METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Computational Chemogenomics

Edited by

J.B. Brown

Life Science Informatics Research Unit, Laboratory of Molecular Biosciences,
Kyoto University Graduate School of Medicine, Kyoto, Japan

Humana Press

Editor
J.B. Brown
Life Science Informatics Research Unit
Laboratory of Molecular Biosciences
Kyoto University Graduate School of Medicine
Kyoto, Japan

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-8638-5 ISBN 978-1-4939-8639-2 (eBook)
https://doi.org/10.1007/978-1-4939-8639-2

Library of Congress Control Number: 2018952357

Cover Illustration: The cover image shows the inhibitory activity of compounds against aromatase, a critical hormone-processing enzyme in many organisms. Each point represents one compound. Green and yellow colors indicate highly weak or micromolar activity, red points represent strong activity, and purple points indicate single-digit nanomolar activity or stronger. Compounds are positioned by relative distance using multi-dimensional scaling. Activity cliffs can be seen where large changes in activity occur between closely spaced compounds, which are often analogs.

This Humana Press imprint is published by the registered company Springer Science+Business Media, LLC part of Springer Nature.
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Preface

This book provides a collection of techniques used in the emerging field of computational chemogenomics. It covers practical processes to execute research and analyses in the field, which is an integration of chemoinformatics, bioinformatics, computer science, statistics, automated pattern recognition and modeling, database usage with data retrieval, and systems integration. Clearly, to master the field of computational chemogenomics requires a considerable variety of knowledge and data processing skills, and this text hopes to get the interested reader acquainted with and capable of many of the practical skills used in the field. The target audience is both those from experimental sciences who are novices to data processing and modeling, and those with computationally oriented backgrounds wishing to engage in this scientific area, which is continually growing and now expected to contribute to industry, academic, and government research projects.

Historically, testing for chemical effects on biological processes, whether at the level of organism response, organ response (e.g., organ toxicity), cellular response (e.g., apoptosis), or individual target protein response in cell lines (e.g., inhibition), has required a large and orchestrated effort; confirmation of chemical purity, preparation of chemicals at a span of concentrations, application of those concentration-specific chemical stocks to the process or target, and precise recording of the outcome have typically been executed and recorded manually. At the same time, methods in genetic manipulation, gene sequence determination, gene expression measurement, and protein expression measurement have similarly required substantial investments in human resources and facilities.

The development of specialized equipment for automated high-content and high-throughput screening as well as parallel automation developments in genetics and proteomics made it possible to have chemical activity data for thousands of compounds instead of hundreds, as well as to expand measurement of gene expression from a few genes to tens, hundreds, or thousands. As a result, the technologies needed to systematically unlock the interface between chemistry and biology on a large scale had arrived. Finally, in 2001, worldwide efforts to create the first draft version of the full human genome were completed, and with such in hand, the stage was set to integrate the technologies for chemistry-biology interface exploration with our newfound knowledge about the genetic underpinnings of human physiology.

Only months after the sequencing of the human genome, the idea of exploring the protein products of a genome from a chemical perspective was proposed, and the term *chemogenomics* was born. This term bears resemblance to two other chemically driven scientific fields, and the reader should be aware of differences in terminology. First, scientists are also often in need of knowing the effect of a chemical on an organism when that organism contains a genetic defect such as a mutation or complete knockout, and this field is known as chemical genetics. Second, scientists may want to understand the functional impact of chemicals on coordinated processes occurring within cells encoded by genomes, for example the multiprotein signaling response to toxic chemicals measured in a variety of organisms. This field, chemical genomics, is more concerned with chemistry and genomics at a systems science level, compared with the chemogenomic focus of chemical modulation of individual proteins.

While research and development based on chemogenomics can be pursued in a variety of ways to ultimately reach project goals, two fundamental directions exist. First is the idea of forward chemogenomics. Much like the idea of forward genetics to identify the genes responsible for a phenotype or disease, forward chemogenomics seeks to identify a set of protein targets to test for chemical modulation in a biological system. Second, reverse chemogenomics is concerned with the identification of compounds which achieve the modulation desired by exerting their effect on the targets identified in the forward chemogenomic analysis. How to achieve these two goals is dictated by the state of the art in experimental methods for chemical and molecular biology research.

While advances in automation to enable chemogenomic-based science were being made, advances were simultaneously being made in computing and computer science-related fields. The CPUs used in workstations and servers were undergoing redesign to support multiple CPU cores, and operating systems and compilers designed to support multicore and multithread programming paved the way to program execution speed-ups of many fold. A key application area for the expanded power was statistics. Where analyses based on large amounts of repeated subsampling or expansive numbers of hypotheses were once prohibitive, they became mainstream and new methods for meta-analyses of results derived from basic statistical procedures gained attention. Leveraging statistical theory and advances in computing was the field of statistical pattern recognition, now commonly referred to as machine learning or artificial intelligence. Algorithms capable of modeling the patterns found in large, nonlinear datasets were shown to have extraordinary versatility, with applications in not only chemical and biological sciences but also physical sciences such as geology and meteorology, and applications in fields outside of natural science such as finance and music.

Hence, science arrived at a new frontier, with the vast quantities of data from automation used to inspire and give rise to chemogenomics, yet with requirements to develop the computing methods and infrastructure needed to harvest chemogenomic experimental results. The computational analyses should make the experimental results intelligible and should result in further hypotheses about living systems that could be validated. Born has been the field of *computational chemogenomics*. Interestingly, though chemogenomics has been driven by high-throughput methods and their computational analyses, accumulated efforts over several decades for structural biology have also contributed large numbers of publicly available three-dimensional crystal structures co-represented by the interaction of compounds with proteins; these now number in the tens of thousands, making structural computational chemogenomics a valuable option in the practitioner's toolbox.

Despite its relatively short history, the impact of computational chemogenomics is already considerably well established. Models for compound-protein interaction in drug discovery are a prominent application, as their ability to predict the interaction of a compound on a panel of targets has large implications for safety profiling, drug lead selection and optimization, and side effect predictions. A highly related application is chemical toxicity screening, which is concerned with chemical dose tolerance or dose lethality, and may incorporate target panel predictions as information to explain toxicity. The field of drug repurposing leverages chemogenomics and computational chemogenomics to suggest new targets for existing and often clinically approved drugs, which then might be applicable to new clinical indications. Still even further, computational chemogenomic methods may contribute to agrochemical sciences, where the organisms and their genomes under study are plants rather than animals. The concept of mining a chemical-protein activity matrix for knowledge discovery and hypothesis generation in agricultural life science is identical.

This volume on methods in computational chemogenomics is organized in a way that can be navigated by the reader in any order they wish. The first major unit covers the presentation of public chemogenomic data resources, where Nanjin *et al.* introduce how to use six different chemogenomics databases that each contain different focal points, and Kim *et al.* present a comprehensive in-depth tutorial on using the PubChem database, arguably the world's largest public chemogenomics information resource. In the second unit, the fundamentals of chemoinformatics, bioinformatics, and chemogenomic data processing are covered. In keeping with the discussion above on the importance of statistics, this unit contains a step-by-step tutorial on processing high-dimensional chemoinformatic data for basic statistical information and correlation in computer representation of compounds. The third unit is focused on techniques to analyze specific proteins or compounds based on their structures. Da Silva and Rognan present a robust workflow for analyzing protein surfaces when structural data is available, Song and Zhang demonstrate how to use resources dedicated to the cataloging and understanding of allosteric binding, Dimova and Bajorath detail methods for looking at the diversity of chemical structures in a large chemogenomic dataset, and Hu and Bajorath give the steps necessary to derive analyses indicating how small changes in scaffold decoration correlate to changes in panels of targets. In the fourth unit, statistical pattern recognition techniques are the focus. Yamanishi provides the reader the fundamental methods and knowledge needed for building custom methods of compound-protein matrix modeling, and Reker and Brown extensively detail the implementation of a new technique used for identifying points in the ligand-target matrix that result in predictive protein family models. The fifth and final unit is concerned with the future of chemogenomics and its application to medical care. Kou *et al.* describe their implementation of a clinical platform to analyze patient genomes and select chemical therapies based on the protein products of potentially altered genes. Jacoby and Brown conclude by discussing what computational chemogenomics has done so far, and what directions is it likely to pursue going forward.

This book is the culmination of many individuals dedicating their time and efforts toward its completion. I wish to express heartfelt thanks to all of the contributing authors, who sacrificed their limited time to describe their protocols in detail. Without their efforts, this book would not be possible. Continuous support by Springer to guide the completion of the book and handle unexpected situations during its development was key, with special thanks to series editors John Walker and Patrick Marton, and coordination by Anna Rakovsky. I also wish to thank colleagues at the Kyoto University Graduate School of Medicine and Kyoto University Hospital who have pushed me to new levels in order to perform chemogenomic research that is not only computationally attractive but equally helpful in translational research. A very special acknowledgement goes to Professor Dr. Jürgen Bajorath of the University of Bonn, who provided essential ideas and advice that played a major role in shaping the organization of the text. I also wish to thank Prof. Dr. Gisbert Schneider, Dr. Anthony Nicholls, Prof. Dr. Shunichi Takeda, and Prof. Dr. Yasushi Okuno for the various wisdoms that they imparted on me over the years of my career. Finally, my most sincere thanks goes to my wife, who accepted uncountable days and nights of canceled plans in order to allow me to concentrate on the completion of this text, as well as my children and my family, for with their understanding and support I draw motivation to push my scientific endeavors to new heights that can benefit society.

Kyoto, Japan ***J.B. Brown***

Contents

Contributors

Rasel Al Mahmud • *Department of Radiation Genetics, Graduate School of Medicine, Kyoto University, Kyoto, Japan*

Jürgen Bajorath • *Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany*

Evan E. Bolton • *Department of Health and Human Services, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*

J.B. Brown • *Life Science Informatics Research Unit, Laboratory of Molecular Biosciences, Kyoto University Graduate School of Medicine, Kyoto, Japan*

Stephen H. Bryant • *Department of Health and Human Services, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*

Viviana Consonni • *Department of Earth and Environmental Sciences, Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Milan, Italy*

Franck Da Silva • *CNRS, LIT UMR 7200, Université de Strasbourg, Strasbourg, France*

Dilyana Dimova • *Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany*

Francesca Grisoni • *Department of Earth and Environmental Sciences, Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Milan, Italy*

Ye Hu • *Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany*

Edgar Jacoby • *Janssen Research & Development, Beerse, Belgium*

Mayumi Kamada • *Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan*

Masashi Kanai • *Department of Therapeutic Oncology, Graduate School of Medicine, Kyoto University, Kyoto, Japan*

Sunghwan Kim • *Department of Health and Human Services, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*

Masaaki Kotera • *Department of Chemical System Engineering, School of Engineering, The University of Tokyo, Tokyo, Japan*

Tadayuki Kou • *Department of Therapeutic Oncology, Graduate School of Medicine, Kyoto University, Kyoto, Japan*

Shigemi Matsumoto • *Department of Therapeutic Oncology, Graduate School of Medicine, Kyoto University, Kyoto, Japan*

Manabu Muto • *Department of Therapeutic Oncology, Graduate School of Medicine, Kyoto University, Kyoto, Japan*

Rifat Ara Najnin • *Department of Radiation Genetics, Graduate School of Medicine, Kyoto University, Kyoto, Japan*

Masahiko Nakatsui • *Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan*
Yasushi Okuno • *Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University, Kyoto, Japan*
Ahsan Habib Polash • *Department of Radiation Genetics, Graduate School of Medicine, Kyoto University, Kyoto, Japan*
Christin Rakers • *Graduate School of Pharmaceutical Sciences, Yoshida-shimoadachicho, Kyoto University, Sakyo-ku, Kyoto, Japan; Graduate School of Science Nagoya University, Nagoya, Japan*
Daniel Reker • *Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA*
Didier Rognan • *CNRS, LIT UMR 7200, Université de Strasbourg, Strasbourg, France*
Benjamin A. Shoemaker • *Department of Health and Human Services, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*
Kun Song • *Department of Pathophysiology, Key Laboratory of Cell Differentiation and Apoptosis of Ministry of Education, Shanghai Jiao-Tong University School of Medicine, Shanghai, China*
Roberto Todeschini • *Department of Earth and Environmental Sciences, Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Milan, Italy*
Yoshihiro Yamanishi • *Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Iizuka, Fukuoka, Japan; PRESTO, Japan Science and Technology Agency, Kawaguchi, Saitama, Japan*
Jian Zhang • *Department of Pathophysiology, Key Laboratory of Cell Differentiation and Apoptosis of Ministry of Education, Shanghai Jiao-Tong University School of Medicine, Shanghai, China*