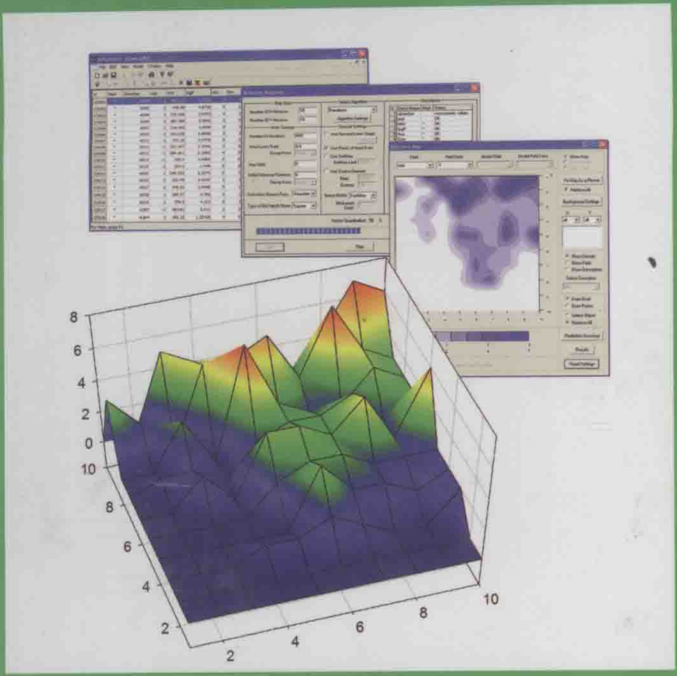
Wiley Series on Technologies for the Pharmaceutical Industry
Sean Ekins, Series Editor

Pharmaceutical Data Mining

APPROACHES AND APPLICATIONS FOR DRUG DISCOVERY



EDITED BY
KONSTANTIN V. BALAKIN

WILEY

PHARMACEUTICAL DATA MINING

Approaches and Applications for Drug Discovery

Edited by

KONSTANTIN V. BALAKIN
Institute of Physiologically Active Compounds
Russian Academy of Sciences

WILEY
A JOHN WILEY & SONS, INC., PUBLICATION

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

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Pharmaceutical data mining : approaches and applications for drug discovery / [edited by] Konstantin V. Balakin.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-19608-3 (cloth)
1. Pharmacology. 2. Data mining. 3. Computational biology. I. Balakin, Konstantin V.
[DNLM: 1. Drug Discovery–methods. 2. Computational Biology. 3. Data Interpretation, Statistical. QV 744 P5344 2010]
RM300.P475 2010
615′.1–dc22

2009026523

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PHARMACEUTICAL DATA MINING

Wiley Series On Technologies for the Pharmaceutical Industry

Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals
Edited by Sean Ekins

Pharmaceutical Applications of Raman Spectroscopy
Edited by Slobodan Šašić

Pathway Analysis for Drućg Discovery: Computational Infrastructure and Applications
Edited by Anton Yuryev

Drug Efficacy, Safety, and Biologics Discovery: Emerging Technologies and Tools
Edited by Sean Ekins and Jinghai J. Xu

The Engines of Hippocrates: From the Dawn of Medicine to Medical and Pharmaceutical Informatics
Barry Robson and O.K. Baek

Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery
Edited by Konstantin V. Balakin

PREFACE

Pharmaceutical drug discovery and development have historically followed a sequential process in which relatively small numbers of individual compounds were synthesized and tested for bioactivity. The information obtained from such experiments was then used for optimization of lead compounds and their further progression to drugs. For many years, an expert equipped with the simple statistical techniques of data analysis was a central figure in the analysis of pharmacological information. With the advent of advanced genome and proteome technologies, as well as high-throughput synthesis and combinatorial screening, such operations have been largely replaced by a massive parallel mode of processing, in which large-scale arrays of multivariate data are analyzed. The principal challenges are the multidimensionality of such data and the effect of "combinatorial explosion." Many interacting chemical, genomic, proteomic, clinical, and other factors cannot be further considered on the basis of simple statistical techniques. As a result, the effective analysis of this information-rich space has become an emerging problem. Hence, there is much current interest in novel computational data mining approaches that may be applied to the management and utilization of the knowledge obtained from such information-rich data sets. It can be simply stated that, in the era of post-genomic drug development, extracting knowledge from chemical, biological, and clinical data is one of the biggest problems. Over the past few years, various computational concepts and methods have been introduced to extract relevant information from the accumulated knowledge of chemists, biologists, and clinicians and to create a robust basis for rational design of novel pharmaceutical agents.

Reflecting the needs, the present volume brings together contributions from academic and industrial scientists to address both the implementation of new data mining technologies in the pharmaceutical industry and the challenges they currently face in their application. The key question to be answered by these experts is how the sophisticated computational data mining techniques can impact the contemporary drug discovery and development.

In reviewing specialized books and other literature sources that address areas relevant to data mining in pharmaceutical research, it is evident that highly specialized tools are now available, but it has not become easier for scientists to select the appropriate method for a particular task. Therefore, our primary goal is to provide, in a single volume, an accessible, concentrated, and comprehensive collection of individual chapters that discuss the most important issues related to pharmaceutical data mining, their role, and possibilities in the contemporary drug discovery and development. The book should be accessible to nonspecialized readers with emphasis on practical application rather than on in-depth theoretical issues.

The book covers some important theoretical and practical aspects of pharmaceutical data mining within five main sections:

- *a general overview of the discipline*, from its foundations to contemporary industrial applications and impact on the current and future drug discovery;
- *chemoinformatics-based applications*, including selection of chemical libraries for synthesis and screening, early evaluation of ADME/Tox and physicochemical properties, mining high-throughput screening data, and employment of chemogenomics-based approaches;
- *bioinformatics-based applications*, including mining the gene expression data, analysis of protein–ligand interactions, analysis of toxicogenomic databases, and vaccine development;
- *data mining methods in clinical development*, including data mining in pharmacovigilance, predicting individual drug response, and data mining methods in pharmaceutical formulation;
- *data mining algorithms, technologies, and software tools*, with emphasis on advanced data mining algorithms and software tools that are currently used in the industry or represent promising approaches for future drug discovery and development, and analysis of resources available in special databases, on the Internet and in scientific literature.

It is my sincere hope that this volume will be helpful and interesting not only to specialists in data mining but also to all scientists working in the field of drug discovery and development and associated industries.

Konstantin V. Balakin

ACKNOWLEDGMENTS

I am extremely grateful to Prof. Sean Ekins for his invitation to write the book on pharmaceutical data mining and for his invaluable friendly help during the last years and in all stages of this work. I also express my sincere gratitude to Jonathan Rose at John Wiley & Sons for his patience, editorial assistance, and timely pressure to prepare this book on time. I want to acknowledge all the contributors whose talent, enthusiasm, and insights made this book possible.

My interest in data mining approaches for drug design and development was encouraged nearly a decade ago while at ChemDiv, Inc. by Dr. Sergey E. Tkachenko, Prof. Alexandre V. Ivashchenko, Dr. Andrey A. Ivashchenko, and Dr. Nikolay P. Savchuk. Collaborations with colleagues in both industry and academia since are also acknowledged. My anonymous proposal reviewers are thanked for their valuable suggestions, which helped expand the scope of the book beyond my initial outline. I would also like to acknowledge Elena V. Bovina for technical help.

I dedicate this book to my family and to my wife.

CONTRIBUTORS

Shandar Ahmad, National Institute of Biomedical Innovation, 7-6-8, Saito-asagi, Ibaraki-shi, Osaka 5670085, Japan; Email: shandar@nibio.go.jp

Munazah Andrabi, National Institute of Biomedical Innovation, Ibaraki-shi, Osaka, Japan; Email: munazah@nibio.go.jp

Jürgen Bajorath, Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany; Email: bajorath@bit.uni-bonn.de

Konstantin V. Balakin, Institute of Physiologically Active Compounds of Russian Academy of Sciences, Severny proezd, 1, 142432 Chernogolovka, Moscow region, Russia; Nonprofit partnership «Orchemed», 12/1, Krasnoprudnaya ul., 107140 Moscow, Russia; Email: balakin@ipac.ac.ru, balakin@orchemed.com

Debra L. Banville, AstraZeneca Pharmaceuticals, Discovery Information, 1800 Concord Pike, Wilmington, Delaware 19850; Email: debra.banville@astrazeneca.com

Andrew Bate, Risk Management Strategy, Pfizer Inc., New York, New York 10017, USA; Department of Medicine, New York University School of Medicine, New York, NY, USA; Departments of Pharmacology and Community and Preventive Medicine, New York Medical College, Valhalla, NY, USA; Email: ajwb@mail.com

Elena V. Bovina, Institute of Physiologically Active Compounds of Russian Academy of Sciences, Severny proezd, 1, 142432 Chernogolovka, Moscow region, Russia; Email: bovina_e@ipac.ac.ru

John Bradshaw, Formerly with Daylight CIS Inc, Sheraton House, Cambridge UK CB3 0AX, UK.

Lyle D. Burgoon, Toxicogenomic Informatics and Solutions, LLC, Lansing, MI USA, P.O. Box 27482, Lansing, MI 48909; Email: burgoonl@txisllc.com

Jeremy S. Caldwell, Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, CA 92121, USA.

Sumit K. Chanda, Infectious and Inflammatory Disease Center, Burnham Institute for Medical Research, La Jolla, CA 92037, USA; Email: schanda@burnham.org

Elizabeth A Colbourn, Intelligensys Ltd., Springboard Business Centre, Stokesley Business Park, Stokesley, North Yorkshire, UK; Email: colbourn@intelligensys.co.uk

Ramona Rad-Curpan, Division of Biocomputing, MSC11 6145, University of New Mexico School of Medicine, University of New Mexico, Albuquerque NM 87131-0001, USA.

Pierre Darlu, INSERM U535, Génétique épidémiologique et structure des populations humaines, Hôpital Paul Brousse, B.P. 1000, 94817 Villejuif Cdedex, France; Univ Paris-Sud, UMR-S535, Villejuif, F-94817, France; Email: darlu@kb.inserm.fr

Matthew N. Davies, The Jenner Institute, University of Oxford, High Street, Compton, Berkshire, RG20 7NN, UK; Email: m.davies@mail.cryst.bbk.ac.uk

Darren R. Flower, The Jenner Institute, University of Oxford, High Street, Compton, Berkshire, RG20 7NN, UK.

Manfred Hauben, Risk Management Strategy, Pfizer Inc., New York, New York 10017 , USA; Department of Medicine, New York University School of Medicine, New York, NY, USA; Departments of Pharmacology and Community and Preventive Medicine, New York Medical College, Valhalla, NY, USA; Email: manfred.hauben@Pfizer.com

Christoph Helma, Freiburg Center for Data Analysis and Modelling (FDM), Hermann-Herder-Str. 3a, 79104Freiburg, Germany; In silico toxicology, Talstr. 20, 79102 Freiburg, Germany; Email: helma@in-silico.de

Yan A. Ivanenkov, Chemical Diversity Research Institute (IIHR), 141401, Rabochaya Str. 2-a, Khimki, Moscow region, Russia; Institute of Physiologically Active Compounds of Russian Academy of Sciences, Severny proezd, 1, 142432 Chernogolovka, Moscow region, Russia; Email: ivanenkov@ipac.ac.ru

Ludmila M. Khandarova, InformaGenesis Ltd., 12/1, Krasnoprudnaya ul., 107140 Moscow, Russia; Email: info@informagenesis.com

Frederick J. King, Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, CA 92121, USA; Novartis Institutes for BioMedical Research, Cambridge, MA 02139, USA.

David J. Livingstone, ChemQuest, Isle of Wight, UK; Centre for Molecular Design, University of Portsmouth, Portsmouth, UK; Email: davel@chemquestuk.com

Paolo Magni, Dipartimento di Informatica e Sistemistica, Universita degli Studi di Pavia, Via Ferrata 1, I-27100 Pavia, Italy; Email: paolo.magni@unipv.it

Andreas Maunz, Freiburg Center for Data Analysis and Modelling (FDM), Hermann-Herder-Str. 3a, 79104 Freiburg, Germany; Email: andreas@maunz.de

Kenji Mizuguchi, National Institute of Biomedical Innovation, 7-6-8, Saito-asagi, Ibaraki-shi, Osaka 5670085, Japan; Email: mizu-0609@kuc.biglobe.ne.jp

Chioko Nagao, National Institute of Biomedical Innovation, 7-6-8, Saito-asagi, Ibaraki-shi, Osaka 5670085, Japan.

Tudor I. Oprea, Division of Biocomputing, MSC11 6145, University of New Mexico School of Medicine, University of New Mexico, Albuquerque NM 87131-0001, USA; Sunset Molecular Discovery LLC, 1704 B Llano Street, S-te 140, Santa Fe NM 87505-5140, USA; Email: toprea@salud.unm.edu

Liliana Ostopovici-Halip, Division of Biocomputing, MSC11 6145, University of New Mexico School of Medicine, University of New Mexico, Albuquerque NM 87131-0001, USA.

Igor V. Pletnev, Department of Chemistry, M.V.Lomonosov Moscow State University, Leninskie Gory 1, 119992 GSP-3 Moscow, Russia; Email: pletnev@analyt.chem.msu.ru

Barry Robson, Global Pharmaceutical and Life Sciences 294 Route 100, Somers, NY 10589; The Dirac Foundation, Everyman Legal, No. 1G, Network Point, Range Road, Witney, Oxfordshire, OX29 0YN; Email: robsonb@us.ibm.com

Raymond C. Rowe, Intelligensys Ltd., Springboard Business Centre, Stokesley Business Park, Stokesley, North Yorkshire, UK; Email: rowe@intelligensys.co.uk

Audrey Sabbagh, INSERM UMR745, Université Paris Descartes, Faculté des Sciences Pharmaceutiques et Biologiques, 4 Avenue de l'Observatoire, 75270 Paris Cedex 06, France; Biochemistry and Molecular Genetics Department, Beaujon Hospital, 100 Boulevard Général Leclerc, 92110 CLICHY Cedex, France; Email: audrey.sabbagh@univ-paris5.fr

Alexey V. Tarasov, InformaGenesis Ltd., 12/1, Krasnoprudnaya ul., 107140 Moscow, Russia; Email: info@informagenesis.com

Andy Vaithiligam, St. Matthews University School of Medicine, Safehaven, Leeward Three, Grand Cayman Island.

Martin Vogt, Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany; Email: martin.vogt@bit.uni-bonn.de

Elizabeth A. Winzeler, Genomics Institute of the Novartis Research Foundation, San Diego, California and The Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA; Email: winzeler@scripps.edu

S. Frank Yan, frank.yan@roche.com

Yingyao Zhou, Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121, USA; Email: yzhou@gnf.org

CONTENTS

PART I

DATA MINING IN THE PHARMACEUTICAL INDUSTRY: A GENERAL OVERVIEW