasavada**, Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA
- Akbar Vaseghi**, Department of Genetics, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran
- Davide Verzotto**, Computational and Systems Biology, Genome Institute of Singapore, Singapore

- Jason T.L. Wang**, Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA
- Emanuel Weitschek**, Department of Engineering, Uninettuno International University, Rome, Italy; Institute of Systems Analysis and Computer Science "A. Ruberti", National Research Council, Rome, Italy
- Renxiang Yan**, Institute of Applied Genomics, College of Biological Science and Engineering, Fuzhou University, Fuzhou, China
- Paul D. Yoo**, Department of Computing and Informatics, Bournemouth University, UK; Centre for Distributed and High Performance Computing, University of Sydney, Sydney, Australia
- Guoxian Yu**, College of Computer and Information Science, Southwest University, Chongqing, China
- Xiuxia Yu**, Department of Computer Science, Georgia State University, Atlanta, GA, USA
- Ziding Zhang**, State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing, China
- Albert Y. Zomaya**, Centre for Distributed and High Performance Computing, University of Sydney, Sydney, Australia

PREFACE

Pattern recognition is the automatic identification of *regularities*, that is, figures, characters, shapes, and forms present in data. Pattern recognition is the core process of many scientific discoveries, whereby researchers detect regularities in large amounts of data in fields as diverse as Geology, Physics, Astronomy, and Molecular Biology. Pattern recognition in biomolecular data is at the core of Molecular Biology research. Indeed, pattern recognition makes a very important contribution to the analysis of biomolecular data. In fact, it can reveal information about shared biological functions of biological macromolecules, that is, DNA, RNA, and proteins, originating from several different organisms, by the identification of patterns that are shared by structures related to these macromolecules. These patterns, which have been conserved during evolution, often play an important structural and/or functional role, and consequently, shed light on the mechanisms and the biological processes in which these macromolecules participate. Pattern recognition in biomolecular data is also used in evolutionary studies, in order to analyze relationships that exist between species and establish whether two, or several, biological macromolecules are *homologous*, that is, have a common biological ancestor, and to reconstruct the phylogenetic tree that links them to this ancestor. On the other hand, with the new sequencing technologies, the number of biological sequences in databases is increasing exponentially. In addition, the lengths of these sequences are large. Hence, the recognition of patterns in such databases requires the development of fast, low-memory requirements and high-performance techniques and approaches. This book provides an up-to-date forum of such techniques and approaches that deal with the most studied, the most important, and/or the newest topics in the field of pattern recognition. Some of these techniques and approaches represent improvements on old ones, while others are completely new. Most of current books on pattern recognition in biomolecular data either lack technical depth or focus on specific, narrow topics. This book is the first overview on techniques and approaches on pattern recognition in biomolecular data with both a broad coverage of this field and enough depth to be of practical use to working professionals. It surveys the most recent developments of techniques and approaches on pattern recognition in biomolecular data, offering enough fundamental and technical information on these techniques and approaches and the related problems, without overloading the reader. This book will thus be invaluable not only

for practitioners and professional researchers in Computer Science, Life Science, and Mathematics but also for graduate students and young researchers looking for promising directions in their work. It will certainly point them to new techniques and approaches that may be the key to new and important discoveries in Molecular Biology.

This book is organized into seven parts: *Pattern Recognition in Sequences*, *Pattern Recognition in Secondary Structures*, *Pattern Recognition in Tertiary Structures*, *Pattern Recognition in Quaternary Structures*, *Pattern Recognition in Microarrays*, *Pattern Recognition in Phylogenetic Trees*, and *Pattern Recognition in Biological Networks*. The 29 chapters, which make up the seven parts of this book, were carefully selected to provide a wide scope with minimal overlap between the chapters so as to reduce duplications. Each contributor was asked to cover review material as well as current developments in his/her chapter. In addition, the choice of authors was made by selecting those who are leaders in their respective fields.

MOURAD ELLOUMI
TUNIS, TUNISIA

COSTAS S. ILIOPOULOS
LONDON, UK

JASON T. L. WANG
NEWARK, USA

ALBERT Y. ZOMAYA
SYDNEY, AUSTRALIA

NOVEMBER 1, 2015

CONTENTS

LIST OF CONTRIBUTORS	xxi
PREFACE	xxvii
I PATTERN RECOGNITION IN SEQUENCES	1
1 COMBINATORIAL HAPLOTYPING PROBLEMS	3
<i>Giuseppe Lancia</i>	
1.1 Introduction / 3	
1.2 Single Individual Haplotyping / 5	
1.2.1 The Minimum Error Correction Model / 8	
1.2.2 Probabilistic Approaches and Alternative Models / 10	
1.3 Population Haplotyping / 12	
1.3.1 Clark's Rule / 14	
1.3.2 Pure Parsimony / 15	
1.3.3 Perfect Phylogeny / 19	
1.3.4 Disease Association / 21	
1.3.5 Other Models / 22	
References / 23	
2 ALGORITHMIC PERSPECTIVES OF THE STRING BARCODING PROBLEMS	28
<i>Sima Behpour and Bhaskar DasGupta</i>	
2.1 Introduction / 28	
2.2 Summary of Algorithmic Complexity Results for Barcoding Problems / 32	
2.2.1 Average Length of Optimal Barcodes / 33	

- 2.3 Entropy-Based Information Content Technique for Designing Approximation Algorithms for String Barcoding Problems / 34
- 2.4 Techniques for Proving Inapproximability Results for String Barcoding Problems / 36
 - 2.4.1 Reductions from Set Covering Problem / 36
 - 2.4.2 Reduction from Graph-Coloring Problem / 38
- 2.5 Heuristic Algorithms for String Barcoding Problems / 39
 - 2.5.1 Entropy-Based Method with a Different Measure for Information Content / 39
 - 2.5.2 Balanced Partitioning Approach / 40
- 2.6 Conclusion / 40
 - Acknowledgments / 41
 - References / 41

3 ALIGNMENT-FREE MEASURES FOR WHOLE-GENOME COMPARISON

43

Matteo Comin and Davide Verzotto

- 3.1 Introduction / 43
- 3.2 Whole-Genome Sequence Analysis / 44
 - 3.2.1 Background on Whole-Genome Comparison / 44
 - 3.2.2 Alignment-Free Methods / 45
 - 3.2.3 Average Common Subword / 46
 - 3.2.4 Kullback–Leibler Information Divergence / 47
- 3.3 Underlying Approach / 47
 - 3.3.1 Irredundant Common Subwords / 48
 - 3.3.2 Underlying Subwords / 49
 - 3.3.3 Efficient Computation of Underlying Subwords / 50
 - 3.3.4 Extension to Inversions and Complements / 53
 - 3.3.5 A Distance-Like Measure Based on Underlying Subwords / 53
- 3.4 Experimental Results / 54
 - 3.4.1 Genome Data sets and Reference Taxonomies / 54
 - 3.4.2 Whole-Genome Phylogeny Reconstruction / 56
- 3.5 Conclusion / 61
 - Author's Contributions / 62
 - Acknowledgments / 62
 - References / 62

4 A MAXIMUM LIKELIHOOD FRAMEWORK FOR MULTIPLE SEQUENCE LOCAL ALIGNMENT

65

Chengpeng Bi

- 4.1 Introduction / 65
- 4.2 Multiple Sequence Local Alignment / 67
 - 4.2.1 Overall Objective Function / 67
 - 4.2.2 Maximum Likelihood Model / 68

- 4.3 Motif Finding Algorithms / 70
 - 4.3.1 DEM Motif Algorithm / 70
 - 4.3.2 WEM Motif Finding Algorithm / 70
 - 4.3.3 Metropolis Motif Finding Algorithm / 72
 - 4.3.4 Gibbs Motif Finding Algorithm / 73
 - 4.3.5 Pseudo-Gibbs Motif Finding Algorithm / 74
- 4.4 Time Complexity / 75
- 4.5 Case Studies / 75
 - 4.5.1 Performance Evaluation / 76
 - 4.5.2 CRP Binding Sites / 76
 - 4.5.3 Multiple Motifs in Helix–Turn–Helix Protein Structure / 78
- 4.6 Conclusion / 80
- References / 81

5 GLOBAL SEQUENCE ALIGNMENT WITH A BOUNDED NUMBER OF GAPS

83

Carl Barton, Tomáš Flouri, Costas S. Iliopoulos, and Solon P. Pissis

- 5.1 Introduction / 83
- 5.2 Definitions and Notation / 85
- 5.3 Problem Definition / 87
- 5.4 Algorithms / 88
- 5.5 Conclusion / 94
- References / 95

II PATTERN RECOGNITION IN SECONDARY STRUCTURES

97

6 A SHORT REVIEW ON PROTEIN SECONDARY STRUCTURE PREDICTION METHODS

99

Renxiang Yan, Jiangning Song, Weiwen Cai, and Ziding Zhang

- 6.1 Introduction / 99
- 6.2 Representative Protein Secondary Structure Prediction Methods / 102
 - 6.2.1 Chou–Fasman / 103
 - 6.2.2 GOR / 104
 - 6.2.3 PHD / 104
 - 6.2.4 PSIPRED / 104
 - 6.2.5 SPINE-X / 105
 - 6.2.6 PSSpred / 105
 - 6.2.7 Meta Methods / 105
- 6.3 Evaluation of Protein Secondary Structure Prediction Methods / 106
 - 6.3.1 Measures / 106
 - 6.3.2 Benchmark / 106
 - 6.3.3 Performances / 107

6.4	Conclusion / 110
	Acknowledgments / 110
	References / 111

7 A GENERIC APPROACH TO BIOLOGICAL SEQUENCE SEGMENTATION PROBLEMS: APPLICATION TO PROTEIN SECONDARY STRUCTURE PREDICTION 114

Yann Guermeur and Fabien Lauer

7.1	Introduction / 114
7.2	Biological Sequence Segmentation / 115
7.3	MSVMpred / 117
7.3.1	Base Classifiers / 117
7.3.2	Ensemble Methods / 118
7.3.3	Convex Combination / 119
7.4	Postprocessing with A Generative Model / 119
7.5	Dedication to Protein Secondary Structure Prediction / 120
7.5.1	Biological Problem / 121
7.5.2	MSVMpred2 / 121
7.5.3	Hidden Semi-Markov Model / 122
7.5.4	Experimental Results / 122
7.6	Conclusions and Ongoing Research / 125
	Acknowledgments / 126
	References / 126

8 STRUCTURAL MOTIF IDENTIFICATION AND RETRIEVAL: A GEOMETRICAL APPROACH 129

Virginio Cantoni, Marco Ferretti, Mirto Musci, and Nahumi Nugrahaningsih

8.1	Introduction / 129
8.2	A Few Basic Concepts / 130
8.2.1	Hierarchy of Protein Structures / 130
8.2.2	Secondary Structure Elements / 131
8.2.3	Structural Motifs / 132
8.2.4	Available Sources for Protein Data / 134
8.3	State of the Art / 135
8.3.1	Protein Structure Motif Search / 135
8.3.2	Promotif / 136
8.3.3	Secondary-Structure Matching / 137
8.3.4	Multiple Structural Alignment by Secondary Structures / 138
8.4	A Novel Geometrical Approach to Motif Retrieval / 138
8.4.1	Secondary Structures Cooccurrences / 138
8.4.2	Cross Motif Search / 143
8.4.3	Complete Cross Motif Search / 146

8.5	Implementation Notes / 149	
8.5.1	Optimizations / 149	
8.5.2	Parallel Approaches / 150	
8.6	Conclusions and Future Work / 151	
	Acknowledgment / 152	
	References / 152	
9	GENOME-WIDE SEARCH FOR PSEUDOKNOTTED NONCODING RNAs: A COMPARATIVE STUDY	155
	<i>Meghana Vasavada, Kevin Byron, Yang Song, and Jason T.L. Wang</i>	
9.1	Introduction / 155	
9.2	Background / 156	
9.2.1	Noncoding RNAs and Their Secondary Structures / 156	
9.2.2	Pseudoknotted ncRNA Search Tools / 157	
9.3	Methodology / 157	
9.4	Results and Interpretation / 161	
9.5	Conclusion / 162	
	References / 163	
III	PATTERN RECOGNITION IN TERTIARY STRUCTURES	165
10	MOTIF DISCOVERY IN PROTEIN 3D-STRUCTURES USING GRAPH MINING TECHNIQUES	167
	<i>Wajdi Dhifli and Engelbert Mephu Nguifo</i>	
10.1	Introduction / 167	
10.2	From Protein 3D-Structures to Protein Graphs / 169	
10.2.1	Parsing Protein 3D-Structures into Graphs / 169	
10.3	Graph Mining / 172	
10.4	Subgraph Mining / 173	
10.5	Frequent Subgraph Discovery / 173	
10.5.1	Problem Definition / 174	
10.5.2	Candidates Generation / 176	
10.5.3	Frequent Subgraph Discovery Approaches / 177	
10.5.4	Variants of Frequent Subgraph Mining: Closed and Maximal Subgraphs / 178	
10.6	Feature Selection / 179	
10.6.1	Relevance of a Feature / 179	
10.7	Feature Selection for Subgraphs / 180	
10.7.1	Problem Statement / 180	
10.7.2	Mining Top- <i>k</i> Subgraphs / 180	
10.7.3	Clustering-Based Subgraph Selection / 181	

10.7.4	Sampling-Based Approaches /	181
10.7.5	Approximate Subgraph Mining /	181
10.7.6	Discriminative Subgraph Selection /	182
10.7.7	Other Significant Subgraph Selection Approaches /	182
10.8	Discussion /	183
10.9	Conclusion /	185
	Acknowledgments /	185
	References /	186

11 FUZZY AND UNCERTAIN LEARNING TECHNIQUES FOR THE ANALYSIS AND PREDICTION OF PROTEIN TERTIARY STRUCTURES

190

Chinua Umoja, Xiaxia Yu, and Robert Harrison

11.1	Introduction /	190
11.2	Genetic Algorithms /	192
11.2.1	GA Model Selection in Protein Structure Prediction /	196
11.2.2	Common Methodology /	198
11.3	Supervised Machine Learning Algorithm /	201
11.3.1	Artificial Neural Networks /	201
11.3.2	ANNs in Protein Structure Prediction /	202
11.3.3	Support Vector Machines /	203
11.4	Fuzzy Application /	204
11.4.1	Fuzzy Logic /	204
11.4.2	Fuzzy SVMs /	204
11.4.3	Adaptive-Network-Based Fuzzy Inference Systems /	205
11.4.4	Fuzzy Decision Trees /	206
11.5	Conclusion /	207
	References /	208

12 PROTEIN INTER-DOMAIN LINKER PREDICTION

212

Maad Shatnawi, Paul D. Yoo, and Sami Muhaidat

12.1	Introduction /	212
12.2	Protein Structure Overview /	213
12.3	Technical Challenges and Open Issues /	214
12.4	Prediction Assessment /	215
12.5	Current Approaches /	216
12.5.1	DomCut /	216
12.5.2	Scooby-Domain /	217
12.5.3	FIEFDom /	218
12.5.4	Chatterjee et al. (2009) /	219
12.5.5	Drop /	219
12.6	Domain Boundary Prediction Using Enhanced General Regression Network /	220
12.6.1	Multi-Domain Benchmark Data Set /	220
12.6.2	Compact Domain Profile /	221

12.6.3	The Enhanced Semi-Parametric Model / 222	
12.6.4	Training, Testing, and Validation / 225	
12.6.5	Experimental Results / 226	
12.7	Inter-Domain Linkers Prediction Using Compositional Index and Simulated Annealing / 227	
12.7.1	Compositional Index / 228	
12.7.2	Detecting the Optimal Set of Threshold Values Using Simulated Annealing / 229	
12.7.3	Experimental Results / 230	
12.8	Conclusion / 232	
	References / 233	
13	PREDICTION OF PROLINE CIS-TRANS ISOMERIZATION	236
	<i>Paul D. Yoo, Maad Shatnawi, Sami Muhaidat, Kamal Taha, and Albert Y. Zomaya</i>	
13.1	Introduction / 236	
13.2	Methods / 238	
13.2.1	Evolutionary Data Set Construction / 238	
13.2.2	Protein Secondary Structure Information / 239	
13.2.3	Method I: Intelligent Voting / 239	
13.2.4	Method II: Randomized Meta-Learning / 241	
13.2.5	Model Validation and Testing / 242	
13.2.6	Parameter Tuning / 242	
13.3	Model Evaluation and Analysis / 243	
13.4	Conclusion / 245	
	References / 245	
IV	PATTERN RECOGNITION IN QUATERNARY STRUCTURES	249
14	PREDICTION OF PROTEIN QUATERNARY STRUCTURES	251
	<i>Akbar Vaseghi, Maryam Faridounnia, Soheila Shokrollahzade, Samad Jahandideh, and Kuo-Chen Chou</i>	
14.1	Introduction / 251	
14.2	Protein Structure Prediction / 255	
14.2.1	Secondary Structure Prediction / 255	
14.2.2	Modeling of Tertiary Structure / 256	
14.3	Template-Based Predictions / 257	
14.3.1	Homology Modeling / 257	
14.3.2	Threading Methods / 257	
14.3.3	<i>Ab initio</i> Modeling / 257	
14.4	Critical Assessment of Protein Structure Prediction / 258	
14.5	Quaternary Structure Prediction / 258	