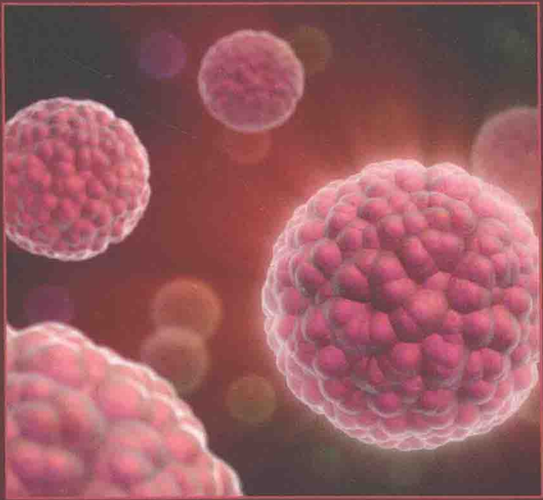
Wiley Series on Bioinformatics: Computational Techniques and Engineering

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

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

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 Blizard 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, Aubiére, 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, Aubiére, 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 Tsanousa, 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

Xiaxia 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