Biomarkers in Disease:
Methods, Discoveries and Applications
Series Editor: Victor R. Preedy

Victor R. Preedy
Vinood B. Patel *Editors*

General Methods in Biomarker Research and their Applications

Springer Reference

Biomarkers in Disease: Methods, Discoveries and Applications

Series Editor

Victor R. Preedy
Department of Nutrition and Dietetics
Division of Diabetes and Nutritional Sciences
Faculty of Life Sciences and Medicine
King's College London
London, UK

In the past decade there has been a sea change in the way disease is diagnosed and investigated due to the advent of high throughput technologies, such as microarrays, lab on a chip, proteomics, genomics, lipomics, metabolomics, etc. These advances have enabled the discovery of new and novel markers of disease relating to autoimmune disorders, cancers, endocrine diseases, genetic disorders, sensory damage, intestinal diseases etc. In many instances these developments have gone hand in hand with the discovery of biomarkers elucidated via traditional or conventional methods, such as histopathology or clinical biochemistry. Together with microprocessor-based data analysis, advanced statistics and bioinformatics these markers have been used to identify individuals with active disease or pathology as well as those who are refractory or have distinguishing pathologies. Unfortunately techniques and methods have not been readily transferable to other disease states and sometimes diagnosis still relies on single analytes rather than a cohort of markers. Furthermore, the discovery of many new markers have not been put into clinical practice, partly because of their cost and partly because some scientists are unaware of their existence or the evidence is still at the preclinical stage. In some cases the work needs further scientific scrutiny. There is thus a demand for a comprehensive and focused evidenced-based text and scientific literature that addresses these issues. Hence the formulation of *Biomarkers in Disease: Methods, Discoveries and Applications*. The series covers a wide number of areas including for example, nutrition, cancer, endocrinology, cardiology, addictions, immunology, birth defects, genetics and so on. The chapters are written by national or international experts and specialists.

Series Titles

1. General Methods in Biomarker Research and Their Applications
2. Biomarkers in Cancer
3. Biomarkers in Cardiovascular Disease
4. Biomarkers in Kidney Disease
5. Biomarkers in Bone Disease
6. Biomarkers in Liver Disease

More information about this series at http://www.springer.com/series/13842

Victor R. Preedy • Vinood B. Patel
Editors

General Methods in Biomarker Research and their Applications

With 203 Figures and 122 Tables

Springer Reference

Editors
Victor R. Preedy
Department of Nutrition and Dietetics
Division of Diabetes and Nutritional Sciences
Faculty of Life Sciences and Medicine
King's College London
London, UK

Vinood B. Patel
Department of Biomedical Sciences
Faculty of Science and Technology
University of Westminster
London, UK

ISBN 978-94-007-7695-1 ISBN 978-94-007-7696-8 (eBook)
ISBN 978-94-007-7697-5 (print and electronic bundle)
DOI 10.1007/978-94-007-7696-8

Library of Congress Control Number: 2015941892

Springer Dordrecht Heidelberg New York London

Printed on acid-free paper

Springer Science+Business Media B.V. Dordrecht is part of Springer Science+Business Media (www.springer.com)

Preface

In the present volume, ***General Methods in Biomarker Research and Their Applications***, we have sections on

- General Aspects: Techniques and Overviews
- Specific Analytes and Their Application
- Pregnancy and Life Events
- Nutrition, Metabolism, and Environmental Health
- Cardiovascular System, Lung, and Kidney
- Brain, Neurology, and Associated Conditions
- Cancer, Immune Function, Inflammation, and Other Conditions
- Further Knowledge

While the Editors recognize the difficulties in assigning particular chapters to particular sections, the book has enormously wide coverage and includes the following areas, analytes, and conditions: high-throughput methods, mass spectrometry, lipidomics, toxicogenomics, pharmacogenomics, personalized medicine, glycome, flow cytometry, creatinine, creatine, paraffin–embedded tissue, macrophage inflammatory protein-1 alpha (MIP-1 alpha)/CCL3, pfetin, pentraxin 3, salivary amylase, urinary hydrogen peroxide, guanylyl cyclase C, isoprostanes, cyclophilin A, oxidative stress, FABP3, fetal membranes, the menopause, nutritional studies, 1-hydroxypyrene, environmental health, pediatric heart surgery, necrosis, myocardial remodeling, serum collagen, galectin-3, natriuretic peptides, heat shock proteins, YKL-40, imaging, hemostatic markers, chronic obstructive pulmonary disease (COPD), klotho, chronic and polycystic kidney diseases, exosomes, depression, psychosis, Parkinson's disease, amyotrophic lateral sclerosis, multiple sclerosis, brain injury, micro-RNAs, S100B, gold nanoparticles, cancer, immunogenic salivary proteins, inflammasome proteins, urinary tract disease, allergic rhinitis, and graft-versus-host disease. Finally, the last chapter is devoted to locating resource material for biomarker discovery and applications. The chapters are written by national or international experts and specialists.

This book is specifically designed for clinical biochemists, scientists, epidemiologists, doctors, and nurses, from students to practitioners at the higher level. It is also designed to be suitable for lecturers and teachers in health care and libraries as a reference guide.

April 2015
London

Victor R. Preedy
Vinood B. Patel

Series Preface

In the past decade, there has been a sea change in the way disease is diagnosed and investigated due to the advent of high-throughput technologies and advances in chemistry and physics, leading to the development of microarrays, lab on a chip, proteomics, genomics, lipomics, metabolomics, etc. These advances have enabled the discovery of new and novel markers of disease relating to autoimmune disorders, cancers, endocrine diseases, genetic disorders, sensory damage, intestinal diseases, and many other conditions too numerous to list here. In many instances, these developments have gone hand in hand with the discovery of biomarkers elucidated via traditional or conventional methods, such as histopathology, immunoassays, or clinical biochemistry. Together with microprocessor-based data analysis, advanced statistics, and bioinformatics, these markers have been used to identify individuals with active disease as well as those who are refractory or have distinguishing pathologies.

Unfortunately, techniques and methods have not been readily transferable to other disease states, and sometimes, diagnosis still relies on a single analyte rather than a cohort of markers. Furthermore, the discovery of many new markers has not been put into clinical practice partly because of their cost and partly because some scientists are unaware of their existence or the evidence is still at the preclinical stage. There is thus a demand for a comprehensive and focused evidence-based text and scientific literature that addresses these issues. Hence the book series ***Biomarkers in Disease: Methods, Discoveries, and Applications***. It imparts holistic information on the scientific basis of health and biomarkers and covers the latest knowledge, trends, and treatments. It links conventional approaches with new platforms. The ability to transcend the intellectual divide is aided by the fact that each chapter has

- *Key Facts* (areas of focus explained for the lay person)
- *Definitions of Words and Terms*
- *Potential Applications to Prognosis, Other Diseases, or Condition*s
- *Summary Points*

The material in ***Potential Applications to Prognosis, Other Diseases, or Condition*s** pertains to speculative or proposed areas of research, cross-transference to

other diseases or stages of the disease, translational issues, and other areas of wide applicability.

The series is expected to prove useful for clinicians, scientists, epidemiologists, doctors and nurses, and also academicians and students at an advanced level.

April 2015 Victor R. Preedy
London

Contents

Volume 1

About the Editors

Victor R. Preedy is a senior member of King's College London (Professor of Nutritional Biochemistry) and King's College Hospital (Professor of Clinical Biochemistry, Honorary). He is attached to both the Diabetes and Nutritional Sciences Division and the Department of Nutrition and Dietetics. He is also founding and current Director of the Genomics Centre and a member of the School of Medicine. Professor Preedy graduated in 1974 with an Honors Degree in Biology and Physiology with Pharmacology. He gained his University of London Ph.D. in 1981. In 1992, he received his Membership of the Royal College of Pathologists, and in 1993, he gained his second Doctoral degree for his contribution to the science of protein metabolism in health and disease. Professor Preedy was elected as a Fellow of the Institute of Biology in 1995 and to the Royal College of Pathologists in 2000. Since then, he has been elected as a Fellow to the Royal Society for the Promotion of Health (2004) and the Royal Institute of Public Health and Hygiene (2004). In 2009, Professor Preedy became a Fellow of the Royal Society for Public Health and in 2012 a Fellow of the Royal Society of Chemistry. In his career, Professor Preedy worked at the National Heart Hospital (part of Imperial College London) and the MRC Centre at Northwick Park Hospital. He has collaborated with research groups in Finland, Japan, Australia, USA, and Germany. He is a leading expert on biomedical sciences and has a long-standing interest in analytical methods and their applications to the study of health and disease. He has lectured nationally and internationally. To his credit, Professor Preedy has published over 500 articles, which includes peer-reviewed manuscripts based on original research, reviews, abstracts, and numerous books and volumes.

Vinood B. Patel is currently a Senior Lecturer in Clinical Biochemistry at the University of Westminster and honorary fellow at King's College London. He presently directs studies on metabolic pathways involved in liver disease, particularly related to mitochondrial energy regulation and cell death. Research is being undertaken to study the role of nutrients, antioxidants, phytochemicals, iron, alcohol, and fatty acids in the pathophysiology of liver disease. Other areas of interest are identifying new biomarkers that can be used for diagnosis and prognosis of liver disease, understanding mitochondrial oxidative stress in Alzheimer's disease, and gastrointestinal dysfunction in autism. Dr. Patel graduated from the University of Portsmouth with a degree in Pharmacology and completed his Ph.D. in protein metabolism from King's College London in 1997. His postdoctoral work was carried out at Wake Forest University Baptist Medical School studying structural–functional alterations to mitochondrial ribosomes, where he developed novel techniques to characterize their biophysical properties. Dr. Patel is a nationally and internationally recognized liver researcher and was involved in several NIH-funded biomedical grants related to alcoholic liver disease. He has edited biomedical books in the area of nutrition and health prevention, autism, and biomarkers and has published over 150 articles. In 2014, he was elected as a Fellow to The Royal Society of Chemistry.

Editorial Advisors

Contributors

Wan Zaidah Abdullah Haematology Department, School of Medical Sciences, Health Campus, Universiti Sains Malaysia, Kubang Kerian, Kelantan, Malaysia

Mehmet Ağirbaşli Department of Cardiology, Marmara University Medical Center, Kadikoy, Istanbul, Turkey

Emin Ozgur Akgul Department of Clinical Biochemistry, School of Medicine, Gulhane Military Medical Academy, Etlik, Ankara, Turkey

Mohanad Al-Sabbagh Department of Oral Health Practice, Division of Periodontology, University of Kentucky College of Dentistry, Lexington, KY, USA

Fawaz Alzaid Institut National de la Santé et de la Recherche Médicale (INSERM) UMRS 1138, Centre de Recherche des Cordeliers (CRC), Paris, France

Ángela Arias Sección de Errores Congénitos del Metabolismo-IBC, Servicio de Bioquímica y Genética Molecular, Hospital Clínic, CIBERER, IDIBAPS, Barcelona, Spain

Timucin Avsar Molecular Biology-Biotechnology and Genetics Research Center, Istanbul Technical University, Istanbul, Turkey

Eiichi Azuma Department of Pediatrics and Cell Transplantation, Mie University Graduate School of Medicine, Tsu, Mie, Japan

Clayton B'Hymer Molecular and Genetics Monitoring Team, Department of Applied Technology, National Institute for Occupational Safety and Health, Taft Laboratories, C-23, Cincinnati, OH, USA

Erik W. Baars University of Applied Sciences Leiden, Leiden, The Netherlands

Louis Bolk Institute, Driebergen, The Netherlands

Beatrix Bahle Roche Diagnostics GmbH, Werk Penzberg, Penzberg, Germany

Petros Bakakos 1st Respiratory Medicine Department, University of Athens Medical School, Athens, Greece

Erin S. Baker Environmental Molecular Sciences Laboratory, and Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

Nagaraja S. Balakathiresan Department of Pathology, Uniformed Services University of the Health Sciences, F. Edward Hébert School of Medicine, Bethesda, MD, USA

Ancha Baranova Center for the Study of Chronic Metabolic Diseases, School of System Biology, George Mason University, Fairfax, VA, USA

Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow, Russia

Ofer Beharier Department of Obstetrics and Gynecology, Faculty of Health Sciences, Soroka University Medical Center, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Ishita Bhavsar Department of Oral Health Practice, Division of Periodontology, University of Kentucky College of Dentistry, Lexington, KY, USA

Manish Bhomia Department of Pathology, Uniformed Services University of the Health Sciences, F. Edward Hébert School of Medicine, Bethesda, MD, USA

Clara Bonanad Servicio de Cardiología, Hospital Clínic Universitari, INCLIVA. Universidad de Valencia, Valencia, Spain

Roberto Bono Department of Public Health and Pediatrics, University of Torino, Torino, Italy

Sarah Bredson The University of Texas Medical Branch, Galveston, TX, USA

Mercè Brunet Pharmacology and Toxicology Laboratory, Centro de Diagnóstico Biomédico, CIBERehd, IDIBAPS, Hospital Clínico de Barcelona, Barcelona University, Barcelona, Spain

Erdinc Cakir Department of Clinical Biochemistry, School of Medicine, Gulhane Military Medical Academy, Etlik, Ankara, Turkey

Daniele Catalucci National Research Council (CNR) Institute of Genetic and Biomedical Research (IRGB) - UOS of Milan, Humanitas Clinical and Research Center, Rozzano, MI, Italy

Raghavendar Chandran Department of Pathology, Uniformed Services University of the Health Sciences, F. Edward Hébert School of Medicine, Bethesda, MD, USA

Biological Sciences Group, Birla Institute of Technology and Science, Pilani, Rajasthan, India

Peter S. Chang Department of Pharmacology and Experimental Therapeutics, Thomas Jefferson University, Philadelphia, PA, USA

Fabian Chaustre Servicio de Cardiología, Hospital Clínic Universitari, INCLIVA. Universidad de Valencia, Valencia, Spain

Kang Jun Cho Department of Urology, School of Medicine, The Catholic University of Korea, Bucheon St. Mary's Hospital, Bucheon, Gyeonggi-do, Republic of Korea

Benjamin J. W. Chow Department of Cardiology, University of Ottawa Heart Institute, Ottawa, ON, Canada

Gianluigi Condorelli National Research Council (CNR) Institute of Genetic and Biomedical Research (IRGB) - UOS of Milan, Humanitas Clinical and Research Center, Rozzano, MI, Italy

Joseph A. Conrad Department of Chemistry, Vanderbilt University, Nashville, TN, USA

Sylvie Cornelie Laboratoire Maladies Infectieuses et Vecteurs: UMR 224 CNRS-IRD-UM1-UM2, Institut de Recherche pour le Développement, Montpellier, France

Juan Carlos de Rivero Vaccari Department of Ophthalmology, Louisiana State University School of Medicine/Ochsner Medical Center, New Orleans, LA, USA

Juan Pablo de Rivero Vaccari The Miami Project to Cure Paralysis, Department of Neurological Surgery, University of Miami Miller School of Medicine, Miami, FL, USA

Monique Dehoux Biochemistry Department, Bichat Hospital, APHP, Paris, France

INSERM, U1152, Paris, France

Johanna K. DiStefano Translational Genomics Research Institute, Diabetes, Cardiovascular and Metabolic Diseases Center, Phoenix, AZ, USA

Souleymane Doucoure Laboratoire Maladies Infectieuses et Vecteurs: UMR 224 CNRS-IRD-UM1-UM2, and Unité de Recherche sur les Maladies Infectieuses Tropicales Emergentes, Institut de Recherche pour le Développement, Campus IRD-UCAD, Dakar, Sénégal

Laboratoire Maladies Infectieuses et Vecteurs: UMR 224 CNRS-IRD-UM1-UM2, Institut de Recherche pour le Développement, Montpellier, France

Pape M. Drame Laboratoire Maladies Infectieuses et Vecteurs: UMR 224 CNRS-IRD-UM1-UM2, and Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, Institut de Recherche pour le DÕveloppement, Montpellier France and National Institutes of Health, Bethesda, Maryland, USA

Fang Fang Department of Pathology, University of Washington, UW Harborview Medical Center R&T Building, Seattle, WA, USA

Vassilios Fanos Department of Pediatrics and Clinical Medicine, Section of Neonatal Intensive Care Unit, Puericutura Institute and Neonatal Section, University of Cagliari, Cagliari, Italy

Jun-ichi Furukawa Laboratory of Medical and Functional Glycomics, Graduate School of Advanced Life Science, and Frontier Research Center for Post-Genome Science and Technology, Hokkaido University, Sapporo, Japan

Claire Gahm Medical School for International Health, Faculty of Health Sciences, Soroka University Medical Center, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Cleveland Heights, OH, USA

Parul Gandhi Division of Cardiology, Massachusetts General Hospital, Boston, MA, USA

Sergio García-Blas Servicio de Cardiología, Hospital Clínic Universitari, INCLIVA. Universidad de Valencia, Valencia, Spain

Judit García-Villoria Sección de Errores Congénitos del Metabolismo-IBC, Servicio de Bioquímica y Genética Molecular, Hospital Clínic, CIBERER, IDIBAPS, Barcelona, Spain

Panagiota Georgiadou 2nd Division of Interventional Cardiology, Onassis Cardiac Surgery Center, Athens, Greece

Jean Guglielminotti Anesthesia Department, Bichat Hospital, APHP, Paris, France

INSERM, UMR1137, IAME, Paris, France

Paridhi Gupta Department of Pathology, Uniformed Services University of the Health Sciences, F. Edward Hébert School of Medicine, Bethesda, MD, USA

Diana Hernández-Romero Department of Cardiology, Hospital Universitario Virgen de la Arrixaca, University of Murcia, Murcia, Spain

Andrea Herold Roche Diagnostics GmbH, Werk Penzberg, Penzberg, Germany

Masahiro Hirayama Department of Pediatrics and Cell Transplantation, Mie University Graduate School of Medicine, Tsu, Mie, Japan

Chunxiu Hu Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, China

Terry Hyslop Department of Pharmacology and Experimental Therapeutics, Thomas Jefferson University, Philadelphia, PA, USA

Kaori Iino Department of Obstetrics and Gynecology, Hirosaki University Graduate School of Medicine, Hirosaki, Aomori, Japan

Hiroshi Ito Department of Cardiovascular Medicine, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan

James L. Januzzi Jr. Division of Cardiology, Massachusetts General Hospital, Boston, MA, USA

Hueiwang Anna Jeng School of Community and Environmental Health, College of Health Sciences, Old Dominion University, Virginia, VA, USA

Amy M. Jimenez UCLA Department of Psychiatry and Biobehavioral Sciences, Desert Pacific Mental Illness Research, Education, and Clinical Center, VA Greater Los Angeles Healthcare System, Los Angeles, CA, USA

Orhun H. Kantarci Department of Neurology, Mayo School of Graduate Medical Education, Mayo Clinic College of Medicine, Rochester, MN, USA

Jens Kastrup Department of Cardiology, Cardiac Catheterization Laboratory 2014, Faculty of Health Sciences, The Heart Centre, Rigshospitale, Copenhagen University Hospital, Copenhagen, Denmark

Itsuro Kazama Department of Physiology I, Tohoku University Graduate School of Medicine, Sendai, Miyagi, Japan

Mahdieh Khosroheidari Translational Genomics Research Institute, Diabetes, Cardiovascular and Metabolic Diseases Center, Phoenix, AZ, USA

Joon Chul Kim Department of Urology, School of Medicine, The Catholic University of Korea, Bucheon St. Mary's Hospital, Bucheon, Gyeonggi-do, Republic of Korea

Andreas D. Kistler Division of Nephrology, University Hospital Zürich, Zürich, Switzerland

Masashi Kitagawa Department of Medicine and Clinical Science, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan

Ran Klein Department of Nuclear Medicine, The Ottawa Hospital, Ottawa, ON, Canada

Yoshihiro Komada Department of Pediatrics and Cell Transplantation, Mie University Graduate School of Medicine, Tsu, Mie, Japan

Tadashi Kondo Division of Pharmacoproteomics, National Cancer Center Research Institute, Tokyo, Japan

Konstantinos Kostikas 2nd Respiratory Medicine Department, University of Athens Medical School, Athens, Greece

Frank Kramer Clinical Sciences/Global Biomarker Strategy and Development, Bayer HealthCare AG, Wuppertal, Germany

Thomas Krüger Institute of Biochemistry I, University Hospital Jena, Jena, Germany

Department of Molecular and Applied Microbiology, Leibniz Institute for Natural Product Research and Infection Biology, Hans Knöll Institute Jena, Jena, Germany

Masayuki Kubo Department of Public Health, Graduate School of Medicine, Dentistry, and Pharmaceutical Sciences, Okayama University, Okayama, Japan

Yasemin Gulcan Kurt Department of Clinical Biochemistry, School of Medicine, Gulhane Military Medical Academy, Etlik, Ankara, Turkey

Michael V. G. Latronico Humanitas Clinical and Research Center, Rozzano, MI, Italy

Jia Li Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, China

Tao Liu Environmental Molecular Sciences Laboratory, and Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

Sabine Lohmann Roche Diagnostics GmbH, Werk Penzberg, Penzberg, Germany

Dan Longrois Anesthesia Department, Bichat Hospital, APHP, Paris, France

Paris Diderot University, Sorbonne Paris Cité, Paris, France

INSERM, U1148, Paris, France

Stelios Loukides 2nd Respiratory Medicine Department, University of Athens Medical School, Athens, Greece

Radha K. Maheshwari Department of Pathology, Uniformed Services University of the Health Sciences, F. Edward Hébert School of Medicine, Bethesda, MD, USA

Hirofumi Makino Department of Medicine and Clinical Science, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan

Alexandra Marie Laboratoire Maladies Infectieuses et Vecteurs: UMR 224 CNRS-IRD-UM1-UM2, Institut de Recherche pour le Développement, Montpellier, France

Francisco Marín Department of Cardiology, Hospital Universitario Virgen de la Arrixaca, University of Murcia, Murcia, Spain

Chie Masatomi Department of Public Health, Graduate School of Medicine, Dentistry, and Pharmaceutical Sciences, Okayama University, Okayama, Japan

Françoise Mathieu-Daudé Laboratoire Maladies Infectieuses et Vecteurs: UMR 224 CNRS-IRD-UM1-UM2, Institut de Recherche pour le Développement, Montpellier, France

Mitsunobu Matsubara Division of Molecular Medicine, Centers for Advanced Research and Translational Medicine, Tohoku University Graduate School of Medicine, Sendai, Miyagi, Japan

Brian Mc Ardle Department of Cardiology, University of Ottawa Heart Institute, Ottawa, ON, Canada

Ramkumar Menon Department of Obstetrics and Gynecology, Division of Maternal-Fetal Medicine Perinatal Research, The University of Texas Medical Branch at Galveston, Galveston, TX, USA

Olga Millán Pharmacology and Toxicology Laboratory, Centro de Diagnóstico Biomédico, CIBERehd, IDIBAPS, Hospital Clínico de Barcelona, Barcelona University, Barcelona, Spain

Craig S. Miller Department of Oral Health Practice, Division of Oral Diagnosis, Oral Medicine, Oral Radiology, University of Kentucky College of Dentistry, Lexington, KY, USA

Hendrik Milting Herz- und Diabeszentrum NRW, Klinik für Thorax- und Kardiovaskularchirurgie, Erich und Hanna Klessmann-Institut für Kardiovaskuläre Forschung und Entwicklung, Universitätsklinikum der Ruhr-Universität Bochum, Bad Oeynhausen, Germany

Yoshiaki Miura S-BIO, Vaupell Inc., Hudson, NH, USA

Hideki Mizunuma Department of Obstetrics and Gynecology, Hirosaki University Graduate School of Medicine, Hirosaki, Aomori, Japan

François Mouchet Laboratoire Maladies Infectieuses et Vecteurs: UMR 224 CNRS-IRD-UM1-UM2, Institut de Recherche pour le Développement, Montpellier, France

Michele Mussap Laboratory Medicine Service, IRCCS AOU San Martino-IST, University-Hospital, Genoa, Italy

Naja Dam Mygind Department of Cardiology, Cardiac Catheterization Laboratory 2014, Faculty of Health Sciences, The Heart Centre, Rigshospitale, Copenhagen University Hospital, Copenhagen, Denmark

Kazufumi Nakamura Department of Cardiovascular Medicine, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan

Emmanuel E. Ndille Laboratoire Maladies Infectieuses et Vecteurs: UMR 224 CNRS-IRD-UM1-UM2, Institut de Recherche pour le Développement, Montpellier, France

Nathalia Noda Nicolau Department of Pathology, Botucatu Medical School, University Estadual Paulista, Botucatu, Sao Paulo, Brazil

Andreas F. M. Nierop Muvara, Leiderdorp, The Netherlands

Julio Núñez Servicio de Cardiología, Hospital Clínic Universitari, INCLIVA. Universidad de Valencia, Valencia, Spain

Keiki Ogino Department of Public Health, Graduate School of Medicine, Dentistry, and Pharmaceutical Sciences, Okayama University, Okayama, Japan

Esteban Orenes-Piñero Department of Cardiology, Hospital Universitario Virgen de la Arrixaca, University of Murcia, Murcia, Spain

Sonia Pajares Sección de Errores Congénitos del Metabolismo-IBC, Servicio de Bioquímica y Genética Molecular, Hospital Clínic, CIBERER, IDIBAPS, Barcelona, Spain

Chin-Hong Pan Division of Occupational Hazards Assessment, Institute of Labor, Occupational Safety and Health, Ministry of Labor, Taipei County 221, Taiwan

Vinood B. Patel Department of Biomedical Sciences, Faculty of Science and Technology, University of Westminster, London, UK

Anne Poinsignon Laboratoire Maladies Infectieuses et Vecteurs: UMR 224 CNRS-IRD-UM1-UM2, Institut de Recherche pour le Développement, Montpellier, France

Jossimara Polettini The University of Texas Medical Branch, Galveston, TX, USA

Amir Pourmoghaddas Physics Department, Department of Cardiology, Carleton University, University of Ottawa Heart Institute, Ottawa, ON, Canada

Victor R. Preedy Department of Nutrition and Dietetics, Division of Diabetes and Nutritional Sciences, Faculty of Life Sciences and Medicine, King's College London, London, UK

Mina Yamazaki Price Department of Nutrition and Dietetics, St George's Healthcare NHS Trust, St George's Hospital, London, UK

David Prvulovic Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, Laboratory of Neurophysiology and Neuroimaging, Johann Wolfgang Goethe University, Frankfurt/Main, Germany

Paolo Racugno Servicio de Cardiología, Hospital Clínic Universitari, INCLIVA. Universidad de Valencia, Valencia, Spain

Rajkumar Rajendram Department of General Medicine and Intensive Care, John Radcliffe Hospital, Oxford, UK

Diabetes and Nutritional Sciences Research Division, School of Medicine, King's College London, London, UK

Roshanna Rajendram School of Medicine, University of Birmingham, Edgbaston, Birmingham, UK

Rupesh Kanchi Ravi Translational Genomics Research Institute, Diabetes, Cardiovascular and Metabolic Diseases Center, Phoenix, AZ, USA

Franck Remoue Laboratoire Maladies Infectieuses et Vecteurs: UMR 224 CNRS-IRD-UM1-UM2, Institut de Recherche pour le Développement, Montpellier, France

Antonia Ribes Sección de Errores Congénitos del Metabolismo-IBC, Servicio de Bioquímica y Genética Molecular, Hospital Clínic, CIBERER, IDIBAPS, Barcelona, Spain

Tiphaine Robert-Mercier Biochemistry Department, Bichat Hospital, APHP, Paris, France

Karin D. Rodland Environmental Molecular Sciences Laboratory, and Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

Valeria Romanazzi Department of Public Health and Pediatrics, University of Torino, Torino, Italy

Noriko Sakano Department of Gerontology Research, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, Japan

Yoshie Sato Graduate School of Health Sciences, Okayama University, Okayama, Japan

Kimio Satoh Department of Cardiovascular Medicine, Tohoku University Graduate School of Medicine, Sendai, Japan

Huub F. J. Savelkoul Cell Biology and Immunology Group, Wageningen University, Wageningen, The Netherlands

Eftihia Sbarouni 2nd Division of Interventional Cardiology, Onassis Cardiac Surgery Center, Athens, Greece

Barbara Schneider Department of Addictive Disorders and Psychiatry, LVR-Klinik Köln, Academic Teaching Hospital of the University of Cologne, Cologne, Germany

Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, Laboratory of Neurophysiology and Neuroimaging, Johann Wolfgang Goethe University, Frankfurt/Main, Germany

Julian Schuster Roche Diagnostics GmbH, Werk Penzberg, Penzberg, Germany

Anuj Sharma Department of Pathology, Uniformed Services University of the Health Sciences, F. Edward Hébert School of Medicine, Bethesda, MD, USA

Hiroaki Shimokawa Department of Cardiovascular Medicine, Tohoku University Graduate School of Medicine, Sendai, Japan

Yasuro Shinohara Laboratory of Medical and Functional Glycomics, Graduate School of Advanced Life Science, and Frontier Research Center for Post-Genome Science and Technology, Hokkaido University, Sapporo, Japan

Aksel Siva Department of Neurology, Cerrahpasa School of Medicine, Istanbul University, Istanbul, Turkey

Richard D. Smith Environmental Molecular Sciences Laboratory, and Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

Chatzikyriakou Sofia 2nd Division of Interventional Cardiology, Onassis Cardiac Surgery Center, Athens, Greece

Tessandra Stewart Department of Pathology, University of Washington, UW Harborview Medical Center R&T Building, Seattle, WA, USA

Hitoshi Sugiyama Department of Chronic Kidney Disease and Peritoneal Dialysis, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan

Kei Takemoto Department of Public Health, Graduate School of Medicine, Dentistry, and Pharmaceutical Sciences, Okayama University, Okayama, Japan

Weida Tong Division of Bioinformatics and Biostatistics, National Center for Toxicological Research (NCTR), FDA, US Food and Drug Administration, Jefferson, AR, USA

Yutaka Tonomura Drug Safety Evaluation, Research Laboratories for Development, Shionogi & Co., Ltd., Toyonaka, Osaka, Japan

Eda Tahir Turanli Molecular Biology and Genetics Department, Science and Letter Faculty, Istanbul Technical University, Istanbul, Turkey

Takeki Uehara Global Project Management Department, Shionogi and Co., Ltd., Kita-ku, Osaka, Japan

Akif Ündar Department of Pediatrics, Penn State Hershey Pediatric Cardiovascular Research Center, Penn State College of Medicine, Penn State Hershey Children's Hospital, Penn State Milton S. Hershey Medical Center, Hershey, PA, USA

Department of Surgery, Penn State Hershey Pediatric Cardiovascular Research Center, Penn State College of Medicine, Penn State Hershey Children's Hospital, Hershey, PA, USA

Department of Bioengineering, Penn State Hershey Pediatric Cardiovascular Research Center, Penn State College of Medicine, Penn State Hershey Children's Hospital, Hershey, PA, USA

Uğur Uygunoğlu Department of Neurology, Cerrahpasa School of Medicine, Istanbul University, Istanbul, Turkey

Mariano Valdés Department of Cardiology, Hospital Universitario Virgen de la Arrixaca, University of Murcia, Murcia, Spain

Silvia Ventura Servicio de Cardiología, Hospital Clínic Universitari, INCLIVA. Universidad de Valencia, Valencia, Spain

Boris Veytsman Center for the Study of Chronic Metabolic Diseases, School of System Biology, George Mason University, Fairfax, VA, USA

Juan Antonio Vílchez Department of Cardiology, Hospital Universitario Virgen de la Arrixaca, University of Murcia, Murcia, Spain

Department of Clinical Analysis, Hospital Universitario Virgen de la Arrixaca, University of Murcia, Murcia, Spain

Vassilis Voudris 2nd Division of Interventional Cardiology, Onassis Cardiac Surgery Center, Athens, Greece

Scott A. Waldman Department of Pharmacology and Experimental Therapeutics, Thomas Jefferson University, Philadelphia, PA, USA

Da-Hong Wang Department of Public Health, Graduate School of Medicine, Dentistry, and Pharmaceutical Sciences, Okayama University, Okayama, Japan

Yuping Wang Division of Bioinformatics and Biostatistics, National Center for Toxicological Research (NCTR), FDA, US Food and Drug Administration, Jefferson, AR, USA

Alexis C. Wong Department of Chemistry, Vanderbilt University, Nashville, TN, USA

David W. Wright Department of Chemistry, Vanderbilt University, Nashville, TN, USA

Chaochao Wu Environmental Molecular Sciences Laboratory, and Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

Guowang Xu Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, China

Halil Yaman Department of Clinical Biochemistry, School of Medicine, Gulhane Military Medical Academy, Etlik, Ankara, Turkey

Jeffrey D. Zahn Department of Biomedical Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

Jing Zhang Department of Pathology, University of Washington, UW Harborview Medical Center R&T Building, Seattle, WA, USA

Part I

General Aspects: Techniques and Overviews

High-Throughput Approaches to Biomarker Discovery and Challenges of Subsequent Validation

1

Boris Veytsman and Ancha Baranova

Contents

Abstract

Recently introduced high-throughput technologies are producing unprecedented volumes of biomedical data available for mining and analysis. The early predictions of the imminent breakthroughs in our understanding of human diseases and making predictive diagnostics easy, however, turned out to be largely over optimistic.

B. Veytsman
Center for the Study of Chronic Metabolic Diseases, School of System Biology, George Mason University, Fairfax, VA, USA
e-mail: borisv@lk.net; bveytsma@gmu.edu

A. Baranova (✉)
Center for the Study of Chronic Metabolic Diseases, School of System Biology, George Mason University, Fairfax, VA, USA

Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow, Russia
e-mail: abaranov@gmu.edu; aancha@gmail.com

V.R. Preedy, V.B. Patel (eds.), *General Methods in Biomarker Research and their Applications*, *Biomarkers in Disease: Methods, Discoveries and Applications*,
DOI 10.1007/978-94-007-7696-8_20

We argue that this situation is not coincidental, but rather is caused by the statistical properties of the data collected. A typical high-throughput biological dataset is deeply imbalanced: the data matrix includes many measured quantities or "levels" in a relatively small number of subjects. Thus, any attempt to analyze these datasets would be undermined by so-called "Dimensionality Curse" that may be solved by removing a majority of variables. The feature selection aimed at increasing the classification power may be done using data mining or correlation-based approaches. In this chapter, both theory-driven and data-driven approaches to deal with complexity in biological systems are discussed in details.

Key Facts of Biomarker Discovery and Validation

The finding of truly novel standalone biomarker with acceptable sensitivity and specificity for the detection of given disease is an extremely rare event.

The requisite traits of sensitivity and specificity are not inherent to the functioning of biological molecules but rather accidental.

Interindividual variability in the baseline levels of biomarkers is an inherent problem for biomarker-based detection of human pathologies.

The problem of relatively low sensitivity and specificity of newly discovered biomarkers is commonly solved by combining them into biomarker panels.

The typical sources for novel biomarkers to be incorporated into the biomarker panel are massive datasets produces by modern pipelines of biomarker discovery collectively known as OMICS approaches.

In many cases, the biomarker panels suffer from relatively low reproducibility of results when tested in independently collected sets of samples.

Typically, the lack of consistency in independently discovered sets of biomarkers is attributed to the differences in profiling technologies, underlying genetic variation in populations of patients, and variability in data normalization and other steps of the data processing.

An extraction of relevant information from the datasets with high dimensionality is a difficult task.

The Biomarkers: the Definition and the Conceptual Shortfall

> . . .while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician.
> Sir Arthur Conan Doyle, "The Sign of the Four" (1890)

Biomarkers are objective indicators of certain, often abnormal, biological states, including pathogenic processes, or pharmacologic responses to a therapeutic intervention. Biomarkers can serve many unique purposes, including screening for early

signs of the disease in community-based settings, confirmation of the diagnoses, monitoring effects of the treatments, or the progression of the disease and prediction of clinical outcomes.

Common perception of "biomarkers" implies that there are some biological molecules, relative concentrations of which may change due to, or in association with, pathogenic process. To date, the quantification of various molecules in biological fluids and tissues remains the primary mean to find novel biomarkers. However, the finding of truly novel standalone biomarker with acceptable sensitivity and specificity for the detection of given disease is an extremely rare event. The ideal molecular marker would be one that is inherently related to the pathogenic process. However, the requisite traits of sensitivity and specificity are not inherent to the functioning of biological molecules but rather accidental. Indeed, from the natural selection standpoint, it is difficult to imagine that these kinds of traits may be supported and improved. The latter is especially true for tumor biomarkers. In tumor-bearing body, any biomarker molecule expressed out of tissue context or overproduced by tumor cell may also moonlight as tumor antigen. Because of that, cells overexpressing the biomarker become a subject of strong negative selection in the microevolutionary process and got eliminated from tumor cell population.

Another inherent problem for biomarker-based detection of human pathologies is interindividual variability in the baseline levels of these biomarkers. Speaking generally, human populations are far from being homogeneous, both in its underlying genetics structure that is known to affect baseline expression of biomarker-encoding genes and in its environmental exposures that influence the prevalence of infraclinical or chronic illnesses in profiled individuals. Well-adapted reference interval is a prerequisite to proper interpretation of biomarker quantification results. However, it seems that in many cases this interval should be adjusted to age, gender, ethnicity, or BMI. Improper classification of laboratory readout as falling within the reference interval may lead to a false negative. The best example of this kind is an inverse correlation of prostate-specific antigen (PSA) and body mass index (BMI) that is further impacted by age (Gray et al. 2004). In obese candidates for curable treatment, i.e., patients in their fifth and sixth decades, the use of proper BMI-PSA adjustment of reference interval results in higher sensitivity in screening that alleviates misleadingly low measured PSA for early biopsy detection of prostate cancer (Hekal and Ibrahiem 2010).

Biomarker Panels

The conventional technique that overcomes the problem of relatively low sensitivity and specificity of newly discovered biomarkers is to combine them into biomarker panels. The logic under the assumption of better multi-analyte performance is as follows. Complex human diseases develop perturb more than one molecular network; if each of these networks would be represented by its own biomarker, the combined panel would be more robust. The typical sources for novel biomarkers to

be incorporated into the biomarker panel are massive datasets produces by modern pipelines of biomarker discovery collectively known as OMICS approaches. In short, these approaches aim at more or less precise quantitative measurement of as many same-class biomolecules as possible. In that, transcriptomics ascertains the mRNAs expressed within given tissue, proteomics – the proteins or, rather, the peptides comprising these proteins and metabolomics – the set of small molecules such as metabolic intermediates, messengers, and other compounds found within a biological sample.

However, in many cases, the biomarker panels suffer from relatively low reproducibility of results when tested in independently collected sets of samples. This is especially true for the mRNA biomarkers identified by microarray experiments. Additionally, when different research groups embark on discovery of biomarkers for the same disease, they rarely arrive on the same list of candidate molecules. In fact, the comparison of the predictive gene lists discovered by different groups revealed very small overlap. A striking example of this kind would provide a mere three-gene overlap between two well-regarded and, in one case, already commercialized, prognostic signatures for breast carcinoma, 76-gene identifier described by Wang et al. (2005), and 70-gene set MammaPrint (van 't Veer et al. 2002).

Typically, the lack of consistency in independently discovered sets of biomarkers is attributed to the differences in profiling technologies, underlying genetic variation in populations of patients, and variability in data normalization and other steps of the data processing. However, it seems that there are larger problems with existing approaches to high-throughput biomarker discovery that could not be shrug off to technical or even biological variation. One recent study showed that even the use of the same dataset may produce more than one gene list, sometimes of equal predictive power (Ein-Dor et al. 2005). The would-be biomarker panels composed of genes within these gene lists correlate with survival and cannot be truly distinguished from one another solely on their performance (that means that there were no true "leader" or "best performing" signature). When the signatures were tested over different subsets of patients, their relative performance scores fluctuated strongly (Ein-Dor et al. 2005). In other words, the robustness of the predictive gene signatures was low, and the membership in a prognostic list was not indicative of the involvement of analyte in the pathophysiology of the underlying disease.

The Perils of Combinatorial Approach to Biomarker Development

To understand the roots of the problems that result from combinatorial approach to discovery and validation of biomarkers, let us consider first the standard framework for diagnostic criteria. We plan to measure some parameter p (say, the serum level of certain biomolecule) that is somehow related to the disease D. Both the patient and the physician expect diagnostic guidance by binary answers: either "yes, you have D" or "no, you do not have this disease." We know that the elevated level of p signifies the disease, so if p is small, then the patient probably is free of D, and if it

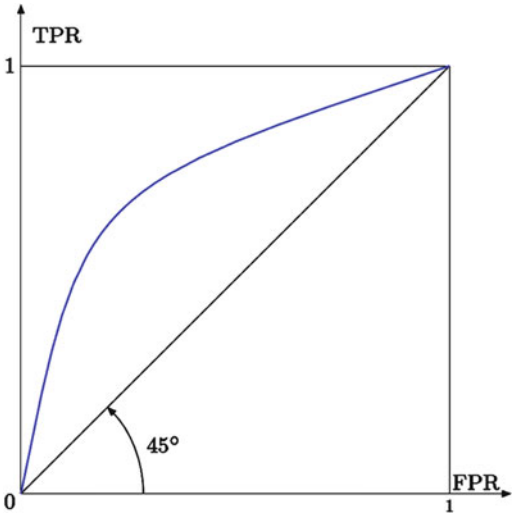

Fig. 1 A typical ROC curve. *TPR*: true-positive rate; *FPR*: false-positive rate

is large, then patient probably suffers from D. That means that continuously distributed levels of the biomarker molecule have to be dichotomized. We can quantify this in the following way: establish some cutoff value p_c such as patients with $p >= p_c$ are diagnosed with D and patients with $P < Pc$ are not.

In this approach there are two kinds of errors: we tell a patient she has not D, while in fact she has (a false negative), or we can tell a patient she has D, while she has not (a false positive). Accordingly we measure sensitivity or true-positive rate (TPR) of our test (one minus the probability to get a false negative) and its specificity (one minus the probability to get a false positive). The probability to get a false positive or false-positive rate (FPR) is therefore 100 % minus specificity.

It is easy to construct a 100 % sensitive test: let us just tell everybody they have D, so we never have a false negative. Conversely, a test that tells everybody, "you are absolutely healthy" has 100 % specificity. The problem is the first test is not specific at all (0 % specificity), and the second one is 0 % sensitive. Returning to the cutoff p_c, our first test corresponds to $p_c = -\infty$, while our second test to $p_c = \infty$. Sometimes the parameter p is defined in such way that $0 \leq p \leq 1$. In this case the 100 % sensitive test corresponds to $p_c = 0$, while the 100 % specific test corresponds to $p_c = 1$.

Obviously, both tests described above are totally impractical. One should prefer to utilize some "reasonable" value of p_c that would simultaneously provide for good specificity and good sensitivity. Of course, now we are stuck with the criteria of "goodness." One of the common approaches is based on the so-called receiver operating curve (ROC), which plots true-positive rate versus false-positive rate (Fig. 1).

To understand the use of this curve, let us consider the following test: suppose that instead of measuring biomarkers, we throw a dice and tell some patients that they have or not have the disease using a random guess. This test would randomly

classify the patients with no regard to their actual health, hence, the fraction of people with the disease would be the same in both groups, and the TPR of this test would be equal to its FPR. The ROC curve for this test is the straight line between the points (0,0) and (1,1) (the 45° line on Fig. 1).

However, clinicians shall hope that biomarker tests would perform better than just throwing a dice. This hope is reflected in expectation that either TPR of the test would be higher than that of the random test at the same FPR or FPR would be lower than that of the random test at the same TPR. In other words, the TPR vs. FPR curve would be drawn above the 45° line on Fig. 1. One can imagine a curve below this line: it describes a truly malicious test, which is worse than the random guess! We are not going to discuss such tests below.

The ROC curve must start in the point (0,0) and end in the point (1,1). On this curve, FPR $= 0$ corresponds to TPR $= 0$ and FPR $= 1$ corresponds to TPR $= 1$. In case of two different tests detecting the same disease, the test with an ROC curve that is completely above the ROC curve for another test is definitely better than the second: for every FPR, we achieved an increase in TPR. This argument is very straightforward and easy to understand. However, real ROC curves may not be convex (i.e., not all straight segments joining two points on the curve lie under it); that means that the test may be redesigned and improved. Indeed, let us choose two points corresponding to the parameter p values equal to p_1 and p_2. Then by randomly selecting either p_1 or p_2 as cutoffs for our prediction, we can obtain all points on the segment connecting these points. If the segment is above the curve connecting the points, this redesigned test is better than the original one. This shows that we need to consider only tests with a convex ROC above the 45° line.

There are two different problems related to the ROC framework. First, how to select the "better" one out of two non-convex tests and, thus, two ROC curves? Second, if we manage to select the "better" test, which cutoff value p_c – or, which is the same thing, which point on the ROC curve, should be chosen as a cutoff?

To solve the first problem, it is customary to compare areas under curve (AUC) defined as the areas between the curve and the 45° line. By convention we say that test A is better than test B if the AUC for test A is greater than that for test B. The ideal test would allow us to choose a cutoff value with 100 % sensitivity and 100 % specificity, so its ROC curve includes the point (0,1). There is only one convex curve between the points (0,0) and (1,1) that includes this point: the combination of two straight segments, one vertical and one horizontal. For this curve the area under the curve is 1/2. On the other hand for the fully random test the area is zero. For any other test AUC is between 0 and 1/2. For these tests, the selection of a cutoff always involves a trade between falsely classifying subjects into diseased or as non-diseased categories. The choice of the cutoff depends on the intended use of biomarker or panel of biomarkers, the population in which it is to be used, and the relative costs of making the error. Essentially, what may be an appropriate cutoff for a particular biomarker used for the screening of susceptible populations may be totally inappropriate when the same test is used to confirm diagnosis made by physician.

The criteria for choosing p_c depend on what exactly do we want to optimize. Dependent on intended application, we may choose to maximize an accuracy of the prediction or to minimize costs associated with false-positive or false-negative outcomes. Thorough review of traditional options can be found in the review by Bartlett et al. (2012) that utilized Alzheimer's disease diagnostics as an example.

The Perils of Feature Selection

Note that in this approach we implicitly assume that we know which biomarker to use for the diagnostics of the disease. In fact, the choice of the proper parameter to be inputted into the model (a candidate biomarker) is a separate and very difficult problem. In some cases, our understanding of the pathogenesis may help: if we know that anemia manifests in the lower count of red blood cells, then the count of these cells is a natural biomarker. In other cases we may try data mining: we can make a panel of putative tests, attempt to validate them all, and choose the one that is closely correlated with the disease. However, this latter approach suffers from the observation bias: every day many researchers attempt to observe some correlations, and only these that were actually observed end up in publications. Thus, when a large number of observations remain not reported, a good correlation might be just a statistical fluke that is due to so-called multiple comparison problem plaguing biomarker research. On the other hand, if a biomarker is not selected as model input, it is "lost" forever as it could not be retrieved later.

It is important to understand that the naïve idea "lets input them all" is not a proper solution. Attempts of to analyze the data with the dimensionality (the number of variable features) higher than the number of individual measurements for each feature may end up in so-called over-fitting of the model. Over-fit models may perfectly deal with the set of samples during the initial analysis, but do not perform in the independently collected sample sets. In fact, if the number of variables is high enough, a good separation of the classes may be achieved even for sets of classifiers chosen randomly (Venet et al. 2011). This problem is widely known as "dimensionality curse," and it is typically solved by removing a majority of variables, a feature selection that increases the classification power (Mayer et al. 2011; Saeys et al. 2007). This feature selection problem is paramount for high-throughput datasets where a researcher cannot intuitively grasp several thousand parameters. To aid an analysis, several algorithms help to identify and interpret the patterns within the data were developed, for example, principal components analysis, clustering, or multidimensional scaling. To develop multiplexed biomarkers tests, the visualization of the data is not required; however, it helps to gain confidence with a particular set of data.

Another way to explain the "curse of dimensionality" is to discuss the sparsity of data in a space of many dimensions.

Consider a panel of N_g biomarkers. We "train" the test on N_p patients. What is the probability that the data for a new, (N_g+1)-st patient are "close" to the data for some of the patients in the training set? To answer this question we need to define a

model of "closeness." Suppose each of our biomarkers varies between -1 and $+1$. We will define the "distance" as a simple Euclidean distance in the N_g-dimensional space and will define the patients "close" if the distance between the corresponding points is less than 1. For simplicity we will further assume that these points are uniformly distributed in the N_g-dimensional hypercube (the calculations for Gaussian distribution are more complex, but give the same result). The volume of the hypercube is 1. If we envelope each point in the training set in a sphere of radius 1, then the total volume of these spheres, not counting overlaps, is $N_p V_s$, where V_s is the volume of a unit ball in the N_g-dimensional sphere, equal to

$$V_s = \frac{\pi^{N_g/2}}{\Gamma\left(N_g/2+1\right)}$$

Γ being the Γ -function. Therefore, the probability is less than

$$P < N_p V_s$$

For a low dimensional space, the right-hand side of this equation is usually above 1. Indeed, two random circles of radius 1 almost always cover a unit square (Fig. 2a). However, the situation completely changes in highly dimensional spaces (Fig. 2b) due to the fact that Gamma function in the denominator of V_s grows much faster than the exponential function. In a 20-dimensional space $V_s = 0.026$, and we need more than three dozen nonoverlapping unit spheres to cover the unit square. For a 100-dimensional space $V_s = 2.4{*}10^{-40}$: there is no way two random points would "resemble" each other.

This means that the probability that a new patient would "resemble" any patient in the training set diminishes with an increase in the number of biomarkers in the panel and vanishes when it reaches the size that is typical for OMICS. For large N_g the volume of the unit ball becomes incredibly small (Fig. 2b). Hence, the

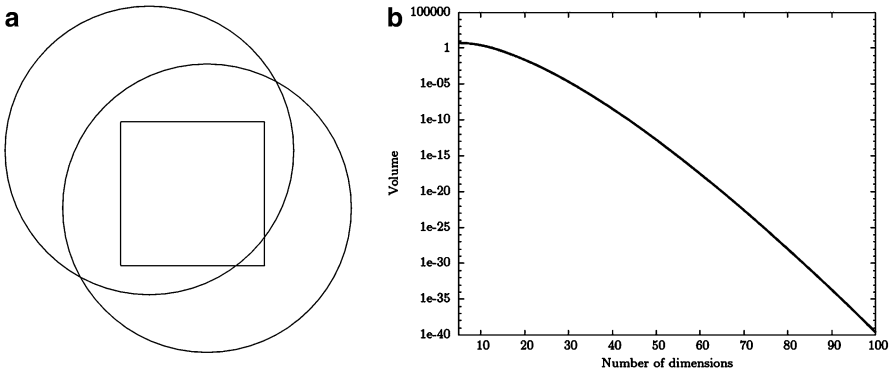

Fig. 2 (**a**) Two random unit circles completely cover unit square; (**b**) volume of a unit sphere in an N_g-dimensional space

probability that a new patient "resembles" any patient in the training set vanishes if the number of biomarkers in the panel is large.

Bayesian Approach to Deal with High Dimensionality

A modern approach to analyze datasets with high dimensionality is based on the Bayesian ideas. In that we start from the prevalence of the disease. If we do not perform any test, the *a priori* probability for a patient to have the disease D is equal to the prevalence (PR), and the *a priori* probability not to have the disease is $1 - \mathrm{PR}$. Suppose we chose the cutoff value p_c. It means that for $p \geq p_c$ we assume the test to be positive, and for $P < Pc$ we assume it to be negative. We can calculate the probability P_p to have the disease if the test is positive (sometimes called positive prediction value). Indeed, TPR and FPR are in fact the conditional probabilities to get the parameter $p \geq p_c$ if the patient has the disease or if she does not. Therefore, according to the general rules of Bayesian estimators Sinay (1992), the *a posteriori* probability to have a disease if the test is positive is

$$P_p = \frac{\mathrm{PR} \cdot \mathrm{TPR}}{\mathrm{PR} \cdot \mathrm{TPR} + (1 - \mathrm{PR}) \cdot \mathrm{FPR}} \tag{1}$$

Similarly the probability to *not* have the disease if the test is negative (*negative prediction value*) is

$$P_n = \frac{(1 - \mathrm{PR}) \cdot (1 - \mathrm{FPR})}{\mathrm{PR} \cdot (1 - \mathrm{TPR}) + (1 - \mathrm{PR}) \cdot (1 - \mathrm{FPR})} \tag{2}$$

We want to increase both positive and negative prediction values. One way to look at this is to associate costs with errors: suppose that the cost of treating the disease when it is in fact absent is c_n and the cost of *not* treating the disease when it is present is c_p. Then we want to minimize the function

$$F = \left(1 - P_p\right) \cdot c_p + (1 - P_n) \cdot c_n \rightarrow \min \tag{3}$$

What happens if instead of one parameter p we have n different parameters $p_1, p_2, p_3, \ldots, p_n$ for each patient? Geometrically this means using an n-dimensional vector p. How can we use this vector for the prediction?

The simplest idea is "flattening" the space. Let $f(\mathbf{p}) = f(p_1, p_2, \ldots, p_n)$ be a scalar function of n variables. Then we can just pretend this is our new parameter and use the one-dimensional theory for making prediction. So we need to find both the function f *and* the optimal cutoff f_c.

When the number of parameters is small, the choice is relatively easy: in the simplest case we just make a linear combination of parameters and choose the parameters in the way that produces the best results for the group of patients with the known diagnosis. This is the *training* of our test. After the training stage we get the optimal combination of measurements to apply to the new patients.

Many questionnaire-based tests use this idea for patients' screening. The medical practitioner fills the response sheet by checking the boxes, one box per symptom. The test instructions say, "A patient has the disease D if she has at least three of six symptoms below." In this case, each symptom is a discrete parameter with only two possible values (1 if symptom is observed, 0 otherwise). In most cases all coefficients in the linear combination are just 1: we count the number of parameters equal to 1. The result is simple and adequate enough for preliminary screening.

The Perils of Multiparametric Datasets Reduction

However, when the number of parameters becomes large, the situation changes dramatically. Suppose we can get expressions of several hundred thousand genes. We know these expression values for many healthy patients and many patients with the disease. Can we combine the expressions result into a predictive expression?

One of the approaches involves data mining: let us look at the measurement results and extract the most predictive combination. However, there is an important limitation for this approach. Information theory tells us that the amount of information we extract from the results obtained on N_p patients is proportional to N_p. A linear formula combining N_g gene expressions has the amount of information proportional to N_g. It means that to generate a reliable test *we need to initially profile many more patients than genes:* $N_p >> N_g$. These simple considerations were corroborated by the calculation by Ein-Dor et al. (2006), which leads to the same sad requirement – namely, thousands of patient's samples to be tested in order to deduce the robust list of biomarkers. In many cases, this luxury cannot be afforded. Even in case of more or less common diseases, like breast carcinoma, the collection of requested amounts of high-quality samples presents a substantial burden. For rare diseases, this approach may be simply not feasible.

However, there are certain techniques that allow to decrease the number of parameters we are about to input into the test. For example, we can measure the individual correlation of each candidate biomarker with the outcome in the training set and then shrink the biomarker list to include only those that have the highest correlations. We can look at the correlations between all the candidate biomarkers, and for each group of highly correlated parameters, leave only one "typical representative." For example, one may remove the expression levels for genes co-regulated by the same transcription factor (and leave the value for this master regulator), or delete all but one mass-spectrometry peaks that represent peptides derived from the same protein (Pyatnitskiy et al. 2010). These and others, even more sophisticated techniques, are reviewed by McDermott et al. (2013).

Let us suppose we successfully dealt with biomarker discovery phase by reducing and then ranking the list of features according to likelihoods they could serve as viable inputs into predictive models. However, the "dimensionality curse" discussed above is eager to produce one more nontrivial problem. Suppose we select two different training sets, both being drawn from the same set of patients

profiled using one or another high-throughput biomarker discovery platform. If the set of discovered biomarkers is robust, we would expect that both training datasets would produce comparable results. An experiment of this kind was performed by Ein-Dor et al. (2005) using a single breast cancer dataset that was analyzed by a single method. However, the training datasets were randomly assembled and different in each analysis trial. The outcome of this study was most frustrating: the resultant sets of biomarkers were not unique; in fact, they were strongly influenced by the subset of patients used for training. In other words, if we start from different groups of patients, we get completely different results.

There is a hope that the situation could be saved by hypothesizing that different sets of biomarkers are not "intrinsically different." Indeed, if two genes belong to the same pathway, then the changes in the expression levels for either of them could be useful as biomarkers reflecting the state of activation in this pathway. In other words, these genes are interchangeable as biomarkers: an anomaly in the expression of any of them signifies a problem with this pathway. This is akin to the typical representative method for highly correlated parameters: it does not matter which parameter from the group is chosen, since the parameters in the group are highly correlated.

If this hypothesis is true, then when we start from different training sets, we get different sets of genes in the tests, but the corresponding pathways must be roughly the same. This prediction is testable and was tested by Drier and Domany (2011). Here the authors took two different biomarker sets proposed for diagnostics of breast cancer. They identified the pathways and calculated the overlap between pathways discovered. As it could be expected, the proliferation pathway was present in both sets, a trivial finding at best. However, the overlap in other pathways was negligible. Hence, the robustness of the traditional techniques to discover reliable biomarkers in high-throughput manner remains very doubtful.

Of course, it is not clear whether the results obtained while using cancer datasets are directly applicable to other diseases. Still the results by Drier and Domany (2011) are disquieting. It seems that our current techniques are dealing with the noise in the samples rather than with the signal. In any case, it is clear that we are dealing with complex biological systems that built upon a multitude of the variables with unknown significance of their individual weights.

Theory-Driven and Data-Driven Approaches to Deal with Complex Systems

Speaking generally, there are two approaches to deal with a complex system: theory driven and data driven. In the first approach, we rely on our understanding of underlying processes to select variables that are most relevant to the process we study. In biological terms, that means that we attempt to discern suitable candidate biomarkers from non-robustly ranked lists of biological molecules by analyzing underlying biological pathways and selecting these most relevant to pathogenesis of the disease we study. Unfortunately, our knowledge of biological processes is far

from being perfect, and what we consider nonoverlapping pathways may turn out to be related, and we may miss suitable biomarker due to incorrectness of our judgment. Additionally, for some diseases we do not have any reliable information, a good example would be a genetic disease for which the causative gene has not been discovered yet. These considerations limit application of theory driven, also known as knowledge-based approaches for biomarker discovery.

In the second one we start with as little preconceptions as possible. Say, ancient physician would add the astrological information to his observations of symptoms. His understanding of the disease included the influence of stars and planets on its course. We, on the other hand, know that stars and planets are not relevant and thus exclude astrological data from the set of our parameters. As evident from above, both the ancient physician and the modern scientist adhered to the theory-driven approach. Their underlying theories were different, though. A purely data-driven approach would be to start with as much data as possible, including astronomical ephemerides, and let the correlations show that the latter are not relevant. At the first glance, this approach is a fallacy, as why should we include the data that we *know* are not relevant. We should bear in mind that the analyses we perform are not without costs, even if these costs are purely computational in its nature. However, data-driven or hypothesis-free approaches are very powerful as they truly do not require any data on intricate ropes that make biological systems tick.

While it is clear that the data-driven approach is indispensable in validating the theories, it is not so straightforward to use it for generating them. In one recent study, the usefulness of hypothesis-free approach was demonstrated for multidimensional mining of global collections of high-throughput public data that integrated, independently correlated, and ranked the data derived from over 4,000 experiments comprising 25,000 signatures (Kupershmidt et al. 2010). In this particular case, the replication of observed correlations across multiple independent datasets allowed researchers to generate a number of meaningful hypotheses concerning the development of brown adipose, a tissue compartment with high relevance to obesity, metabolic syndrome, and other human pathologies.

In short, to formulate a meaningful hypothesis that is relevant to a complex system, we need a huge amount of data. As discussed above, information theory tells us that the number of samples should be much greater than the dimensionality of the system. For biomarkers a sample is a patient, and dimensionality is the number of candidate genes. This means that data-driven approach requires huge training sets with thousands of patients (Ein-Dor et al. 2006). The bootstrapping methods of prefiltering the data cannot solve this problem. The situation is similar to that in thermodynamics: one can make very sophisticated thermal machines, but their efficiency still cannot exceed the theoretical limit set by the laws of thermodynamics. In the same way, while we can improve the performance, fundamental laws of information theory do not allow us to get meaningful conclusions about thousands of genes based on the data from hundreds of patients. This means that to get robust predictions we cannot use data alone: we must add some assumptions about underlying biological processes and blend them with the

data (McDermott et al. 2013). The quality of these assumptions is an important issue. There is a significant hope that it will improve with an accumulation of biological data and its subsequent interpretation.

To overcome these problems, many practical tests, the biomarker-based tests, combine laboratory measurements of certain analytes with demographic or other physically scorable parameters, for example, age, ethnicity, BMI, or the blood pressure. However, the heterogeneity of the dataset provides additional challenges. These new parameters may be highly correlated with the candidate biomarkers, and these correlations must be accounted for in the analysis as selection biases. For example, the probability that a person would seek medical help is closely correlated with social status, age, and often with ethnicity. Thus, demographic factors may provide misleading clues. As a general rule, the performance of good biomarker shall be consistent across genders and ethnic groups.

Conclusion

Harnessing the power of high throughput is widely used for the discovery of the next generation of biomarkers. Mining of various "omics" profiles also holds a significant promise to improve our understanding of the biology of health and disease. However, the road to this bright and shiny future is full of statistical traps that may preclude an extraction of relevant information from the datasets with high dimensionality. Those who embark on this journey should be aware of perils.

Summary Points

- This chapter focuses on the common pitfalls in biomarker discovery and validation.
- Biomarker panels suffer from relatively low reproducibility of results when tested in independently collected sets of samples.
- Proper application of ROC curves allows maximizing accuracy of the prediction or minimizing costs associated with false-positive or false-negative outcomes.
- Attempts of to analyze the data with the dimensionality (the number of variable features) higher than the number of individual measurements for each feature may end up in so-called over-fitting of the model.
- Complex biological systems are built upon a multitude of the variables with unknown significance of their individual weights.
- "Dimensionality curse" is typically solved by removing a majority of variables. This feature selection increases the classification power. Feature selection may be done using data mining or correlation-based approaches.
- Theory-driven and data-driven approaches to deal with complexity in biological systems are discussed.

Acknowledgment The authors express gratitude to the general support provided by College of Science, George Mason University, a State Contract 14.607.21.0098 dated November 27th, 2014 (Ministry of Science and Education, Russia) and by the Human Proteome Scientific Program of the Federal Agency of Scientific Organizations, Russia.

References

Bartlett JW, Frost C, Mattsson N, Skillbäck T, Blennow K, Zetterberg H, Schott JM. Determining cut-points for Alzheimer's disease biomarkers: statistical issues, methods and challenges. Biomark Med. 2012;6(4):391–400.

Drier Y, Domany E. Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? PLoS One. 2011;6(3):e17795. doi:10.1371/journal.pone.0017795. http://dx.doi.org/10.1371%2Fjournal.pone.0017795

Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 2005;21(2):171–8.

Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A. 2006;103(15):5923–8.

Gray MA, Delahunt B, Fowles JR, Weinstein P, Cookes RR, Nacey JN. Demographic and clinical factors as determinants of serum levels of prostate specific antigen and its derivatives. Anticancer Res. 2004;24:2069–72.

Hekal IA, Ibrahiem E. Obesity-PSA relationship: a new formula. Prostate Cancer Prostatic Dis. 2010;13(2):186–90.

Kupershmidt I, Su QJ, Grewal A, Sundaresh S, Halperin I, Flynn J, Shekar M, Wang H, Park J, Cui W, Wall GD, Wisotzkey R, Alag S, Akhtari S, Ronaghi M. Ontology-based meta-analysis of global collections of high-throughput public data. PLoS One. 2010;5(9):e13066. doi:10.1371/journal.pone.0013066. http://dx.doi.org/10.1371%2Fjournal.pone.0013066

Mayer G, Heinze G, Mischak H, Hellemons ME, Heerspink HJ, Bakker SJ, de Zeeuw D, Haiduk M, Rossing P, Oberbauer R. Omics-bioinformatics in the context of clinical data. Methods Mol Biol. 2011;719:479–97.

McDermott JE, Wang J, Mitchell H, Webb-Robertson BJ, Hafen R, Ramey J, Rodland KD. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. Expert Opin Med Diagn. 2013;7(1):37–51.

Pyatnitskiy M, Karpova M, Moshkovskii S, Lisitsa A, Archakov A. Clustering mass spectral peaks increases recognition accuracy and stability of SVM-based feature selection. J Proteomics Bioinform. 2010;3:048–54. doi:10.4172/jpb.1000120.

Saeys Y, Inza I, Larraaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507–17.

Sinay YG. Probability theory, an introductory course. Berlin/New York: Springer; 1992.

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530–6.

Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol. 2011;7(10):e1002240. doi:10.1371/journal.pcbi.1002240. http://dx.doi.org/10.1371%2Fjournal.pcbi.1002240

Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005;365(9460):671–9.