Analyzing Network Data in Biology and Medicine

An Interdisciplinary Textbook for Biological, Medical, and Computational Scientists

The increased and widespread availability of large network data resources in recent years has resulted in a growing need for effective methods for their analysis. The challenge is to detect patterns that provide a better understanding of the data. However, this is not a straightforward task because of the size of the datasets and the computer power required for the analysis. The solution is to devise methods for approximately answering the questions posed and these methods will vary depending on the datasets under scrutiny. This cutting-edge text introduces biological concepts and biotechnologies producing the data, graph and network theory, cluster analysis and machine learning, before discussing the thought processes and creativity involved in the analysis of large-scale biological and medical datasets, using a wide range of real-life examples. Bringing together leading experts, this text provides an ideal introduction to and insight into the interdisciplinary field of network data analysis in biomedicine.

Nataša Pržulj is Professor of Biomedical Data Science at University College London (UCL) and an ICREA Research Professor at Barcelona Supercomputing Center. She has been an elected academician of The Academy of Europe, Academia Europaea, since 2017 and is a Fellow of the British Computer Society (BCS). She is recognized for designing methods to mine large real-world molecular network datasets and for extending and using machine learning methods for integration of heterogeneous biomedical and molecular data, applied to advancing biological and medical knowledge. She received two prestigious European Research Council (ERC) research grants, Starting (2012–2017) and Consolidator (2018–2023), and USA National Science Foundation (NSF) grants among others. She is a recipient of the BCS Roger Needham Award for 2014. She was previously an Associate Professor (Reader, 2012–2016) and Assistant Professor (Lecturer, 2009–2012) in the Department of Computing at Imperial College London and an Assistant Professor in the Computer Science Department at University of California Irvine (2005–2009). She obtained a PhD in Computer Science from University of Toronto in 2005.

Analyzing Network Data in Biology and Medicine

An Interdisciplinary Textbook for Biological, Medical, and Computational Scientists

Edited and authored by

NATAŠA PRŽULJ
Professor of Biomedical Data Science, Computer Science Department, University College London
ICREA Research Professor at Barcelona Supercomputing Center

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314-321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/bionetworks
DOI: 10.1017/9781108377706

First published 2019

Printed and bound in Great Britain by Clays Ltd, Elcograf S.p.A.

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Pržulj, Nataša, editor.
Title: Analyzing network data in biology and medicine : an interdisciplinary textbook for biological, medical and computational scientists / edited by Nataša Pržulj, University College London.
Description: Cambridge, United Kingdom ; New York, NY : Cambridge University Press, 2019. | Includes bibliographical references.
Identifiers: LCCN 2018034214 | ISBN 9781108432238 (hardback : alk. paper)
Subjects: LCSH: Medical informatics–Data processing. | Bioinformatics.
Classification: LCC R858 .A469 2019 | DDC 610.285–dc23
LC record available at https://lccn.loc.gov/2018034214

ISBN 978-1-108-43223-8 Paperback

Additional resources for this publication at www.cambridge.org/bionetworks

To my loving family: Cvita, Bogdan, Nina, Sofia, and Laurentino.
And to my best friend, Vesna.

Contents

Contributors

Nicolás Alcaraz
The Bioinformatics Centre Section for RNA and Computational Biology, University of Copenhagen, Copenhagen, Denmark

Cristian Arriaga-Canon
CONACyT-Instituto Nacional de Cancerología, Mexico

Omer Basha
Department of Clinical Biochemistry and Pharmacology, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Karsten Borgwardt
Machine Learning and Computational Biology Lab, Department of Biosystems Science and Engineering, Basel, ETH Zurich, Switzerland
Swiss Institute of Bioinformatics, Basel, Switzerland

Pisanu Buphamalai
CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria

Alberto Cacciola
Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Dresden, Germany
Brain bio-inspired computing (BBC) lab, IRCCS Centro Neurolesi "Bonino Pulejo," Messina, Italy, Department of Biomedical, Dental Sciences and Morphological and Functional Images, University of Messina, Italy

Michael Caldera
CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria

Carlo Vittorio Cannistraci
Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Dresden, Germany
Brain bio-inspired computing (BBC) lab, IRCCS Centro Neurolesi "Bonino Pulejo," Messina, Italy

Laura Contreras-Espinosa
Universidad Nacional Autonoma de Mexico (UNAM), Mexico

Thomas Gaudelet
Department of Computer Science, University College London, London, UK

Rodrigo González-Barrios
Instituto Nacional de Cancerología, Mexico

Lissania Guerra-Calderas
Instituto Nacional de Cancerología, Mexico

Anne-Christin Hauschild
Krembil Research Institute, Toronto Western Hospital, Toronto, Canada, Department of Pharmacogenetics Research, Center for Addiction and Mental Health, Toronto, Canada

Idan Hekselman
Department of Clinical Biochemistry & Pharmacology, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Igor Jurisica
Krembil Research Institute, Toronto Western Hospital, Toronto, Canada
University of Toronto, Toronto, Canada

Marko Klobučar
University of Rijeka, Department of Biotechnology, Centre for High-Throughput Technologies, Rijeka, Croatia

Rok Košir
Institute of Biochemistry, Faculty of Medicine, University of Ljubljana
BIA Separations CRO, Labena Ltd, Ljubljana, Slovenia

Max Kotlyar
Krembil Research Institute, Toronto Western Hospital, Toronto, Canada

Sandra Kraljević Pavelić
University of Rijeka, Department of Biotechnology, Centre for High-Throughput Technologies, Rijeka, Croatia

Dolores Kuzelj
University of Rijeka, Department of Biotechnology, Centre for High-Throughput Technologies, Rijeka, Croatia

Luis G. Leal
Department of Life Sciences, Imperial College London, UK
Supported by a President's PhD Scholarship from Imperial College London

Felipe Llinares-López
Machine Learning and Computational Biology Lab, Department of Biosystems Science and Engineering, Basel, ETH Zurich, Switzerland
Swiss Institute of Bioinformatics, Basel, Switzerland

Noël Malod-Dognin
Department of Computer Science, University College London, London, UK

Jörg Menche
CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria

Tijana Milenković
Department of Computer Science and Engineering, Eck Institute for Global Health, and Interdisciplinary Center for Network Science and Applications (iCeNSA), University of Notre Dame, Notre Dame, Indiana, USA

John H. Morris
Department of Pharmaceutical Chemistry, University of California San Francisco, USA

Felix Müller
CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria

Alessandro Muscoloni
Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Dresden, Germany

Khalique Newaz
Department of Computer Science and Engineering, Eck Institute for Global Health, and Interdisciplinary Center for Network Science and Applications (iCeNSA), University of Notre Dame, Notre Dame, Indiana, USA

Chiara Pastrello
Krembil Research Institute, Toronto Western Hospital, Toronto, Canada

Krešimir Pavelić
Juraj Dobrila University of Pula, Pula, Croatia

Nataša Pržulj
ICREA Research Professor at Barcelona Supercomputing Center, Barcelona, Spain; Professor of Biomedical Data Science at Computer Science Department, University College London, London, UK

Richard Röttger
Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

Marisol Salgado-Albarrán
Instituto Nacional de Cancerología, Mexico

Moran Sharon
Department of Clinical Biochemistry and Pharmacology, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Ernesto Soto-Reyes
Natural Science Department, Universidad Autónoma Metropolitana-Cuajimalpa (UAM-C), Mexico

Sam F. L. Windels
Department of Computer Science, University College London, London, UK

Esti Yeger-Lotem
Department of Clinical Biochemistry and Pharmacology, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel National Institute for Biotechnology in the Negev, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Preface

We are witnessing tremendous changes in the world around us. Technological advances are impacting our lives and increasing our ability to measure things. They are yielding an astounding harvest of data about all aspects of life that form large systems of diverse interconnected entities. We are beginning to utilize the data systems to improve our understanding of the world and find solutions to some of the foremost challenges.

One such challenge is to better understand biological phenomena and apply the newly acquired understanding to improve medical treatments and outcomes. Even at the level of a cell, we are far from fully understanding the processes that we measure by genomic, epigenomic, transcriptomic, proteomic, metabolomic, metagenomic, and other "omic" data. All these different data types measure different aspects of the functioning of a cell. As these observational data grow, it is increasingly harder to analyze them and understand what they are telling us about the cell, not only due to their sizes, but also their complexities. It is not only the biology that we need to understand, which is being measured, but also the ways to abstract these complex data systems by using mathematical models that make the data amenable to computational analyses. In addition, we need to comprehend the computational challenges coming from the theory of computing, which teach us about the problems that we can efficiently and exactly solve by using computers, and about those that we cannot. Furthermore, we need to put all this biology, mathematics, and computing jointly in use by the medical sciences if we are to contribute to personalizing treatments and improving our health.

This textbook provides a resource for training upper level undergraduate students, graduate students, and researchers in this multidisciplinary area. The goal is to enable them to understand these complex issues and undertake independent research in this exciting, emerging field. The textbook presents the material in a way understandable to researchers of diverse backgrounds. Exercises are provided at the end of each chapter to put the learned material into practice. The solutions to exercises are also provided for lecturers on www.cambridge.org/bionetworks.

The textbook material is carefully chosen to start from basics and lead to more advanced concepts in a succession of chapters that build on the previous ones. The book first introduces the complex genomic and epigenomic data related to diseases and risk prediction along with the main machine learning, bioinformatics and other methods used in this domain (Chapters 1 and 2). Then it introduces the widely adopted mathematical models of graphs (networks) and the basic theory needed to understand the tools constructed for analyzing complex omics network data (Chapter 3). A very important and widely studied omics network is that of physical interactions between proteins in a cell. Hence, the biotechnologies producing these data are surveyed in Chapter 4, the quality of the data is discussed and major public databases containing the data are introduced. An introduction into methods for advanced analyses of these data is given in Chapter 5.

The textbook proceeds with the basics of machine learning commonly used to analyze network data. First, it introduces a key methodology of unsupervised

learning, cluster analysis (Chapter 6) and the applications of it in this interdisciplinary area. Then it proceeds with the basics of machine learning for data integration (Chapter 7) and advanced topics in machine learning for biomarker discovery (Chapter 8).

Just as aligning genetic sequences has revolutionized our biological and medical understanding, aligning molecular networks is expected to have similar groundbreaking impacts. This important topic is addressed and network alignment methods introduced in Chapter 9. The field of network medicine is introduced in Chapter 10. Methodology for elucidating genotype-to-phenotype relationships via analyses of human tissue-specific interactomes is presented in Chapter 11. Another important interconnected network is that of neurons in our brain. The basics of network neuroscience are presented in Chapter 12. Finally, a description of how the material presented in the textbook can be put to practice by using a major software package for analyzing network data, Cytoscape, and a major protein interaction database, STRING, are presented in the last two chapters.

I hope you will find this textbook a good resource for getting you started with doing research in this exciting and inspiring multidisciplinary area. I wish you enjoyable learning!

Nataša Pržulj