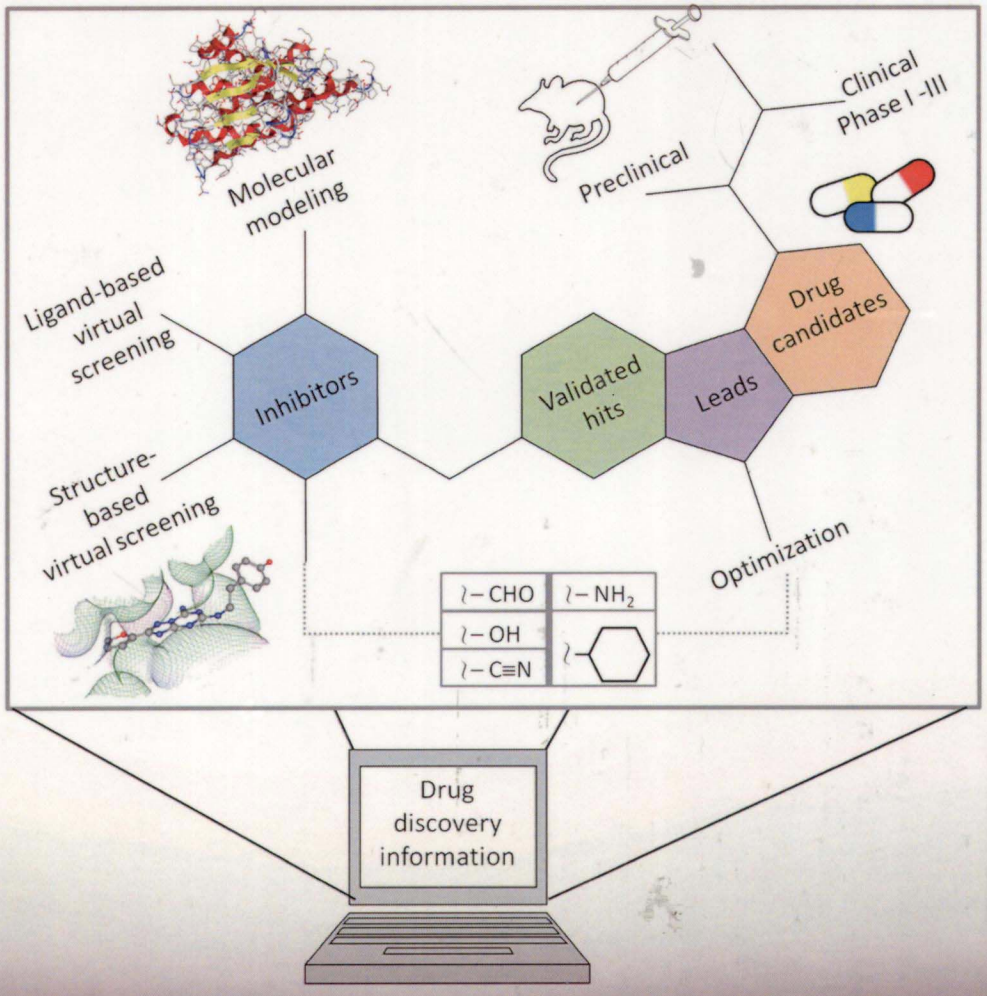
Chemoinformatics for Drug Discovery

Edited by Jürgen Bajorath



WILEY

CHEMOINFORMATICS FOR DRUG DISCOVERY

Edited by

JÜRGEN BAJORATH

WILEY

Cover design: John Wiley & Sons, Inc.
Cover image: Designed and provided by Dr. Dagmar Stumpfe

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

Library of Congress Cataloging-in-Publication Data

Chemoinformatics for drug discovery / edited by Jürgen Bajorath.
pages cm
Includes index.
ISBN 978-1-118-13910-3 (cloth)
1. Cheminformatics. 2. Drug development–Data processing. 3. Pharmacy informatics.
I. Bajorath, Jürgen, editor of compilation.
RS418.C482 2014
615.1′9–dc23

2013018927

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CHEMOINFORMATICS FOR DRUG DISCOVERY

PREFACE

Chemoinformatics: From methods and models to pharmaceutical applications

Chem(o)informatics is a relatively young and still evolving discipline, although some of its scientific origins can be traced back at least five decades. It continues to be challenging to clearly define chemoinformatics as a scientific field. Essentially, chemoinformatics uses algorithms and computational methods, often adapted from computer science, to organize and process chemical data, analyze and predict structure–property relationships of small molecules, and design compounds. Although chemoinformatics is not confined to questions and tasks that are relevant for pharmaceutical research, this field has firm roots in drug discovery. In fact, when the term chemoinformatics was first introduced in the literature in 1998 (Brown FK. Chemoinformatics: What is it and how does it impact drug discovery. *Ann. Rep. Med. Chem.* 1998;33:375–384), there was a strong focus on drug discovery research—and this has been a characteristic of this field ever since. Accordingly, the study of biological activities of chemical compounds and analysis of their structure–activity relationships (SARs) are hallmarks of chemoinformatics as we understand it today. As a consequence, methodological boundaries between chemoinformatics, computational chemistry, and drug design are rather fluid. In more specific terms, chemoinformatics has been defined to cover a wide range of scientific topics, from chemical data collection, management, and analysis to the exploration of SARs and prediction of compound activity or *in vivo* properties (Bajorath J. Understanding chemoinformatics: A unifying approach. *Drug Discov. Today* 2004;9:13–14). The scientific diversity of the field is high (Warr WA. Some trends in chem(o)informatics. *Meth. Mol. Biol.* 2011;672:1–37) and likely to even further increase, given the advent of research disciplines such as chemical biology or nanoscience, for which concepts from chemoinformatics are also relevant. Despite the presence of fluid scientific boundaries, characteristic features of chemoinformatics include its large-scale character (i.e., very large numbers of compounds and activity data are processed and analyzed) and its dual purpose of generating computational infrastructures and predictive models or data mining methods. Given its roots, another characteristic feature of chemoinformatics is that many important developments have originated from

pharmaceutical environments, in addition to research carried out in academia. It is evident that the pharmaceutical industry is the place where the need for chemoinformatics technologies and experts has been and continues to be the greatest. One should also note that the chemoinformatics literature is dominated by reports of computational methods and benchmark investigations, rather than practical applications. This is not very surprising, given that the majority of pharmaceutical applications are a part of drug discovery campaigns and hence proprietary (at the least for the duration of a discovery project). However, there clearly is a need to evaluate and better understand what chemoinformatics can actually accomplish in practical drug discovery situations. This need is not sufficiently met by the current scientific literature.

Having briefly introduced chemoinformatics as a scientific discipline, I should address the question why this book was originally planned and ultimately written. What was the prime motivation? Different from other currently available textbooks on chemoinformatics (there are not many), this book was envisioned to mostly (but not exclusively) focus on practical applications of chemoinformatics approaches in pharmaceutical research, hence addressing the need referred to earlier. It was intended to bring together leading experts from the pharmaceutical industry and selected academic institutions to describe the practice of chemoinformatics, illustrate the interplay between academic and pharmaceutical research, and showcase collaborations. Among others, key questions for authors included: What does chemoinformatics mean to you? How is it applied in your specific research environment? How does chemoinformatics contribute to pharmaceutical research? What works? What does not? Hence, special emphasis was put on expert views and experience values that might reflect the "true" impact of chemoinformatics approaches in drug discovery. In addition, a few selected methodological concepts were considered to further expand the spectrum of the presentations.

The 15 chapters presented herein include contributions from major pharmaceutical companies, a leading software firm, and several academic groups. They also cover collaborative efforts between academia and pharma. The chapters are arranged to follow a conceptual path from the description of methods and models to drug discovery applications and the design of chemoinformatics infrastructures. Hence, they span a wide range of topics.

Chapter 1 by W. Patrick Walters from Vertex presents a practical guide to the generation and evaluation of predictive models. It emphasizes common pitfalls in model building and assessment and shows how to avoid them. Many practical examples are provided including source code, which results in an instructive and much needed contribution. In Chapter 2 by Ajay Jain of the University of California at San Francisco, computational methods and models are considered from a principal point of view. The argument is made—and well supported—that the success of computational models often depends on the incorporation of sound physical principles (termed physical reality), although their consideration inevitably also introduces approximations. A number of well-selected methodological examples are presented.

Chapter 3 by Gerald M. Maggiora of the University of Arizona and collaborators of the Mayo Clinic and the Torrey Pines Institute for Molecular Studies reports the

adaptation of a new approach for chemoinformatics, that is, rough set theory, and discusses opportunities of this approach for drug discovery applications. In Chapter 4, Kiyoshi Hasegawa of the Chugai Pharmaceutical Company and Kimito Funatsu of the University of Tokyo also introduce new methodology. Their collaborative effort describes the application of the bimodal partial least-squares regression technique to analyze compound activity data by taking both ligand and target representations into account. Furthermore, in Chapter 5, Anthony Nicholls and Brian Kelley of OpenEye Scientific Software investigate search characteristics of different types of two-dimensional fingerprints, which are among the most popular molecular representations for chemical similarity searching and ligand-based virtual screening. Nicholls and Kelley pay particular attention to the way molecular similarity relationships are accounted for by different fingerprint representations and analyze how similarity assessment might be biased by fingerprints having high or low chemical resolution. On the basis of their findings, differences in search characteristic between fingerprints of alternative design can be rationalized. Practical implications of these results and possible methodological extensions are also discussed. Chapter 6, a contribution from our research group, further expands on ligand-based virtual screening, puts the approach into scientific context, and presents a critical assessment of practical virtual screening applications. Then, Meir Glick and colleagues of the Novartis Institutes for Biomedical Research, the authors of Chapter 7, describe a variety of applications of Bayesian modeling methods in drug discovery. Bayesian methods currently are among the most popular chemoinformatics approaches for compound classification, activity prediction, and target assignment. The topics discussed in this contribution include the analysis of phenotypic screening data and the prediction of off-target effects of drugs.

The contributions described thus far largely focus on approaches for the identification and characterization of active compounds. Once new active molecules have been identified, early-phase drug discovery projects transition into the hit-to-lead and lead optimization phases. Chapter 8 by Darren Green of GlaxoSmithKline and Matthew Segall of Optibrium Ltd. presents a thoughtful account of the evolution of lead optimization strategies and illustrates how different chemoinformatics concepts are adapted to aid in the optimization process. This contribution is very well complemented by Chapter 9 that reports on lead optimization collaborations between academia and the pharmaceutical industry. This work involved Valerie Gillet and Peter Willett of the University of Sheffield and George Papadatos et al. of GlaxoSmithKline. Here, the use of compound arrays for lead optimization is the major topic. A variety of chemoinformatics approaches have been employed to aid in the design of compound arrays and analyze progress made over time in lead optimization projects. This contribution also illustrates practical constraints involved in data assembly that affect medicinal chemistry projects and often work against a systematic and timely application of computational methods during lead optimization. In Chapter 10, Hans Matter and colleagues of Sanofi-Aventis further extend the lead optimization theme. They present a thorough and extensively referenced review of chemoinformatics methodologies for the analysis and prediction of SARs and demonstrate how such approaches have specifically been adapted for in-house applications.

The chapter also contains a discussion of methods to transfer SARs from one chemical series to another, which is a topic of high interest in medicinal chemistry.

The optimization of leads and generation of clinical candidates is a complex multiparametric process in which *in vivo* compound characteristics such as absorption, distribution, metabolism, extraction, and toxicology (ADMET) properties are as important as compound potency and specificity. The following two contributions address these issues. In Chapter 11, Karl-Heinz Baringhaus et al., also of Sanofi-Aventis, discuss how different types of computational ADMET models are generated and present a case study in which a model of human liver microsomal lability (a measure of metabolic instability of compounds) was derived for in-house use. Then, in Chapter 12, Scott Boyer and colleagues of AstraZeneca further expand the discussion of ADMET models with a focus on toxicology assessment. Their contribution also highlights the critically important role primary *in vivo* data play for predictive model building, given their sparseness and expected error margins. Both contributions cover a wide range of chemoinformatics methodologies for the derivation of ADMET models. With a concluding discussion of data delivery and communication issues, Chapter 12 also represents a transition point to another important thematic section of the book.

The contributions described thus far introduce scientific concepts, derive increasingly complex prediction models, and illustrate how such models are practically applied in drug discovery. As such, they represent a major category of chemoinformatics approaches in pharmaceutical research, that is, modeling and prediction of various compound properties. Another major category includes the design and implementation of computational infrastructures and information systems that is equally important for drug discovery as data mining and predictive modeling. In fact, pharmaceutical research environments heavily rely on the availability of specialized database structures and information systems to enable data warehousing with consistent deposition, distribution, access, and use across an organization. For large pharmaceutical companies, these requirements represent challenging tasks. The last two contributions in this book address these challenges. In Chapter 13, Nils Weskamp et al. describe how comprehensive chemoinformatics and database structures have been designed and implemented at Boehringer-Ingelheim. Here, it becomes clear that data archiving and handling is only a part of the equation—it is equally important to provide general access to modeling tools to, for example, analyze high-throughput screening data or characterize SARs. This presents considerable challenges for chemoinformaticians because such computational tools must not only be generated or adopted but also be made accessible to nonexpert users in the form of automated and easy-to-use workflows. In addition, results must be communicated in an intuitive and interpretable manner. Furthermore, in Chapter 14, Michael S. Lajiness and Thomas R. Hagadone of Eli Lilly and Company discuss lessons learned from over three decades of design and implementation of different generations of chemoinformatics systems for pharmaceutical research. These investigators are among the pioneers in building and maintaining such computational infrastructures in different company-specific environments. Their contribution illustrates how such systems have evolved, and continue to evolve, as computational resources and requirements rapidly change and data volumes and drug discovery demands further increase. On the basis of their

long experience, Lajiness and Hagadone comment on a number of practical aspects associated with system design that should be taken into consideration to ensure quality, accessibility, and utility of chemoinformatics infrastructures in drug discovery settings.

The book begins with chemoinformatics methodology and so it ends. To close the circle, in the final chapter (Chapter 15), José L. Medina-Franco of the Torrey Pines Institute for Molecular Studies and Gerald M. Maggiora of the University of Arizona describe foundations of molecular similarity analysis, one of the central themes in chemoinformatics. The evaluation and quantification of molecular similarity as an indicator of activity similarity is at the core of many chemoinformatics methods and an intensely investigated research topic to this date, conceptually linked to the design and navigation of chemical feature spaces.

Taken together, the contributions in this book highlight—from different points of view—key issues for the practice of chemoinformatics. The initial goals of this book project were quite ambitious and potential complications were expected. On the one hand, it was anticipated that it might be difficult for researchers in academia to present studies that are of high practical relevance for drug discovery; on the other hand, that it might be even more difficult for many investigators in the pharmaceutical industry to elaborate on details of their chemoinformatics work, given the proprietary nature of most of their projects. However, the chapters in this book have clearly exceeded initial expectations. Hence, I am very grateful to all authors who have spent their time and efforts to put together these excellent contributions! Without their early commitment and dedication, this project would not have been possible.

The contents of the book should be of interest to experts and practitioners in this field as well as to newcomers; there will be interesting materials for individuals with different motivations and levels of experience. Many of the questions that were initially asked have been answered in different ways and from different perspectives, which is highly desirable—after all, authors should have the last word.

Last but not least, given the critical expert views presented in this book and its practical drug discovery orientation, it is hoped that this publication will represent another important step forward in further defining and supporting chem(o)informatics as a scientific discipline at the interface between chemistry, computer science, and drug discovery.

JÜRGEN BAJORATH

CONTRIBUTORS

ERNST AHLBERG, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

SAMUEL ANDERSSON, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

JÜRGEN BAJORATH, Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany

KARL-HEINZ BARINGHAUS, R&D, LGCR, Structure, Design and Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

BERND BECK, Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

MICHAEL BIELER, Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

SCOTT BOYER, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

LARS CARLSSON, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

KIMITO FUNATSU, Department of Chemical System Engineering, University of Tokyo, Tokyo, Japan

VALERIE J. GILLET, Information School, University of Sheffield, Sheffield, UK

MEIR GLICK, Novartis Institutes for BioMedical Research, Cambridge, MA, USA

DARREN V. S. GREEN, GlaxoSmithKline Medicines Research Centre, Stevenage, Herts, UK

STEFAN GÜSSREGEN, R&D, LGCR, Structure, Design and Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

PETER HAEBEL, Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

THOMAS R. HAGADONE, Eli Lilly and Company, Indianapolis, IN, USA

KIYOSHI HASEGAWA, Chugai Pharmaceutical Company, Kamakura Research Laboratories, Kamakura, Kanagawa, Japan

CATRIN HASSELGREN, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

GERHARD HESSLER, R&D, LGCR, Structure, Design and Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

AJAY N. JAIN, Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA

BRIAN KELLEY, OpenEye Scientific Software, Inc., Santa Fe, NM, USA

PETER S. KUTCHUKIAN, Novartis Institutes for BioMedical Research, Cambridge, MA, USA

MICHAEL S. LAJINESS, Eli Lilly and Company, Indianapolis, IN, USA

EUGEN LOUNKINE, Novartis Institutes for BioMedical Research, Cambridge, MA, USA

CHRISTOPHER N. LUSCOMBE, GlaxoSmithKline, Medicines Research Centre, Stevenage, UK

GERALD M. MAGGIORA, College of Pharmacy and BIO5 Institute, University of Arizona, Tucson, AZ, USA; Translational Genomics Research Institute, Phoenix, AZ, USA

HANS MATTER, R&D, LGCR, Structure, Design and Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

IAIN M. MCLAY, The Open University, Cardiff, UK

JOSÉ LUIS MEDINA-FRANCO, Torrey Pines Institute for Molecular Studies, Port St. Lucie, FL, USA

NATHALIE MEURICE, Mayo Clinic, Scottsdale, AZ, USA

DANIEL MUTHAS, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

THORSTEN NAUMANN, R&D, LGCR, Structure, Design and Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

ANTHONY NICHOLLS, OpenEye Scientific Software, Inc., Santa Fe, NM, USA

TOBIAS NOESKE, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

GEORGE PAPADATOS, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, UK

JOACHIM PETIT, Mayo Clinic, Scottsdale, AZ, USA

STEPHEN D. PICKETT, GlaxoSmithKline, Medicines Research Centre, Stevenage, UK

FRIEDEMANN SCHMIDT, R&D, LGCR, Structure, Design and Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

MATTHEW SEGALL, Optibrium Ltd., Cambridge, UK

JONNA STÅLRING, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

DAGMAR STUMPFE, Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany

ANDREAS TECKENTRUP, Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

W. PATRICK WALTERS, Vertex Pharmaceuticals Incorporated, Cambridge, MA, USA

ALEXANDER WEBER, Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

NILS WESKAMP, Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

PETER WILLETT, Information School, University of Sheffield, Sheffield, UK

CONTENTS