Bioinformatics and Biomarker Discovery

"omic" data analysis for personalized medicine

Francisco Azuaje

Public Research Centre for Health (CRP-Santé), Luxembourg

Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine is designed to introduce biologists, clinicians and computational researchers to fundamental data analysis principles, techniques and tools for supporting the discovery of biomarkers and the implementation of diagnostic/prognostic systems.

The focus of the book is on how fundamental statistical and data mining approaches can support biomarker discovery and evaluation, emphasizing applications based on different types of "omic" data. The book also discusses design factors, requirements and techniques for disease screening, diagnostic and prognostic applications.

Readers are provided with the knowledge needed to assess the requirements, computational approaches and outputs in disease biomarker research. Commentaries from guest experts are also included, containing detailed discussions of methodologies and applications based on specific types of "omic" data, as well as their integration.

- Covers the main range of data sources currently used for biomarker discovery
- Puts emphasis on concepts, design principles and methodologies that can be extended or tailored to more specific applications
- Offers principles and methods for assessing the bioinformatic/biostatistic limitations, strengths and challenges in biomarker discovery studies
- Discusses systems biology approaches and applications
- Includes expert chapter commentaries to further discuss relevance of techniques, summarize biological/clinical implications and provide alternative interpretations

Cover design: Gary Thompson

ISBN 978-0-470-74460-4

9 780470 744604

WILEY-BLACKWELL
www.wiley.com/wiley-blackwell

Bioinformatics and Biomarker Discovery

"Omic" Data Analysis for Personalized Medicine

Francisco Azuaje
Public Research Centre for Health (CRP-Santé), Luxembourg

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

This edition first published 2010, © 2010 by John Wiley & Sons, Ltd.

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

Registered office: John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex,
PO19 8SQ, UK

Other Editorial Offices:
9600 Garsington Road, Oxford, OX4 2DQ, UK
111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. 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 or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloguing-in-Publication Data

Azuaje, Francisco.
Bioinformatics and biomarker discovery : "omic" data analysis for
personalized medicine / Francisco Azuaje.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-74460-4
1. Biochemical markers. 2. Bioinformatics. I. Title.
[DNLM: 1. Computational Biology. 2. Biological Markers. 3.
Genomics–methods. 4. Statistics as Topic. QU 26.5 A997b 2010]
R853.B54A98 2010
610.285–dc22

2009027776

ISBN: 978-0-470-74460-4

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

Set in 10/12 Times by Thomson Digital, Noida, India
Printed in Singapore by Markono Print Media Pte Ltd

1 2010

Bioinformatics and Biomarker Discovery

To my family:
Alayne
Nelly and Francisco José
Nelytza, Oriana and Valentina

此为试读，需要完整PDF请访问：www.ertongbook.com

Author and guest contributor biographies

Francisco Azuaje has more than fifteen years of research experience in the areas of computer science, medical informatics and bioinformatics. His contributions have been reflected in several national and international research projects and an extensive publication record in journals, conference proceedings and books. Dr Azuaje is a Senior Member of the IEEE. He held a lectureship and readership in computer science and biomedical informatics at Trinity College Dublin, Ireland, and at the University of Ulster, UK, from January 2000 to February 2008. He is currently leading research in translational bioinformatics and systems biology approaches to prognostic biomarker development at the Laboratory of Cardiovascular Research, CRP-Santé, Luxembourg. He has been a member of the editorial boards of several journals and scientific committees of international conferences disseminating research at the intersection of the physical and computer sciences, engineering and biomedical sciences. He is an Associate Editor of the IEEE Transactions on Nanobioscience and BioData Mining. Dr Azuaje co-edited the books: *Data Analysis and Visualization in Genomics and Proteomics*, *Artificial Intelligence Methods and Tools for Systems Biology*, and *Advanced Methods and Tools for ECG Data Analysis*. He is currently a Section Editor of the *Encyclopaedia of Systems Biology*.

Guest contributor biographies

Guest commentary on chapter 4

Ana Dopazo holds a PhD in Molecular Biology and has worked in the field of gene expression analysis for more than 16 years, including periods in the USA, Germany and Spain, both in academia and in private companies. She currently heads the Genomics Unit at the CNIC (Centro Nacional de Investigaciones Cardiovasculares) in Madrid. The CNIC Genomics Unit is dedicated to providing high-quality genomic technology as a key element in the expansion of our knowledge of genomes, mainly in the context of translational cardiovascular research. The Unit has extensive experience in the study of transcriptomes by means of DNA microarrays, and the group's current array-based studies include genome-wide gene (mRNA) and microRNA expression analysis and whole-genome microarray differential gene expression analysis at the exon-level. The Unit's expertise in array-based transcriptome analysis encompasses all steps required by these approaches, including experimental design, sample preparation and processing, and statistical data analysis.

Guest commentary on chapter 5
Haiying Wang received a PhD degree on artificial intelligence in biomedicine from the University of Ulster, Jordanstown, UK, in 2004. He is currently a lecturer in the School of Computing and Mathematics at the University of Ulster. His research interests include knowledge engineering, data mining, artificial intelligence, XML, and their applications in medical informatics and bioinformatics. Since 2000, he has published more than 50 publications in scientific journals, books and conference proceedings related to the areas at the intersection of computer science and life science.

Huiru Zheng (IEEE member) is a lecturer in the Faculty of Engineering at the University of Ulster, UK. Dr Zheng received a BEng degree in Biomedical Engineering from Zhejiang University, China in 1989, an MSc degree in Information Processing from Fuzhou University, China in 1992, and a PhD degree on data mining and Bioinformatics from the University of Ulster in 2003. Before she joined the University of Ulster, she was working in Fuzhou University, China, as an Assistant Lecturer (1992), Lecturer (1995) and Associate Professor (2000). Her research interests include biomedical engineering, medical informatics, bioinformatics, data mining and artificial intelligence. She has over 80 publications in journals and conferences in these areas.

Guest commentary on chapter 6
Kenneth Bryan graduated from Trinity College Dublin with a degree in Microbiology in 2001. He attained a Graduate Diploma in IT 2002 at Dublin City University before returning to Trinity College to complete a PhD in Machine Learning/Bioinformatics in 2006 which chiefly focused on bicluster analysis of microarray gene expression data. During 2006–2008 Dr Bryan worked as a post-doctoral researcher in the Machine Learning group in the Complex and Adaptive Systems Laboratory (CASL) in University College Dublin in a number of areas including semi-supervised classification of gene expression data, feature selection in metabolomics data and adapting bioinformatics metrics to alternative domains. In 2008 Dr Bryan joined the Cancer Genetics group at the Royal College of Surgeons, Ireland and is currently carrying out research into molecular events that lead to the development and progression of paediatric cancers, particularly Neuroblastoma.

Guest commentary on chapter 7
Zhongming Zhao received his PhD degree in human and molecular genetics from the University of Texas Health Science Centre at Houston, USA in 2000. He also received three MSc degrees in genetics (1996), biomathematics (1998), and computer science (2002). After completion of his Keck Foundation postdoctoral fellowship, he became an assistant professor of bioinformatics in the Virginia Commonwealth University, USA, in August 2003. He became an associate professor in the Department of Biomedical Informatics, Vanderbilt University, and Chief Bioinformatics Officer in

Vanderbilt-Ingram Cancer Center, USA in 2009. His research interests are bioinformatics and systems biology approaches to studying complex diseases (data management, integration, gene ranking, gene features and networks, etc.); genome-wide or large-scale analysis of genetic variation and methylation patterns; microRNA gene networks; comparative genomics; and biomedical informatics. He has published more than 50 papers in these areas. He served as editorial board member in six journals and program committee member and session chair in nine international conferences including WICB'06, BMEI'08, ICIC'08, IJCBS'09, and SSB'09. He received several awards, including the Keck Foundation Post-doctoral Fellowship (twice: 2002, 2003), White Magnolia Award (2006), NARSAD Young Investigator Award (twice, 2005, 2008) and the best paper award from the ICIC'08 conference.

Guest commentary on chapter 8
Yves Moreau is a Professor of Engineering at the University of Leuven, Belgium. He holds an MSc in Engineering from the Faculté Polytechnique de Mons, Belgium and an MSc in Applied Mathematics from Brown University, RI, where he was a Fulbright scholar. He holds a PhD in Engineering from the University of Leuven. He is co-founder of two spin-offs of the University of Leuven: Data4s (www.norkom.com) and Cartagenia (www.cartagenia.com), the last one being active in clinical genetics. His research focuses on the application of computational methods in systems biology towards the understanding and modulation of developmental and pathological processes in constitutional disorders. Thanks to a unique collaboration with the Centre for Human Genetics, University Hospitals Leuven, his team develops an integrative computational framework for supporting genetics research from patient to phenotype to therapy. From a methodological point of view, his team develops methods based on statistics, probabilistic graphical models, and kernel methods for such analyses, with an emphasis on heterogeneous data integration and the development of computational platforms that are directly useful to biologists.

Guest commentary on chapter 10
Gary B. Fogel is Chief Executive Officer of Natural Selection, Inc. (NSI) in San Diego, California. He joined NSI in 1998 after completing a PhD in biology from the University of California, Los Angeles, with a focus on the evolution and variability of histone proteins. While at UCLA, Dr Fogel was a Fellow of the Centre for the Study of Evolution and the Origin of Life and earned several teaching and research awards. Dr Fogel's current research interests focus on the application of computational intelligence methods to problems in biomedicine and biochemistry, such as gene expression analysis, gene recognition, drug activity/toxicity prediction, structure analysis and similarity, sequence alignment, and pattern recognition. Dr Fogel is a senior member of the IEEE and member of Sigma Xi. He currently serves as Editor-in-Chief for *BioSystems*, and as an associate editor for *IEEE Transactions on Evolutionary Computation* and *IEEE Computational Intelligence Magazine*. He co-edited a volume on *Evolutionary Computation in Bioinformatics*, published in 2003 (Morgan Kaufmann) and co-edited *Computational Intelligence in Bioinformatics*, published in 2008 (IEEE Press). Dr Fogel serves as conference chair for the 2010 IEEE Congress on Evolutionary Computation (http://www.wcci2010.org) held as part of the IEEE World Congress on Computational Intelligence.

Guest commentary on chapter 10
Riccardo Bellazzi is Associate Professor of Medical Informatics at the Dipartimento di Informatica e Sistemistica, University of Pavia, Italy.

He teaches Medical Informatics and Machine Learning at the Faculty of Biomedical Engineering and Bioinformatics at the Faculty of Biotechnology of the University of Pavia. He is a member of the board of the PhD in Bioengineering and Bioinformatics of the University of Pavia.

Dr Bellazzi is Past-Chairman of the IMIA working group of Intelligent Data Analysis and Data Mining, program chair of Medinfo 2010, the world conference on Medical Informatics and of the AIME 2007 conference; he is also part of the program committee of several international conferences in medical informatics and artificial intelligence. He is a member of the editorial board of *Methods of Information in Medicine* and of the *Journal of Diabetes Science and Technology*. He is affiliated with the American Medical Informatics Association and with the Italian Bioinformatics Society. His research interests are related to biomedical informatics, comprising data mining, IT-based management of chronic patients, mathematical modelling of biological systems and bioinformatics. Riccardo Bellazzi is author of more than 200 publications on peer-reviewed journals and international conferences.

Acknowledgements

I thank my wife, Alayne Smith, for continuously helping me to succeed in personal and professional challenges. Her patience and understanding were essential to allow me to overcome the many obstacles encountered during the development of this project. The love and teachings given to me by my parents, Nelly and Francisco José, have been the greatest sources of support and inspiration for accomplishing my most valued contributions and aspirations. I thank my sister, Nelytza, and my nieces, Oriana and Valentina, for teaching me great lessons of personal strength, determination and compassion in the face of adversity.

Highly esteemed colleagues: Ana Dopazo, Haiying Wang, Huiru Zheng, Kenneth Bryan, Zhongming Zhao, Yves Moreau, Riccardo Bellazzi and Gary Fogel, enriched this project through the contribution of commentaries to accompany some of the chapters. I also thank them for their advice and corrections that allowed me to improve the content and presentation of this book.

I appreciate the support I have received from Fiona Woods and Izzy Canning, Project Editor and Publishing Assistant respectively, at John Wiley & Sons. I thank production staff at John Wiley & Sons for their assistance with book formatting and cover design. I also appreciate the help and advice from Andrea Baier during the early stages of this project. I thank Poirei Sanasam, at Thomson Digital, for management support during final production stage.

I thank all those colleagues and students, who over the years have helped me to expand my understanding of science and education. In particular, I express my affection for my school teachers, university mentors and friends in my homeland, Venezuela. Their experiences and generosity have greatly influenced my love for scientific knowledge and research.

Preface

Biomarkers are indicators of disease occurrence and progression. Biomarkers can be used to predict clinical responses to treatments, and in some cases they may represent potential drug targets. Biomarkers can be derived from solid tissues and bio-fluids. Also they can refer to non-molecular risk or clinical factors, such as life-style information and physiological signals. Different types of biomarkers have been used in clinical practice to detect disease and predict clinical outcomes.

Advanced laboratory instruments and computing systems developed to decipher the structure and function of genes, proteins and other substances in the human body offer a great variety of imperfect yet potentially useful data. Such data can be used to describe systems and processes with diverse degrees of accuracy and uncertainty. These limitations and the complexity of biomedical problems represent natural obstacles to the idea of bringing new knowledge from the laboratory to the bedside.

The greatest challenge in biomarker discovery is not the discovery of powerful predictors of disease. Nor is it the design of sophisticated algorithms and tools. The greatest test is to demonstrate its potential relevance in a clinical setting. This requires strong evidence of improvements in the health or quality of life of patients. This also means that potential biomarkers should stand the challenge of independent validations and reproducibility of results.

Advances in this area have traditionally been driven at the intersection of the medical and biological sciences. Nevertheless, it is evident that current and future progress will also depend on the combination of skills and resources originating from the physical and computational sciences and engineering. In particular, bioinformatics and computational biology have the mission to bring new capacities and possibilities to understand and solve problems.

The promise of new advances based on the synergy of these disciplines will also depend on the growth and maturation of a new generation of researchers, managers and policy makers. This will be accomplished only through new and diverse training opportunities, ranging from pre-college, through undergraduate and post-graduate, to post-doctoral and life-long education.

此为试读，需要完整PDF请访问：www.ertongbook.com

One of the crucial challenges for bioinformaticians and computational biologists is the need to continuously accumulate a great diversity of knowledge and skills. Moreover, despite the fact that almost everyone in the clinical and biological sciences would agree on the importance of computational research in translational biomedical research, there are still major socio-cultural obstacles that must be overcome. Such obstacles mirror the complexity and speed of unprecedented changes in technology, scientific culture and human relations.

Bioinformaticians and computational biologists have a mission that goes beyond the provision of technical support or the implementation of standard computing solutions. Their mission is to contribute to the generation and verification of new knowledge, which can be used to detect, prevent or cure disease. In the longer term, this may result in a more effective fight against human suffering and poverty. This demands from us a continuous improvement of skills and changes in attitude. Skills and attitudes that can prepare us to cooperate and lead in this endeavour.

This book aims to support efforts in that direction. It represents an attempt to introduce readers to some of the crucial problems, tools and opportunities in bioinformatics and biomarker research. I hope that its content will at least serve to foster new conversations between and within research teams across disciplines, or even to help to recognize new value and purpose of ongoing interactions.

Contents

1 Biomarkers and bioinformatics

This chapter discusses key concepts, problems and research directions. It provides an introduction to translational biomedical research, personalized medicine, and biomarkers: types and main applications. It will introduce fundamental data types, computational and statistical requirements in biomarker studies, an overview of recent advances, and a comparison between 'traditional' and 'novel' molecular biomarkers. Significant roles of bioinformatics in biomarker research will be illustrated, as well as examples of domain-specific models and applications. It will end with a summary of expected learning outcomes, content overview, and a description of basic mathematical notation to be used in the book.

1.1 Bioinformatics, translational research and personalized medicine

In this book, the term bioinformatics refers to the design, implementation and application of computational technologies, methods and tools for making 'omic' data meaningful. This involves the development of information and software resources to support a more open and integrated access to data and information. Bioinformatics is also used in the context of emerging computational technologies for modelling complex systems and informational patterns for predictive purposes. This book is about the discovery of knowledge from human molecular and clinical data through bioinformatics. Knowledge that represents 'biomarkers' of disease and clinically-relevant phenotypes.

Another key issue that this book addresses is the 'translational' role of bioinformatics in the post-genome era. Translational research aims to aid in the transformation of biological knowledge into solutions that can be applied in a clinical setting. In addition,

Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine Francisco Azuaje

© 2010 John Wiley & Sons, Ltd.

此为试读，需要完整PDF请访问：www.ertongbook.com