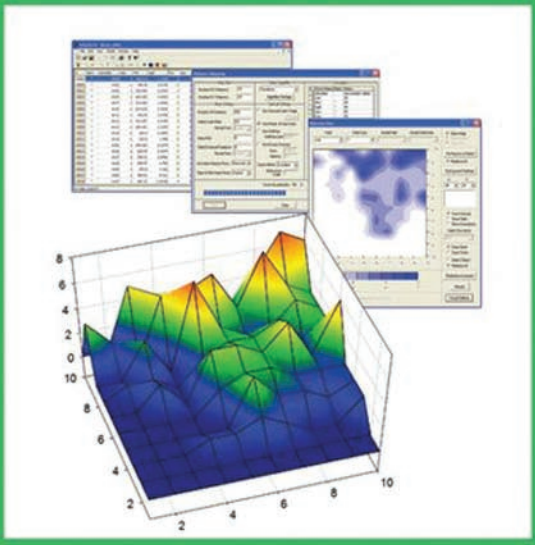
Wiley Series on Technologies for the Pharmaceutical Industry
Sean Ekins, Series Editor

Pharmaceutical Data Mining

APPROACHES AND APPLICATIONS FOR DRUG DISCOVERY



EDITED BY
KONSTANTIN V. BALAKIN

WILEY

PHARMACEUTICAL DATA MINING

Wiley Series On Technologies for the Pharmaceutical Industry

Sean Ekins, Series Editor

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

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

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

CONTENTS

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

PART I

DATA MINING IN THE PHARMACEUTICAL INDUSTRY: A GENERAL OVERVIEW

1

A HISTORY OF THE DEVELOPMENT OF DATA MINING IN PHARMACEUTICAL RESEARCH

David J. Livingstone and John Bradshaw

Table of Contents

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

1.1 INTRODUCTION

From the earliest times, chemistry has been a classification science. For example, even in the days when it was emerging from alchemy, substances were put into classes such as "metals." This "metal" class contained things such as iron, copper, silver, and gold but also mercury, which, even though it was liquid, still had enough properties in common with the other members of its class to be included. In other words, scientists were grouping together things that were related or similar but were not necessarily identical, all important elements of the subject of this book: data mining. In today's terminology, there was an underlying data model that allowed data about the substances to be recorded, stored, analyzed, and conclusions drawn. What is remarkable in chemistry is that not only have the data survived more than two centuries in a usable way but that the data have continued to leverage contemporary technologies for its storage and analysis.

In the early 19th century, Berzelius was successful in persuading chemists to use alphabetic symbols for the elements: "The chemical signs ought to be letters, for the greater facility of writing, and not to disfigure a printed book" [1]. This Berzelian system [2] was appropriate for the contemporary storage and communication medium, i.e., paper, and the related recording technology, i.e., manuscript or print.

One other thing that sets chemical data apart from other data is the need to store and to search the compound structure. These structural formulas are much more than just pictures; they have the power such that "the structural formula of, say, p-rosaniline represents the same substance to Robert B. Woodward say, in 1979 as it did to Emil Fischer in 1879" [3]. As with the element symbols, the methods and conventions for drawing chemical structures were agreed at an international level. This meant that chemists could record and communicate accurately with each other, the nature of their work.

As technologies moved on and volumes of data grew, chemists would need to borrow methodology from other disciplines. Initially, systematic naming of compounds allowed indexing methods, which had been developed for text handling and were appropriate for punch card sorting, to deal with the explosion of known structures. Later, graph theory was used to be able to handle structures directly in computers. Without these basic methodologies to store the data, data mining would be impossible.

The rest of this chapter represents the authors' personal experiences in the development of chemistry data mining technologies since the early 1970s.

1.2 TECHNOLOGY

When we began our careers in pharmaceutical research, there were no computers in the laboratories. Indeed, there was only one computer in the company and that was dedicated to calculating the payroll! Well, this is perhaps a slight exaggeration. A Digital Equipment Corporation (DEC) PDP-8 running in-

house regression software was available to one of us and the corporate mainframes were accessible via teleprinter terminals, although there was little useful scientific software running on them.

This was a very different world to the situation we have today. Documents were typed by a secretary using a typewriter, perhaps one of the new electric golf ball typewriters. There was no e-mail; communication was delivered by post, and there was certainly no World Wide Web. Data were stored on sheets of paper or, perhaps, punched cards (see later), and molecular models were constructed by hand from kits of plastic balls. Compounds were characterized for quantitative structure–activity relationship (QSAR) studies by using lookup tables of substituent constants, and if an entry was missing, it could only be replaced by measurement. Mathematical modeling consisted almost entirely of multiple linear regression (MLR) analysis, often using self-written software as already mentioned.

So, how did we get to where we are today? Some of the necessary elements were already in existence but were simply employed in a different environment; statistical software such as BMDP, for example, was widely used by academics. Other functionalities, however, had to be created. This chapter traces the development of some of the more important components of the systems that are necessary in order for data mining to be carried out at all.

1.3 COMPUTERS

The major piece of technology underlying data mining is, of course, the computer. Other items of technology, both hardware and software, are of course important and are covered in their appropriate sections, but the huge advances in our ability to mine data have gone hand in hand with the development of computers. These machines can be split into four main types: mainframes, general-purpose computers, graphic workstations, and personal computers (PCs).

1.3.1 Mainframes

These machines are characterized by a computer room or a suite of rooms with a staff of specialists who serve the needs of the machine. Mainframe computers were expensive, involving considerable investment in resource, and there was thus a requirement for a computing department or even division within the organizational structure of the company. As computing became available within the laboratories, a conflict of interest was perceived between the computing specialists and the research departments with competition for budgets, human resources, space, and so on. As is inevitable in such situations, there were sometimes "political" difficulties involved in the acquisition of both hardware and software by the research functions.

Mainframe computers served some useful functions in the early days of data mining. At that time, computing power was limited compared with the requirements of programs such as ab initio and even semi-empirical quantum chemistry packages, and thus the company mainframe was often employed for these calculations, which could often run for weeks. As corporate databases began to be built, the mainframe was an ideal home for them since this machine was accessible company-wide, a useful feature when the organization had multiple sites, and was professionally maintained with scheduled backups, and so on.

1.3.2 General-Purpose Computers

DEC produced the first retail computers in the 1960s. The PDP-1 (PDP stood for programmable data processor) sold for $120,000 when other computers cost over a million. The PDP-8 was the least expensive general-purpose computer on the market [4] in the mid-1960s, and this was at a time when all the other computer manufacturers leased their machines. The PDP-8 was also a desktop machine so it did not require a dedicated computing facility with support staff and so on. Thus, it was the ideal laboratory computer. The PDP range was superseded by DEC's VAX machines and these were also very important, but the next major step was the development of PCs.

1.3.3 Graphic Workstations

The early molecular modeling programs required some form of graphic display for their output. An example of this is the DEC GT40, which was a monochrome display incorporating some local processing power, actually a PDP-11 minicomputer. A GT40 could only display static images and was usually connected to a more powerful computer, or at least one with more memory, on which the modeling programs ran. An alternative lower-cost approach was the development of "dumb" graphic displays such as the Tektronix range of devices. These were initially also monochrome displays, but color terminals such as the Tek 4015 were soon developed and with their relatively low cost allowed much wider access to molecular modeling systems. Where molecular modeling was made generally available within a company, usually using in-house software, this was most often achieved with such terminals.

These devices were unsuitable, however, for displaying complicated systems such as portions of proteins or for animations. Dedicated graphic workstations, such as the Evans and Sutherland (E&S) picture systems, were the first workstations used to display the results of modeling macromolecules. These were expensive devices and thus were limited to the slowly evolving computational chemistry groups within the companies. E&S workstations soon faced competition from other companies such as Sun and, in particular, Silicon Graphics International Corporation (SGI). As prices came down and computing performance went up, following Moore's law, the SGI workstation became

the industry standard for molecular modeling and found its way into the chemistry departments where medicinal chemists could then do their own molecular modeling. These days, of course, modeling is increasingly being carried out using PCs.

1.3.4 PCs

IBM PCs or Apple Macintoshes gradually began to replace dumb terminals in the laboratories. These would usually run some terminal emulation software so that they could still be used to communicate with the large corporate computers but would also have some local processing capability and, perhaps, an attached printer. At first, the local processing would be very limited, but this soon changed with both the increasing sophistication of "office" suites and the usual increasing performance/decreasing price evolution of computers in general. Word processing on a PC was a particularly desirable feature as there was a word processing program running on a DEC VAX (MASS-11), which was nearly a WYSIWYG (what you see is what you get) word processor, but not quite! These days, the PC allows almost any kind of computing job to be carried out.

This has necessarily been a very incomplete and sketchy description of the application of computers in pharmaceutical research. For a detailed discussion, see the chapter by Boyd and Marsh [5].

1.4 DATA STORAGE AND MANIPULATION

Information on compounds such as structure, salt, melting point, molecular weight, and so on, was filed on paper sheets. These were labeled numerically and were often sorted by year of first synthesis and would be stored as a complete collection in a number of locations. The data sheets were also microfilmed as a backup, and this provided a relatively faster way of searching the corporate compound collection for molecules with specific structural features or for analogues of compounds of interest. Another piece of information entered on the data sheets was an alphanumeric code called the Wiswesser line notation (WLN), which provided a means of encoding the structure of the compound in a short and simple string, which later, of course, could be used to represent the compound in a computer record. WLN is discussed further in a later section.

Experimental data, such as the results of compound screening, were stored in laboratory notebooks and then were collated into data tables and eventually reports. Individual projects sometimes used a system of edge-notched cards to store both compound and experimental information. Figure 1.1 shows one of these edge-notched cards.

Edge-notched cards were sets of printed cards with usually handwritten information. Along the edge were a series of holes, which could be clipped to

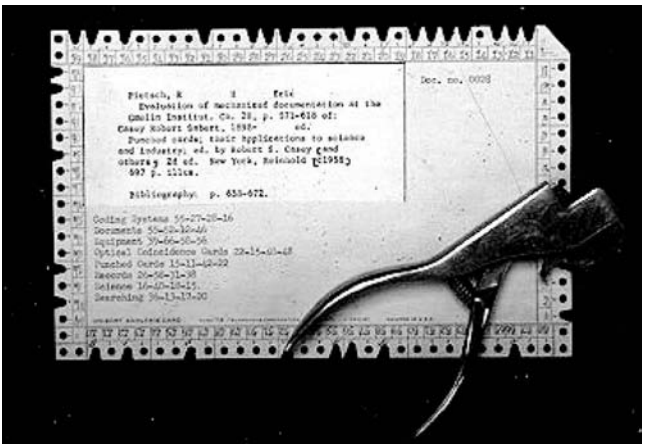


Figure 1.1 Edge-notched card and card punch.

form a notch. Each of these notches corresponded to some property of the item on the card. Which property corresponded to which notch did not matter, as long as all cards in a project used the same system. Then, by threading a long needle or rod through the hole corresponding to a desired property and by lifting the needle, all the cards that did *not* have that property were retained on the needle and were removed. (Note this is a principle applied to much searching of chemical data—first remove all items that could not possibly match the query.) The cards with a notch rather than a complete hole fall from the stack. Repeating the process with a single needle allows a Boolean "and" search on multiple properties as does using multiple needles. Boolean "or" search was achieved by combining the results of separate searches [6]. This method is the mechanical equivalent of the bit screening techniques used in substructure searching [7]. The limitations of storing and searching chemical information in this way are essentially physical. The length of the needle and the dexterity of the operator gave an upper limit to the number of records that could be addressed in a single search, although decks of cards could be accessed sequentially. There was no way, though, that all of the company compound database could be searched, and the results of screening molecules in separate projects were effectively unavailable. This capability would have to wait until the adoption of electronic databases.

1.5 MOLECULAR MODELING

Hofmann was one of the earliest chemists to use physical models to represent molecules. In a lecture at the Royal Society in 1865, he employed croquet balls as the atoms and steel rods as the bonds. To this day, modeling kits tend to use the same colors as croquet balls for the atoms. In the 1970s, models of

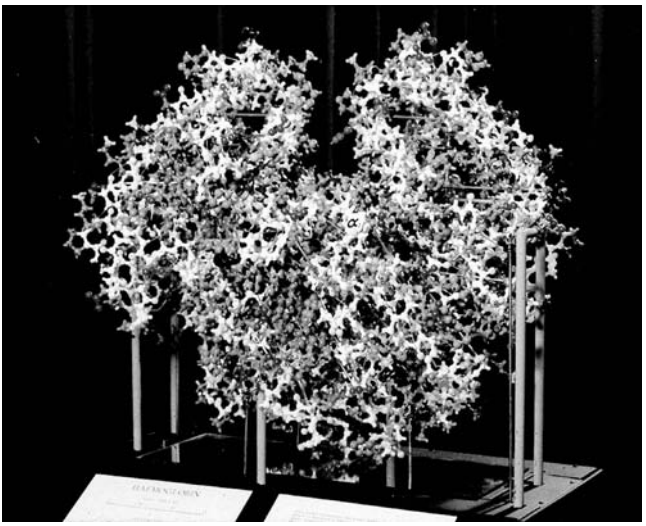


Figure 1.2 Physical model of hemoglobin in the deoxy conformation. The binding site for the natural effector (2,3-bisphosphoglycerate) is shown as a cleft at the top.

small molecules or portions of proteins used in the research laboratories were physical models since computer modeling of chemistry was in its infancy. An extreme example of this is shown in Figure 1.2, which is a photograph of a physical model of human hemoglobin built at the Wellcome research laboratories at Beckenham in Kent. This ingenious model was constructed so that the two α and β subunits were supported on a Meccano framework, allowing the overall conformation to be changed from oxy- to deoxy- by turning a handle on the base of the model. To give an idea of the scale of the task involved in producing this model, the entire system was enclosed in a perspex box of about a meter cube.

Gradually, as computers became faster and cheaper and as appropriate display devices were developed (see Graphic Workstations above), so molecular modeling software began to be developed. This happened, as would be expected, in a small number of academic institutions but was also taking place in the research departments of pharmaceutical companies. ICI, Merck, SKF, and Wellcome, among others, all produced in-house molecular modeling systems. Other companies relied on academic programs at first to do their molecular modeling, although these were soon replaced by commercial systems. Even when a third party program was used for molecular modeling, it was usually necessary to interface this with other systems, for molecular orbital calculations, for example, or for molecular dynamics, so most of the computational chemistry groups would be involved in writing code. One of the great advantages of having an in-house system is that it was possible to add any new technique as required without having to wait for its implementation by a software company. A disadvantage, of course, is that it was necessary

to maintain the system as changes to hardware were made or as the operating systems evolved through new versions. The chapter by Boyd gives a nice history of the development of computational chemistry in the pharmaceutical industry [8].

The late 1970s/early 1980s saw the beginning of the development of the molecular modeling software industry. Tripos, the producer of the SYBYL modeling package, was formed in 1979 and Chemical Design (Chem-X) and Hypercube (Hyperchem) in 1983. Biosym (Insight/Discover) and Polygen (QUANTA/CHARMm) were founded in 1984. Since then, the software market grew and the software products evolved to encompass data handling and analysis, 3-D QSAR approaches, bioinformatics, and so on. In recent times, there has been considerable consolidation within the industry with companies merging, folding, and even being taken into private hands. The article by Allen Richon gives a summary of the field [9], and the network science web site is a useful source of information [10].

1.6 CHARACTERIZING MOLECULES AND QSAR

In the 1970s, QSAR was generally created using tabulated substituent constants to characterize molecules and MLR to create the mathematical models. Substituent constants had proved very successful in describing simple chemical reactivity, but their application to complex druglike molecules was more problematic for a number of different reasons:

- It was often difficult to assign the correct positional substituent constant for compounds containing multiple, sometimes fused, aromatic rings.
- Missing values presented a problem that could only be resolved by experimental measurement, sometimes impossible if the required compound was unstable. Estimation was possible but was fraught with dangers.
- Substituent constants cannot be used to describe noncongeneric series.

An alternative to substituent constants, which was available at that time, was the topological descriptors first described by Randic [11] and introduced to the QSAR literature by Kier and Hall [12]. These descriptors could be rapidly calculated from a 2-D representation of any structure, thus eliminating the problem of missing values and the positional dependence of some substituent constants. The need for a congeneric series was also removed, and thus it would seem that these parameters were well suited for the generation of QSARs. There was, however, some resistance to their use.

One of the perceived problems was the fact that so many different kinds of topological descriptors could be calculated and thus there was suspicion that relationships might be observed simply due to chance effects [13]. Another objection, perhaps more serious, was the difficulty of chemical interpretation. This, of course, is a problem if the main aim of the construction of a QSAR

is the understanding of some biological process or mechanism. If all that is required, however, is some predictive model, then QSARs constructed using topological descriptors may be very useful, particularly when calculations are needed for large data sets such as virtual libraries [14,15].

One major exception to the use of substituent constants was measured, whole-molecule, partition coefficient (log *P*) values. The hydrophobic substituent constant, π, introduced by Hansch et al. [16], had already been shown to be very useful in the construction of QSARs. The first series for which this parameter was derived was a set of monosubstituted phenoxyacetic acids, but it soon became clear that π values were not strictly additive across different parent series, due principally to electronic interactions, and it became necessary to measure π values in other series such as substituted phenols, benzoic acids, anilines, and so on [17]. In the light of this and other anomalies in the hydrophobic behavior of molecules, experimental measurements of log *P* were made in most pharmaceutical companies. An important resource was set up at Pomona College in the early 1970s in the form of a database of measured partition coefficients, and this was distributed as a microfiche and computer tape (usually printed out for access) at first, followed later by a computerized database. Figure 1.3 shows a screen shot from this database of some measured values for the histamine H2 antagonist tiotidine.

The screen shot shows the Simplified Molecular Input Line Entry System (SMILES) and WLN strings, which were used to encode the molecular struc-

```
SMILES      CNC(NCCSCc1csc(N=C(N)N)n1)=NC#N                          1
  MOLFORM     C10H16N8S2                                             5
  WLN         T5N CSJ BNUYZZ E1S2MYM1&UNCN                           6
    LOCAL NAME  TIOTIDINE                                            7
    LOGP        0.67                                                 8
    SOLV PAIR   Octanol
    REFERENCE   Livingston,D. & Hill,A., Wellcome Research, U.K.,
                Private Communication
    FOOTNOTE    Not ion-corrected.
    ...  2      Borate buffer
    SELECTED    *
    pH          9.2
    LOGPSTAR    0.67                                                 9
    LOGP        0.68                                                10
    SOLV PAIR   Octanol
    REFERENCE   Taylor, P., Ici Pharmaceuticals, Private Communication
More:
```

Figure 1.3 Entry from the Pomona College log *P* database for tiotidine.

ture (see later) and two measured log *P* values. One of these has been selected as a log *P* "star" value. The "starlist" was a set of log *P* values that were considered by the curators of the database to be reliable values, often measured in their own laboratories. This database was very useful in understanding the structural features that affected hydrophobicity and proved vitally important in the development of the earliest expert systems used in drug research—log *P* prediction programs. The two earliest approaches were the fragmental system of Nys and Rekker [18], which was based on a statistical analysis of a large number of log *P* values and thus was called reductionist, and the alternative (constructionist) method due to Hansch and Leo, based on a small number of measured fragments [19]. At first, calculations using these systems had to be carried out by hand, and not only was this time-consuming but for complicated molecules, it was sometimes difficult to identify the correct fragments to use. Computer programs were soon devised to carry out these tasks and quite a large number of systems have since been developed [20,21], often making use of the starlist database.

Theoretical properties were an alternative way of describing molecules, and there are some early examples of the use of quantities such as superdelocalizability [22] and Ehomo [23,24]. It was not until the late 1980s, however, that theoretical properties began to be employed routinely in the creation of QSARs [25]. This was partly due to the increasing availability of relatively easy-to-use molecular orbital programs, but mostly due to the recognition of the utility of these descriptors. Another driver of this process was the fact that many pharmaceutical companies had their own in-house software and thus were able to produce their own modules to carry out this task. Wellcome, for example, developed a system called PROFILES [26] and SmithKline Beecham added a similar module to COSMIC [27]. Table 1.1 shows an early example of the types of descriptors that could be calculated using these systems.

Since then, the development of all kinds of descriptors has mushroomed until the situation we have today where there are thousands of molecular properties to choose from [29,30], and there is even a web site that allows their calculation [31].

The other component of the creation of QSARs was the tool used to establish the mathematical models that linked chemical structure to activity. As already mentioned, in the 1970s, this was almost exclusively MLR but there were some exceptions to this [32,33]. MLR has a number of advantages in that the models are easy to interpret and, within certain limitations, it is possible to assess the statistical significance of the models. It also suffers from some limitations, particularly when there are a large number of descriptors to choose from where the models may arise by chance [13] and where selection bias may inflate the values of the statistics used to judge them [34,35]. Thus, with the increase in the number of available molecular descriptors, other statistical and mathematical methods of data analysis began to be employed [36]. At first, these were the "regular" multivariate methods that had been developed and

TABLE 1.1 An Example of a Set of Calculated Properties (Reproduced with Permission from Hyde and Livingstone [28])

Calculated Property Set (81 Parameters, 79 Compounds)	
Whole-molecule properties	
"Bulk" descriptors	M.Wt., van der Waals' volume, dead space volume, collision diameter, approach diameter, surface area, molar refraction
"Shape" descriptors	Moment of inertia in *x*-, *y*-, and *z*-axes; principal ellipsoid axes in *x, y*, and *z* directions
Electronic and energy descriptors	Dipole moment; *x, y*, and *z* components of dipole moment; energies (total, core–core repulsion and electronic)
Hydrophobicity descriptors	Log *P*
Substituent properties	
For two substituents	Coordinates (*x, y*, and *z*) of the center, ellipsoid axes (*x, y*, and *z*) of the substituent
Atom-centered properties	
Electronic	Atom charges and nucleophilic and electrophilic superdelocalizability for atom numbers 1–14
Shape	Interatomic distances between six pairs of heteroatoms

applied in other fields such as psychology, but soon other newer techniques such as artificial neural networks found their way into the molecular design field [37]. As with any new technique, there were some problems with their early applications [38], but they soon found a useful role in the construction of QSAR models [39,40].

This section has talked about the construction of QSAR models, but of course this was an early form of data mining. The extraction of knowledge from information [41] can be said to be the ultimate aim of data mining. (See edge-notched cards above.)

1.7 DRAWING AND STORING CHEMICAL STRUCTURES

Chemical drawing packages are now widely available, even for free from the web, but this was not always the case. In the 1970s, chemical structures would be drawn by hand or perhaps by using a fine drawing pen and a stencil. The first chemical drawing software package was also a chemical storage system called MACCS (Molecular ACCess System) produced by the software company MDL, which was set up in 1978. MDL was originally intended to offer consultancy in computer-aided drug design, but the founders soon realized that their customers were more interested in the tools that they had developed

```
9  8  0  0  0  0  0  0  0  0  1 V2000
   12.3345   -6.8066    0.1462 C   0  0  0  0  0  0  0  0  0  0  0  0
   13.6295   -6.0048   -0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   14.7418   -6.8848    0.0703 O   0  0  0  0  0  0  0  0  0  0  0  0
   12.1293   -6.9337    1.1167 H   0  0  0  0  0  0  0  0  0  0  0  0
   12.4445   -7.6994   -0.2908 H   0  0  0  0  0  0  0  0  0  0  0  0
   11.5844   -6.3102   -0.2908 H   0  0  0  0  0  0  0  0  0  0  0  0
   13.7202   -5.3346    0.7366 H   0  0  0  0  0  0  0  0  0  0  0  0
   13.6625   -5.5368   -0.8831 H   0  0  0  0  0  0  0  0  0  0  0  0
   15.3830   -6.9423   -0.6949 H   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0  0  0  0
  3  2  1  0  0  0  0
  4  1  1  0  0  0  0
  5  1  1  0  0  0  0
  6  1  1  0  0  0  0
  7  2  1  0  0  0  0
  8  2  1  0  0  0  0
  9  3  1  0  0  0  0
```

Figure 1.4 Connection table for ethanol in the MDL mol file format.

for handling chemical information and so MACCS was marketed in 1979. MDL may justly be regarded as the first of the cheminformatics software companies.

MACCS allowed chemists to sketch molecules using a suitable graphics terminal equipped with a mouse or a light pen [42] and then to store the compound in a computer using a file containing the information in a format called a connection table. An example of a simple connection table for ethanol is shown in Figure 1.4. The connection table shows the atoms, preceded in this case by their 3-D coordinates, followed by a list of the connections between the atoms, hence the name. The MACCS system stored extra information known as keys, which allowed a database of structures to be searched rapidly for compounds containing a specific structural feature or a set of features such as rings, functional groups, and so on. One of the problems with the use of connection tables to store structures is the space they occupy as they require a dozen or more bytes of data to represent every atom and bond. An alternative to connection tables is the use of line notation as discussed below.

1.7.1 Line Notations

Even though Berzelius had introduced a system that allowed chemical elements to be expressed within a body of text, there was still a need to show the structure of a polyatomic molecule. Structural formulas became more common, and the conventions used to express them were enforced by international committees, scientific publications, and organizations, such as Beilstein and Chemical Abstracts. However, there were two areas where the contemporary technology restricted the value of structural formula.

First, in published articles, printing techniques often separated illustrative pictures from the text so authors attempted to put the formula in the body of the text in a line format. This gave it authority, as well as relevance to the surrounding text. Once you move away from linear formulas constrained to read left to right by the text in which they are embedded, you need to provide a whole lot of information like numbering the atoms to ensure that all the readers get the same starting point for the eye movement, which recognizes the structure. So linear representations continued, certainly as late as 1903, for structures as complicated as indigo [43]. Even today we may write C_6H_5OH. It has the advantage of being compact and internationally understood and to uniquely represent a compound, which may be known as phenol or carbolic acid in different contexts.

Second, organizations such as Beilstein and Chemical Abstracts needed to be able to curate and search the data they were holding about chemicals. Therefore, attempts were made to introduce systematic naming. So addressing the numbering issues alluded to above. Unfortunately, different organizations had different systematic names (Chemical Abstracts, Beilstein, IUPAC), which also varied with time so you needed to know, for instance, which Collective Index of Chemical Abstracts you were accessing to know what the name of a particular chemical was (see Reference 44 for details). The upside for the organization was that the chemical names, *within the organization*, were standard so they could use the indexing and sorting techniques already available for text to handle chemical structures. With the advent of punched cards and mechanical sorting, the names needed to be more streamlined and less dependent on an arbitrary parent structure, and thus there was a need for a linear notation system that could be used to encode any complex molecule.

Just such a system of nomenclature, known as WLN, had been invented by William Wisswesser in 1949 [45]. WLN used a complex set of rules to determine how a molecule was coded. A decision had to be made about what was the parent ring system, for example, and the "prime path" through the molecule had to be recognized. WLN had the advantage that there was only one valid WLN for a compound, but coding a complex molecule might not be clear even to experienced people, and disputes were settled by a committee. Even occasional users of WLN needed to attend a training course lasting several days, and most companies employed one or more WLN "experts." An example of WLN coding is shown below:

6-dimethylamino-*4-phenylamino*-naphthalene-2-sulfonic acid;

the WLN is

L66J *BMR&* DSWQ **INI&1**.

Here the four sections of the WLN have been separated by spaces (which does not happen in a regular WLN string) to show how the four sections of the

sulfonic acid, indicated by regular text, italic, underline, and bold, have been coded into WLN.

Beilstein, too, made a foray into line notations with ROSDAL, which required even more skill to ensure you had the correct structure. The corresponding ROSDAL code for the sulfonic acid above is

1=-5-=10=5,10-1,1-11N-12-=17=12,3-18S-19O,18=20O,18=21O, 8-22N-23,22-24.

Despite the complexity of the system and other problems [46], WLN became heavily used by the pharmaceutical industry and by Chemical Abstracts and was the basis for CROSSBOW (Computerized Retrieval Of StructureS Based On Wiswesser), a chemical database system that allowed susbstructure searching, developed by ICI pharmaceuticals in the late 1960s.

A different approach was taken by Dave Weininger, who developed SMILES in the 1980s [47,48]. This system, which required only five rules to specify atoms, bonds, branches, ring closures, and disconnections, was remarkable easy to learn compared to any other line notation system. In fact it was so easy to learn that "SMILES" was the reaction from anyone accustomed to using a line notation system such as WLN when told that they could learn to code in SMILES in about 10 minutes since it only had five rules. One of the reasons for the simplicity of SMILES is that coding can begin at any part of the structure and thus it is not necessary to determine a parent or any particular path through the molecule. This means that there can be many valid SMILES strings for a given structure, but a SMILES interpreter will produce the same molecule from any of these strings.

This advantage is also a disadvantage if the SMILES line notation is to be used in a database system because a database needs to have only a single entry for a given chemical structure, something that a system such as WLN provides since there is only one valid WLN string for a molecule. The solution to this problem was to devise a means by which a unique SMILES could be derived from any SMILES string [49]. Table 1.2 shows some different valid SMILES strings for three different molecules with the corresponding unique SMILES.

Thus, the design aims of the SMILES line notation system had been achieved, namely, to encode the connection table using printable characters but allowing the same flexibility the chemist had when drawing the structure and reserving the standardization, so the SMILES could be used in a database system, to a computer algorithm. This process of canonicalization was exactly analogous to the conventions that the publishing houses had instigated for structural diagrams. Thus, for the sulfonic acid shown earlier, a valid SMILES is **c1ccccc1Nc2cc(S(=O)(=O)O)cc3c2cc(N(C)C)cc3** and the unique or canonical SMILES is **CN(C)c1ccc2cc(cc(Nc3ccccc3)c2c1)S(=O)(=O)O**.

It was of concern to some that the SMILES canonicalizer was a proprietary algorithm, and this has led to attempts to create another linear representation,

TABLE 1.2 Examples of Unique SMILES

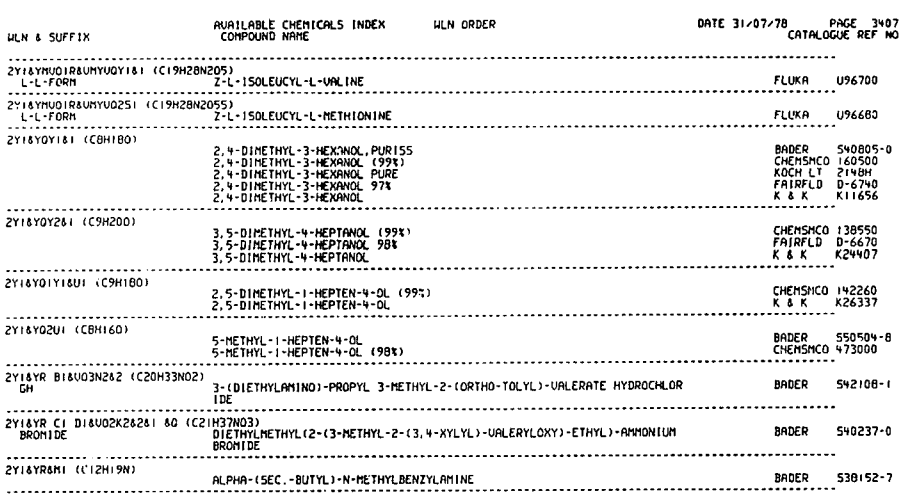
CH_3CH_2OH (1), CH_2=$CHCH_2CH$=$CHCH_2OH$ (2), 4-Cl-3Br-Phenol (3)		
Compound	SMILES	Unique SMILES
1	OCC	CCO
1	CC(O)	CCO
1	C(O)C	CCO
2	C=CCC=CCO	OCC=CCC=C
2	C(C=C)C=CCO	OCC=CCC=C
2	OCC=CCC=C	OCC=CCC=C
3	OC1C=CC(Cl)=C(Br)C=1	Oc1ccc(Cl)c(Br)c1
3	Oc1cc(Br)c(Cl)cc1	Oc1ccc(Cl)c(Br)c1
3	c(cc1O)c(Cl)c(Br)c1	Oc1ccc(Cl)c(Br)c1

International Chemical Identifier (InChI), initially driven by IUPAC and NIH (for details, see Reference 50).

1.8 DATABASES

Nowadays, we take databases for granted. All kinds of databases are available containing protein sequences and structures, DNA sequences, commercially available chemicals, receptor sequences, small molecule crystal structures, and so on. This was not always the case, although the protein data bank was established in 1971 so it is quite an ancient resource. Other databases had to be created as the need for them arose. One such need was a list of chemicals that could be purchased from commercial suppliers. Devising a synthesis of new chemical entities was enough of a time-consuming task in its own right without the added complication of having to trawl through a set of supplier catalogs to locate the starting materials. Thus, the Commercially Available Organic Chemical Intermediates (CAOCI) was developed. Figure 1.5 shows an example of a page from a microfiche copy of the CAOCI from 1978 [51]. The CAOCI developed into the Fine Chemicals Directory, which, in turn, was developed into the Available Chemicals Directory (ACD) provided commercially by MDL.

The very early databases were simply flat computer files of information. These could be searched using text searching tools, but the ability to do complex searches depended on the way that the data file had been constructed in the first place, and it was unusual to be able to search more than one file at a time. This, of course, was a great improvement on paper- or card-based systems, but these early databases were often printed out for access. The MACCS chemical database system was an advance over flat file systems since this allowed structure and substructure searching of chemicals. The original MACCS system stored little information other than chemical structures, but a combined data and chemical information handling system (MACCS-II) was soon developed.



AVAILABLE CHEMICALS INDEX WLN ORDER DATE 31/07/78 PAGE 3407

WLN & SUFFIX	COMPOUND NAME	CATALOGUE	REF NO
2Y1&YMUQ1R&UMYUQY1&1 (C19H28N2O5) L-L-FORM	Z-L-ISOLEUCYL-L-VALINE	FLUKA	U96700
2Y1&YMUQ1R&UMYUQ25I (C19H28N2O5S) L-L-FORM	Z-L-ISOLEUCYL-L-METHIONINE	FLUKA	U96680
2Y1&YQY1&I (C8H18O)	2,4-DIMETHYL-3-HEXANOL,PURISS	BADER	540805-0
	2,4-DIMETHYL-3-HEXANOL (99%)	CHEMSMCO	160500
	2,4-DIMETHYL-3-HEXANOL PURE	KOCH LT	2148H
	2,4-DIMETHYL-3-HEXANOL 97%	FAIRFLD	D-6740
	2,4-DIMETHYL-3-HEXANOL	K & K	K11656
2Y1&YQY2&I (C9H20O)	3,5-DIMETHYL-4-HEPTANOL (99%)	CHEMSMCO	138550
	3,5-DIMETHYL-4-HEPTANOL 98%	FAIRFLD	D-6670
	3,5-DIMETHYL-4-HEPTANOL	K & K	K24407
2Y1&YQ1Y1&U1 (C9H18O)	2,5-DIMETHYL-1-HEPTEN-4-OL (99%)	CHEMSMCO	142260
	2,5-DIMETHYL-1-HEPTEN-4-OL	K & K	K26337
2Y1&YQ2U1 (C8H16O)	5-METHYL-1-HEPTEN-4-OL	BADER	550504-8
	5-METHYL-1-HEPTEN-4-OL (98%)	CHEMSMCO	473000
2Y1&YR B1&UQ3N2&2 (C20H33NO2) GH	3-(DIETHYLAMINO)-PROPYL 3-METHYL-2-(ORTHO-TOLYL)-VALERATE HYDROCHLORIDE	BADER	542108-1
2Y1&YR C1 D1&UQ2K2&2&1 &Q (C21H37NO3) BROMIDE	DIETHYLMETHYL(2-(3-METHYL-2-(3,4-XYLYL)-VALERYLOXY)-ETHYL)-AMMONIUM BROMIDE	BADER	540237-0
2Y1&YR&M1 (C12H19N)	ALPHA-(SEC.-BUTYL)-N-METHYLBENZYLAMINE	BADER	538152-7
2Y1&YR&N1&U1N2&2 (C18H30N2O) GH	2-DIETHYLAMINO-N-METHYL-N-(2-METHYL-1-PHENYLBUTYL)-ACETAMIDE HYDROCHLORIDE	BADER	542234-7
2Y1&YR&UN2S1 (C15H23NOS)	3-METHYL-N-(2-(METHYLTHIO)-ETHYL)-2-PHENYLVALERAMIDE	BADER	538360-0
2Y1&YR&UN1&1 (C14H21NO)	2-PHENYL-N,N,3-TRIMETHYLVALERAMIDE	BADER	542120-0
2Y1&YR&UN2&2 (C16H25NO)	N,N-DIETHYL-3-METHYL-2-PHENYLVALERAMIDE	BADER	542317-3
2Y1&YR&UQY1&R D1 (C21H26O2)			

Figure 1.5 Entry (p. 3407) from the available chemical index of July 1978.

The great advance in database construction was the concept of relational databases as proposed by E.F. Codd, an IBM researcher, in 1970 [52]. At first, this idea was thought to be impractical because the computer hardware of the day was not powerful enough to cope with the computing overhead involved. This soon changed as computers became more powerful. Relational databases are based on tables where the rows of the table correspond to an individual entry and the columns are the data fields containing an individual data item for that entry. The tables are searched (related) using common data fields. Searching requires the specification of how the data fields should be matched, and this led to the development, by IBM, of a query "language" called Structured Query Language (SQL).

One of the major suppliers of relational database management software is Oracle Corporation. This company was established in 1977 as a consulting company, and one of their first contracts was to build a database program for the CIA code named "oracle." The adoption of a relational database concept and the use of SQL ensured their success and as a reminder of how they got started, the company is now named after that first project.

About 10 years ago, Oracle through its cartridges [53], along with other relational database providers such as Informix with its DataBlades [54], allowed users to add domain-specific data and search capability to a relational database. This is a key step forward as it allows chemical queries to be truly

integrated with searches on related data. So, for instance, one can ask for "all compounds which are substructures of morphine which have activity in test1 > 20 and log P < 3 but have not been screened for mutagenicity, and there is >0.01 mg available." The databasing software optimizes the query and returns the results. These technologies, while having clear advantages, have not been taken up wholesale by the pharmaceutical industry. Some of this is for economic reasons, but also there has been a shift in the industry from a hypothesis-testing approach, which required a set of compounds to be preselected to test the hypothesis [55], to a "discovery"-based approach driven by the ability to screen large numbers of compounds first and to put the intellectual effort into analyzing the results.

1.9 LIBRARIES AND INFORMATION

In the 1970s, each company would have an information (science) department whose function was to provide access to internal and external information. This broad description of their purpose encompassed such diverse sources as internal company reports and documents, the internal compound collection, external literature, patents both in-house and external, supplier's collections, and so on. Part of their function included a library that would organize the circulation of new issues of the journals that the company subscribed to, the storage and indexing of the journal collection and the access, through interlibrary loans, of other scientific journals, books, and information. Company libraries have now all but disappeared since the information is usually delivered directly to the scientist's desk, but the other functions of the information science departments still exist, although perhaps under different names or in different parts of the organization. The potential downside to this move of chemical information from responsibility of the specialists is that there is a loss of focus in the curation of pharmaceutical company archives. Advances in data handling in other disciplines no longer have a channel to be adapted to the specialist world of chemical structures. The scientist at his/her desk is not likely to be able to influence a major change in company policy on compound structure handling and so will settle for the familiar and will keep the status quo. This could effectively prevent major advances in chemical information handling in the future.

1.10 SUMMARY

From the pen and paper of the 19th century to the super-fast desktop PCs of today, the representation of chemical structure and its association with data has kept pace with evolving technologies. It was driven initially by a need to communicate information about chemicals and then to provide archives, which could be searched or in today's terminology "mined." Chemistry has

always been a classification science based on experiment and observation, so a tradition has built up of searching for and finding relationships between structures based on their properties. In the pharmaceutical industry particularly, these relationships were quantified, which allowed the possibility of predicting the properties of a yet unmade compound, totally analogous to the prediction of elements by Mendeleev through the periodic table. Data representation, no matter what the medium, has always been "backward compatible." For instance, as we have described, for many pharmaceutical companies, it was necessary to be able to convert legacy WLN files into connection tables to be stored in the more modern databases. This rigor has ensured that there is a vast wealth of data available to be mined, as subsequent chapters in this book will reveal.

REFERENCES

1. Berzelius JJ. Essay on the cause of chemical proportions and some circumstances relating to them: Together with a short and easy method of expressing them. *Ann Philos* 1813;2:443–454.
2. Klein U. Berzelian formulas as paper tools in early nineteenth century chemistry. *Found Chem* 2001;3:7–32.
3. Laszlo P. *Tools and Modes of Representation in the Laboratory Sciences*, p. 52. London: Kluwer Academic Publishers, 2001.
4. Web page of Douglas Jones of the University of Iowa. Available at http://www.cs.uiowa.edu/~jones/pdp8/.
5. Boyd DB, Marsh MM. *Computer Applications in Pharmaceutical Research and Development*, pp. 1–50. New York: Wiley, 2006.
6. Wikipedia contributors. Edge-notched card. *Wikipedia, The Free Encyclopedia*. Available at http://en.wikipedia.org/w/index.php?title=Edge-notched_card&oldid=210269872 (accessed May 12, 2008).
7. Weininger D, Delany JJ, Bradshaw J. *A Brief History of Screening Large Databases*. Available at http://www.daylight.com/dayhtml/doc/theory/theory.finger.html#RTFToC77 (accessed May 12, 2008).
8. Boyd DB. *Reviews in Computational Chemistry*, Vol. 23, pp. 401–451. New York: Wiley-VCH, 2007.
9. Richon AB. An early history of the molecular modeling industry. *Drug Discov Today* 2008;13:659–664.
10. Network Science web site and links thereon. Available at http://www.netsci.org/Science/Compchem/.
11. Randic M. On characterization of molecular branching. *J Am Chem Soc* 1975;97:6609–6615.
12. Kier LB, Hall LH. *Molecular Connectivity in Chemistry and Drug Research*. New York: Academic Press, 1976.
13. Topliss JG, Edwards RP. Chance factors in studies of quantitative structure-activity relationships. *J Med Chem* 1979;22:1238–1244.

14. Huuskonen JJ, Rantanen J, Livingstone DJ. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur J Med Chem* 2000;35:1081–1088.
15. Livingstone DJ, Ford MG, Huuskonen JJ, Salt DW. Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure. *J Comput Aided Mol Des* 2001;15:741–752.
16. Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 1962;194:178–180.
17. Fujita T, Iwasa J, Hansch C. A new substituent constant, π, derived from partition coefficients. *J Am Chem Soc* 1964;86:5175–5180.
18. Nys GC, Rekker RF. Statistical analysis of a series of partition coefficients with special reference to the predictability of folding of drug molecules. Introduction of hydrophobic fragmental constants (f values). *Eur J Med Chem* 1964;8:521–535.
19. Hansch C, Leo AJ. *Substituent Constants for Correlation Analysis in Chemistry and Biology*, pp. 18–43. New York: Wiley, 1979.
20. Livingstone DJ. Theoretical property predictions. *Curr Top Med Chem* 2003;3:1171–1192.
21. Tetko IV, Livingstone DJ. *Comprehensive Medicinal Chemistry II: In Silico Tools in ADMET*, Vol. 5, pp. 649–668. Elsevier, 2006.
22. Yoneda F, Nitta Y. Electronic structure and antibacterial activity of nitrofuran derivatives. *Chem Pharm Bull Jpn* 1964;12:1264–1268.
23. Snyder SH, Merril CR. A relationship between the hallucinogenic activity of drugs and their electronic configuration. *Proc Nat Acad Sci USA* 1965;54:258–266.
24. Neely WB, White HC, Rudzik A. Structure-activity relations in an imidazoline series prepared for their analgesic properties. *J Pharm Sci* 1968;57:1176–1179.
25. Saunders MR, Livingstone DJ. *Advances in Quantitative Structure-Property Relationships*, pp. 53–79. Greenwich, CT: JAI Press, 1996.
26. Glen RC, Rose VS. Computer program suite for the calculation, storage and manipulation of molecular property and activity descriptors. *J Mol Graph* 1987;5:79–86.
27. Livingstone DJ, Evans DA, Saunders MR. Investigation of a charge-transfer substituent constant using computer chemistry and pattern recognition techniques. *J Chem Soc Perkin 2* 1992;1545–1550.
28. Hyde RM, Livingstone DJ. Perspectives in QSAR: Computer chemistry and pattern recognition. *J Comput Aided Mol Des* 1988;2:145–155.
29. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Mannheim: Wiley-VCH, 2000.
30. Livingstone DJ. The characterisation of chemical structures using molecular properties—A survey. *J Chem Inf Comput Sci* 2000;40:195–209.
31. Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone DJ, Ertl P, Palyulin VA, Radchenko EV, Makarenko AS, Tanchuk VY, Prokopenko R. Virtual Computational Chemistry Laboratory. Design and description. *J Comput Aided Mol Des* 2005;19:453–463. Available at http://www.vcclab.org/.

32. Hansch C, Unger SH, Forsythe AB. Strategy in drug design. Cluster analysis as an aid in the selection of substituents. *J Med Chem* 1973;16:1217–1222.
33. Martin YC, Holland JB, Jarboe CH, Plotnikoff N. Discriminant analysis of the relationship between physical properties and the inhibition of monoamine oxidase by aminotetralins and aminoindans. *J Med Chem* 1974;17:409–413.
34. Livingstone DJ, Salt DW. Judging the significance of multiple linear regression models. *J Med Chem* 2005;48:661–663.
35. Salt DW, Ajmani S, Crichton R, Livingstone DJ. An Improved Approximation to the estimation of the critical F values in best subset regression. *J Chem Inf Model* 2007;47:143–149.
36. Livingstone DJ. Molecular design and modeling: Concepts and applications. In: *Methods in Enzymology*, Vol. 203, pp. 613–638. San Diego, CA: Academic Press, 1991.
37. Aoyama T, Suzuki Y, Ichikawa H. Neural networks applied to structure-activity relationships. *J Med Chem* 1990;33:905–908.
38. Manallack DT, Livingstone DJ. Artificial neural networks: Application and chance effects for QSAR data analysis. *Med Chem Res* 1992;2:181–190.
39. Manallack DT, Ellis DD, Livingstone DJ. Analysis of linear and non-linear QSAR data using neural networks. *J Med Chem* 1994;37:3758–3767.
40. Livingstone DJ, Manallack DT, Tetko IV. Data modelling with neural networks—Advantages and limitations. *J Comput Aided Mol Des* 1997;11:135–142.
41. Applications of artificial neural networks to biology and chemistry, artificial neural networks. In: *Methods and Applications Series: Methods in Molecular Biology*, Vol. 458. Humana, 2009.
42. http://depth-first.com/articles/2007/4 (accessed May 20, 2008).
43. Bamberger E, Elger F. Über die Reduction des Orthonitroacetophenons- ein Betrag zur Kenntis der ersten Indigosynthese. *Ber Dtsch Chem Ges* 1903;36: 1611–1625.
44. Fox RB, Powell WH. *Nomenclature of Organic Compounds: Principle and Practice*. Oxford: Oxford University Press, 2001.
45. Wisswesser WJ. How the WLN began in 1949 and how it might be in 1999. *J Chem Inf Comput Sci* 1982;22:88–93.
46. Bradshaw J. Introduction to Chemical Info Systems. Available at http://www.daylight.com/meetings/emug02/Bradshaw/Training/ (accessed May 12, 2008).
47. Weininger D. Smiles 1. Introduction and encoding rules. *J Chem Inf Comput Sci* 1988;28:31–36.
48. SMILES—A Simplified Chemical Language. Available at http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed May 25, 2008).
49. Weininger D, Weininger A, Weininger JL. SMILES 2. Algorithm for generation of unique SMILES notation. *J Chem Inf Comput Sci* 1989;29:97–101.
50. http://www.InChI.info/ (accessed May 25, 2008).
51. Walker SB. Development of CAOCI and its use in ICI plant protection division. *J Chem Inf Comput Sci* 1983;23:3–5.
52. Codd EF. A relational model of data for large shared data banks. *Commun ACM* 1970;13:377–387.

53. De Fazio S. Oracle8 Object, Extensibility, and Data Cartridge Technology. Available at http://www.daylight.com/meetings/mug98/DeFazio/cartridges.html (accessed June 5, 2008).
54. Anderson J. Taking Advantage of Informix DataBlade Technology. Available at http://www.daylight.com/meetings/mug98/Anderson/datablades.html (accessed June 5, 2008).
55. Bradshaw J. *Chronicles of Drug Discovery*, Vol. 3, pp. 45–81. Washington DC: ACS, 1993.

2

DRUG GOLD AND DATA DRAGONS: MYTHS AND REALITIES OF DATA MINING IN THE PHARMACEUTICAL INDUSTRY

BARRY ROBSON AND ANDY VAITHILIGAM

Table of Contents

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

2.1 THE PHARMACEUTICAL CHALLENGE

2.1.1 A Period of Transition

Healthcare and the pharmaceutical industry are going through a period of change and rapid evolution. *Evidence-based medicine* (EBM) [1] has been for some years promoted to replace physicians' textbooks, personal experience,

and anecdotal evidence by latest well-founded knowledge and statistical validation. In its most modern form, it represents a move to more *personalized medicine* [2] based on new patient data capture technologies and information technology (IT), becoming *information-based medicine*, a further evolution of EBM primarily centered on the electronic medical record [3]. In the movement toward personalized medicine, physicians will rely on the integration of phenotypic with genotypic data and on the identification of *patient cohorts* as subpopulations that are defined by shared genomic characteristics in particular.

Correspondingly, the biopharmaceutical industry is moving from its "blockbuster model" to a new *stratified medicine* or "nichebuster" model [4] based on *biomarkers*. In one white paper [5], it has been pointed out that this research and development (R&D) approach will require further serious efforts to address R&D productivity issues. There will be potential longer-term benefits [6] at more imminent cost [7]. The term biomarker has been coined for any parameter that helps distinguish a patient and actual or potential disease states. Though many authors include classical clinical descriptors such as gender, age, weight, blood pressure, and "blood work" results in that term, it is particularly used in regard to the new genomic [8–10] data, although also increasingly with medical imaging data [11]. The general idea is not simply to achieve better diagnosis but, in addition, to use the biomarkers as the clues for the best possible drug, meaning both at the level of pharmaceutical R&D and in regard to the physician's selection from currently marketed products.

At the same time, the growth of general understanding of bodily function and drug action in molecular terms offers hope that the emerging physicochemical principles can be applied not only to rationalize the relevance of individual molecular characteristics of genomic cohorts and even individual patients but also to exploit understanding of the mechanisms that determine how drug action is affected by them, that is to say, to develop methods with predictive power for more personalized drug design and for therapy selection. In general, we would like an engineering-quality level of understanding about the flow of information from the individual patient DNA to the patients' equally unique mixes of health and disease, as influenced by their own unique lifestyles and environments. To achieve this in atomic detail with the same capability as when a computational engineer architects a customized IT system, or when a structural engineer designs a bridge to specific needs and environment, is still probably a long way off. But we can at least try to ensure the best tools for analysis of empirical observations and to see what predictions can be made from them along with the use of such mechanistic understanding as we have available. From some perspectives, this is a critical first step with a potential for a considerable return on effort.

2.1.2 The Dragon on the Gold

A previous review [12] by one of the authors (B. Robson) is entitled "The Dragon on the Gold: Myths and Realities for Data Mining in Biotechnology

using Digital and Molecular Libraries," is a forerunner of the present review, and addresses analogous issues to those of the present paper, but now for the pharmaceutical industry. The gold relates to the *relevant* knowledge inherent in huge heaps of data, while the dragon relates to the "universe's protection" against human access to that knowledge by the combinatorial problems that arise when attempting to extract knowledge from high-dimensional data.

A reason for starting with the biotechnology industry was that many mathematical issues for data mining digital data also apply to molecular libraries in the form of nucleic acid libraries, notably expression arrays. It is insightful to consider this to highlight the relative difficulty that a pharmaceutical company (as opposed to a biotechnology company) faces. A query in IT is analogous to partial binding between nucleic acids, a fact that incidentally is of interest for a potential bio*nanotechnology* approach to storage and data mining [12]. The G–C, A–T, and A–U base pair binding has an effect remarkably like digital storage and information recovery, albeit in base 4 "quits," not base 2 "bits." Binding and recognition at the nucleic acid level even has a therapeutic role. For example, the potential use of small nuclear RNA interference (snRNAi) polynucleotides or simply RNAi depends on the particular use and administration of a competitive interference agent.

As a bonus, the biotechnology industry also has the benefit of the ability to use biological systems to select peptides and proteins and even program antibodies for specific binding. Continuing the RNA world example, antibodies against *small nuclear ribonucleoproteins* (snRNPs) may be useful in developing biotechnological products that target that complex in systemic lupus erythematosus patients. The required protein–nucleic acid interactions are far more complex than the digital code of binding between nucleic acid and nucleic acid via base pairing, but, despite advances in antibody engineering, there is for the most part no great need to have intimate and 100% appreciation of the physicochemical details at the point of binding. The immune system of an animal does most of the critical "design work" through the immune system process of selection and maturation and does an outstanding job (despite attempts to reproduce the effect with receptor-induced chemical assembly or phage display). The simple interaction code here is (immune system)–(target), which magically translates through the mysteries of nature to (antibody)–(target). It is "simply" required to harvest, clean up, and check quality to Food and Drug Administration (FDA) standards. If biotechnology's herding of antibodies can be compared to herding cattle, then to paraphrase the theme song of the old television cowboy series *Rawhide*, "Don't try to understand them, just rope them in and brand them." The analogy is not inappropriate considering that fields of appropriately immunized cows or sheep can do the job of the R & D division and of the first stage production plant.

Compared with the biotechnology industry, the pharmaceutical industry has it tougher, though the prizes are typically greater. It is constantly looking for drugs that are *small* organic molecules. Building a specific binding surface to order is much harder than in the case of nucleic acids and antibodies, though

this is outweighed by the benefits of success, notably the ability to market the therapeutic agent in a long-life, orally administrable pill form. A promising ability to bind and/or to activate a receptor comes mainly from discovery by selection, i.e., systematic screening of large numbers of compounds in bioassays, selective optimization of known compounds on new pharmacological targets, chemical modification of an existing lead or drug, or virtual docking (combined with attempts at rational *design* of drugs from knowledge of biomolecular mechanism and theoretical chemistry). *Recognition* information, i.e., molecular, down-to-atomic details about the drug and its protein target and the interactions between them, is a complex matter and the holy grail for rational development. This is even though screening compounds blindly against receptors, cells, and tissues, and so on, in large robotically controlled laboratories currently continues with relative success and hence popularity irrespective of a lack of detail at the atomic level [13]. Naturally, the same *general* physicochemical information principles that apply to the interaction between nucleic acids must obviously inevitably still apply to molecular recognition encoded into the interacting surfaces of ligands and proteins, the important basis of pharmacology. However, the latter interactions are more complexly encoded at the level of van der Waals, coulomb, and solvent-dependent interactions in a nonsimple way, whereas the same forces in the Crick–Watson base pairing allow us to address nucleic acid interaction at a kind of "digital storage level" as mentioned above.

In the pharmaceutical industry, the role of information inherent in nucleic acid is emerging strongly through the discipline of pharmacogenomics and personalized medicine, with focus on genomic biomarkers. Nonetheless, while laboratory detection of biomarkers depends on simple complementary pairing of nucleic acids, the final phenotypic effect is rarely conceivable in simple terms. Even for those genetic differences that most directly affect health and disease, it is necessary to consider a complicated interplay of many biomarkers, i.e., genomic and other factors. Although some thousands of genetic diseases are known due to single or very few base pair changes and single biomarkers can by themselves govern phenotypic effects such as eye color, these have become the exception rather than the rule. The complexity of human physiology makes it intrinsically unlikely that any one new biomarker would alone explain or reveal a particular clinical diagnosis. For the time being, we can assume the unknown variables together with the known genotypic information by identifying the phenotypic affect on a particular individual in terms of clinical affect. In practice, it is essential to reduce the dimensionality of the description to explore if all of the variables within this profile are strongly "in play." Typically, a large subset will remain strong players. In this regard, not only identification of, but also the relationship between, many biomarkers within a cohort will most often be more representative in elucidating a particular diagnosis. However, in addition to a bewildering abundance of useful data, the broad lens of the methods used captures many incidental biomarkers, which may be harmless distinguishing (ideotypic)

or irrelevant distinguishing features, distracting the medical professionals and encouraging the need for even more tests [14]. While many of the challenges can be met by sophisticated engineering in IT, the fundamental challenge posed by many parameters is a more fundamental and formidably mathematical one [15]. The huge quantity of new parameters describing patients pushes existing data analytic tools [16] to the limit, stimulating the quest for heuristic approaches [17] and even new mathematical strategies [18].

2.1.3 The Pharmaceutical Industry Is an Information Industry

Drucker, a leading management thinker of the 20th century, described the pharmaceutical industry as an information industry [19]. He noted that the value of the medicine lays not in the production and distribution of the final product, which is a relatively negligible cost, but in the knowledge from years of data sifting and R & D. He considered that the hierarchy from data through information to knowledge is the discovery and application of relationships, patterns, and principles between each stage [20,21]. It may be added that discovery in areas of little or known prior expectation is closely related to at least the R & D part of *innovation* [5]; otherwise, it would be merely the that testing of prior hunches and hypotheses constitutes *validation* (of those hunches and hypotheses). The challenge of innovation is that it is much easier to find that of which we have some knowledge than that of which we are completely ignorant. A theme of the present review is that the pharmaceutical industry process is indeed a sifting and processing of a very broad base of information to specific knowledge, and that classical statistics with its emphasis on formulating and testing hypotheses is *not* well suited to meet the challenges of pharmacology and biomedicine in the "post-genomic era."

Until recently, medicinal chemists have typically started with one or several lead compounds, and then utilized an optimization process to turn lead compounds into clinical candidates. This implies a restriction that can prohibit true discovery as discussed below, and in any event it looks like the "low-lying fruits" are running out. What does seem to be generally agreed is that most pharmaceutical R & Ds in 2007 start with a hypothesis picking a specific molecular target (meaning here not the type of chemistry to be achieved for drug molecule but the molecule in the body with which it will interact to trigger the required effect). This is usually a protein (more rarely DNA or RNA, and even more rarely saccharides, lipids, etc.). The idea is to try and realize four primary goals:

1. a drug molecule that affects the molecular target,
2. a method to deliver this drug to the protein target,
3. a choice of chemistry or vehicle that maintains effective drug concentration in the patient for a desired time, and
4. minimal interaction with other protein targets (avoidance of adverse effects).

Whatever the means by which the initial leads may be selected, there is still a needle-in-haystack problem. Each iteration usually includes a decision point concerning which molecules to make next: with D decision points each involving m molecules, there are D^m possible molecules to be synthesized and/or tested, which could often easily mean there are at least a billion. Ackoff and Emery [20,21] argue that the relevant knowledge to rationalize a path through this process constitutes *efficiency of choice* [20,21] and consists of *possession of facts* (or awareness of a state of affairs) and *possession of skills* [20,21]. The former, they consider, consists of *ontology*, i.e., about what entities exist and what statements about them are true; the latter is about how we can obtain knowledge about facts in the world and how can we ascertain their reliability. From a data analytic perspective, this classification is somewhat misleading. The two key issues are ontology and associations, which map to *universal* and *existential* qualification in higher-order human logic known as the predicate calculus (PC). Both at least most generally can be discovered from data analysis and both initially at least may involve uncertainty, though there is a tendency to assign ontology *a priori* as a self-evident or preset classification (taxonomy) of things based on human judgment, and association analysis as a more research skill-driven process of relationship discovery (in this process, strong associations may also emerge as an ontological relationship). As discussed below, the number of rules or guidelines that may *empirically* be extracted from N parameter data and that relate to ontologies and associations may be at least of the order of 2^N where extracted rules cannot necessarily be deduced from simpler rules and vice versa.

No route nor stage in the journey seems to avoid the huge numbers of options to consider. Though the ultimate challenge here is mathematical, physics instead could have been blamed. The inexorable laws of thermodynamics hold even broader sway across the industries. The automotive industry is ultimately an energy industry, progressing by respecting the laws of thermodynamics, not assuming that one may get more energy out than you put in, and not by working flat out to make the perpetual motion machine.

The pharmaceutical industry as an information industry is also thus by definition a *negative entropy* industry. It too is bound by the laws of thermodynamics. The pharmaceutical industry cannot ignore entropy (which is information with change of sign). As the information content of a pharmaceutical company increases, its entropy by definition decreases… all other factors being equal! Key rules here are that you cannot have all the information you could use without someone or something somehow working hard to get it, you rarely get it all, and, compared with energy that is conserved (though because of entropy, degradable to heat), information that was almost in your hands can easily get lost forever.

2.1.4 Biological Information

There has always been proof of concept that biologically active chemical agents for humans can be discovered. It is represented by the biochemistry

and molecular biology of all humans who have ever lived on Earth. The bad news is that getting the implied information took some 3.5 billion years of running a trial-and-error-based evolutionary-genetic algorithm running with very high parallelism to attain these human beings. That is perhaps a quarter of the age of the entire universe, depending on current estimates.

The numbers arise as follows. Since humans are to a first approximation rather similar, and since proteins including enzymes largely dictate the chemistry of the biological molecules in humans, then in principle the information content of all the biological molecules can be estimated through the proteome of a human by reference to the genes coding for them starting with a single human. For example, if in the human genome there are, say, 40,000 such genes and each had, say, an information content of 250 bits, then the total information content of man is 8×10^6 bits. We can go beyond the uniformity approximation. Since about two in every 1000 base pairs vary between humans in their DNA, we can estimate about 10^7–10^8 bits in the human race of recent history. The topic is of interest to astrophysicists since even higher information content is not beyond interstellar transmission. Their revised estimates indicate that between 10^{13} and 10^{14} bits per human should suffice to specify "genetic information, neuronal interconnections and memory, and process or initialization information" [22].

There are two interesting numbers here: the amount of information in a human, 10^{13}–10^{14} bits, and the evolutionary rate at which it was generated. In 3.5 billion years, it works out to imply roughly 10 bits/s or more based on the numbers discussed. That compares quite favorably with a clinical trial that, on the basis of accepting or rejecting a drug, implies 1 bit of information gained for a compound over what the FDA estimated in 2001 as an average 2 years for the overall process. It is a sobering thought that random mutation and natural selection seems to be some 100 times faster. But of course this is misleading on several fronts. For one thing, there is no shortage of scientific data to help remove randomness from the overall drug selection process, and the generation rate will soon be dwarfing the process of natural evolution.

2.1.5 The Available Information in Medical Data

Capturing massive amounts of data from patients can help us get to the new chemical entity that represents a trial candidate, as well as understanding its action during trials. Such scientific data that are currently being generated and are potentially relevant to a patient in clinical trials and the patient in the physician's office are often called *translational science* or *translational research*. Medical imaging alone is or soon will be producing many petabytes (1 byte = 8×10^{15} bits) worldwide, with new imaging modalities pumping out as much as 13 GB/s per device (though it is significantly reduced by considering resolution required on a local and on-demand basis). Other sources of biomedical information, from genomics and proteomics, and including human expertise and information in the form of medical text, add signifi-

cantly to the still considerable load. All this can, in principle, be stored and transmitted (and the latter may be much more problematic because of the bandwidth issues). Note that artificial storage is not so efficient: DNA could store at approximately 1 bit/nm^3, while existing routine storage media require 10^{12} nm^3 to store 1 bit. However, the universe has allocated the human race a lot more space to play with for artificial storage than has been allowed to the tiny living cell.

However, the universe has allocated the human race a lot more space to play with for artificial storage (say, soon some 10^{17} bits) than has been allowed to the tiny living cell (10^{10}–10^{13} bits). The trick is in using this artificial data. It is still 10^{17} bits that have to be sifted for relevance. It is not information "in the hand" but rather more like the virtual reservoir that evolution has tapped in its trial-and-error process. Looking at the above numbers and information rates suggests that normal processing, but using trial and error to sift the data, would demand some 8 billion years. Clearly, a strategy is required, and one that leaves little room for fundamental error, which will collapse some of it to a trial-and-error basis, or worse by pointing us in wrong directions.

An emerging dilemma for the physician reflects that for the drug researcher. In fact, we are fast approaching an age when the physician will work hand in hand with the pharmaceutical companies, every patient a source of information in a global cohort, that information being traded in turn for patient and physician access to the growing stockpile of collective wisdom of best diagnoses and therapies. But in an uncertain world that often seems to make the role of the physician as much as an art as a science, physicists are not surprised that medical inference has always been inherently probabilistic. The patient is a very complex open system of some 10^{28} atoms interacting with a potentially accessible environment of perhaps 10^{35}–10^{43} atoms. Just within each human, then, there are thus roughly 10^{15} atoms mostly behaving unpredictably for every bit of information that we considered relevant above. There are many hidden variables, most of which will be inaccessible to the physician for the long foreseeable future. Balanced against this, the homeostatic nature of living organisms has meant that they show fairly predictable patterns in time and space. Thus so far, there have been relatively rigid guidelines of best practice and contraindications based on the notion of the more-or-less average patient, even if their practical application on a case-by-case basis still taxes daily the physician's art.

Ironically, however, while the rise of genomics and proteomics substantially increases the number of medical clues or biomarkers relevant to each patient, and so provides massive amounts of personal medical data at the molecular level, it brings the physician and researcher closer to the atomic world of uncertainty and underlying complexity. It demands a probabilistic approach to medical inference beyond the medical textbooks, notably since the development of disease and the prognosis of the patient based on the molecular data are often inherently uncertain. Importantly, the high dimensionality of the data includes many relevant features, but also variations and abnormal

features that may be harmless or otherwise irrelevant. They are poised to overwhelm the physician, increase the number of tests, and escalate clinical costs, thus imminently threatening rather than aiding the healthcare system [14].

2.1.6 The Information Flow

The useful information that is available in biomedicine is best understood as a flow, which it is pretty much the same in any discipline. We shall define useful information as that which leads to an actionable decision with a required beneficial outcome.

$$\text{Data} \rightarrow \text{structured data} \rightarrow \text{rules} \rightarrow \text{inference} \rightarrow \text{decisions} \rightarrow \text{beneficial outcome}$$

In an overview of what follows, a brief comment may be made on this sequence. *Data* (or "raw data") should be necessarily qualified as *accessible data* and should ideally be in a s*tructured data* form suitable for analysis. In business and industry generally, roughly some 95% it is not. In medicine, medical text and medical images well exemplify unstructured forms. Explicitly or perhaps implicitly in an analysis procedure, conversion to at least a transient structured form is required.

This structured form is then transformed into a set of elemental statements about associations and correlations, above indicated as *rules*, which express the content in a succinct way suitable for *inference*. However, classically, the rules step has been represented by statistical analysis, with inference and decision making left to the human expert based on the results. There are numerous tools that have of course been developed to analyze data, and these obviously remain of interest. The probability theory [23] underlines classical statistics [24–26]. Of particular interest here, because of the high dimensionality of clinical with genomic and proteomic data, is multivariate analysis [27–32]. Dimensional reduction techniques such as multidimensional scaling [33] and principal coordinate analysis are essentially clustering (and by implication dendrogram or "tree") methods that reveal useful patterns in data in fewer dimensions while preserving the rank order of distances or the distances themselves, respectively. There are several pharmaceutical and biotechnological applications. For example, multidimensional scaling in conjunction with structure–activity data seems very useful for identification of active drug conformers [34–37].

Less classically, direct use of information (as opposed to probability)-based methods seems well suited to the automation of the above sequence, which is, after all, an information flow. Information theory has already long been recognized as of value in inference from rules, and the decision process based on that inference [38,39], whence it is closely related to decision theory [40,41]. Application of information theory in commercial methods of data mining for the rules, i.e., *empirical rule generation*, as the first step has been less common, though it is the approach taken by one of the authors (B. Robson) [18,42,43]

and applied to 667,000 patient records [44]. Because the method is somewhat less orthodox, it is worth stating that it has its roots in the theory of expected information [44] and in the subsequent widely used application as the Garnier-Osguthorpe-Robson (GOR) method [45] for data mining protein sequences. Widely cited and used since its publication 1978, the latter had some 109,000 Google hits on Robson GOR protein in September 2007. The "rules" here were basically rules in the same sense as in subsequent data mining efforts, though then known as the "GOR parameters," and concerned the relationships between amino acid residues and their conformation in proteins. The difficulty was that the GOR method and its rules took advantage of and was "hard wired" to the chemistry and biology of protein structure. In effect, the more recent papers [18,42,43] developed a more general data mining approach where there is no imposition on what the rules are about, except for a choice of plug-in cartridges, which customized to particular domains such as clinical data.

A simple example of such a rule may be that if a patient is tall, he will be heavy. This illustrates that rules are not in general 100% guaranteed to be correct. Rather, rules will, in general, be associated with a quantity (*weight*) expressing uncertainty in an uncertain world, even if some or many of them, such as "if the patient is pregnant, the patient is female," emerge as having a *particular* degree of certainty of 100% and may constitute ontology in the sense of "All A are B." In the abovementioned rule generation methods [18,42–44], the probabilistic weight was actually an estimate of the information available to the observer, reflecting both the strength of the relationships and the amount of data available for estimating them (a natural and formal combination; see below). Weights will be discussed in several contexts in what follows.

As in a large study of patient data [44], the rules themselves represented the end of the road as far as basic research is concerned, with the important qualification that they were automatically fed to medical databases such as PubMed to ascertain how many hits were associated with the rule. Some (3–4%) had few or no hits and represented potential new discoveries to be further investigated. The significance of subsequent inference is that it allows for the fact that rules are not independent; indeed many weak positive and negative rules with topics in common like patient weight may add up to a strong weight of evidence regarding that topic. Rules interact to generate further rules within an inference process without further information except for certain established *laws of inference* used in logical and probabilistic argument. It may be noted in passing that this is more easily said than done because some of the laws of higher-order logic required for much inference, such as syllogistic reasoning, are not well agreed upon in the matter of handling uncertainty. When focus is on a specific decision or a set of decisions as opposed to general discovery, there is funneling or selection, focusing on the domain of relevant rules. A *decision* is in that sense an inference step out of many possible inference steps. To choose the appropriate decision, one must consider what exactly *beneficial* means. In medicine, conveniently, we can characterize this in terms

of *outcomes* and specifically a sense of enhancement in the well-being of a patient in particular and of the population in general. Of course, well-being is a somewhat fuzzy and not invariant concept, but then so is the sense of lack of well-being in the first place; fuzziness and, conversely, distinguishability are some of the recurrent issues that are important to deal with at several points.

Though it seems odd at first, the files containing extracted rules can *in principle* be much larger (though in practice this is currently rarely so) than the files including the raw data analyzed. That does not mean that information is created, but that there is an overhead price to pay in putting data in a more knowledge-related form, which is appropriate for inference. The important notion is that these rules may be used to some fundamental underlying principles comparable to laws of nature. The explosion potentially occurs because relations between things are rendered specific in terms of, behind the scenes at least, combinatorial mathematics. This can be glimpsed by stating that in studying the relationships in a mere four items, A, B, C, and D, the relationships to be explored are (A, B) (B, C), (C, D), (A, B, C), (A, B, D), (B, C, D), (A, C, D), and (A, B, C, D). The consequent "combinatorial explosion" as the number of items is increased is considerable. It is at least 10^{30} for 100 items, still an incredibly small number of items for, for example, a patient record including genomic and proteomic data and image data. This makes the discovery of relevant rules difficult and computationally expensive and represents the "dragon" protecting the discovery of the gold of knowledge therein.

There may also be more rules generated in the inference process, in the sense of logical or essentially probabilistic interim or final deductions from the data-mined rules. For example, in the PC, the syllogisms generate a further rule, which can follow from two given rules. One may say that the increase of information *available to the researcher* is inevitable because it is necessarily so that these interim or final rules are unexpected or at least are hidden from consideration, else why acquire an *inference engine* software that performs the inference process? That accepted, then the data mining process, as a combinatorial expansion of the description of the relationship between things in the raw data, can be considered a part of inference, which is another reason why data mining and inference cannot be divorced. The feature that dictates the severity of combinatorial expansion is not the explicit information content of the whole file in terms of bits, but rather, the width of the data, reflecting the number of parameters to consider, not the depth of data, reflecting the sample size. The terms width and depth comes particularly from the concept of archives of analyzable records discussed below, the width of the record representing the number of parameters and the depth representing the number of records. Width makes analysis more challenging; depth makes it statistically more reliable, and there is a relationship in that increasing width demands increasing the depth to obtain statistical significance. The information is a logarithmic function of aspects arising from data and hence rises only as the logarithm of the depth. The information in terms of the actual rule content rises proportionally, however, to the width, this representing an explosive

increase. The width as number of parameters represents the true *complexity* of any analogous problem in both the colloquial and mathematical sense, as follows.

2.1.7 The Information in Complexity

The difficulty is the above "The Dragon on the Gold," but at least it is quantifiable: we can certainly know our enemy. The difficulty of discovery increases as a power function, x^c, where x is normally at least 2 depending on the nature of the data, and where the power c is the *degree of complexity* of the data being examined. Above it was discussed in terms of the "number of items." More precisely, it relates to the number of distinct *descriptors* (or *attributes*) that characterize what we want to be considered in our knowledge about a system. Complexity is thus mathematically the dimensionality of the problem, and a dimension is any kind of descriptor described in the next section, which becomes the *atomic object of analysis*, i.e., the basic indivisible component. It can be related directly to information content I of data D, viz,

$$I(D) = c \cdot \log(x). \tag{2.1}$$

Comparison with the above discussion of information would mean that log to the base 2 is used giving bits (binary units). Probably more frequently in data mining and in data analytics, log to the base e is used giving nats (natural units): simply multiply by 1.4427… to get bits. Where x is invariant in a study, x could be used as the basis of the logarithm, in which case $I(D) = c$ and *is* the complexity. An example of x is in fitting or in deducing statistical parameters to any specified error (on a scale from 0 to 1) that we are willing to tolerate. That error is $1/x$. Hence, we can write

$$I(D) = -c \cdot \log(\text{error}). \tag{2.2}$$

This useful way of writing the role of complexity can be used fairly generally as relating to the error with which one wishes to work when studying data, whenever one can formulate the study in terms of the error. Note that irrespective of the complexity c, by Equation 2.2, studying data without any error (error = 0) will require infinite information, which is not a recommended strategy, while studying data with total error (error = 1) will require and imply zero information, which is not a very interesting strategy.

2.1.8 The Datum, Element of Information

When discussing the information content of a human being, the numbers obtained relate to storage *capacity*, much as one would talk about a hard drive or the amount of RAM memory in a computer. A measure of information that is capable of imparting knowledge is information *about* something. Unlike

a computer, the associated information content does not come written on (or in a manual for) that something, but we must form opinions, statistically rigorous or otherwise, based on multiple occurrences of it.

The basic something that is collected for analysis, say the *datum*, is variously called an entry, item, observation, measurement, parameter, quality, property, event, or state. When discussing matters like patient records, the term entry is usually used. When discussing more abstract matters, and by analogy with quantum mechanics (QM) and statistical mechanics, the term *state* is frequently used, perhaps even when it may be that the measurable valuable of the property of a state is the intended meaning. An example of a datum is the weight of a patient.

In the most general definition of a biomarker, a biomarker is simply a datum and vice versa, though often the term "biomarker" is reserved for genomic, proteomic, image, and clinical laboratory data for a patient.

Structured data mining (in contrast to *unstructured data mining*, which addresses text and images) places emphasis not only on the datum but also on the *record*. The patient record, including specifically the clinical trial patient record, is an excellent example. The patient or the arbitrary unique patient identifier is a kind of true underlying state, analogous to an eigenvalue in QM, of which there may be many observable properties or qualities over time. The datum represents such properties or qualities and corresponds to an entry on the record for that patient, such as patient name and identification (if the record is not "anonymized"), date of birth, age, ethnic group, weight, laboratory work results, outcomes of treatment, and so forth. They are observables of that patient. In the complicated world of *data analytics*, including data mining, it is good news that in many respects, the above clinical examples of a datum all describe a form that can, for present purposes, all be treated in the same way, as discussed in the following section. Better still, anything can have a record. A molecule can also have a record with entries on it, for example, indicating a molecular weight of 654. When considering theoretical aspects related to prior belief and its impact on statistics, then even more generally, a record is any kind of data structure that contains that entry, even if it is only a transient repository like the short-term working memory in our heads. The terms *observation* or *measurement* do imply a distinction as something that is done before placing it in a record, as the moment it is found that the patient weight is 200 lb. However, for analysts of other peoples' data, and for present purposes, it only comes into existence when we get our hands on a record and inspect it: that is an observation of a sort for the data analyst.

The set of records is an archive and the order of records in it is immaterial except of course when they are separated into specific cohorts or subcohorts, in which case each cohort relates to a distinct archive wherein the order in each cohort is immaterial. Perhaps contrary to the reader's expectations, the entries on each record can be rendered immaterial with respect to order on the record, as discussed below, though a meaning can be attached to entries

that occur more than once on the record, as, say, multiple measurements with error (see below). In contrast, records cannot recur twice. Even if the record is anonymized, it has an implied unique index (analogous to eigenvalue), which may simply be its arbitrary position in the list of records that comprises the archive. A duplicate entry such as *Hispanic* on different records is, however, considered the same state; it is just that it is *associated with* a particular patient, implicitly or explicitly in each record. Each occurrence is an *incidence* of that state. More importantly, an entry is *associated with* all of the other entries on a record. *Association analysis*, which quantifies that association as a kind of statistical summary over many records, is a key feature of data mining, both structured and unstructured.

Above all things, a datum is an observable, ideally based unambiguous state as in physics, though with the following two caveats (providing redundant information is removed in subsequent inference). First, a *degenerate* state, such that the blood pressure is greater than a specified level, is allowable, whereas it is not in the world of QM. Second, states that show degrees of distinguishability (from none to complete distinguishability) are allowed. As the above examples imply, the observable may be qualitative or quantitative. If it is qualitative and distinguishable by recurrence, it is *countable*; if it is quantitative, it is *measurable*. The counting implied in countable is typically over the analogous state. Because states can be degenerate, a range of values, e.g., blood hemoglobin, can be used to represent a state, e.g., the state of being in the normal range for hemoglobin, and can be counted. Measurements that relate to the same state *distinguishable by recurrence* can also be counted. An *event* can also be considered as the appearance of a state or measurable value of a property of that state distinguishable by recurrence, ideally qualified or "time stamped" at a moment of time or a range of time.

A measurement may not yield the same value twice or more due to *error*. An error is a process such that the measured values are random when applied to the same state or what is considered the same state, but are random in a way such that the mean square difference between measurements in an indefinitely large sampling set of measured values is not considered significantly different from that for many subsequent sampling sets of an indefinitely large number of measured values. A state that shows continuity in time but with a change in the measurable value of a property of it that is not attributable to error is not strictly the same state but represents *an evolution* of the previous state. However, a state that represents an evolution of a state or shows multiple occurrences at the same time may be *held* to be the same state in an elected context, even if there are means to distinguish it outside that context. This is such that we may, for example, consider the patients in a cohort as subjected to repeated measurements on the same state and ameliorable to statistical analysis based on the concept of error in observation, even though there are means to distinguish those patients. The model here is that measurements on different states are treated to represent repeated measurements on the same state with error, in which case the notion of the normal (Gaussian,

bell curve) distribution applies until proven otherwise. The mean or average value is the *expectation* or *expected value* of the measured property, and the variance in the values from that expectation is a function of the magnitude of the error, specifically the mean square value. In practice, sometimes with the same raw data, account is taken of patient differences. For example, pharmacogenomics requires us to distinguish patients by their genomic characteristics, and if that is done, only patients with the same selected genomic features are treated as the same state (see below).

Countable states can be counted with one or more other states, so that the number of times that they occur together as opposed to separately is known. This usually means incrementing by one *counter function* $n(A)$ for any state A when encountering a state on a further record, and also $n(A \ \& \ B)$, $n(B \ \& \ F)$, $n(A \ \& \ B \ \& \ C)$, $n(C \ \& \ F \ \& \ Y \ \& \ Z)$, and so on for all combinations of states with which it is associated on the record encountered. The functions with more than one argument, such as $n(A \ \& \ B)$ and $n(A \ \& \ B \ \& \ C)$, represent the counting of *concurrences* of, here, A and B, and A, B, and C, respectively. Combinatorial mathematics reveals that there are 2^N such counter functions to be considered for a record of N entries, though one usually writes $2^N - 1$ since one of these relates to the potential empty record and hence null entry. Because duplicate entries on a record can have meaning as discussed above, the counter function would be incremented n times for n duplicate entries. When the value of the counter function is greater than zero, the occurrence of the state such as A or the concurrence of states such as A and B indicates that the states are *existentially qualified*, which means that the specified state exists or the specified states can coexist. For example, in terms of the PC discussed below, one can say that "Some A are B" and "Some B are A," *some* meaning at least one. Computationally, that may be the first time that a counter function is created to handle those arguments (why waste space creating variables otherwise?); hence, from a programming perspective, they are not of zero value but are undefined, which data mining interprets as zero.

The number of concurrences observed as indicated by the final or latest value of a counter function with more than one argument is a raw measure of some degree of *association* between the states, here "degree of" meaning that that as well as a tendency to occur together (positive association), random occurrence (zero association), and a tendency to avoid each other (negative association) are all degrees of association. A crude measure with these features, which can be thought of as associated primarily with the $n(A \ \& \ B)$ counter function, is the ratio $N \times n(A \ \& \ B)/[n(A) \times n(B)]$, where N is the normalizing total amount of appropriate data. The value of the logarithm of this measure may be positive, zero, or negative relates to the notion of positive, zero, and negative association. As noted above, because states can show degeneracy, continuous values can be partitioned into states (e.g., low, normal, and high values in clinical laboratory measurements) and can be counted, including other states. Association can thus be applied to both qualitative and quantitative data.

A related idea to association but applying only to quantitative data is that of a common trend in variance between lists of values, i.e., *intervariance*, *covariance*, or *multivariance*. But furthermore, because states can show degeneracy *and* degrees of distinguishability, the results of intervariance between values could be expressed in a fuzzy set approach so that the result looks analogous to the case when the values are partitioned into two states above and below a value, say, a mean value. Essentially, a Pearson correlation coefficient (which lies on the range $-1 \ldots +1$) is rescaled by the number of values analyzed in such a way that the values for very strong positive or negative correlation cover the same range as the true association values. Since that aspect is "rigged," by "looks analogous" is basically only meant that a positive correlation reflects a positive association, a zero correlation reflects a zero association, and a negative correlation reflects a negative association, though data that reflect a strong linear regression will also show a strong correlation between the values from association interpretations and corresponding values from the corresponding covariance interpretations.

Also, we will take here the position that even continuous data, like a cardiogram, can be decomposed into datum elements for analysis. If a Fourier analysis is applied implying that the information is captured as a wave, the parameters of that wave still each represent a datum. It is true that much data can appear in forms that have various degrees of structure *by virtue of their interrelationships*, having a graphic structure or representing arrays like medical images or lists (such as biosequences, entities on a spreadsheet, or relational database), or data types called sets and collections. However, these distinctions are an illusion to the extent that each datum in such data can be represented in a form that can meaningfully stand alone (see next section).

2.1.9 Data and Metadata

Any datum (entry) is susceptible to data analysis because of its relationship to other data. Whether qualitative or quantitative, as far as the product of data mining is concerned, it is always a countable entity by definition even if in practice we postulate it and it is never seen at all. Nothing prevents nonetheless inserting rules into data-mined output that are based on human expertise and have appropriate weights representing human confidence or degree of belief, and as described below, this can be a parameter combined with the count to obtain weights for data-mined rules too.

Even when it is countable does not mean that it is in a useful form. Even the basic datum can have a composite form enhancing its utility by enlarging on its meaning. In the form $age := 63$, then it is $age := 63$, which is the item, *age* is called metadata, and 63 is the data value (parameter value). Metadata is indicated by the symbol :=, which we can consider as an operator meaning "is metadata of." Optionally, there may or may not be metadata. Hence, other examples as actual plausible data items are *male*, *Asian*, $height := 6\,\text{ft}$, *systolic BP* $:= 125$, $weight := > 200\,\text{lb}$, $Rx :=$ *chloramphenicol*,

outcome:=*infection_eradicated*. Though less commonly used in practice in the same context, there may be higher-order metadata, as in animal:= vertebrate:=primate:=human:=patient_#65488, which reveals the relation to ontology or taxonomy, i.e., classification of things. Incidentally, one can of course with the use of brackets write a taxonomic tree, but for the present purpose, the descriptor relates to just one path from a selected point, as from the trunk of the tree to one particular selected leaf node. Though above these descriptors were described as *atomic*, it is clear that operations could be applied to them and this could, for example, take place in inference. However, the above fundamental form represents the state in which they come to data mining, and they are typically immutable for the duration of that process.

The input for structured data mining can come in variety of formats, but very often as comma-separated value (CSV) files, which are interchangeable with Excel and Lotus spreadsheets, as well as relational databases such as Oracle and DB2. The records relate to patients, chemical compounds, and so on, and the first or zero row, i.e., the column heads, is typically the metadata. These and more complex inputs such as graphs are really better classified in a more fundamental way, however, since many of them are essentially the same thing. In contrast, a graph structure for data relationships can be placed on a spreadsheet where each row of a spreadsheet represents, say, a node followed by its input and output arcs to other nodes, but the implied data structure is fundamentally different.

It is clear that the above way of treating a composite datum provides a universal description into which more structured data can potentially be rendered, even if the result of that rendering is as banal as Column_6:=smoker, Pixel_1073:=1, or Base_Pair_10073:=G. This theme can now be expanded upon. To begin, note that, typically, *structured records* of maximum interest in current data mining may be classified into

1. *Graphs*, in which data appear as nodes on a graph and are structured in their relationships by the arcs connecting them. They are harder to handle for data mining input since some self-consistent fragmentation of the network into maximally useful and logically sensible input chunks is required. Indeed it is best to think of this kind of data as a step in unstructured data analysis of which further kinds of further analysis transform the data into the following forms, 2–6. However, probabilistic semantic nets (concept maps, etc.) in which nodes are nouns or noun phrases connected by arcs representing verbs, prepositions, and so on, may themselves be the ultimate inference structure of the future.
2. *Trees*. Easier to handle for data mining input are trees, in which all items are nodes on branches going back to a common root. They lend themselves to the extended higher-order metadata description such as A:=B:=C:=D, and to ontological systems for holding data, notably XML.

3. *Lists*, such as biological sequences, spreadsheet rows, and relational data entries, in which a specific order has meaning for descriptors. An image might be included here, as an array, i.e., in general a multidimensional list, of pixels. A vector or matrix of *discrete* elements is thus a generalization of a list, and so in principle is a continuous distribution, i.e., of *indiscrete* elements, since by one means or another, it can be rendered as discrete data including data that are parameters of a distribution.
4. *Sets*, in which descriptors can appear in any order but only once or not at all.
5. *Collections or bags*, in which descriptors can occur in any order but now more than once (or once or not at all).
6. *Partially distinguishable item collections*. Since an item can be counted more than once in a collection, the issue arises as to the extent to which they are really distinct. If A occurs twice or more and is not counted more than once, it reflects the fact that they are considered identical, i.e., redundant duplications, and we are back to the set. If they are all counted, then they are *distinguishable by recurrence*, and measurements become repeated measurements that happen to be identical, to be taken into account in the statistics. Between these two, there are potentially intermediate degrees of distinguishability that can be discovered as strong relationships by a first pass of data mining. Then the degree of distinguishability entered in a second pass.

The closer to the top of the list, the more rigid is the structure specification. Nonetheless, that is an illusion and, transformed properly from one to the other, the information content is equivalent. Consistent with Equation 2.3 and the associated discussion, the notation used here has abolished the distinctions of graphs, lists, sets, and collections by making collections (also known as bags) the general case. We know that G is the 100th item in a DNA sequence (a list) because we now write Base_100 := G. At the very worst in a spreadsheet without specified metadata, we can always write, e.g., Column_26 := yes. In consequence also, original data could be a mix of the above types 1–4 and could be converted to a collection as the *lingua franca* form. Such mixed data are not unstructured, but are merely of mixed structure, providing the entries in each structure class of 1–4 are clearly indicated as such.

How does one build or chose such a composite datum? It is not always so easy. First, we specify a general principle of notation introduced informally above. In much of this review, it is found that it is convenient to use A, B, C, and so on, to stand generally for a datum for whatever structured form it is in, much as mathematicians use x to stand for any number. Occasionally, to avoid cumbersome use of subscript indices where they would be abundant in equations, it is important to recognize that B immediately follows A, and so on, i.e., $A = X_1$, $B = X_2$, and so on, and certainly A, B, C,… Z means all the data that there are for consideration, not just 26 of them (the number of letters

in the alphabet). Each of *A*, *B*, *C*, and so on, can stand for, for example, hydropathy of a molecule, the gender of a patient, the ethnic group, the height, weight, systolic blood pressure, an administered drug, a clinical outcome, and so on. At the point of structured data mining, a symbol such as *A* will be potentially a composite datum such as $E := F$. Prior to that, however, the *A*, *B*, may not have yet come together to form such a composite structure, e.g., prior to the reading of text. The most general approach for managing the *A*, *B*, *C*,… is to assume that all items are potentially data values, not just metadata, and then to discover ontological relations such as $A := B$ or $A := C := F$ by unstructured data mining, being in part the process that defines which is the metadata. Where all nodes on a graph have unique names, one may note that one may find $B := A$ where *A* is always associated with *B*, though not the converse, suggesting that "All *A* are *B*." Unstructured data analysis is not confined to ontology. In other instances, the fact that *C* is merely sometimes associated with *B* does not imply an ontological relationship. From the perspective of higher-order logic, an association is an *existential* relationship, e.g., "Some *A* are *B*," while an ontological relationship is a *universal* one, e.g., "All *A* are *B*."

This building or choosing process for a composite datum is not always so easy. The human brain appears to handle concepts as a kind of concept map or *semantic net*, which is a graph that is used in a way that can handle uncertainty, e.g., probabilities. Representing and utilizing such a structure as efficiently as does a human is a holy grail of artificial intelligence and actually of data mining for rules and drawing inference from them. In the interim, in the absence of fulfillment of that goal, defining and mining composite data (with metadata) in the best way can pose conceptual challenges that are practical matters. A descriptor can be a specific path through the graph represented by metadata of various orders (not just first order) such as molecule := pharmaceutical := antibiotic := sulfonamide. A record could be transformed to a collection form with items representing several such paths as descriptors (and thus separated by &s). The data mining then "merely" has to extract data leading to a terminal leaf node item such as "sulfonamide" to identify a descriptor. In this case, one is told or assumes that the structure is purely ontological (specifically, taxonomic). But there is, regarding the semantic net that the human brain somehow holds, more than one way to relate it to a practical graph for data analysis. One might have substance_abuse := legal_substance_abuse := tobacco := cigarettes := emphysema. In such a case, one must extract indirectly linked combinations including, for example, substance_abuse := tobacco, and worse still, need to recognize it as analogous to simpler useful entries in isolation such as smoker := yes.

To paraphrase the above, thinking about data and metadata in the above way provides a flexible, though not traditional, way to think about proceeding. Once a composite datum is constructed including any metadata and higher-order metadata (involving several := symbols), it represents one of the data in the a *bag* or *collection* form. To some extent, the data mining can be con-

veniently phased: the data, whether structured, unstructured like images and text, or both, are converted first to bag form, and then analysis proceeds again starting with that form. The first phase is not considered much here because it is (arguably) starting to fall in the realm of unstructured data analysis and specifically analysis of written text. We are interested in the next step, considering what do we do with structured data.

2.1.10 Rule Weights

As discussed above, a rule represented by one datum or many *may* be associated with a weight. In the big picture, a rule is *always* explicitly or implicitly associated with a weight, which may be an implied logical value of "true" (or some other measure consistent with truth) if no further statement is made. That is to say, if someone flatly states a rule, we may assume that he or she is attaching truth to it. For example one says, "The bank is closed on Sunday," without having to be as elaborate as "That the bank will be closed on Sunday has a logical value of 'true'." Used in logical reasoning, however, one may also allow for a value that relates to falsity so that the rule may be refuted and used as a variable in a chain of reasoning. With uncertainty, there are intermediate values. The most commonly appreciated measure that ranges continuously from absolute falsity (taken as 0) to absolute truth (taken as 1 is of course probability).

2.2 PROBABILITIES, RULES, AND HYPOTHESES

Classical statistics also starts early with structured data. The statistician's collection sheets are highly structured, clearly identifying independent (e.g., patient age) and dependent variables (e.g., blood pressure as a function of age) and ordering them if appropriate.

2.2.1 Semantic Interpretation of Probabilities

Classical statistics being routed in probability theory is interested in the probability of any datum A. How it considers and calculates the values of forms $P(\ )$ is discussed later below, but for probability theory in general, it is a reasonably intuitive reflection of probability as used in colloquial speech, not least in regard to laying bets. There are also coincident, conjoint, or compound probabilities such as $P(A\ \&\ B)$, which in everyday conversation might be paralleled by speaking the chances that A and B are seen together, or the extent to which A and B are two qualities or quantities describing a common thing. There may be several symbols, A, B, C, and so on, as in $P(A\ \&\ B\ \&\ C)$; their number (here 3) is in fact the complexity of that probability expression, in the sense of the word complexity used above. The data as a whole will in

general be of much higher complexity, so $P(A \,\&\, B \,\&\, C)$ is just one facet of it. Probability *functions of states* like those above are basically associations and hence quantify *existential* or "some" statements probabilistically. There is, however, a case for using *probability ratios* $P(A \,\&\, B)/[P(A)\ P(B)]$ expressing departure from randomness, because observing just one of few coincidences of A and B may not be meaningful as indicating "Some A are B," perhaps representing experimental errors. In which case, for complexity 3 or more, there are complications. For example, note that $P(A \,\&\, B \,\&\, C)/[P(A)\ P(B)\ PC]$ for quantifying existential statements is not necessarily the same quantity as $P(A \,\&\, B \,\&\, C)/[P(A \,\&\, B)\ P(C)]$ and that $P(A \,\&\, B \,\&\, C)/[P(A)\ P(B \,\&\, C)]$ and $P(A \,\&\, B \,\&\, C)/[P(A \,\&\, C)\ P(B)]$ can be different again. The correct perspective is arguably that there should be associated distinguishing existential statements expressing departure from what kind of prior expectation [3], analogous to issues in defining the free energy of the ABC molecular complex relative interacting molecules A, B, and C. There are also conditional probabilities such as $P(A \mid B)$, equal to $P(A \,\&\, B)/P(B)$ when B is defined, countable, and exists, such that $P(A)$ is greater than zero. These requirements are generally the case when $P(A \,\&\, \mathrm{B})$ also satisfies them, and then $P(A) = \Sigma_X P(A \,\&\, X)$ for all possible X that may exist. An analogy exists with the above discussion on random association in that if $P(A \mid B)$ indicates "All B are A," i.e., a universal or "All" statement, just one single observation of A not being B, again perhaps an error, can break its validity and make "Some A are B" the appropriate semantic interpretation [3]. There are in fact ways of treating this problem for both Some and All statements by adding the caveat "for all practical purposes" to the statement, raising then the argument that perhaps we should take the square root of that probability, i.e., make the probability larger, because we hold a strong belief in a weaker statement [3]. Semantically, this constitutes a *hedge* on a statement, as in "A is fairly large" compared with "A is large." Nonetheless, the general sense in human thought and conversation seems to be that one observation in a trillion that A is B justifies less the Some statement than that one observation in a trillion of B not being A invalidates the All statement.

2.2.2 Probability Theory as Quantification of Logic

Boole published binary logic as "the laws of thought," so one should be able to drill deeper with this more rigorous perspective. The probability theory is actually a quantification of binary logic, say, with functions of states such as $L(A \,\&\, B)$, which can only take the value 0 (false) or 1 (true). $P(A)$ would be a quantification of statements like $L(A) = 1$, which can be interpreted as a statement that A exists. The probability theory thus handles uncertainty, i.e., intermediate values. $L(A \,\&\, B) = 1$ is an existential statement that A and B exist and coexist, i.e., and $P(A \,\&\, B)$ quantifies the extent to which $L(A \,\&\, B) = 1$. If meaning can be attached to $L(A \mid B)$, it is the *universal* statement

that "All B are A." Existential and universal statements form the core of a higher-order logic called the PC, which goes back to the ancient Greeks. Interestingly, there is no widely agreed quantification of that, handling uncertainty in ontology combined with that for associations, in the same kind of sense that probability theory is a quantification of binary logic, though some obvious methodologies follow from the following statements. This lack of agreement is a handicap in the inference to be deduced from data-mined rules.

2.2.3 Comparison of Probability and Higher-Order Logic Perspective Clarifies the Notions of Hypotheses

Statements like "Refute the null hypothesis" sound scientifically compelling but do not mean much more than "Get rid of that notion that we don't like." Clearer statements have a more elaborate *higher-order logic* structure. PC is one example of a higher-order logic because we can write nested things like, $L[L(A \mid B) = 1 \;\&\; L(B \;\&\; C)] = 1] = 1$ (All B are A, some B are C, so some A are C, an example of a *syllogism*). For a statement like this, all the values of 1 reflect that the syllogism is *valid*, not necessarily true, as in the sense of *given that* $L(A \mid B) = 1$ and $L(B \;\&\; C)] = 1$, then $L[L(A \mid B) = 1 \;\&\; L(B \;\&\; C)] = 1] = 1$ (but $L(A \mid B)$ and $L(B \;\&\; C)$ may not actually equal 1). In that sense, it is useful to consider them, and certainly the inner terms, as hypotheses, hunches, or postulates or propositions, which do not actually have to be the case, reserving L for actual empirical truth (maybe T instead of L for "truth of," or R for "reality" would be better than L). Then one writes H in place of L as, e.g., $H(A \mid B)$. Despite the above comments on quantification, it is certainly meaningful to build quantified examples, e.g., as $P[H(A \mid B) = 1 \mid L(A \mid B)] = 1]$, meaning the probability that the hypothesis that "All B are A" takes a truth value of 1 when it is empirically true, a semantic overkill which statisticians use (or ought to, see below) in the form contracted to $P(H_+ \mid D)$. This is the probability of the *positive hypothesis* $H(A \mid B) = 1$ being true given data D, which means that $L(A \mid B) = 1$. Alternatively, there is $P(H_- \mid D)$, the probability of the *negative hypothesis* $H(A \mid B) = 0$ being consistent with data D, which actually means here $L(A \mid B) = 0$. In practice, as analyzed below, the preference in classical statistics is to use the probability of the *null hypothesis*. By analogy to the above, this would be $P(H_0 \mid D)$, which ought by that name and 0 subscript to be the type of hypothesis associated with $H(A \mid B) = 0$. One hopes the probability is low so that the hypothesis can be rejected. Actually, it relates something like $H(A' \mid B) = 1$, where A' is variously some most expected state or the most boring, or even the most costly, and such that $H(A' \mid B) = 1$ implies something much closer to $H(A \mid B) = 0$ than $H(A \mid B) = 1$. This unsatisfactory account of a state is discussed below. As it happens, things are even more tortuous because it is $P(D \mid H_0)$, which is used in classical statistics, a matter also discussed below.

2.2.4 Pharmaceutical Implications

For simplicity, we start off with the *positive hypothesis* H_+. This seems reasonable, and arguably it is the basis of the inference process that often goes on (and should go on) at least qualitatively behind the scenes in R & D before a more classically framed statistical report is produced. After all, in any scientific paper about drug action or in a project of drug R & D, the *hope* that a new drug will work is H_+. The probability of that being true prior to generating or seeing any hard data is the *prior probability* $P(H_+)$. With the data subsequently considered, the probability of the hypothesis typically changes for better or worse, to the *posterior probability*, which is the conditional probability $P(H_+ \mid D) = P(H_+ \ \& \ D)/P(D)$. The vertical bar again means "conditional on."

As indicated above, the writing of $P(H_+)$ and $P(H_+ \mid D)$ is really shorthand because inference involving hypotheses has a more complex higher-order structure. For easy comprehension, this will be framed in terms of more specific examples and need not drill down quite as far, at least in terms of symbolic overkill, as the previous section. It still requires a significant elaboration. In the pharmaceutical industry, even the so-called prior probability $P(H_+)$ typically really corresponds to a conditional probability such as P(drug works | drug X & disease Y), a probability which as written is thus really of complexity 3. An even greater complexity may emerge as important for proper analysis. Notably, while the above notation should convey adequate sense, something such as Pr[P(disease Y at time $T + t$ = false | drug X given at time T & disease Z at time T = true) > 0.9 | D] is implied. By analogy with the discussion above on higher-order logic, this new P implies a higher-order inference process, and here a higher-order probability theory. In practice, Pr signifies a *probability distribution* (*probability density*) or one value on such, and X, Y, T, and t are all variables underlying that density, as follows.

2.2.5 Probability Distributions

In reality, no one but a mathematician interested in statistical methodology wants a probability *distribution*. It implies uncertainty (perhaps even uncertainty about degrees of uncertainty about specific things) and ultimately that we can at best *expect* rather than *rely* on something (and perhaps not expect with any great reliability). From a Bayesian perspective (see below), it can represent a spread of different degrees of belief in the observer's brain as opposed to a firm, judiciously held opinion (which would appear as a single spike—a so-called delta function—if the distribution perspective is still taken). This all reflects ignorance. A distribution arises instead of discrete points because we are, for example, pooling studies on many clinical trial patients. Here is impossible to completely know and control the observation system and process: there are experimental errors. Even if we could, we cannot know and, with some 10^{10}–10^{13} bits of information capacity to potentially worry

about, may perhaps never be able to know all the features and mechanism of each patient and environment.

Challenges can arise even in a single dimension, say, along a single parameter like height in a population. For example, even when there is an inkling of the genes involved, the genes can of course interact to express phenotype in a complex way with each other and with the environment. It is thus difficult to set up the many detailed conditional probabilities; conditional, that is, on each relevant factor, which if finely conditional enough would be a sharp spike (a delta function). Without complete resolution onto different conditions, one might at least hope that separate peaks would be seen to inspire the hunt for appropriate conditions. Unfortunately, that does not always happen. There are, for example, not just two variations of a single gene that would make patients five foot tall and six foot tall, and no other height, but many. All unknown factors may effectively appear as a random influence when taken together, and a typical distribution is thus the *normal* (Gaussian, bell curve) distribution, the basis of the *z*- and *t*-tests.

Worse still, there is no reason *a priori* to expect that the ideal distribution based on many factors can be adequately expressed on a one-dimensional axis. In two dimensions, the probability peaks will look like hills on a cartographer's contour map, yet hills seen in perspective from the roadway on the horizon as a one-dimensional axis can blur into an almost continuous if ragged mountain range profile. And if we use a two-dimensional map to plot the positions of currants in a three-dimensional bun (or berries in a blueberry muffin), the picture will be confusing: some currants would overlap to look like a bigger, more diffuse, and perhaps irregularly shaped currant, and even those that remain look distinct and may look closer than they really are. With big enough currants, they can look like a normal distribution describing one big fuzzy currant. Many parameters as dimensions arise in *targeted medicine* where we wish to consider the application to specific cohort populations, and rather similarly *personalized medicine* where attention directs toward drug selection for a specific patient in the clinic. Keeping the first notation (the one without *T* and *t*) for relative simplicity, a typical probability of interest, at least in the researcher's head, elaborates to

$$
\begin{aligned}
&P(H_{+} \mid D) = P(\text{drug works} \mid \text{drug } X \ \& \ \text{disease } Y \\
&\& \ \text{clinical record feature 1} \ \& \ \text{clinical record feature 2} \ \& \ldots. \\
&\& \ \text{lifestyle feature 1} \ \& \ \text{lifestyle feature 2} \ \& \ldots. \\
&\& \ \text{genomic feature 1} \ \& \ \text{genomic feature 2} \ \& \ldots. \\
&\& \ \text{proteomic feature 1 and proteomic feature 2} \ \& 11 \\
&\& \ \text{D}).
\end{aligned}
\tag{2.3}
$$

Supposing that there is enough data for all this, which is as yet uncommon, the distribution seen in the full number of dimensions may have a potentially complicated shape (including, in many dimensions, complicated topology).

The question is, when we have knowledge only of fewer dimensions (relevant parameters), when is the shape described real?

2.3 PATTERN AND NECESSITY

2.3.1 Mythological Constellations Can Appear in Projection

In many displays of data, the data may be very distinct, typically well-separated points. That may of course be because there is simply not enough sampling. However, there may be some very crisp measurements or collections of measurements that we know are functions of several parameters, but not all the relevant parameters are known. Only the perception of three dimensions of space, plus one more of time (and many more properties of the light from the stars), allows scientists to build a true picture of the universe and to deduce the underlying physical relationships of the stars in the night sky. The two-dimensional view of the night sky yields the constellations that are mostly artifacts of perspective, reflecting the worldview of the observers. Many of the northern hemisphere were seen by the ancient civilizations, and are mythological in the traditional meaning of the word: flying horses and supernatural beings. However, even the stars that are contained within a constellation can vary with culture: the Chinese constellations are different from the European. The Southern Constellations were mostly named in modern times by European seamen and scientists and include the ship's keel, the compass, the clock, the pump, and the microscope.

In the quest for real patterns, valid techniques exist for reducing such multi-dimensional data into fewer dimensions. Principal coordinate analysis seeks to do so while preserving with minimal stress some metric of distance between the points, while multidimnsional scaling preserves the rank order of the such metrics [4]. Both can produce meaning patterns, for example, clusters and clusters of clusters, and so on. With such a result, and with all points envisaged as the intersection of branches with a horizontal cross section through a tree, that tree may be deduced and may reveal genuine ontological relationships. However, this cross section may not produce an evenly spread or random distribution of points, such that many objects such as a circle (with many points clustered round the circumference) or part circle (crescent moon shape) may also appear. Occasionally, more angular shapes like triangles may emerge. While the dimensions into which reduction occurs may be arbitrary, they may not happen to correspond to the real parameters, or they may simply represent axes at non-right angles to the dimensions representing the real parameters. Correlating the shapes with the parameters describing the original points can yield genuine, if sometimes surprising relationships with physical meaning. The principle has been applied to drug design based on analysis of predicted conformers and in regard to protein structure analysis [5–7].

However, while such dimensional reduction approaches are valid, the golden rule should perhaps be that persuasive patterns based on a viewpoint

that implies a *projection* like stars in the sky should always be suspect. The Ramsey theory [8,9] studies the conditions under which order *must* appear. In particular, how many elements of some structure must there be to guarantee that a particular property will hold? Consider the night sky as a graph of n vertices (the stars) and each vertex is connected to every other vertex by an *edge* (a line in the pictorial rendering of a constellation as an image). Now color every edge randomly green or red. Imagine that the ancient Chinese happened to pick constellations corresponding to the green lines, and observers of the ancient Middle East picked those correpsonding to the red. How large must n be in order to ensure that there we see *either* a green triangle *or* a red triangle? It turns out that the answer is just 6, and the different triangles emerged with equal probability. A common party puzzle of the same structure is this: what is the minimum number of people at a party such that there are either three people who are all mutual acquaintances (each one knows the other two) or mutual strangers (each one does not know either of the other two). The answer is again six.

2.3.2 The Hunger for Higher Complexity

To avoid the above and other problems, there is of course, or should be, a hunger. Throughout, an important bottom line is that expansion of human knowledge is reflected by our ability to *increase the complexity of probability terms* or other measures of uncertainty that we are able to *quantify*. Each term, A, B, C,… that makes up the complexity of a probability, $P(A \,\&\, B \,\&\, C,\ldots)$, represents a dimension. A rule of high complexity can be very strong, yet rules of lower dimensionality like $P(A \,\&\, B)$ may not be deducible from it and vice versa. That there were no pregnant males is not deducible from the abundance of male patients and the abundance of female patients (see below). This is not always the case. In a study of some kind of metric distances between using multidimensional scaling, principal coordinate analysis, and clustering and other techniques, things can produce meaningful patterns and relationships. But that depends on the nature of the system under study and is not in general true.

Obviously, a critical factor in that is the amount of data available. The sparseness of data points increases as the number of dimensions, that is, the number of parameters represented, and hence with the complexity of any rule associated with a probability $P(A \,\&\, B \,\&\, C \,\&\ldots)$. This means of course that we have much less data to deduce any n-dimensional probability distribution from $P(A \,\&\, B \,\&\, C \,\&\, D \,\&\, E)$ than from $P(A \,\&\, B \,\&\, C)$. The number of possible potentially interesting combinations $P(A \,\&\, B)$, $P(A \,\&\, D \,\&\, H \,\&\, Q)$, and so on, rises as at least approximately 2^N for N parameters A, B, C,…. Data thus run out fast. Many thousands of complexity 2 and complexity 3 rules, mostly known but many new, came from an analysis of 667,000 patient records. Yet many rules of complexity just 4 and especially 5 might be represented by a single observation, if any. That said, a few strong rules of much higher com-

plexity can show up. Nonetheless, there is always a level of analysis plowing into higher dimensions, which, in principle, can contain data and the tendency to overreach the interpretation of the sparse data encountered.

2.3.3 Does Sparseness of Data Breed Abundance of Pattern?

At first inspection, the answer is no (but see Discussion and Conclusions). When data is sparse, it at least *tends* to look more random, in the sense that a true pattern distinct from randomness will only emerge as data build up. We tend to look forward to, for example, the beautiful and smooth normal curve that will one day emerge from our ragged bar chart that currently looks more like the Manhattan sky line. The reliability of our statistical summaries assuming the normal curve is the right choice, and the convergence of our bar chart to it, rises as $\sqrt{N}$ the amount of data, a consideration taken into account more robust *t*-test making it more robust than the *z*-test in utilizing the normal curve model. In gathering data to plot a normal distribution, there may be many modes that appear, meaning that several values will be the same or similar. But our dreams of convergence to that curve reflect our expectation that the normal curve is the correct underlying model. For distributions in general, such modes in the raw data may, but may well not, survive to be ultimately perceived as the true modes of a multimodal (i.e., non-normal) probability distribution.

So our occasional initial assumption that we might adequately pool data into a single dimension may be too optimistic. In any event, whether or not a multidimensional description is reached or there from the outset, increasing the number of dimensions increases the opportunities for greater separation between points. In many dimensions, rogue outlying points due to experimental error and representing a rare probability of belonging to a cluster (while physically, chemically, or biologically entitled to belong to it nonetheless) tend to lie at greater Euclidean distances when that distance is in more dimensions. This can be distracting to visual analysis, attracting too much attention to it. Now it may be countered that the Ramsey theory does lead to increased chances that we might read too much into them as these sparse data are encountered. The Ramsey theory does indeed mean that we will tend to find irrelevant patterns in any data, and this presents a particular lure to the unwary when there is not enough data to be convergent to true distributions. But in another sense, the Ramsey theory runs in the opposite direction. It predicts that more elaborate patterns will emerge as the number of data points *increases*, and that the number of them rises explosively.

2.3.4 Sparse Data Can in Context Be Strong Data When Associated with Contrary Evidence

High dimensionality is not the only cause of sparse data in certain specific circumstances, and there can be a strong pattern of sorts by absence of obser-

vations. This applies to *negative associations*. Obviously, noting an unexpected large black hole in a starry sky will be significant—hopefully indicating just a cloud! The case where there is just one dimension, a marked local gap or gaps in data may be equally significant. However, with two or more dimensions, it is also true that the hole represents less data than we would expect on the basis of the projections onto to the axes. In other words, data may be sparser in the volume or hypervolume in many dimensions than it ought to be, based on the data for fewer dimensions. A negative association expressed most simply means, for example, that $P(A \ \& \ B \ \& \ C)$ is much lower than we would expect on chance bases, as calculated by $P(A) \times P(B) \times P(C)$ and $P(A) \times P(B \ \& \ C)$ and $P(A \ \& \ B) \times P(C)$ and $P(A \ \& \ C) \times P(B)$. The first of these is the projection on three one-dimensional axes, the others on one axis and the implied plane formed by the two remaining axes. As with the "black hole," strong negative associations ("pregnant males") in the limit mean that the events linked by & in the probability measure never show up at all. That does *not* mean that there is inadequate data to support the implied negative association rule. The weight of such a rule is *strengthened* by the fact that $P(A \ \& \ B \ \& \ C)$ seems to be zero or close to it as well as by a large value of $P(A \ \& \ B \ \& \ C)$ recalculated on the above bases of random association, say, as $P(A) \times P(B) \times P(C)$. In the above, notice that there is strong data, a kind of prior data, of lower dimensionality, that sets an expectation of something. That it does not occur is *evidence to the contrary*.

2.4 CONTRARY EVIDENCE

2.4.1 Lack of Contrary Evidence Breeds Superstition and Mythology

In the next section, it is discussed how prior opinions, assuming them to be rational and judiciously considered, can formally dominate over weak data. Following up the preceding sections and of interest in the present section is the matter of how, when there is no such strong prior opinion and no evidence to the contrary, the sparse data itself can set dominant prior probability distributions in the mind of the observer as far as what a fuller study would reveal, if possible. Since the true pattern is not reflected but there is a strong random element, and importantly there is a lack of control study and contradictory evidence, there is a greater opportunity for mythology and superstition.

Information theory points out that in any decision, the information for a hypothesis should be supplemented by subtracting the evidence for the contrary hypothesis; in probability theory, that implies the use of the ratio $P(H_+ \mid D)/P(H_- \mid D)$, while in the fullest forms of decision theory, there is a further combined ratio $\$(H_- \mid D)/\$(H_+ \mid D)$, which relates to the cost of the consequences associated with accept and refuting hypotheses [19]. We can also, guided by the Bayesian approach discussed below, elaborate our decision equation to include $P(H_+)/P(H_-)$, representing our prior view, which we may

well believe is rational and judicious. However, there does seem to be a fundamental difference between any prior sense of cost and a prior sense about the truth of hypotheses *per se*, and much human behavior seems to be implicitly linked to the cost issue, as follows.

The problem is that it typically does not cost too much to be superstitious. The effect of the weakness of the data for the $P(H_+ \mid D)/P(H_- \mid D)$ ratio combined with a weak prior ratio $P(H_+)/P(H_-)$ and a strong sense about the $\$(H_- \mid D)/\$(H_+ \mid D)$ ratio can be dramatic. That magic amulet seems to work well in protecting you from wild bears in the streets of Manhattan, and how insignificant is the cost of the failure of that lucky rabbit's foot compared to the potential benefits of discovering a blockbuster drug? Are there those among us who, on a honeymoon or before an important business meeting, would still avoid hotel room 13, all other rooms being equally satisfactory? And if just one friend has a total remission from aggressive pancreatic cancer on trying two herbal remedies, a high dose of aspirin, and moving to a high altitude on a vegetarian diet, who would not be tempted to try the same combination if found in the same medical position? Not only is there an absence of contradictory data, but an appreciation that even if not all these factors are relevant, why take a chance on the potentially terrible consequences of missing one? In many statistical approaches, such a therapy scenario is still a maximum likelihood or represents the expectation in the absence of contrary data from control studies. It is just not a big maximum likelihood and not a very reliable expectation, a deficiency swamped by the fact that a decision process often deliberately, and almost always subjectively, includes a component, which is the ratio of the cost of success and the cost of failure.

2.5 SOME PROBLEMS OF CLASSICAL STATISTICAL THINKING

2.5.1 Statistical Myth 1: Classical Statistics Is Objective against the Yardstick of Bayesian Thinking, Which Is Subjective

The analysis of raw data gives us probabilities about the data, not the hypothesis, via, for example, the binomial or multinomial function of the number of times that actually count things. Probabilities imply degrees of certainty or uncertainty or more precisely information that are relative, however, and in this case it is necessarily a probability conditional on the data D that is examined. Hence, what is directly obtained is the *likelihood* $P(D \mid H_+)$, not the probability $P(H_+ \mid D)$. According to bishop Bayes publishing in 1763, this is no problem because we can calculate the latter from the above likelihood using the definition of quantifiable conditional probabilities, i.e.,

$$P(H \mid D) = P(D \mid H) \times P(H)/P(D). \tag{2.4}$$

The term $P(D)$ is a bit difficult to grasp in some respects. If there are just two states, the positive and the complimentary negative hypothesis,

$P(D) = P(H_+ \ \& \ D) + P(H_- \ \& \ D)$. More generally, we can sum overall well-defined nonoverlapping states. This is implied by considering $1/P(D)$ as the multiplicative term required for normalization, i.e., such that all appropriate posterior probabilities add up to one. Loosely speaking, then, it can allow us to "soak up" the meaning of $P(D)$ to make the final interesting probabilities make sense.

Bayes' point, however, was that $P(H_+)$ is an ever-present factor that does not disappear just because it is ignored, and in modern times we saw no such "soaking up" liberties applied to $P(H_+)$. It expresses prior belief without seeing data D, a potentially highly variable determining factor. Notably, recalling that the P's should really be expressed as distributions on underlying variables, a strong prior degree of belief represented as a prior distribution $P(H_+)$ that is far from flat can swamp the contribution of the distribution of $P(D \mid H_+)$ to that of $P(H_+ \mid D)$. *Classical* statistics with its probabilities rooted in observations, measurements, and counting, hates the notion of a probability, $P(H_+)$, which can exist without data D.

Since classical statistics ignores the prior probability, and yet it is there behind the scenes, it is effectively out of control. Its presence is active, if hidden, inside the recipes that the classical statisticians develop for statistical tests, and the recommended modes of use. Its problems (as data D becomes sparse) are reflected in much disagreement between classical statisticians about what is an appropriate level of data to justify analysis or to avoid pooling to increase it. By recognizing the beast, Bayesian statistics seeks to control and either to exploit the prior probabilities where appropriate, as in the obvious case where there is prior data or just influential common sense, or to reduce the element of subjectivity where it is not appropriate. It has not itself, however, escaped from some discussion about what a distribution representing no prior knowledge should look like. The so-called Dirichlet choice, on mathematical grounds, is a binomial (or analogous multinomial) $P^{-1}(1 - P)^{-1}$. That cannot even be integrated (i.e., it is "improper"), though of course the posterior probability arising from it and the likelihood can be integrated.

Many theorists have argued that *any* conditional probabilities are axiomatic, and that there is always some kind of condition. In practice, there is always a kind of *implicit* data as prior data D^*, which relates to prior data: it is just that it is rendered most tangible by reference to the researcher's mind, which can be rendered explicit by querying the researcher. So what we really mean by $P(H_+)$ is $P(H_+ \mid D^*)$, and what we really mean by $P(D \mid H_+)$ is $P(D \mid H_+ \ \& \ D^*)$. The contribution of D^* cannot magically go away in the answer any more than the implication of $P(H_+)$ could—we are not free to discard contributions in which we truly believe. Used to reformulate the Bayes equation, it can be seen that one is really discussing the consequences of *upgrading* D to D^*:

$$P(H_+ \mid D \ \& \ D^*) = P(D \mid H_+ \ \& \ D^*) \cdot P(H_+ \ \& \ D^*) / P(D \ \& \ D^*). \quad (2.5)$$

So what kind of data is D^*? It can be well-founded prior information. For example, statistical predictions of polypeptide conformation could use as a prior probability the fact that proline is relatively fixed in conformation because of the constraint implied by the proline ring. At its least tangible in drug design, this data D^* will most likely reflect the pharmaceutical spirit of the times, anecdotal evidence inferred from streetwise chat in the underground network of researchers. At its most tacit, it could be simply this morning's scientific news. Somewhere in between is the "hot tip" at the conference bar. D^* has no role in classical statistics nor in submissions to the FDA, but (1) it does reflect decisions in the mind of the observer, sometimes appropriately so, and most reputably, (2) it could well have a role in a decision support system where it would be assigned a quantitative role on the basis of expert opinion.

2.5.2 Statistical Myth 2: H_o, the Null Hypothesis

For reasons discussed in the next section, classical statistical testing does not address H_+. Rather, it directly employs the null hypothesis, the one we wish to falsify in order to give a fighting chance that we are "on to something" in R & D. Unfortunately, it is not simply a negative hypothesis, H_-, like "drug does not work." $P(H_+ \,\&\, D) + P(H_0 \,\&\, D) = 1$ is not true. H_0 as a general concept has a level of fuzziness reflected by diversity of interpretation that truly borders on making it mythological. As generally defined, H_o is the expected, the norm, the establishment view, or simply the costly and risky alternative, in such a way that the onus is on the researcher to reject that view. "Dull" is in fact a term commonly used to explain the null hypothesis in a U.S. Medical Licensing Examination (USMLE) course to budding U.S. physicians. By the basic textbook definitions, the null hypothesis typically does not actually exist in any specific context, or ought not to, since it is not a well-defined state. In contrast, the hoped-for positive hypothesis can, as long as it is crisply defined.

What seems to be historically behind the null hypothesis is that there is sometimes potentially a multiplicity of states with probabilities adding up to 1, the specific state space of the null hypothesis being defined in some way, which minimizes the expectations of the establishment while maximizing the extent to which the hypothesis is dull. Importantly moreover, the remaining set of states being complementary to the null hypothesis represents the alternate hypothesis H_1 that the drug does work in one of some possible variety of ways. This set H_1 contains H_+. H_1 is classically held to be what a statistical hypothesis test is set up to establish.

In pharmaceutical R & D, the null hypothesis typically predicts "no difference" situations, e.g., that the rates of symptom relief in a sample of patients who received a placebo and in a sample who received a medicinal drug will be equal. Rejection of it allows one to make the alternative that the rates *did* differ. It does not prove that the drug worked, though intellectually it gives us greater confidence in this alternative hypothesis. A null hypothesis is of

course only useful if in principle it is possible to calculate the probability of observing relevant data D. In general, it is harder to be precise about the probability of getting D if the alternative hypothesis were true. Concerns about the power of statistical tests to distinguish genuine "difference" and "no difference" situations in large samples have led to suggested redefinitions; some are attempts to resolve the confusion between *significant* and *substantial*. That is to say, large enough samples are even more likely to be able to indicate the presence of differences, though they may be small.

It all seems intellectually unsatisfactory anyway, when the alternative hypothesis is suspected to be true at the outset of the experiment, making the null hypothesis the opposite of what the experimenter actually believes.

2.5.3 Statistical Myth 3: Rejection and the Falsifiability Model

The modern justification of classical statistics is that the formulation, testing, and rejection of null hypotheses is *methodologically consistent* with the *falsifiability model* of scientific discovery formulated by Karl Popper. It is widely held to apply to most kinds research. Hence, the FDA and the pharmaceutical industry follow or are formally supposed to follow the decision process advocated by classical statistics; that is,

$$P(D \mid H_o) < \alpha. \tag{2.6}$$

H_o is the *null hypothesis*, a hypothesis that we wish to falsify. Strictly speaking, we should somewhat better write, with some variation on the range part according to the question of interest, $P(D \mid H_o)$ as $P\ f\ x\ f\ X \mid H_o)$, where P is now known as the p value. This is, say, the probability that any selected value such as systolic blood pressure x from a blood pressure-reducing drug will equal or exceed a specified X given the null hypothesis is true. X would typically be the *clinically worthwhile effect* that we wish to achieve, in which case the minimum amount of data (sample size) required is said to be 16 (standard deviation)^2 by common agreement. The argument for using $P(D \mid H_o) = P(x\ f\ X \mid H_o)$ to simply illustrate the points of interest below is that "$x\ f\ X$" relates *indirectly* to the data D as follows. x is actually a range of values determined by the data; that range and all the remaining values in the data presume here a normal curve and that we do not have to see all the data in detail, just the statistical summary based on the mean and standard deviation (z-test) or the standard error (t-test). By *statistical summary* is meant that the common z- and t-tests give formulas for z and t to obtain standard values that can be addressed in a table to look up P; they essentially imply a result *as if* one has processed the normal curve such that the mean is always zero and the standard deviation 1 (z-test) or standard error 1 (t-test). The data D directly implied by "$x\ f\ X$" is thus a symbolic representation of data assuming the normal distribution and the fact that the mean and the standard deviation (or standard error) are parameters that adequately describe it.

It is at least a simple matter to take $P(x < X \mid H_o) = 1 - P(x\ f\ X \mid H_o)$ or other ranges of interest (cf. "two-tailed testing") as the p value, if we are interested in them. The net effect of the above example is that we are hoping to see the lowest possible probability that the systolic blood pressure will fall below a specified value if, in effect, the drug is useless. If that is a sufficiently small probability, the null hypothesis is *unlikely* to be true; so, it is *likely* that the drug did do something useful in lowering blood pressure. "Likely" and "unlikely" are of course matters of degree. Actually, the modern trend in research is just to quote this p value. The "$>\alpha$" part comes in because it is often necessary to make a decision to act on the result in clinical practice or business, where α is 0.95 or occasionally 0.9 are arbitrary but agreed *decision points*.

This is manifestly complicated, with several pitfalls. Many objections have been raised (see, for example, references to classical hypothesis testing, null hypothesis, and p value in Wikipedia), of which only the most relevant ones will be raised below. Despite the invoked authority of Karl Popper to justify the classical statistical approach, he was somewhat quoted out of context since he was essentially considering positive and negative evidence in a large and complex world. Related notions, such as there are so many potential things that might be but are not, do crop up in data mining and negative associations, and hence contrary evidence, as a combinatorial difficulty despite their equal importance *a priori*. Hence, associations occurring less than they should are not generally reported in *unstructured* data mining (e.g., text and image analysis), which studies an open system representing the larger real world. The output would be huge. However, the fact is that in structured data analytics and experimental designs which will employ it, we are free to choose our highly controlled micro-universe of study, and for example, design tests in which the drugs work in the way we are testing. Apart from the fact that any positive rule can usually be reexpressed in a form that renders it as a negative one (as in associations of emphysema with smokers and nonsmokers, respectively), the negative associations between disease and genomic propensity to disease and between disease and preventative therapy are the associations of interest in preventative medicine.

2.5.4 Statistical Myth 4: The Value of $P(D \mid H_o)$ Is Interesting

Note that, when pressed, everyone appears to agree that it is $P(H_+ \mid D)$, the probability of the *positive hypothesis* that the drug works given good data D, which is interesting, *in principle*. This is the *alternative hypothesis* to the null hypothesis in classical statistics jargon, or would be if the null hypothesis were better defined. Especially, the critically ill patient is even less likely to be interested in the probability with which you might get your tediously boring data given some vaguely opposite hypothesis in which the patient has by definition the least possible interest. Including D^*, an acceptance/rejection criterion, would be

$$P(H_+ \mid D \ \& \ D^*) \ f \ a, \tag{2.7}$$

where a is again an inevitably arbitrary but nonetheless agreed threshold probability. This can be converted to the analogue of Equation 2.6. Since $P(H_+ \mid D \ \& \ D^*) + P(H_- \mid D \ \& \ D^*) = 1$, we can also write

$$P(H_- \mid D \ \& \ D^*) < 1 - a, \tag{2.8}$$

that is, cognizant of Bayes' Equation 3,

$$P(D \mid H_- \ \& \ D^*) \cdot P(H_- \ \& \ D^*) / P(D \ \& \ D^*) < 1 - a, \tag{2.9}$$

so

$$P(D \mid H_- \ \& \ D^*) < \alpha. \tag{2.10}$$

2.5.5 Statistical Myth 5: The Value of $P(H_o \mid D)$ Is Interesting

It is indeed a bit more interesting than $P(D \mid H_o)$. However, it of little use to classical statisticians, in practice. They are not allowed to use it, for the following reasons. First, they are not allowed to accept the concept of prior probability $P(H_o)$ and hence they cannot use Bayes' equation above to get $P(H_o \mid D)$. And moreover, second, rejecting the null hypothesis $P(D \mid H_o)$ says very little about the likelihood $P(H_o \mid D)$ that the null is true. Actually, classical statistics considers probabilities as the result of counting things in *infinitely* large data, so it is not quite so clear why it has the nerve to quibble with the notion of probability when there is no data! It suggests of course abandoning counting-based notions of probability in favor of degrees of belief, which is what Bayesian statisticians do.

2.5.6 Statistical Myth 6: Rejecting the Null Hypotheses Is a Conservative Choice

It certainly is if one wishes to satisfy the FDA, and so on, but this does not have to be assumed lightly for internal R & D. The willingness of the pharmaceutical companies to use rejection of the likelihood of the null hypothesis is said to reside in the fact that it is a conservative choice. More generally, the rejection of the null hypothesis is said to mean that the onus is on the more general researcher to make a especially strong case to overthrow the establishment view. This concept has been important to the pharmaceutical industry: since drugs are expensive to develop and can carry both medical and commercial risks, having to make a strong case seems attractive too. Importantly, though not often stated this way in statistics books, it seems as if this would be a powerful filter in selecting from a bewildering array of opportunities in

early preclinical studies. However, as to whether it is a conservative choice, the classical decision point is equivalent to rejecting if

$$P(D \mid H_{-} \,\&\, D^{*}) < a \cdot P(D \mid D^{*}) / P(H_{-} \mid D^{*}). \tag{2.11}$$

Refuting the negative hypothesis is thus

$$\alpha < a \cdot P(D \,\&\, D^{*}) / P(H_{-} \,\&\, D^{*}) = P(D \mid D^{*}) / P(H_{-} \mid -D^{*}). \tag{2.12}$$

We may deduce that the final decision criterion depends on the relative probabilities of obtaining the hard data D to the hypothesis given the prior more subjective data D^{*}. Again, we want to refute the contrary of the positive hypothesis. Evidently, α is not a constant of fixed meaning, so whether it is a conservative choice depends on the value of $P(D \mid D^{*})/P(H_{-} \mid D^{*})$. In some cases, we can reject the negative hypothesis (or the null hypothesis) when they are virtually certain to be true. If observations contraindicate the null hypothesis, it means either (1) the null hypothesis is false or (2) certain relevant aspects of data occur very improbably, say, overall that there is a low $P(D)$. This gives confidence in the falsity of the null hypothesis, a confidence that rises in proportion to the number of trials conducted. Accepting the alternative to the null hypothesis does not, however, prove that the idea that predicted such a difference is true. There could be differences due to additional factors not recognized by theory.

2.6 DATA MINING VERSUS THE MOLECULE WHISPERER PRIOR DATA *D**

2.6.1 The Two-Edged Sword

In the pharmaceutical industry, D^{*} is a two-edged sword. A hot tip from the "horse whisperer" at the racecourse may occasionally work in your favor, but no one regards anecdotal evidence rooted in vaguely defined sources as a general strategy. Having a good implicit data D^{*} in the mind of the researcher will direct, as if through a powerful filter, toward answers in a problem of incredible complexity. However, a manager may feel that a researcher was not innovative in a special if rather unforgiving sense. He or she already had the hunch that (1) some *particular* H_{+} is interesting and that (2) it represented a reasonable bet for investigation in the sense that the prior probability P(drug works | drug X & disease Z), held before examining any data D, is at least not too low. Recall that the data D^{*} will most likely reflect things like the pharmaceutical spirit of the times, anecdotal evidence including streetwise chat in the underground network of researchers, or simply this morning's scientific news. Given that, the novelty and uniqueness are in question. Even if the origin of D^{*} lies within the confines of a particular company, it does carry

some aspect of needless use of resource for "reinventing the wheel" when discovery is paramount. In fact, under the enticing burden of D^*, the more the researcher is judicious and correct in the hypotheses selected, that means that the hard data D does not make too much difference to D^*, and hence the data analysis does not make too much difference to the positive hypothesis.

2.6.2 Types of Data Mining Reflect Types of Measure

The various data mining techniques differ primarily by *normalization*, though sometimes in the more general sense of ensuring that all probabilities add to 1. For simplicity, we shall neglect D^* here, bringing it back in the next section with an example from a specific data mining approach. In many cases, normalization means what the raw numbers such as $n(A \,\&\, B \,\&\, C)$ are divided by. The parameter $n(A \,\&\, B \,\&\, C)$ is the number of times that A, B, and C are observed together. The following examples also use $A \,\&\, B \,\&\, C$ and can be extended to an indefinite number of states A, B, C, D,…, which may be variously described as states, events, objects, observations, measurements, properties, parameters, or descriptors, among other things. These are not conceptually all the same thing, though it is interesting that with few modifications, they can be treated statistically in exactly the same way. $n(A \,\&\, B \,\&\, C)$ can be written as $n(A, B, C)$, which highlights the fact that A, B, and C are dimensions, but the nature of logical & is that this comes to the same thing.

2.6.2.1 Pattern Recognition This "simply" notes that A, B, and C are observed to occur together more than k times, where k is at least 2:

$$n(A \,\&\, B \,\&\, C) \; f \; k. \tag{2.13}$$

This does allow for efficient identification of events of high complexity, though not a strong negative association when the pattern does not occur (bust statistically should). Assuming $n(A \,\&\, B \,\&\, C)$ are reasonable large, a method can also report a probability, i.e.,

$$P(A \,\&\, B \,\&\, C) = n(A \,\&\, B \,\&\, C)/N. \tag{2.14}$$

N is the total amount of data, which is not quite as simple as it sounds. In data mining, it can often be stated more accurately as the number of (patient, chemical compound, etc.) records mined. However, that is not in general true because, although it is rarely considered, A, B, and C may be not completely distinguishable. The three major degrees of distinguishability correspond to the maximum number of times that, say, A can be observed on a record and how they are counted. If A can only occur once on a record (i.e., the record is a *set*), it is said to be *fully distinguishable*. If it can occur more than once and can be counted each time it is seen (i.e., the record is a *bag* or *collection*), it is *indistinguishable except by recurrence*. If it can occur more than once on

a record and is counted only once, it is *fully indistinguishable*. Intermediate cases can occur if the definition is deliberately fuzzy or simply prone to uncertainty. Similar arguments in distinguishing or in occurrences of A on a record, i.e., "comparing" A with A, equally apply to "comparing" A with B or C, and so on, since in theory B, for example, may turn out to be A. Even if A is fully distinguishable from B, it does not mean that A and B are mutually exclusive. N is not in general $n(A) + n(B) + n(C) + \ldots$ because A, B, and C may not be mutually exclusive, e.g., female patients, overweight patients, and high blood pressure patients. If we can treat A, B, and C as parameters each with a number of possible values such as 1, 2, 3,…, then $n(A \,\&\, B \,\&\mathrm{C}) = \Sigma_B \Sigma_C \, n(A \,\&\, B \,\&\, C)$, $n(A) = \Sigma_B \Sigma_C \, n(A \,\&\, B \,\&\, C)$, and $N = \Sigma_A \Sigma_B \Sigma_C \, n(A \,\&\, B \,\&\, C)$.

2.6.2.2 Predictive Analysis This is common and generates conditional probabilities that quantify statements such as "All B & C are A." For reasonably large $n(A \,\&\, B \,\&\, C)$, this may be written as

$$P(A \mid B \,\&\, C) = P(A \,\&\, B \,\&\, C)/P(B \,\&\, C) = n(A \,\&\, B \,\&\, C)/n(B \,\&\, C). \quad (2.15)$$

Though conditional probabilities like the above can be applied to quantifications of semantic IF statements, the problem is that they make no statement about the probability that A would have occurred anyway.

2.6.2.3 Association Analysis Association analysis does take the above into account by means of an association ratio. By analogy to Equation 2.16, there is an association ratio that, when $n(A \,\&\, B \,\&\, C)$ is sufficiently large, is

$$\begin{aligned} K(A\,;\, B \;\&\; C) &= P(A \mid B \,\&\, C)/P(A) \\ &= P(A \,\&\, B \,\&\, C)/[P(A)\,P(B \;\&\; C)] \\ &= N \cdot n(A \,\&\, B \,\&\, C)/[\boldsymbol{n}(A)\,n(B \,\&\, C)]. \end{aligned} \quad (2.16)$$

The notation is a glimpse at a more general language. It is to be understood throughout all the above that there may be more states such as $D, E, F, \ldots$ and that any state can also be expressed as a conditional; that is, there may be one of more states written to the right of a "|" symbol. Also, there can be more than one ";." So, for example,

$$\begin{aligned} K(A\,;\; B \mid C) &= P(A \mid B \,\&\, C)/P(A \,\&\, C) \\ &= P(A \,\&\, B \,\&\, C)/[P(A \,\&\, C)\,P(B \,\&\, C)] \end{aligned} \quad (2.17)$$

and

$$\begin{aligned} K(A\,;\, B\,;\, C) &= K(A\,;\, B \;\&\; C)\,K(B\,;\, C) \\ &= P(A \,\&\, B \,\&\, C)/[P(A)\,P(B)\,P(C)] \\ &= N^2 \cdot n(A \,\&\, B \,\&\, C)/[n(A)\,n(B)\,n(C)], \end{aligned} \quad (2.18)$$

which perhaps makes clearer that K is analogous to a chemical equilibrium constant.

2.6.2.4 Mutual Information Analysis This is actually association analysis in an information theoretic form. It is in essence the use of logarithms of K measures defines the limit of infinitely large data general logarithms plus some further sophistication. Analogous to Equations 17–19, there are corresponding forms of *Fano's mutual information*,

$$\begin{aligned} I(A\ ;\ B\,\&\,C) &= \log_e[P(A\,|\,B\,\&\,C)/P(A)] \\ &= \log_e P(A\,\&\,B\,\&\,C) - \log_e P(A) - \log_e P(B\,\&\,C) \\ &= \log_e[N\cdot n(A\ \&\ B\ \&\ C)/[n(A)n(B\,\&\,C)]], \end{aligned} \tag{2.19}$$

$$\begin{aligned} I(A\ ;\ B\,|\,C) &= \log_e[P(A\,|\,B\,\&\,C)/P(A\,|\,C)] \\ &= \log_e P(A\,\&\,B\,\&\,C) - \log_e P(A\,\&\,C) - \log_e P(B\,\&\,C) \\ &= \log_e[N\cdot n(A\,\&\,B\,\&\,C)/[n(A\,\&\,C)n(B\,\&\,C)]], \text{ and} \end{aligned} \tag{2.20}$$

$$\begin{aligned} I(A\ ;\ B\ ;\ C) &= \log_e[P(A\,|\,B\,\&\,C)/P(A)] \\ &= \log_e P(A\,\&\,B\,\&\,C) - \log_e P(A) - \log_e P(B) - \log_e P(C) \\ &= \log_e[N^2\cdot n(A\,\&\,B\,\&\,C)/[n(A)n(B)n(C)]]. \end{aligned} \tag{2.21}$$

2.6.2.5 Atomic Rule Data Mining This is a term used by the authors and colleagues that seems to coincidentally turn up in discussions with other research teams. It implies simply that in just a particular definition of information measure (or analogous K measure) such that after the rules are generated, one focuses on mining those, and other rues can be deduced from them, as opposed to attempting to generate them in the data mining from the outset. The advantage of the form in which all "&" are replaced by ";" (conditionals placed after "|" may remain) is that it is more "atomic," i.e., other forms can be deduced from such. For example,

$$I(A\ ;\ B\,\&\,C) = I(A\ ;\ B\ ;\ C) - I(B\ ;\ C). \tag{2.22}$$

Note how the information theory representation forms can be used in an algebra of inference and in predictive methods, based on rules obtained by data mining. For example,

$$I(A\ ;\ B\,\&\,C\,\&\,D\,\&\ldots) = I(A\ ;\ B) + I(A\ ;\ C\,|\,B) + I(A\ ;\ D\,|\,B\,\&\,C) + \ldots \tag{2.23}$$

A convenient concept is that neglecting complex terms at some arbitrary point to the right neglects more complex, high-dimensional, and sparse data, so allowing estimates in the absence of data. So-called Bayesian networks (more properly, conditional probability networks) are analogous exponential forms

of such equations. What is not explicitly represented in the "Bayes network" implies analogous neglect.

2.6.2.6 Inclusion of Complementary Information This produces information expressions awfully close to those of *decision theory*, or the equivalent probability ratios of most formulations of that subject. It involves a form of further information in inference that is so fundamental that it need not be included in inference but built into measures from the outset. To include the fullest possible information in regard to, say, A, it may be noted that information for the complementary state $\sim A$ is information against A, i.e., negative information for A. The information measure uses the colon (:) to indicate use of two alternative states in this way.

$$\begin{aligned} I(A :\sim A\ ;\ B\ ;\ C\ ;\ \ldots) &= I(A = 1\ ;\ B\ ;\ C\ ;\ \ldots) - I(A = 1\ ;\ B\ ;\ C\ ;\ \ldots) \\ &= \log_e [n(A\ \&\ B\ \&\ C)] - \log_e [n(\sim A\ \&\ B\ \&\ C)] - \\ &\quad \log_e [n(A)] + \log_e [n(\sim A)] \end{aligned} \tag{2.24}$$

Again this assumes reasonably large $n(A\ \&\ B\ \&\ C)$ and $n(\sim A\ \&\ B\ \&\ C)$, and again $\sim A$ means "not A," i.e., the complementary state to A. Equation 2.15 is a perfect equation for a perfect (classical statistics?) world of indefinitely large amounts of data to define the probabilities. It says nothing about information given levels of data, however, and it should. If there are several data, the information should conform to Equation 2.15 with probabilities P evaluated for very large numbers of observations. But if we have no data, we have no information (except that in prior belief). Between those two extremes, the information should rise as the amount of information available to us rises as data increases.

2.6.2.7 Zeta Theory [12,18,19] Proposed by one of the authors (B. Robson), zeta theory allows a more general strategy by controlling a parameter, s, as described below. It links to number theory in order to develop powerful mining algorithms and to characterize the combinatorial explosion of rules. Most importantly for present purposes, it "normalizes" according to the amount of data in a natural way. Expected values of the information measures so that the amounts of data are taken into account. Clearly there is a problem with Equation 2.21 if $n(A\ \&\ B\ \&\ C)$ and $n(\sim A\ \&\ B\ \&\ C)$ are not sufficiently large. In fact, if they are both zero, $I(A\colon \sim A\ ;\ B\ ;\ C\ ;\ldots) = -\log_e [n(A)] + \log_e [n(\sim A)]$. In other words, we obtain information about the relation between A and B and C even without data concerning that relation, which is unreasonable. The above measures are not "normalized" or adjusted in any kind of way according to the amount of data, yet we know that if there is no data, there is no information. The adjustment comes naturally on theoretical grounds, as discussed below, as an expectation of the information on data D, i.e., $E[I(A\ ;\ B\ ;\ C\ ;\ldots) \mid D]$. This expectation converges to $E[I(A\ ;\ B\ ;\ C\ ;\ldots) \mid D]$ as $n(A\ \&\ B\ \&\ C)$ increases:

$$E[I(A\ ;\ B\ ;\ C\ ;\ \ldots)\,|\,D] = \zeta(s=1,\ o[A\ \&\ B\ \&\ C\ \&\ \ldots]) - \zeta(s=1,\ e[A\ \&\ B\ \&\ C\ \&\ \ldots]) \quad (2.25)$$

and so also

$$E[I(A :\sim A\ ;\ B\ ;\ C\ ;\ \ldots)\,|\,D] = \zeta(s=1,\ o[A\ \&\ B\ \&\ C\ \&\ \ldots]) - \zeta(s=1,\ e[A\ \&\ B\ \&\ C\ \&\ \ldots]) - \zeta(s=1,\ o[\sim A\ \&\ B\ \&\ C\ \&\ \ldots]) + \zeta(s=1,\ e[\sim A\ \&\ B\ \&\ C\ \&\ \ldots]). \quad (2.26)$$

Here $o[A\ \&\ B\ \&\ C] = n[A\ \&\ B\ \&\ C]$, the observed frequency (number of occurrences of) A and B and C, and so on, all together, and $e[A\ \&\ B\ \&\ C]$) is the expected frequency:

$$e[A\ \&\ B\ \&\ C\ \&\ \ ]) = n[A]n[B]n[C]N^{-t+1}, \quad (2.27)$$

where N is the total amount of data as discussed above, and there are t terms A, B, C… Though it looks complicated, this is the same expected frequency as in the chi-square test. For present purposes, we may adequately define ζ for values of s of 1 and greater

$$\zeta(s,n) = 1 + 2^{-s} + 3^{-s} + 4^{-s} + \ldots + n^{-s}, s\ f\ 1. \quad (2.28)$$

However, using much more complicated expressions, it can be defined for all real and even complex numbers (including an imaginary component) of s and n. Note that Equation 2.27, which gives a possible value for n, is most generally a real, not integer, value, and that more general definitions of z can handle this. For data mining, interpolation from integer values of n in Equation 2.28 suffices. The interesting case for real values of $s = 0$ and greater covers several lines of text [42]. Though $s = 1$ defines information, other values relate to moments of information (square, etc.) and other measures of surprise. For example, with a sufficiently large value of s (in the tens or hundreds is an adequate approximation), the zeta function converges to a binary logic value, 0 or 1.

2.6.3 Including *D**

Consider first that the *new* probability as a consequence of the data is on the basis of something closer to classical likelihood testing (but focusing on the positive hypothesis as in this case the compliment of the null hypothesis):

$$P(D\,|\,\text{drug works \& drug } X \text{ \& disease } Y \text{ \& } D^*) = P(\text{drug works \& drug } X \text{ \& disease } Y \text{ \& } D \text{ \& } D^*) \,/\, P(\text{drug works \& drug } X \text{ \& disease } Y \text{ \& } D^*). \quad (2.29)$$

This is typically close to 1 because the researcher was focusing on a particular hypothesis such that his D^* was already relatively strong and D was consistent with it. Rather, one is interested in the mutual information

$$\begin{aligned} &I(\text{drug works; drug } X \,\&\, \text{disease } Y \mid D \,\&\, D^*) \\ &\quad = \log P(\text{drug works} \,\&\, \text{drug } X \,\&\, \text{disease } Y \,\&\, D \,\&\, D^*) - \\ &\qquad \log P(\text{drug works} \,\&\, D \,\&\, D^*) - \\ &\qquad \log P(\text{drug } X \,\&\, \text{disease } Y \,\&\, D \,\&\, D^*). \end{aligned} \tag{2.30}$$

Here the focus is on the information concerning the *mutual information* between *drug works* and *drug X & disease Y* (how much information one gives about the other) not the correlation with the data, albeit that the informational value is conditional on D and D^*. The ";" symbol indicates what the mutual information is between. It is symmetrical; we could interchange the parts on either side of the ";" (but must leave any conditional factors after the "|"). That the drug works is here H_+. In practice, there is generally at least one further piece of information that cannot be ignored, the information, if any, that the drug does not work, here H_-. Formally, this is included even though information for H_- is simply and conveniently one minus that for H_+. This contrary information is negative information for the positive hypothesis, and should be subtracted, viz,

$$\begin{aligned} &I(\text{drug works : drug does not work; drug } X \,\&\, \text{disease } Y \mid D \,\&\, D^*) \\ &\quad = \log P(\text{drug works} \,\&\, \text{drug } X \,\&\, \text{disease } Y \,\&\, D \,\&\, D^*) - \\ &\qquad \log P(\text{drug works} \,\&\, D \,\&\, D^*) - \\ &\qquad \log[1 - P(\text{drug works} \,\&\, \text{drug } X \,\&\, \text{disease } Y \,\&\, D \,\&\, D^*)] + \\ &\qquad \log[1 - P(\text{drug works} \,\&\, D \,\&\, D^*). \end{aligned} \tag{2.31}$$

Note that log P(drug X & disease Y & D & D^*) has disappeared due to cancellation.

How is D^* quantified? The zeta theory (see below) makes that easy and leads naturally to the ability to add numbers, say, **a** and ~**a**, which represent our prior degree of belief relating to $n[A \,\&\, B \,\&\, C \,\&\, \ldots]$ and $n[\sim A \,\&\, B \,\&\, C \,\&\, \ldots]$ in a Bayesian sense. They come from the parameters of an implied prior binomial or multinomial distribution. They express belief in terms of a sense of virtual frequencies of observation, the sense that we would expect to see $A \,\&\, B \,\&\, C \,\&\, \ldots$ **a** times in the study, on the basis of D^*.

$$\begin{aligned} &E[I(A :\sim A \,;\, B \,;\, C \,;\, \ldots) \mid D \,\&\, D^*] \\ &\quad = \zeta(s = 1,\ o[A \,\&\, B \,\&\, C \,\&\, \ldots] + \mathbf{a}) - \zeta(s = 1,\ e[A \,\&\, B \,\&\, C \,\&\, \ldots] + \mathbf{a'}) \\ &\qquad - \zeta(s = 1,\ o[\sim A \,\&\, B \,\&\, C \,\&\, \ldots] + \sim\mathbf{a}) + \zeta(s = 1,\ e[\sim A \,\&\, B \,\&\, C \,\&\, \ldots] + \sim\mathbf{a'}). \end{aligned} \tag{2.32}$$

If **a** = **a′** and ~**a** = ~**a′** are distinct values, this is the "quench choice"; it takes a stricter view and demands that we have to obtain more data to get the same

amount of information. Having parameters different in all four terms will obviously give a nonzero information value when all *o* and *e* are zero, quantifying that prior knowledge. Expert human judgment would assign the values on the view that this is the same strength of information *as if* the numbers had been real numbers of observations. Optionally, choosing a prior probability $P(A \,\&\, B \,\&\, C \,\&) \times N$ where N is the total amount of data or strength analogous to a given amount of data, allows estimates based on a sense of proportional probabilities and the amount of data available.

Strictly speaking, since such parameters as **a** are themselves dependent on the data, they should be written as $\mathbf{a}[A \,\&\, B \,\&\, C \,\&\, \ldots]$. Also, $\mathbf{a}\,[A \,\&\, B \,\&\, C \,\&\, \ldots]$ is subject to the same constraints as $n[A \,\&\, B \,\&\, C \,\&\, \ldots]$. The $n[A]$, $n[A \,\&\, B]$, and $n[A \,\&\, B \,\&\, C]$,… are not generally the same value as each other, and there is a *marginal sum* constraint when we reduce dimensionality such that $n[A] = \Sigma_X\, n[A \,\&\, X]$, $n[A \,\&\, B \,\&\, X] = \Sigma_X\, n[A \,\&\, B \,\&\, X]$, and so on. This still applies replacing all the n by **a**. However, this most seriously becomes an issue when information terms of different complexity are added together in inference. Moreover, there is (at least arguably) a philosophical position that the parameters like **a** relate to *absolutely prior information* and need not be subject to these considerations in the same way as more tangible early data based on *n*. For example, the probability theorist Dirichlet proposed what is equivalent to using $\mathbf{a} = -1$ as relating to the absolutely prior probability distribution [45], a choice that many purist theoreticians still insist on today. As long as there is no real counting involved, *D** can be considered of the absolutely prior class, and fixed value parameters can be used. This is not, however, a universally held opinion in analogous other contexts. It becomes a difficult conceptual point if it is argued that only gut feeling, or arcane matters of mathematical best choice to represent zero prior information, can represent the absolutely prior case. *D** is a little more tangible than these.

2.6.4 *D** and the Filtering Effect

Filtering for gold is good, providing the easy-to-spot lumps that are caught are not worthless rock and the value is not in the massive amounts of gold dust that flushes way. None of the above considerations avoid the luring effect of *D**; it merely makes its involvement clearer and allows a means to quantify its effect at least based on human judgment. Let us allow that this nonclassical approach was what the researcher favors and that he or she constructs some such test accordingly. The information gained may be much higher, but the particular hypothesis being tested still relates to a particular hypothesis, say, H_+ (1) = "drug works & drug *X* & disease *Y*." From the larger perspective, the researcher perhaps unwittingly used a heuristic: he or she prefiltered the larger mass of data available by focusing on H_+ (1). The good news for the industry and the bad news for the overworked researcher

is that there are in fact $H_+(1), H_+(2), H_+(3), \ldots H_+(2^N - 1)$ hypotheses related to the N probabilities of increasing complexity with which N things show up and, importantly, the probabilities of the $2^N - N - 1$ combinations in which they show up. For $N = 100$, $2^N - N - 1$ is circa 10^{30}. Hidden in this data, like needles in a haystack, are many events, associations, and correlations that relate to many hypotheses, some much weaker, some much stronger than $H_+(1)$.

2.6.5 No Prior Hunch, No Hypothesis to Test

As a consequence of the above arguments, the underlying special nature of data mining is underlined. Data mining is not the testing of a hypothesis. That would be in database terms highly analogous to a directed query to a database, asking "Is this true?" or "Does this occur?" Data mining is *undirected* and *unsupervised*. Of course it is easy to add a focus, with, for example, command specifiying things that are interesting and not interesting, even down to a specific hypothesis test or query. But ideally, it seeks to obey the command "Find me everything interesting, irrespective of any prior views or data and without focus on the relationships between *particular* things." This ambitious aim does not of itself get round the fact that finding everything of potential interest is very difficult to do and is sometimes astronomically impossible if there is no guidance at all. Hitting upon what is interesting is still a matter of entropy, a combinatorial explosion problem representing the "dragon on the gold."

2.6.6 Good Data Mining Is Not Just Testing Many Randomly Generated Hypotheses

Even if the core parts of a data mining program look like iteration over many arbitrary hypotheses, the code overall, and its effect, is much larger than the sum of those parts. If it were not, we would simply run out of project time, patience, or computer power long before all the space was covered for discovery (i.e., generating all the possible rules), and there would be an arbitrary focus on what was examined first. Thus, data mining really implies many algorithms to try and enhance genuine discovery.

At its rawest, data mining has no sense of what is interesting, or even new, to the researcher. It has no sense of physics, chemistry, or biology. It reports or should report the surprising absence of pregnant males with equal enthusiasm to relationships implying a potential cure for cancer. Of course, well-founded prior judgment is not excluded, and it is useful to have "interesting" and importantly "not interesting" commands, ideally with a probabilistic element rather than being commands set in stone, so that related matters do not avoid discovery. However, these imply the presence and application of

*D**. They draw data mining back toward classical hypothesis testing. This is sometimes a good thing, but it runs the risk of the dangers as well as the benefits posed by *D**.

There is much that can be done with more general heuristic algorithms that are relatively "*D** free" and yet restrict the early calculations and search to where it counts [16–18]. Most abundant of these is based on the amount of data, related to "the level of support for a rule." Where there is inadequate data, why bother to calculate? But inadequate data does *not* mean, however, that $n(A \,\&\, B \,\&\, C \,\&\, \ldots)$ counts as a small number of observations. Consider the situation that thousands of female patients taking, say, a cholesterol-lowering drug X for a month never get pregnant. Here n(female & pregnant & drug X) equals zero, yet the effect is very significant indeed.

The information theoretic approach makes the situation clearer. In the huge number of potential rules above, most are not likely to be rules in the everyday colloquial sense. Some will contain little information: their probability is close to what we might expect on a chance basis. On the face of it, we would have to look at every possible rule to calculate that. However, rules could be avoided from further consideration where there is enough data to obtain reliable rules. The value can be positive or negative, and that it is close to zero implies no information. Thus, the algorithm would typically be to halt calculation where early it is detected that information greater than $+x$ or less than $-x$ cannot be obtained. The rule for this is again not that $n(A \,\&\, B \,\&\, C \,\&\, \ldots)$ are below a critical number, but that the terms of lower complexity are below a critical number. Moreover, we may start with the terms of least complexity $n(A)$ working up to $n(A \,\&\, B)$, $n(A \,\&\, B \,\&\, C)$…, which are inevitably always smaller values. When that complexity falls below a critical value for any subset of the parameters in the full set $A \,\&\, B \,\&\, C \,\&\, \ldots$ of interest, we may halt.

Note that this impact of data has nothing to do with estimates of probabilities $P(A)$, $P(A \,\&\, B)$, and so on, *per se*, since this conveys nothing about the levels of data involved. We might get the same probability (depending on the estimate measure) by taking a subset of just one thousandth of the overall data.

A direct measure of information including the level of data is philosophically sound and feasible. It measures the information in a system that is available to the observer. On such grounds, the real form of interest having the above properties dependent on data levels arises naturally. It is an *expectation* calculated by integration over Bayesian probabilities given the data [18,12,20]. This was used in the GOR method in bioinformatics [20], which was based on several preceding studies including Robson's expected theory of information [21]. In the latter study, the integration of information functions $\log_e P$ is made over the probability distribution $\Pr(P \mid)\, dP$, where $\Pr[P \mid D]$ is given by Bayes' equation as $\Pr[D \mid P]\Pr(P)/P(D)$. Consider also that what we imply by $\Pr(P)$ is really the estimate or expectation $E(P \mid D)$ of an underlying P "out there in nature," conditional on data D, say, $D = [n(A), n(B), n(A \,\&\, B)]$. It means

that the estimate arises only by considering cases when here $D = [n(A), n(B), n(A\ \&\ B)]$.

The integration over information measures is similar. The most general way to write it is to simply state $I(P)$ as some function of P, viz,

$$\begin{aligned} E(I \mid D) &= \int_{0,\ldots 1} I(P) \Pr(P \mid D) \cdot \mathrm{dP} \\ &= \int_{0,\ldots 1} I(P) [\Pr(D \mid P) \times \Pr(P) / \Pr(D) \mathrm{dP}. \end{aligned} \tag{2.33}$$

Equation 2.18 is in some respects the most complete because it includes not only A but the contrary information in $\sim A$. In fact, whatever terms A, $\sim A$, and B, C, D are implied in $P(\ )$ introduced either with the use of "&" or the conditional bar "|," we can write

$$\begin{aligned} E(I \mid D) &= \int_{0,\ldots 1} [\log_e(P) - \log_e(1-P)] \Pr(P \mid D) \cdot \mathrm{dP} \\ &= \int_{0,\ldots 1} [\log_e(P) - \log_e(1-P)] \Pr(D \mid P) \times \Pr(P) / \Pr(D) \mathrm{dP} \end{aligned} \tag{2.34}$$

and

$$\begin{aligned} E(I \mid D) &= \int_{0,\ldots 1} \log_e(P) \Pr(P \mid D) \cdot \mathrm{dP} - \int_{0,\ldots 1} \log_e(1-P)] \Pr(P \mid D) \cdot \mathrm{dP} \\ &= \int_{0,\ldots 1} \log_e(P) \Pr(D \mid P) \times \Pr(P) / \Pr(D) \mathrm{dP} - \\ &\quad \log_e(1-P)] \Pr(D \mid P) \times \Pr(P) / \Pr(D) \mathrm{dP}. \end{aligned} \tag{2.35}$$

Since the log terms are separable, we may focus on

$$\begin{aligned} E(I \mid D) &= \int_{0,\ldots 1} \log_e(P) \Pr(P \mid D) \cdot \mathrm{dP} \\ &= \int_{0,\ldots 1} \log_e(P) \Pr(D \mid P) \times \Pr(P) / \Pr(D) \mathrm{dP}. \end{aligned} \tag{2.36}$$

The "plug-in" point in the above for actually introducing the counted numbers is the *likelihood*, whence one must be specific about the arguments of P. It is a binomial, or in general multinomial, function of the number of observations of something (or joint occurrences of something, say, n. The integration then yields

$$E(I \mid D) = \zeta(s = 1, n) + C. \tag{2.37}$$

Here z is the incomplete Riemann zeta function discussed above, actually more general than the complete one, which implies $n = \infty$.

When n becomes indefinitely large,

$$\zeta(s = 1, n) - \gamma \to \log_e(n), n \to \infty \tag{2.38}$$

$$E(I \mid D) = \zeta(s = 1, n) - \gamma, n \to \infty. \tag{2.39}$$

Constant $\gamma = 0.5772156649\ldots$ is the Euler–Mascheroni constant. In fact, there seems no interesting case in data mining yet noted where one zeta function is not subtracted from another, so the constant C always cancels, as when we wrote for Equation 2.27

$$E[I(A:{\sim}A\ ;\ B\ ;\ C\ ;\ \ldots)\,|\,D] = \zeta(s=1,\ o[A\ \&\ B\ \&\ C\ \&\ \ldots]) - \zeta(s=1, e[A\ \&\ B\ \&\ C\ \&\ \ldots]) - \zeta(s=1,\ o[{\sim}A\ \&\ B\ \&\ C\ \&\ \ldots]) + \zeta(s=1,\ e[{\sim}A\ \&\ B\ \&\ C\ \&\ \ldots]).$$

2.7 INFERENCE FROM RULES

2.7.1 Rule Interaction

Rules may not stand in isolation, but when they have a common parameter, A, B, C, and so on, they may interact. Negative rules about, say, A may cancel positive rules about A, while many weak rules about A could add to a strong weight of evidence about A.

2.7.2 It Is Useful to Have Rules in Information Theoretic Form

The methods of one of the authors [12,18,42,43] use Fano's mutual information measures such as $I(A\ ;\ B)$ are weights about the rule $(A\ ;\ B)$. However, the output of other data mining and data analytic methods in general can be converted to this form. They describe the degree of *association* (positive, zero, or negative) of A and B, and on a fuzzy set argument we can also express *correlations* in the same way, as arising from a multivariate analysis of trends between quantities [42]. These, however, may be said to differ in *type*, and more types can be defined. The information theoretic approach is convenient because all rules irrespective of type and complexity (say, $I(A\ \&\ B\ ;\ C\ \&\ D\ \&\ E)$, which is of complexity 5) can be co-ranked in the same list from large positive down to large negative.

2.7.3 A PC under Uncertainty

It would be nice if the inference method using such rules could handle a quantitative form of the PC in which both statements such as "All A are B" and "Some A are B" can be quantified to express uncertainty and can be represented in a common measure. Importantly, this would allow handling of *uncertain ontology*. It would also be nice if that measure can perform such inference in both directions of reasoning. Conceptually, $P(A\ |\ B)$ quantifies "All B are A" in that P("All B are A") = 1 and P("No B are A") = 0 for absolute certainty that it is true, and P("All B are A") = 0 and P("No B are A") = 1 for absolute certainty that it is not true.

In moving into the intermediate range of uncertainty, however, there are necessary complications. $\sqrt{P}\,(A\,|\,B)$ is held to be a quantification of "All B are

A for all practical purposes" and $\sqrt{P}\,(B \mid A)$ is held to be a quantification of "All A are B for all practical purposes." The arguable significance of "for all practical purposes" is that statements in PC are too strict to handle uncertainty. Just one observation of A and B together on a record for a patient or a molecule would otherwise convert "No A are B," i.e., P("Some A are B") = 0 to "Some A are B" with P("Some A are B") = 1, while just one case where A occurs without B would flip P("All A are B") = 1 to P("All A are B") = 0. The square root is a "hedge," i.e., a common practice in automated inference founded on set theory to take the square root to strengthen a probability if a weaker statement is made. Note that $\sqrt{P} > P$ except at $P = 0$ and $P = 1$.

This taking the square root may or may not be considered a rather forced device because that is exactly what is needed to bring everyday inference usefully into line with the amplitudes of QM as follows. Perhaps so, but it is at least a happy coincidence.

2.7.4 Borrowing from Dirac

To embrace the above considerations, one of the authors (B. Robson) proposed such a method called quantitative predicate calculus (QPC) [46] based on Dirac's system of inference [47] in QM [48–50]. Though this method may be read in the spirit of an example as just one possible method of inference, it has two entwined considerations. The first is that it can be made sufficiently general to represent (or critique) many other inference methods. The second is that it ought to be best practice, if realized correctly. It is intriguing that QM and Dirac's system is supposed to be universally applicable, not just to the world of the very small. Indeed it has been applied by cosmologists to the entire universe. We can, however, promise some changes in what follows, for the familiar everyday world of human experience. Manifestly, for example, the patient cannot be alive and dead with the same or different weights at the same time, in what QM calls a superposition of states. It is, however, a valid description expressing uncertainty in everyday inference when we do not see the evidence directly or when it is a prediction for the future (a distinction not made in certain languages like Mayan!). The difference seems to be that in the world of molecules and smaller things, such superposition actually exists now (if that has any meaning in QM) and, as much as we can tell, in the past. We can use that principle, for example, to calculate the molecular orbitals and energies of molecules and their conformations and interactions in the course of drug design.

The description of the QPC can be rephrased as making fundamental use of the constants

$$\mathfrak{t} = (1 + \boldsymbol{h})/2 \quad \text{and} \tag{2.40}$$

$$\mathfrak{t}^* = (1 - \boldsymbol{h})/2, \tag{2.41}$$

where in QM

$$\boldsymbol{hh} = -1, \tag{2.42}$$

in which case $\boldsymbol{h} = I = \sqrt{(-1)}$, and in everyday inference in the everyday world of human experience,

$$\boldsymbol{hh} = 1, \tag{2.43}$$

a so-called *split complex number* or hidden root of 1.

The basic idea is that Dirac's basic element [50] of QM inference, the bra-ket $\langle A \mid B \rangle$, can be expressed in terms of linear algebra as

$$\langle A \mid B \rangle = \mathfrak{u}\sqrt{P}\,(A \mid B) + \mathfrak{u}^*\sqrt{P}\,(B \mid A). \tag{2.44}$$

Actually, even more fundamental in QM are the basic indefinite vectors, the bra $\langle A|$ and the ket $|B\rangle$, and while relevant to inference (notably *incomplete inference*, on incomplete data [40]), they are not needed here. Note that $\langle A \mid B \rangle$ is in QM a *probability amplitude* and has analogy with probability $P(A \mid B)$ (while containing $P(B \mid A)$ too), and the case for two states, events, or measurements, and so on. A and B are easily extensible to more such, D, E, and so on, as in the above probability measures. The constants $\mathfrak{u}$ and $\mathfrak{u}^*$ imply a normalization for inference that confines values to a particular part of the complex plane, but this is not itself considered a fundamental difference from QM and indeed values not so confined can be defined by use of some operator $\boldsymbol{O}$ as in the QM notation $\langle \mathrm{A} \mid \boldsymbol{O} \mid \mathrm{B} \rangle$. They also imply that

$$\langle A \mid B \rangle = \langle B \mid A \rangle^*, \tag{2.45}$$

where * indicates taking the complex conjugate, i.e., flips the sign of the part of the value proportional to $\boldsymbol{h}$ (i.e., the imaginary part, in QM). This represents *conditionality reversal* and allows inference to be performed in both directions. Depending on interpretation, it may also represent *causality reversal* or *time reversal*.

The $P(A \mid B)$ and $P(B \mid A)$ are the required PC quantifications for universal qualification, to do with ontology (classification, taxononomy, metadata), and semantic forms of IF, i.e., $P(B \mid A)$ also quantifies "If A then B."

The existential qualification related to "Some A are B" and describing the degree of association of A and B is inherent, however, in the same equation (Eq. 2.41, which may be rewritten as

$$\langle A \mid B \rangle = [c(\langle A|) + c(|B\rangle)]\, e^{\frac{1}{2}I(A\,;\,B)}, \tag{2.46}$$

where the $I(A\,;\,B)$ is Fano's mutual information describing the degree of association of A and B and which can be obtained directly from data mining, and the weights c are

$$c(\langle A|) = \mathfrak{i}\sqrt{P}\,(A) \quad \text{and} \tag{2.47}$$

$$c(|B\rangle) = \mathfrak{i}^*\sqrt{P}\,(B). \tag{2.48}$$

For finite amounts of data, which is always the case, the above may be expressed in terms of the Riemann zeta function discussed above, and a prior degree of belief α. Strictly speaking, these relate to an expectation $\langle A \mid \zeta[D] \mid \mathrm{B}\rangle$, but can be operationally regarded as the more general solution for $\langle A \mid B\rangle$ on finite as well as on infinite data.

$$c(\langle A|) = \mathfrak{i}\sqrt{\left[e^{[\zeta(s=1,o[A]+(1-\alpha)N)-\zeta(s=1,N)]}+1-\alpha\right]} \tag{2.49}$$

$$c(|B\rangle) = \mathfrak{i}^*\sqrt{\left[e^{[\zeta(s=1,o[B]+(1-\alpha)N)-\zeta(s=1,N)]}+1-\alpha\right]} \tag{2.50}$$

Actually, one should write $\alpha[A]$ and $\alpha[B]$ since the value can vary with the state, event, measurement, and so on, and may be needed to satisfy certain marginal constraints (see above) that apply to the frequencies themselves, such that $\alpha[A]$ is not independent of $\alpha[A \,\&\, C]$, for example. The information part related to existential qualification can also readily encompass α in terms of adding to observed o and expected e frequencies and corresponding virtual frequency $(1 - \alpha)N$ where N is here the total amount of data:

$$I(A\,;\,B) = e^{\frac{1}{2}[\zeta(s=1,o[A\,\&\,B]+(1-\alpha[A\,\&\,B])N)-\zeta(s,\,e[A\,\&\,B]+(1-\alpha[A\,\&\,B])N)]}. \tag{2.51}$$

In more realistic inference, A of course may be exchanged with B or by other simple, e.g., C, D, or conjugate states, e.g., F & G & H states. What matters is whether the state A, and so on, referred to are associated with the bra side or the ket side of $\langle A \mid B\rangle$; hence, $c(\langle A|)$ and $c(|B\rangle)$. They may, however, be interconverted via the complex conjugate as described above.

Now inference net $\boldsymbol{N}$ can be built up as in QM. For example, the inference net

$$\langle A \mid B\rangle\langle B \mid C\rangle(\langle C \mid D\rangle\langle D \mid E\rangle,\langle C \mid F\rangle\langle F \mid G\rangle\langle G \mid I$$

defines three chains $\langle A \mid B\rangle\langle B \mid C\rangle$, $\langle C \mid D\rangle\langle D \mid E\rangle$, and $\langle C \mid F\rangle\langle F \mid G\rangle\langle G \mid H\rangle$ meeting at a fork node, C, and it is convenient to think of this as defining an operator and with some mathematical liberties so write

$$\langle A|\boldsymbol{N}|E\,\&\,H\rangle = \langle A \mid B\rangle\langle B \mid C\rangle(\langle C \mid D\rangle\langle D \mid E\rangle,\langle C \mid F\rangle\langle G \mid H\rangle) \tag{2.52}$$

Multiplication follows the rules analogous to complex multiplication but such that $\boldsymbol{hh} = 1$ in the QPC. This may be used to embrace any Bayesian net [40]. The above relates to logical AND: the OR case can be defined by addition essentially as in QM. Finally, there is nothing to stop use of forms such as $\langle A \,\&\, B \mid B \,\&\, C \,\&\, D\rangle$. This may be used to embrace any Bayesian

net [40]. For such (as discussed in a manuscript in preparation by the present authors), the use of normal complex numbers but replacing $\langle A \mid B \rangle$ by an expected value of information $\langle A \mid \boldsymbol{I} \mid B \rangle$, and then using addition such as $\langle A \mid \boldsymbol{I} \mid B \rangle + \langle B \mid \boldsymbol{I} \mid C \rangle$, is equivalent.

2.8 CLINICAL APPLICATIONS

Much of the above is naturally general, and data mining is by no means of interest only to the pharmaceutical industry. However, as described above, many of the above more unusual techniques have been motivated specifically to meet the challenges of medical data [18,42–44], including genomic [42] and proteomic medical [43] data, and to perform clinical inference [46]. Of interest here to the pharmaceutical industry are clinical trials, and the notion of the larger population of patients as a global cohort to feedback information about the outcomes and contraindications for already marketed drugs, including genomic and other biomarker information. As such, there is a need for systems that prepare patient data from a variety of legacy record forms and that feed the data in a unified form to R & D workflows including data mining, modeling of patient polymorphic protein targets, and so forth. Since clinical examples have been the primary examples used above, it is mainly necessary to add how these data are ultimately linked back to patient source data and to the role that data mining plays in an R & D workflow. The safest statement is to say that these can be quite varied, so IT solutions should allow for that.

One difference between records for patients and for chemical compounds is that the former is much more strongly affected by law and guidelines for best practice, imposed from without on any R & D. Data mining and the system that embeds it needs to take account of laws that require compliance. It would be desirable to have built-in tools not only to ensure patient privacy but also for conditional, fine-grained consent from the patient about what R & D can be done with the data. This needs some explanation. Current U.S. regulations basically deny any rights to a patient on what research is done if the patient ceases to be a human being. Federal law 45CFR46.102.f supporting the Health Insurance Portability and Accountability Act (HIPAA) regulations seeks to promote research through privacy, but, in effect, it reads that a de-identified human being is not a human being, and thus the patient has no rights over the data. In practice, that also tends to mean de-identified and/or dead. However, not everyone agrees with that, and there is no reason why IT should not add the ability for patients to have say on use of their data beyond the basic regulations. Then a great deal of data can be used that would be discarded under blanket consent.

Whereas many patients say at first that they do not care if they are de-identified, consent for research requires them to be well informed. To make the patient more aware, it is not hard to construct a list of some 40 possible R & D uses that give many patients course to rethink: uses for military pur-

poses, research involving conscious animals, R&D by tobacco companies, uses in context with religious principles, ethical issues of the stem cell type, and so forth, plus the desire (or not) to be informed of risks detected for the patient in the course of the R&D.

Though current legislation is well intended to restore the balance of solidarity (pooling data for common good) against autonomy (patient self-interest), significant documentation and intuitive reasoning in bioethics support the above kinds of concern. The patient already also has rights in any event notably in relation to living will and patient's bill of rights. Hence, the law may be ready for change. One reason is that the regulations are not well consistent at all levels anyway. In an interview by one of the authors with Cynthia L. Hahn, research privacy officer at The Feinstein Institute for Medical Research, Manhasset, NY, startling differences showed up between the federal definition of a human subject and the state definition, notably,

- Federal Law 45CFR46.102.f: "Human subject means a living individual…"
- New York State (NYS) (Article 24A, Section 2441): "Human subject" shall mean any individual…"
- The Federal Law overrides in case of doubt, but if the State Law conflicts, many extenuating circumstances may be allowed according to weight.
- The North Shore Long Island Jewish Health System extrapolates that in New York, research regulations apply to both the living and the deceased.
- However, NYS rules actually exclude epidemiological studies which the federal rules do not and the HIPAA does not apply to the information of the deceased, except again NYS does not create that distinction.
- Federal law regarding legally authorized representatives defers to state law, and state law makes no mention of research.
- Most International Rugby Boards have structured policies relating to the above that await legal challenge.

To encompass many scientific, technical, and compliance considerations is not trivial. One example effort, again an arguably good example because it embraces many other approaches, is the "genomic messaging system" (GMS) or more generally, clinical messaging system [51,52] developed by one of the authors and colleagues. It could be regarded as a storage and transmission protocol in the style of a specialized data exchange protocol as, for example, HTTP, but specialized for biomedical applications. However, it has much more elaborate capabilities. It translates HL7, other XML documents, and other legacy records into a lingua franca language called GMSL, from which the same or other format documents, or IT-driven R&D workflows, can be constructed. It thus obtained great interest in the press from XML.org and a variety of healthcare IT journals. Nonetheless, its flexibility has made it hard

for many to grasp. Apart from a rich set of encryption, compliance, DNA and protein management, various data porting types including medical images and software, and workflow control tools, the language is also interesting in that it can carry XML, can be carried in XML, or can be a combination of both. In addition, it follows the rules for writing DNA and protein sequences with extensions represented by the automatic and interactive annotation tools, and the compliance and workflow tools, so that it could be actually written into DNA sequences. It is an interesting concept that one could synthesize or clone a DNA sequence that can readily carry the language with sequencing and appropriate interpretation software. This bizarre notion, plus the fact that it is a concise language that is ("under the hood") machine coded in which every bit (meaning unit of information) counts, makes it of theoretical interest in studying the amounts of information in the interaction between patients and healthcare systems.

As one may imagine from the above, a degree of popularity of GMS on the Internet has been outweighed by lack of implementation in the healthcare field. Indeed GMS was intended to provide appropriate concepts and several novel concepts rather than a specific solution, so it remains at time of writing a research code. But all of the above capabilities and others not mentioned here (see above references) touch on issues that are relevant to feeding real clinical data to data mining as well as computational biology tools. The notion that each gene or other data is protected by nested passwords ensures compliance especially with patient wishes on what research can be carried out. Any system like GMS *and* the subsequent data mining must have the ability to preserve privacy, yet, for those patients who wish it, must have the ability to report back to the physician if the research shows the patient to be at risk. In general, this means a unique identifier assigned to the patient after de-identification, the key for which is held only by the physician and/or patient.

2.9 MOLECULAR APPLICATIONS

2.9.1 Molecular Descriptors

For a patient record, the datum was identified as a biomarker, at least if one adopts the most general description of the latter. For the molecule record, it is a *molecular descriptor*. "The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment" [53]. Common examples are chemical formulas, molecular weight, molecular dimensions, net charge, polarity, polarizability, dipole moments, hydrophobicity/hydrophilicity, melting and boiling points, and numbers of rotameric forms. In addition, there are pharmaceutically important biological data such as applications, efficacy, and toxicity, delivery and transport, renal clearance and active half-life data. Ease and cost of synthesis can also be appropriate data.

2.9.2 Complex Descriptors

There are also more complex and detailed gross descriptions of the molecular field including multipole (as opposed to simple dipole) descriptions, and the detailed spatial distribution of the van der Waals and electrostatic surface of the molecule. Typically, molecule representations may be divided into *molecular field description*s and more idealized and schematic *scaffolds* or *frameworks*, the latter often being the intended binding compliment to a similar representation framework for the binding site. There are also clustering representations of the classifications of the large number of conformers of flexible molecules with their relative energies.

"More complex" is to be taken in the sense of complexity used above. These cannot usefully be represented as a single rule relating to A, but require A & B & C &…. For optimal treatment in the same manner as described above, each of these will represent a separate column. Associations and correlations between columns *in that set* are typically not of interest since it is known that they are associated conceptually by relating to the same compound. It is not hard to set a powerful data mining tool to avoid calculating associations within the set, though relatively few methods seem set up to do that. Generally speaking, the complexity (number of states A & B, etc.) is of the order of 5–50, and up to a 100 or so in rare cases. The situation for analysis is thus challenging, but, roughly speaking, no worse than consideration of the number of parameters affecting complicated disease states such as cardiovascular diseases and certain cancers.

2.9.3 Global Invariance of Complex Descriptors

Though data preparation and in general preanalytic curation lies outside present scope, some comments on general concepts is important. In the above more complex cases, the parameters within a set have to be prepared in a way that shows *global invariance*; that is, they make sense in isolation when compared with descriptors outside that set. For example, a partial charge q with coordinates (x, y, z) is only meaningful in the *reference frame* represented by the complex set describing the molecule. A simple step in the right direction would be use of polar coordinates with rotation–translation superposition, though in practice more elaborate schemas based on solid ellipses and so forth are desirable. Also, taking all possible sets of four properties like charge and assignation, the distances between them, as well as the handedness of the implied tetrahedron, is helpful, ideally described as a single state of which there may be many thousands. Clearly, the number of molecular records should exceed this by at least a factor of 10 and ideally much more. This is most appropriate in the case of a scaffold or framework representing the molecule. The general idea is to make the description as meaningful as possible to data mining by stating that property A is located *at* a relative location where the latter has global not just local significance and has the fewest para-

meters possible. In practice, these methods are likely to work best in the present state of the art when confined to a drug series, i.e., with related chemistry.

A seeming alternative approach to global invariance comes to the same thing. Globally meaningful descriptors are generated at run time. An approach used in the data mining techniques of one of the authors [42,43] is that a datum can be a *variable* that ultimately gets replaced by a meaningful value. That means for the most part that entries can make reference (1) to further data such as graph data, a medical image, or a complex molecular descriptor, and (2) a pointer to a program or mode of use of a program that returns a (numeric or text) value to the datum, such that having a tangible value, it can now be data mined. This can be done prior to the main data mining exercise in a prerun, or at run time). Since the value returned should be as a single scalar quantity or text string for each entry on a record, several entries may be needed to return the full set of information. For example, when a graph is processed by polynomial regression to a polynomial of degree 6 (up to the sixth power), seven entries on a record (the zeroth term, i.e., intercept, plus those six x, x^2,…x^6) will have the above pointers to the external data and programs. This could be achieved in other more elaborate ways, but the above fits rather nicely into the simple theoretical-based schema of the method. The above example is insightful because it is typical that in the general case, the lower-order and/or stronger terms be data mined first, whether directly represented by polynomial coefficients or not. For example, the intercept can represent the basal level of a metabolite in a patient without administration of a drug, while the first-order term shows how the metabolite increases with the dosage. In the case of complex molecular descriptors, the process of converting the datum variable to a constant can of course be time consuming.

There is a special natural affinity of thermodynamic and statistical-mechanical processes with data mining principles *and* with the notion of a global descriptor that is simple but of global power. The underlying mathematics of data capture, storage, query, and mining issues applies not only to artificial electronic systems but to molecular systems [12]. The binding of a chemical compound and biological response at one target against many is a form of query subject to the same combinatorial and information theoretic considerations. In particular, it includes the concept of "partial match" (fit) to many targets, and the capability to understand the number of potential partial matches (interestingly, again using the Riemann zeta function) [12]. This opens the opportunity to link the data analytic to the biophysical aspects of the problem. Ultimately, it is the free energy of the drug at the target *in the active conformation of both*, meaning of the drug–target complex, which is responsible for the *primary* effect [54], and in this one may include the binding at false and/or toxicity generating targets. This is an information measure and specifically a mutual association information measure, $I(D\ ;\ P)$, an "ultimate descriptor" that relates to $-\Delta F^*$, which may be considered as a kind of mutual information measure with units in terms of the Boltzmann kT component. This is here best considered as the free energy for drug species and protein

species forming an activated complex, say, relative to isolated molecule conformer ground states D and P, but forming complex D^*P^*. Individually, D^* and P^* are the active, effect-triggering, conformational states of drug and protein [54]. Thus, $I(D\ ;\ P)$ has a special meaning as output, and the free energy of binding to a named target to an input descriptor has a special affinity with it. Ultimately, the key considerations can, as a first pass, be phrased in terms of the above free energy. Even effects of risk and cost can ultimately be included insomuch as the discussed above information theory is part of the broader field of decision theory. However, exploitation of such considerations has not, to the authors' knowledge, been made so far.

2.9.4 Peptide and Protein Structures

The biotechnology industry has it somewhat easier in that biosequences can be directly represented as lists for analysis. In other words, for each record is a peptide or protein sequence rendered ultimately as Residue[1]:=A, Residue[2]:=E, Residue[3]:=V, Residue[4]:=V, and so on. This makes data mining ideal for protein engineering. However, peptides developed as peptidomimetics and then with progressive development to a nonpeptide compound do represent a valid route in the pharmaceutical industry. In addition, one may make *staggered segment* choices like Residues[1–10]:= AEVVQLNATW, Residues[2–11]:=EVVQLNATWC, Residues[2–11]:= EVVQLNATWC, and so on. The general method of rendering sequences for mining is included in the data mining of one of the authors [42,43], with a specific study for enhancement of enzymic activity for a biotechnology company [43]. It is not surprising that these methods should work well with the particular software since as mentioned above, it has its roots in the widely used GOR method [45,55]. Thus, these mining techniques may be considered here as that method with the physical, biological, and chemical constraints removed. In particular, the analysis need not be confined to the effects of residues on the conformation of a residue 8 distant (the strongest effect, because of the local effects of secondary structure formation). Rather, rules for tertiary structure from the point of view of residue conformations and the effect on biological activity should emerge if there is appropriate and sufficient data.

Above, for design of small compounds suitable for tablet drugs, it was stated that "In practice, these methods are likely to work best in the present state of the art when confined to a drug series, i.e. with related chemistry." This is also true in a protein engineering study, where typically the records analyzed all relate to each other by homology, and may be natural proteins and/or proteins from previous engineering studies [43]. However, there is nothing to stop an application to all known protein sequences as in the GOR method, whence conformation prediction is likely to be the prime target unless there are, for example, specific descriptors about, say, enzymic activity or immunological effect. Clearly, such studies will require a great deal of data,

but bearing in mind that the early forms of GOR method [45] had to work with a handful of sequences of known conformation, there is now public access to roughly 700,000 sequences of which about half can be related in part to protein domains of known conformation by virtue of more than 16,000 or more protein structures known by X-ray crystallography or nuclear magnetic resonance (NMR). One important consequence of the GOR method is its use of the "#" function, which may now be identified with the Riemann zeta function and is well suited to handling low amounts of data. In data mining, protein sequences, and in general, there is always some complex data of interest for which the data are sparse.

2.9.5 Mining Systems Biology Input and Output

Often data mining can be considered as mining the input and output of a black box. For example, the above GOR methods was in fact specifically defined as an attempt to define the transform T in $\{S\} = T\{R\}$, where R is the amino acid residue sequence and $\{S\}$ is the corresponding sequence of residue states. Similar notions can be applied to the complex system represented by a patient. By extension, these notions also be applied potentially to the input and output of simulations in systems biology, including simulations supported by animal studies and linked to human clinical trials. When analyzing a simulation, however, there is nothing to stop sampling internal variables of this system.

2.10 DISCUSSION AND CONCLUSIONS

This review has focused on general principles that will hopefully inspire some thoughts, and perhaps the revision of some thoughts, in the analysis of the abundance of data available for analysis in the pharmaceutical industry.

A great deal could be written on specific techniques such as time series analysis, though in general, the above principles apply. Time series analysis is very much like time stamping data with an added function to interpolate effect. As with any parametric statistical method, this adds further powerful information, providing the functional form is right. Clustering and data reduction techniques have been significantly if relatively briefly mentioned. Perhaps they play a deeper role in data mining, but whether one considers them best as matters of processing input, output, or even intermediate steps is at present a matter of taste. The notion of dimensional reduction closely relates to the extent to which a rule, A & B & C & D & … can be reduced to simpler rules such as A & B, A & C, A & IB & D), and so forth. The ability to deduce this is in the natural domain of data mining, by inspection of the output rules and their weights.

Though medicinal chemistry remains a mainstay of the pharmaceutical industry, its conditioning to deal with specific patients or at least cohorts of them is profoundly impacted by the move to personalized medicine. Each

patient has his or her unique spectrum of delivery issues, target properties, responses to target activation, and toxicity due to unwanted targets.

> I believe we are moving into a remarkable and powerful new era in medicine and particularly in prescription drugs. I'd refer to it as an era of personalized medicine. During the next decade, the practice of medicine will change dramatically through genetically based diagnostic tests and personalized, targeted pharmacologic treatments that will enable a move beyond prevention to preemptive strategies. (M. Leavitt, pers. comm.)

This demands that there will ultimately be a need for a fundamental linkage between data mining patient records and data mining molecular records. At its most complete, this would, in principle, be to some extent a joint data, to the extent that each patient record shows personal differentiating biomolecular detail and specific outcomes to molecules of certain properties. Though this sounds lavish, it is not beyond imminent and future storage capacity, though bandwidth, aggregation, and distribution is an issue. Perhaps, for example, secure software robots roaming cyberspace and sending back relevant findings as rules will be required.

The key bottleneck lies in the "needle-in-a-haystack" hunt implied in processing that data, and particularly in a balance. This balance lies in the discovery of new unexpected relationships of enormous pharmaceutical worth, balanced against a preconstructed or reconstructed focus (D^*) to make an investigation tractable in reasonable time.

At present, no embracing rationale exists for addressing that balance. One may imagine that some stochastic element with weighting allows a focus while having an appropriate decreasing probability of discovery for progressively less relevant associations and correlations. To allow an appropriate distribution for this, with new focuses on areas of potential interest and capability to a company, the world would of course benefit from a pooling of new discoveries into a database from which all might draw. At present, pragmatic commercial considerations limit (though do not eliminate all of the time) such a global communal activity.

REFERENCES

1. Elstein AS. On the origins and development of evidence-based medicine and medical decision making. *Inflamm Res* 2004;53:S184–S189.
2. Robson B, Garnier J. The future of highly personalized health care. *Stud Health Technol Inform* 2002;80:163–174.
3. Shortliffe EH, Perrault LE, Wiederhold G, Fagan LM (Eds.) *Medical Informatics: Computer Application in Health Care and Biomedicine*. New York: Springer, 2000.
4. Trusheim MR, Berndt ER, Douglas FL. Stratified medicine: Strategic and economic implications of combining drugs and clinical biomarkers. *Nature Reviews* 2007;6:287–293.

5. Arlington S, Barnett S, Hughes S, Palo J. Pharma 2010: The Threshold of Innovation. Available at http://www.ibm.com/industries/healthcare/doc/content/resource/insight/941673105.html?g_type=rhc (accessed December 9–12, 2002).
6. DiMasi JA. The value of improving the productivity of the drug development process: Faster times and better decisions. *Pharmacoeconomics* 2002:20:1–10.
7. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: New estimates of drug development costs. *J Health Econ* 2003;22:151–185.
8. Katz R. Biomarkers and surrogate markers: an FDA perspective. *NeuroRx* 2004;1(2):189–195.
9. Gobburu JV. Biomarkers in clinical drug development. *Clin Pharmacol Ther* 2009;86(1):26–27.
10. Lesko LJ, Atkinson AJ Jr. Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: Criteria, validation, strategies. *Annu Rev Pharmacol Toxicol* 2001;41:347–366.
11. Hehenberger M, Chatterjee A, Reddy U, Hernandez J, Sprengel J. IT Solutions for imaging biomarkers in bio-pharmaceutical R&D. *IBM Sys J* 2007;46:183–198.
12. Robson B. The dragon on the gold: Myths and realities for data mining in biotechnology using digital and molecular libraries. *J Proteome Res* 2004:3:1113–1119.
13. Stockwell BR. Exploring biology with small organic molecules. *Nature* 2004 ;432(7019):846–854.
14. Kohane IS, Masys DR, Altman RB. The Incidentalome: A threat to genomic medicine. *J Am Med Assoc* 2006;296:212–215.
15. Donoho DL. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. Lecture to the American Mathematical Society "Math Challenges of the 20th Century." 2000. Available at http://www.stat-stanford.edu/~dononho/ (accessed August 8, 2000).
16. Cabena P, Hee Choi H, Soo Kim I, Otsuka S., Reinschmidt J, Saarenvirta G. Intelligent Miner for Data Applications Guide. Available at http://publib-b.boulder.ibm.com/Redbooks.nsf/ (accessed April 2, 1999).
17. Tang C, Zhang A. An Iterative Strategy for Pattern Discovery in High Dimensional Data Sets. 2002. Available at www.cse.buffalo.edu/DBGROUP/bioinformatics/papers/CIKM02.pdf (accessed November 11, 2002).
18. Robson B. Clinical and pharmacogenomic data mining: 3. Zeta theory as a general tactic for clinical bioinformatics. *J Proteome Res* 2005;4:445–455.
19. Drucker PF. *Innovation and Entrepreneurship*. New York: Harper and Row, 1985
20. Ackoff RL, *Emery FE. On Purposeful Systems*. London: Travistock, 1972.
21. Ackoff RL. "From Data to Wisdom," Presidential Address to ISGSR, June 1988. *J Appl Sys Anal* 1989;16:3–9.
22. Reupka WA. Efficiently coded messages can transmit the information content of a human across interstellar space. IAF, 41st International Astronautical Congress, Dresden, Germany, 1990.
23. Whittle P. *Probability*. London: Library of University Mathematics, Penguin Books, 1970.
24. Fischer RA. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1941.

25. Aitken AC. *Statistical Mathematics*. Edinburgh: Oliver and Boyd, 1945.
26. Wilks SS. *Mathematical Statistics. Section 7.7*. New York: Wiley, 1962.
27. Cooley WW, Lohnes PR. *Multivariate Data Analysis*. New York: John Wiley & Sons, Inc., 1971.
28. Dunteman GH. *Introduction to Multivariate Analysis*, Thousand Oaks, CA: Sage Publications, 1984.
29. Morrison DF. *Multivariate Statistical Methods*. New York: McGraw-Hill, 1967.
30. Overall JE, Klett CJ. *Applied Multivariate Analysis*. New York: McGraw-Hill, 1972.
31. Mardia KV, Kent JT, Bibby JM. *Multivariate Analysis*. London: Academic Press, 1979.
32. Tabachnick BG, Fidell LS. *Using Multivariate Statistics*. New York: *Harper Collins College Publishers*, 1996.
33. French S, Robson B. What is a conservative substitution? *J Mol Evol* 1983;19: 171–175.
34. Robson B, Finn PW. Rational design of conformationally flexible drugs. *ATLA J* 1984;11:67–78.
35. Baris C, Griffiths EC, Robson B, Szirtes T, Ward DJ. Studies on two novel TRH analogues. *Br J Pharmacol* 1986;87:173S.
36. Ward DJ, Griffiths EC, Robson B. A conformational study of thyrotrophin releasing hormone 1. Aspects of importance in the design of novel TRH analogues. *Int J Pept Protein Res* 1986;27:461–471.
37. Robson B, Garnier J. *Introduction to Proteins and Protein Engineering. B*. Amsterdam: Elsevier Press, 1986.
38. Kullback S. *Information Theory and Statistics*. New York: Wiley, 1959.
39. Fano R. *Transmission of Information*. New York: Wiley, 1961.
40. Savage LJ. *Recent Advances in Information and Decision Processes*, pp. 164–194. New York: Macmillan, 1962.
41. Goode H. *Recent Developments in Information and Decision Processes*, pp. 74–76. New York: Macmillan, 1962.
42. Robson B. Clinical and pharmacogenomic data mining. 1. The generalized theory of expected information and application to the development of tools. *J Proteome Res* 2003;2:283–301.
43. Robson B, Mushlin R. Clinical and pharmacogenomic data mining. 2. A simple method for the combination of information from associations and multivariances to facilitate analysis, decision and design in clinical research and practice. *J Proteome Res* 2004;3:697–711.
44. Mullins IM, Siadaty MS, Lyman J, Scully K, Carleton T, Garrett W, Miller G, Muller R, Robson B, Apte C, Weiss S, Rigoustsos I, Platt D, Cohen S, Knaus WA. Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Comput Biol Med* 2006;36:1351–1377.
45. Robson B. Analysis of the code relating sequence to conformation in globular proteins: Theory and application of expected information. *Biochem J* 1974; 141:853–867.

46. Robson B. The new physician as unwitting quantum mechanic: Is adapting Dirac's inference system best practice for personalized medicine, genomics and proteomics? *J Proteome Res* 2007;6:3114–3126.
47. Dirac AM. *The Principles of Quantum Mechanics*. Oxford: Oxford University Press, 1930.
48. Parks AD, Hermann MA. Dirac networks: An approach to probabilistic inference based on the Dirac algebra of quantum mechanics. NSWCDD(/TR-1/103/(2002025 080); Dahlgren Division naval Surface Warfare Center, 2001.
49. Chester M. *Primer of Quantum Mechanics*. New York: Dover Publications, 2003.
50. Messiah A. *Quantum Mechanics*. New York: Dover Publications, 1999.
51. Robson B, Mushlin R. Genomic messaging system for information-based personalized medicine with clinical and proteome research applications. *J Proteome Res* 2004;3:930–948.
52. Robson B, Mushlin R. The genomic messaging system language including command extensions for clinical data categories. *J Proteome Res* 2005;4:275–299.
53. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH, 2000.
54. Robson B, Garnier J. *Introduction to Proteins and Protein Engineering*. Amsterdam: Elsevier Press, 1986.
55. Garnier J, Rosbon B, Osguthorpe DJ. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 1978;120:97–120.

3

APPLICATION OF DATA MINING ALGORITHMS IN PHARMACEUTICAL RESEARCH AND DEVELOPMENT

KONSTANTIN V. BALAKIN AND NIKOLAY P. SAVCHUK

Table of Contents

3.1 INTRODUCTION

Data mining can be defined as the extraction of significant, previously unknown, and potentially useful information from data. In areas other than the life sciences and health care, the industries that actively use different data mining approaches include marketing, manufacturing, the financial industry,

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

government, engineering, and many others. In general, these industries all have massive amounts of information accumulated in special databases. In order to maximize the usefulness of this information, the mentioned industries apply specific computational approaches to discover specific patterns and trends from the data and to make predictions. Today, data mining is a huge industry providing a wide array of products and services that help obtain, generate, and analyze large quantities of data.

For the pharmaceutical industry inundated with truly massive amounts of chemical, biological, and clinical data, sophisticated data mining tools employing a variety of conceptually different methodologies are of vital importance. The specific applications range from a wide number of advanced chemoinformatics- and bioinformatics-based approaches to employment of toxicogenomic and chemogenomic databases, data mining in pharmacovigilance, predicting individual drug response, analysis of individual and population clinical trial results, and so on. Somewhat conditionally, the pharmaceutical data mining applications can be classified according to Figure 3.1.

The technological aspects of data mining, underlying algorithms, and software tools are described in a wide number of excellent publications [1–3] as well as throughout this book. In general, data mining uses a variety of machine learning approaches and also statistical and visualization methodologies to discover and represent knowledge in a form that can be easily understood by a human researcher. The objective is to extract as much relevant and useful information from data sets as possible. This chapter outlines different applications of data mining approaches in contemporary pharmaceutical R & D process, with some illustrations representing the authors' personal experiences in chemical data mining.

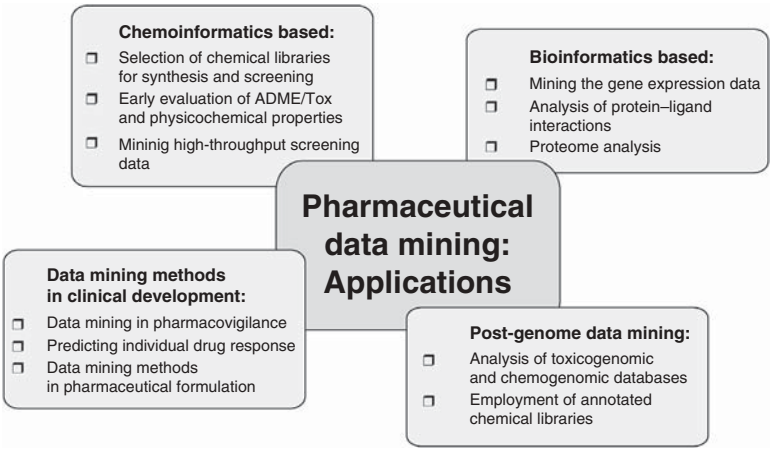


Figure 3.1 Main areas of application of data mining algorithms in contemporary pharmaceutical research and development. ADME/Tox = absorption, distribution, metabolism, and excretion/toxicity.

3.2 CHEMOINFORMATICS-BASED APPLICATIONS

Current chemoinformatics-based applications of data mining represent a large, well-developed group of technologies aimed in the rationalization of research efforts at early drug discovery stages. Somewhat conditionally, they can be divided into three specific subgroups, depicted in Figure 3.1 and discussed below.

3.2.1 Analysis of HTS Data

With the advent of high-throughput synthesis and screening technologies, simple statistical techniques of data analysis have been largely replaced by a massive parallel mode of processing, in which many thousands of molecules are synthesized, tested, and analyzed. As a result, the complete analysis of large sets of diverse molecules and their structural activity patterns has become an emerging problem. Hence, there is much current interest in novel computational approaches that may be applied to the management, condensation, and utilization of the knowledge obtained from such data sets. Among them, the modern data mining approaches to processing HTS data and developing biological activity models represent an important cluster of technologies that provide a functional interface between real and virtual screening programs [4–7].

The analysis of HTS data has many challenges and typical problems: (1) dramatic imbalance between the number of active and inactive compounds; (2) structure–activity relationship (SAR) data recovered during HTS analysis inevitably involve threshold and nonlinear effects; (3) large amounts of random or systematic measurement errors, noisy nature of the data, which can cause significant false positive and false negative levels; (4) real chemical databases usually have strong local clustering in the descriptor space; (5) potent compounds belonging to different chemotypes may be acting in different ways in the same assay; as a result, different mechanisms might require different sets of descriptors within particular regions of the descriptor space to operate, and a single mathematical model is unlikely to work well for all mechanisms. Because of these and other complexities discussed in literature [8,9], traditional statistical analysis methods are often ineffective in handling HTS analysis problems and tend to give low accuracy in classifying molecules. To meet these challenges and to open the way for the full exploitation of HTS data, sophisticated data mining methods and specialized software are required. The ultimate goal of the research efforts in this field is to develop smart and error-tolerant ways to measure and interpret raw HTS data and to transform them into a knowledge of target–ligand interactions.

Several comprehensive reviews describe contemporary approaches to HTS data mining [7,10,11]. Chapter 7 of this book describes theoretical and practical aspects of a knowledge-based optimization analysis (KOA) algorithm and

illustrates its applications in high-throughput data mining at several different stages of the drug discovery process.

To address specific issues associated with high-throughput data mining, in the last decade, there has been a significant increase in the number of industrial software tools that aim to provide complex solutions for the analysis of HTS results (Table 3.1). These tools, typically integrated in multifunctional chemoinformatics platforms, can be used for HTS quality control, data visualization, clustering, classification, generation of SAR models, and integration with genomic data.

As an example, Leadscope (Leadscope, Inc.) offers a portfolio of advanced solutions useful in the analysis of HTS data. The software package was used for the solution of several practical HTS data mining tasks (for example, see references 12 and 13). The procedure of HTS data analysis in Leadscope comprises three major phases. At the first phase, the primary screening set is filtered to identify and/or to remove undesirable compounds based on physical properties, the presence of toxic or reactive groups, or more subtle distinctions based on "drug likeness." Phase 2 seeks to identify local structural neighborhoods around active compound classes and includes similar inactive compounds. One of the used algorithms representing a combination of recursive partitioning and simulated annealing methods consistently identifies structurally homogeneous classes with high mean activity. Phase 3 is the analysis of SARs within local structural neighborhoods. The local neighborhoods are structurally homogeneous and include both active and inactive compounds. These tools comprise R-group analysis, macrostructure assembly, and building local prediction models.

The rapid growth of the integrated program tools for HTS data analysis reflects the increasing need in sophisticated technologies that open the way for the full exploitation of HTS data and meet the associated challenges.

In addition to finding active compounds among those screened, it would be very useful to know how to find additional active compounds without having to screen each compound individually. Sequential HTS (also known as recursive screening and progressive screening) screens compounds iteratively for activity, analyzes the results using data mining approaches, and selects a new set of compounds for the next screening based on what has been learned from the previous screens. The purpose of this iterative process is to maximize information about ligand–receptor interactions by using high-throughput screening and synthesis technologies to ultimately minimize early-stage discovery costs. Several cycles of screening appeared to be more efficient than screening all the compounds in large collections [14–16]. In most of the reported examples of the practical application of this technology, compound selection during these iterative cycles is driven by rapid SAR analyses using recursive partitioning techniques. Blower et al. [17] studied the effects of three factors on the enrichment ability of sequential screening: the method used to rank compounds, the molecular descriptor set, and the selection of the initial training set. The primary factor influencing recovery rates was the method of selecting the initial training set. Because structure–activity information is

TABLE 3.1 Chemoinformatics Software (with Focus on HTS Data Mining)

Name	Developer	Features
Leadscope	Leadscope http://www.leadscope.com/	Software for end-to-end analysis of HTS data sets; includes special tools for filtering, clusterization (RPSA method, hierarchical agglomerative clustering), and SAR analysis
Screener	GeneData AG, Switzerland http://www.genedata.com/	An integrated tool for comprehensive analysis of HTS data, including quality control, standardization, compound classification, biological-chemical evaluation and pharmacological classification
SARNavigator	Tripos http://www.tripos.com/	Suite of HTS data analysis tools, including computation of molecular descriptors, SAR analysis of scaffolds and R-group fragments, visualization, and QSAR modeling
QuaSAR-Binary	Chemical Computing Group Inc. http://www.chemcomp.com/	HTS data analysis tool based on the binary QSAR approach
ClassPharmer	Bioreason, Inc.	Suite of programs for HTS data analysis, including data normalization, classification (based on adaptively grown phylogenetic-like trees), and SAR extraction
HTSview	The Fraunhofer Institute for Algorithms and Scientific Computing http://www.scai.fraunhofer.de/	Program for interactive analysis and visualization of HTS data and extraction of SAR (definition of "biophores")
GoldenHelix	GoldenHelix http://www.goldenhelix.com/	Program for molecular data analysis based on recursive partitioning
BioAssay HTS and BioSAR Browser	CambridgeSoft Corporation http://www.cambridgesoft.com/	Suite of programs for HTS data analysis, including quality control, visualization, data mining, and SAR extraction
DecisionSite	Spotfire, Inc. http://www.spotfire.com/	HTS data analysis program with powerful quality control system, interactive data visualization, and integrated web browser
Accord HTS	Accelrys http://www.accelrys.com/	HTS data management system, including data analysis and visualization

RPSA = recursive partitioning and simulated annealing.

incrementally enhanced in intermediate training sets, sequential screening provides significant improvement in the average rate of recovery of active compounds when compared to noniterative selection procedures.

3.2.2 Target-Specific Library Design

Data mining methods used for correlation of molecular properties with specific activities play a significant role in modern drug discovery strategies. Since the current drug discovery paradigm states that mass random synthesis and screening do not necessarily provide a sufficiently large number of high-quality leads, such computational technologies are of great industrial demand as a part of a virtual screening strategy [18–20]. The most typical application of such algorithms includes development of predictive models that can further be used for selection of screening candidates from chemical databases.

Among a big variety of data mining algorithms used for the design of target-directed libraries, the artificial neural networks (ANNs) are about to become *de facto* standard [21]. ANN is relatively easy to use yet is a powerful and versatile tool. However, despite these clear advantages, ANN suffers from some drawbacks discussed in literature, such as a "black box" character of ANN, which may hamper analysis and interpretation of developed models, and may result in possible overfitting (i.e., ability to fit for training data noise rather than actual data, which results in poor generalization). Support vector machine (SVM) algorithm [22] (see Chapter 15 of this book) represents a useful alternative, at least as powerful and versatile as ANNs. In the last decade, SVM approach has been used in various areas, from genomics and face recognition to drug design. The researchers at ChemDiv tested SVM as a classification tool in several drug discovery programs and found it typically outperforming ANNs [23,24].

As an illustration, SVM algorithm was used for selection of compounds for primary and secondary screening against *abl* tyrosine kinase [24]. A set of 1249 known tyrosine kinase (TK) ligands from different classes was used as a positive training set, TK(+), and a set of over 8592 compounds, representing over 200 various nonkinase activities, was considered a negative training set, TK(−). The training set compounds represented late-stage (pre)clinical candidates and marketed drugs. A total of 65 molecular descriptors were calculated, which encode lipophilicity, charge distribution, topological features, and steric and surface parameters. The redundant variables were removed using sensitivity analysis, and the resulting eight molecular descriptors were used for generation of the SVM classification model. For the model validation, an internal test set (25% of the entire training database) was used. Curves 1 and 2 (Fig. 3.2) show the distributions of calculated SVM scores for compounds in TK(+) and TK(−) internal test sets, correspondingly. With the threshold score 0.4, the model correctly classified up to 70% of TK(+) and 80% of TK(−) compounds.

Then we carried out a wet laboratory experimental validation of the developed model via high-throughput screening of 5000 compounds from ChemDiv's

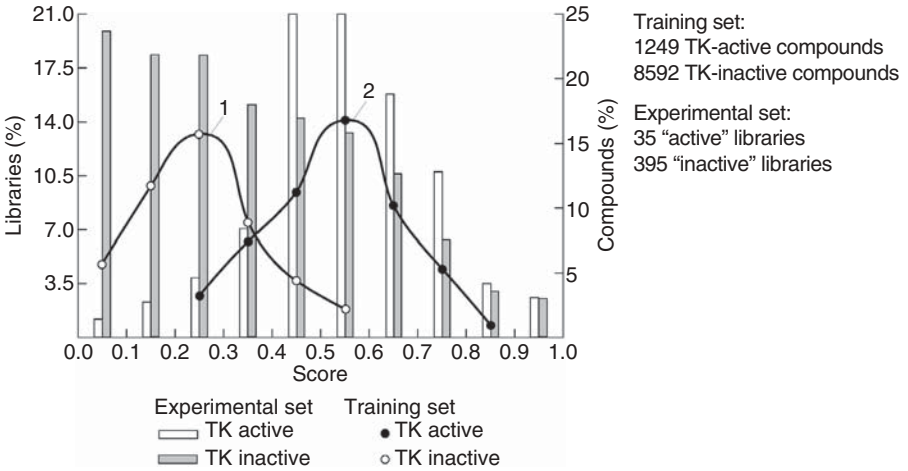


Figure 3.2 SVM score distribution for the training set compounds (curves 1 and 2) and for the experimental database screened against *abl* kinase (I. Okun and K.V. Balakin, unpublished data).

corporate compound database against *abl* kinase (I. Okun and K.V. Balakin, unpublished data). The total experimental databases consisted of 430 small congeneric combinatorial subsets (libraries) typically represented by 5–15 compounds. Based on experimental results, all these libraries were divided into two categories, "active" and "inactive." The "active" subsets were defined as having at least one active compound (35 combinatorial subsets in total); "inactive" subsets had no active compounds (395 combinatorial subsets in total). This categorization model is reasonable for distinguishing between two categories of chemotypes: (1) "active" chemotypes suitable for further development (for example, via quantitative structure activity relationships [QSAR] modeling, SAR library generation, and further optimization) and (2) chemotypes with few or no "actives" for which no effective development can be anticipated. For each combinatorial subset, an average SVM score was calculated. It was observed that the developed SVM model was able to discriminate between "active" and "inactive" libraries (histograms on Fig. 3.2). Although there is a certain overlap between SVM score regions of active and inactive libraries, these distributions are clearly distinct, thus indicating a good discrimination power of the trained network. We observe a significant enrichment in the high-scoring regions with TK inhibitor chemotypes: the portion of active chemotypes in the high-scoring regions is almost an order of magnitude higher than in the initial database. The developed model permits early evaluation of the protein kinase inhibition potential of small molecule combinatorial libraries. It can also be used as an effective *in silico* filtering tool for limiting the size of combinatorial selections.

3.2.3 Assessment of ADME/Tox and Physicochemical Properties

Binding to the target protein is only part of the process of drug discovery, which requires molecules that are readily synthesizable with favorable molecular properties or what has been termed "drug likeness" [25–28]. Drug likeness studies are an attempt to understand the chemical properties that make molecules either successful or possibly expensive clinical failures. Similarly, ADME/Tox properties are recognized alongside therapeutic potency as key determinants of whether a molecule can be successfully developed as a drug. As a result, *in silico* assessment of such properties of compounds is one of the key issues at early drug discovery stages. To address this need, many different data mining approaches and tools were developed for prediction of key ADME/Tox and physical-chemical parameters.

Thus, human intestinal absorption (HIA) and blood–brain barrier (BBB) permeability are the major issues in the design, optimization, and selection of candidates for the development of orally active and/or central nervous system–active pharmaceuticals [29–37]. In general, the molecular properties affecting HIA and BBB penetration via passive diffusion mechanisms are well understood, and the reported models adequately describe this phenomenon. However, while much effort continues to be expended in this field with some success on existing data sets, perhaps the most pressing need at this time is for considerably larger, high-quality sets of experimental data and for an effective data mining algorithm to provide a sound basis for model building [38].

There are also many complex properties related to ADME/Tox that have been modeled in silico using data mining approaches. Thus, molecular clearance, which is indicative of elimination half-life that is of value for selecting drug candidates, has been modeled using ANNs and multivariate statistical techniques [39]. Another complex property is the volume of distribution that is a function of the extent of drug partitioning into tissue versus plasma; along with the plasma half-life, it determines the appropriate dose of a drug, and there have been several attempts at modeling this property [40–43]. The plasma half-life and the integral plasma protein binding have been modeled with Sammon and Kohonen maps using data for 458 drugs from the literature and several molecular descriptors [42]. Metabolic effects, CYP450-mediated drug–drug interactions and other complex ADME/Tox-related phenomena were also modeled using data mining techniques [44–47].

Combinations of different computational models for ADME are applicable to the selection of molecules during early drug discovery and represent an approach to filtering large libraries alongside other predicted properties. With the addition of further data, it is likely that even more significant models can be generated. At present, the different methods we have used could be combined and used in parallel as a consensus modeling approach to perhaps improve the predictions for external molecules (for example, see reference 47).

Several comprehensive reviews elucidate the current state in the developments of predictive ADME/Tox models based on data mining approaches

[48–52]. While considerable progress has been achieved in ADME predictions, many challenges remain to be overcome. It was argued that the robustness and predictive capability of the ADME models are directly associated with the complexity of the ADME property [53]. For the ADME properties involving complex phenomena, such as bioavailability, the *in silico* models usually cannot give satisfactory predictions. Moreover, the lack of large and high-quality data sets also greatly hinders the reliability of ADME predictions.

Early assessment of the potential toxicity of chemical compounds is another important issue in today's drug discovery programs. There are numerous limitations that affect the effectiveness of the early toxicity assessment, which can create a significant bottleneck in the drug discovery process. The ability to predict the potential toxicity of compounds based on the analysis of their calculated descriptors and structural characteristics prior to their synthesis would be economically beneficial when designing new drugs. Although various *in silico* algorithms of toxicity prediction have been reported [54,55], in general, the quantitative relationship between the toxicity of structurally diverse compounds and their physicochemical/structural properties has proved to be an elusive goal due, in part, to the complexity of the underlying mechanisms involved. Chapter 5 discusses data mining approaches and tools for the prediction of the toxic effects of chemical compounds.

Prediction of key physicochemical properties of chemical compounds is another serious problem in modern drug discovery, which can be addressed using different data mining approaches. In particular, solubility of chemical compounds represents a highly important issue critically influencing the success of early drug discovery projects [56–58]. As one practical example from the authors' experience, clusterization based on structural fingerprints can be used for discrimination between soluble and insoluble compounds in dimethyl sulfoxide (DMSO). Poor DMSO solubility represents a serious problem for large-scale automatic bioscreening programs, and several computational models have been developed for the prediction of this property (reviewed in reference 57). In particular, in 2004, we have described a computational approach based on the Kohonen self-organizing map (SOM) algorithm and a large proprietary training database consisting of 55,277 compounds with good DMSO solubility and 10,223 compounds with poor DMSO solubility [59]. The developed model was successful in classification of DMSO well-soluble and poorly soluble compounds from the internal test set.

Later, we have performed a nonhierarchical clusterization of the same database using the Jarvis–Patric method [60] based on the nearest-neighbor principle. Daylight fingerprints were used as molecular descriptors. A total of 68,124 structures (93% of the whole database) were clustered yielding 3409 clusters (average cluster size of 20 compounds). Graphically, the results of the clusterization procedure are presented in Figure 3.3. We have obtained three principal types of clusters: (1) clusters with domination of soluble compounds, (2) intermediate type in which soluble and insoluble compounds are present

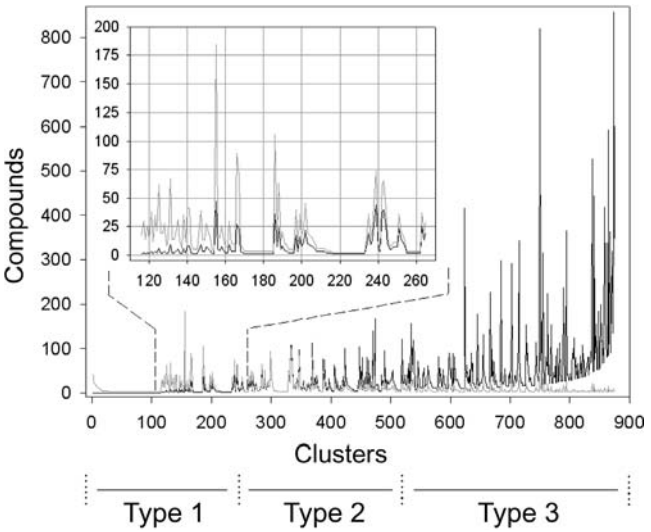


Figure 3.3 Irregularity of distribution of DMSO well-soluble and poorly soluble compounds across the training data set (K.V. Balakin and Y.A. Ivanenkov, unpublished data).

in approximately equal proportion, and (3) clusters with domination of insoluble compounds. The obtained data suggest that for clusters of the first and third types, the Jarvis–Patrick clusterization provides a reasonable and computationally inexpensive tool for the classification of compounds based on their DMSO solubility. However, there is a need in further SAR refinement for the clusters of type 2, since clusterization based on structural fingerprints only does not lead to discrimination between these categories of compounds.

In order to classify compounds within clusters of this type, we used a special data dimensionality reduction algorithm, Sammon nonlinear mapping (NLM) [61]. Figure 3.4 depicts Sammon maps of two large clusters, which contain congeneric compounds of the shown general structures. On each map, well-soluble compounds (black crosses) occupy regions distinctly different from the areas of location of poorly soluble compounds (gray triangles). The map is based on the eight calculated molecular descriptors, the same used for generation of SOM [59]. Similar results were obtained for most of the other individual clusters of the second type. Obviously, in this case, physicochemical rather than substructural determinants play a key role in DMSO solubility.

The ability to optimize different molecular parameters (such as target-specific activity and ADME/Tox-related and physicochemical properties) in a parallel fashion is a characteristic feature of many chemoinformatics-based data mining methods. In this case, we have a multiobjective optimization

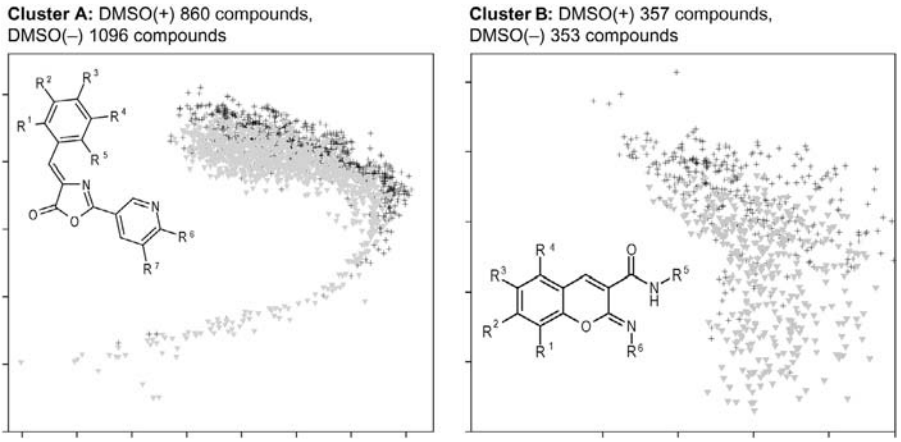


Figure 3.4 Sammon map classifications of well-soluble (black crosses) and poorly soluble (gray triangles) compounds in DMSO from two congeneric compound sets (K.V. Balakin and Y.A. Ivanenkov, unpublished data).

problem, which has become a topic of growing interest over the last decade in the pharmaceutical industry. The general idea of multiobjective optimization is to incorporate as much knowledge into the design as possible. Many factors can be taken into consideration, such as diversity, similarity to known actives, favorable physicochemical and ADME/Tox profile, cost of the library, and many other properties. Several groups have developed computational approaches to allow multiobjective optimization of library design [62,63]. One method developed by researchers from 3-Dimensional Pharmaceuticals employs an objective function that encodes all of the desired selection criteria and then identifies an optimal subset from the vast number of possibilities [63]. This approach allows for the simultaneous selection of compounds from multiple libraries and offers the user full control over the relative significance of a number of objectives. These objectives include similarity, diversity, predicted activity, overlap of one set of compounds with another set, property distribution, and others.

3.3 BIOINFORMATICS-BASED APPLICATIONS

A plethora of bioinformatics-based applications is focused on sequence-based extraction of specific patterns or motifs from genomes and proteomes and also on specific pattern matching. Thus, as an essential part of bioinformatics-based applications, microarray analysis technologies have become a powerful technique for simultaneously monitoring the expression patterns of thousands of genes under different conditions. The principal goal is to identify groups of genes that manifest similar expression patterns and are activated by similar

conditions. A general view of data mining techniques used in gene expression analysis is presented in Chapter 8. Biological interpretation of large gene lists derived from high-throughput experiments, such as genes from microarray experiments, is a challenging task. A wide number of publicly available high-throughput functional annotation tools, such as those listed in Table 3.2, partially address the challenge.

TABLE 3.2 Bioinformatics Software (with Focus on Gene Expression Data Analysis)

Software	Web Site	Description
BioWeka	http://www.bioweka.org	A popular and freely available framework that contains many well-known data mining algorithms. This software allows users to easily perform different operations with bioinformatics data, such as classification, clustering, validation, and visualization, on a single platform
DAVID Functional Annotation Tool Suite	http://david.abcc.ncifcrf.gov/summary.jsp	A component in the DAVID Bioinformatics Resources (http://david.niaid.nih.gov/) for biological interpretation of large gene lists derived from high-throughput experiments, such as genes from microarray experiments
SIGMA	http://sigma.bccrc.ca/	A publicly available application to facilitate sophisticated visualization and analysis of gene expression profiles
MIAMExpress	http://www.ebi.ac.uk/miamexpress/	An annotation tool at the European Bioinformatics Institute (EBI) database
Gene Traffic	http://www.stratagene.com/	A microarray data management and analysis software
Ipsogen Cancer Profiler	http://www.ipsogen.com/	A bioinformatics system composed of Discovery Software tools and of ELOGE database, utilized to identify transcriptional signatures belonging to each cancer type

TABLE 3.2 ***Continued***

Software	Web Site	Description
BioArray Software Environment (BASE)	http://base.thep.lu.se/	A web-based open source microarray database and analysis platform
GeneData Expressionist™	http://www.genedata.com/products/expressionist/	A computational system that efficiently processes gene expression data generated by high-throughput microarray technologies
GeneDirector	http://www.biodiscovery.com/index/genedirector	An image and data analysis platform with Oracle database capability to enhance microarray discovery
Multiconditional Hybridization Intensity Processing System (MCHIPS)	http://www.dkfz-heidelberg.de/mchips/	A system for microarray data warehousing and microarray data analysis
maxd	http://www.bioinf.man.ac.uk/microarray/maxd/	A data warehouse and visualization environment for genomic expression data
Genowiz™	http://www.ocimumbio.com/web/default.asp	A comprehensive multiplatform package for tracking and analyzing Gene Expressions data
TeraGenomics™	http://www.teragenomics.com/	A scalable, high-performance data warehousing solution for analyzing and sharing Affymetrix® GeneChip® data
TM4	http://www.tm4.org/	The TM4 suite of tools consist of four major applications, Microarray Data Manager (MADAM), TIGR_Spotfinder, Microarray Data Analysis System (MIDAS), and Multiexperiment Viewer (MeV), as well as a Minimal Information about a Microarray Experiment (MIAME)-compliant MySQL database
GenStat	http://www.vsn-intl.com/genstat/gsprod_details.htm	A comprehensive statistics system for gene data analysis
GeneMaths XT	http://www.applied-maths.com/genemaths/genemaths.htm	A complete and professional software package for microarray analysis

The concept of genome mining for novel natural product discovery [64] promises to provide new bioactive natural compounds. This approach exploits the huge and constantly increasing quantity of DNA sequence data from a wide variety of organisms that is accumulating in publicly accessible databases. Using computational sequence comparison tools, genes encoding enzymes likely to be involved in natural product biosynthesis can be readily identified in genomes. This information can be exploited in a variety of ways in the search for new bioactive natural products.

Data mining of biomedical data has boosted the post-genome target discovery, which is one of the key steps in the biomarker and drug discovery pipeline to diagnose and fight human diseases. A recent review explicates various data mining approaches and their applications to target discovery with emphasis on text and microarray data analysis [65].

Proteomic studies involve the identification as well as the qualitative and quantitative comparison of proteins expressed under different conditions, and the elucidation of their properties and functions, usually in a large-scale, high-throughput format [66,67]. The high dimensionality of data generated from these studies requires the development of improved bioinformatics tools and data mining approaches for efficient and accurate data analysis of biological specimens from healthy and diseased individuals.

Protein–ligand interactions are crucial in many biological processes with implications to drug targeting and gene expression. The nature of such interactions can be studied by analyzing local sequence and structure environments in binding regions in comparison to nonbinding regions. With an ultimate aim of predicting binding sites from sequence and structure, such methods, described in Chapter 9, constitute an important group of data mining approaches in the field of bioinformatics.

Immunoinformatics (Chapter 11) is a new discipline in the field of bioinformatics that deals with specific problems of the immune system. As interest in the vaccine sector grows, the highly empirical disciplines of immunology and vaccinology are on the brink of reinventing themselves as a quantitative science based on data delivered by high-throughput, post-genomic technologies. Immunoinformatics addresses problems such as the accurate prediction of immunogenicity, manifest as the identification of epitopes or the prediction of whole protein antigenicity. Application of such methods will greatly benefit immunology and vaccinology, leading to the enhanced discovery of improved laboratory reagents, diagnostics, and vaccines.

To date, while the developed data mining tools and approaches for bioinformatics-based applications are extremely useful and have been employed in hundreds of research projects, the development of other effective data mining algorithms, as additional components to the already existing programs, will improve the power of investigators to analyze their gene and protein sequences from different biological angles.

3.4 POST-GENOME DATA MINING

Post-genome data mining tends to combine techniques and data sources employed by chemoinformatics and bioinformatics. We are now witnessing rapid development of new methods for mining the chemical genomics data based on the integration of these important disciplines.

The effective identification and optimization of high-quality pharmaceutical leads across diverse classes of therapeutic targets can be based on the systematic analysis of structural genomics data [68,69]. In particular, annotated compound libraries have emerged as an interesting phenomenon in drug discovery in the post-genomic era [70]. The underlying strategy behind the selection of these biased libraries is to bring together information pertaining to the relationships between biological targets, respective small molecule ligands and their biological functions in a single knowledge management platform. Specific issues discussed in Chapter 6 include chemogenomics databases, annotated libraries, homology-based ligand design, and design of target-specific libraries, in the context of G protein-coupled receptor (GPCR)-targeted drug design.

For example, a collection of properly characterized ligands covering a diverse set of mechanisms of action can be an extremely useful tool to probe disease pathways and to identify new disease-associated targets belonging to well-validated target families within these pathways. A method was reported for testing many biological mechanisms and related biotargets in cellular assays using an annotated compound library [71]. This library represents a collection of 2036 biologically active compounds with 169 diverse, experimentally confirmed biological mechanisms and effects. These compounds were screened against A549 lung carcinoma cells, and subsequent analysis of the screen results allowed the determination of 12 previously unknown mechanisms associated with the proliferation of the studied carcinoma cells.

Rapid growth of researches demonstrating the value of chemogenomic libraries in drug discovery triggered rapid growth of supporting industrial technological solutions, such as BioPrint® [72,73] and DrugMatrix® [74]. Another recent example is GLIDA, a novel public GPCR-related chemical genomics database that is primarily focused on the correlation of information between GPCRs and their ligands [75]. The database is integrated with an *in silico* screening module based on statistical machine learning of the conserved patterns of molecular recognition extracted from comprehensive compound–protein interaction data.

The combination of HTS and genome data analysis provides novel opportunities in drug design. The achievements in genome researches allow for establishing the relationships between ligands and targets, and thus offer the potential for utilizing the knowledge obtained in the screening experiments for one target in lead finding for another one. In particular, activity profiles based on parallel high-throughput assays can be used to generate the ligand–

target arrays. Using such arrays, subtle correlations between gene expression and cell sensitivity to small molecule compounds can be identified. For example, in-depth investigations of compound mode of action and side effects can be conveniently provided by analysis of cellular gene expression patterns and their modification by small molecule compounds [76]. In another work [77], 60 cancer cell lines were exposed to numerous compounds at the National Cancer Institute (NCI) and were determined to be either sensitive or resistant to each compound. Using a Bayesian statistical classifier, it was shown that for at least one-third of the tested compounds, cell sensitivity can be predicted with the gene expression pattern of untreated cells. The gene expression patterns can be related not just to the drugs as entities but to particular substructures and other chemical features within the drugs [78]. Using a systematic substructure analysis coupled with statistical correlations of compound activity with differential gene expression, two subclasses of quinones were identified whose patterns of activity in the NCI's 60-cell line screening panel correlate strongly with the expression patterns of particular genes. The researchers from GeneData (GeneData AG, Basel, Switzerland; http://www.genedata.com/) developed a software for the analysis of HTS data integrated with gene expression databases. On the basis of this integrated system, hypotheses about possible biotargets for the analyzed hits can be generated.

Despite tremendous efforts of computational chemists, effective prediction of toxic effects remains an elusive goal. This is due, in part, to the fact that drug candidates generally target multiple tissues rather than single organs and result in a series of interrelated biochemical events. These challenges are best addressed through data collection into a well-designed toxicogenomic database. Successful toxicogenomic databases serve as a repository for data sharing and as resources for data mining that pave the way to effective toxicity prediction. Chapter 10 describes the existing toxicogenomic databases and approaches to their analysis.

It can be envisaged that a meaningful integration of chemical and biological data with advanced methods of data analysis will significantly facilitate the future efforts of the drug discovery community directed to efficient discovery of leads across diverse classes of biological targets.

3.5 DATA MINING METHODS IN CLINICAL DEVELOPMENT

The past decade has seen a significant increase in the number of reported applications of data mining tools being used in clinical development.

Thus, in a search for the personalized therapies, the researchers actively used various pharmacogenomic data mining approaches to identify a genetic marker, or a set of genetic markers, that can predict how a given person will respond to a given medicine. A significant challenge for pharmacogenetic researchers is therefore to identify and to apply appropriate data mining methods for finding such predictive marker combinations. Chapter 13 of this

book describes how data mining tools can be used for finding such combinations, with the main focus on methods based on partitioning the data along a tree with various optimization criteria, methods based on combinatorial procedures searching for the best combination of input genetic variables as predictive of the phenotype of interest, and neural network methods that attempt to classify phenotype by training successive layers through an activation function.

Another interesting area of application of data mining tools is the development of pharmaceutical formulation. These applications include, for instance, immediate and controlled release tablets, skin creams, hydrogel ointments, liposomes and emulsions, and film coatings. Traditionally, formulators used statistical techniques such as a response surface methodology to investigate the design space. However, for the complex formulations, this method can be misleading. As a result, more advanced data mining techniques have been actively implemented during the last decade in the field. Among them are ANNs, genetic algorithms, neurofuzzy logic, and decision trees. Chapter 14 of this book reviews the current state of the art and provides some examples to illustrate the concept.

Possible benefits associated with wide application of data mining in pharmaceutical formulation include effective and rapid analysis of available data sets, effective exploration of the total design space, irrespective of its complexity, ability to accommodate constraints and preferences and to generate simple rules intuitively understandable for human researchers. Business benefits are primarily associated with enhancement of product quality and performance at low cost.

The principal concern of pharmacovigilance is the detection of adverse drug reactions (ADRs) as soon as possible with minimum patient exposure. A key step in the process is the detection of "signals" using large databases of clinical information; such an analysis directs safety reviewers to associations that might be worthy of further investigation. In the last decade, several health authorities, pharmaceutical companies, and academic centers are developing, testing, and/or deploying various data mining tools to assist human reviewers. For example, since 1998, Bayesian Confidence Propagation Neural Network (BCPNN) data mining has been in routine use for screening of the World Health Organization (WHO) adverse reaction database, Vigibase [79]. The identification of drug/ADR combinations that have disproportionately high reporting relative to the background of all reports constitutes the first quantitative step in the Uppsala Monitoring Centre (UMC) signal-detection process. A computerized system for drug/ADR signal detection in a spontaneous reporting system (SRS) has recently been developed in Shanghai [80]. This system is very useful for post-marketing surveillance on both chemical medicine and Chinese traditional medicine.

Data mining analyses for the purposes of pharmacovigilance are usually performed on existing databases such as those exemplified in Table 3.3. The necessary size of the data set depends on the data quality, the background

TABLE 3.3 Examples of Databases Used for Data Mining for the Purposes of Pharmacovigilance

Type of Database	Example	Reference
Spontaneous reporting database	WHO Uppsala Monitoring Centre	[79]
	FDA's spontaneous reporting database, Silver Spring, MD, USA	[82]
Prescription event monitoring database	Drug Safety Research Unit, Southampton, UK	[83]
Large linked administrative database	Medicaid, Baltimore, MD, USA	[84]
Clinical trial databases	Cardiovascular clinical trial database	[85]
	U.S. Vaccine Adverse Event Reporting System	[86]

frequency of the event, and the strength of the association of the event with the drug. However, for even moderately rare events, large databases are required. The characteristics of the different large databases are discussed elsewhere [81].

One of the commonly used data mining algorithms involves disproportionality analysis that projects high-dimensional data onto two-dimensional (2×2) contingency tables in the context of an independence model. For example, this algorithm was used to compare reporting frequencies of hepatic adverse events between PEGylated and non-PEGylated formulations of active medicinal compounds in SRSs [87]. As a further illustration, data mining of an adverse event database was used to assist in the identification of hypothermia associated with valproic acid therapy and adjunctive topiramate [88]. Two statistical data mining algorithms, proportional reporting ratios (PRRs) and multi-item gamma Poisson shrinker (MGPS), were applied to an SRS database to identify signals of disproportionate reporting (SDRs) [89]. The analysis reveals the potential utility of data mining to direct attention to more subtle indirect drug adverse effects in SRS databases that as yet are often identified from epidemiological investigations.

Chapter 12 of this book discusses the evaluation, potential utility, and limitations of the commonly used data mining algorithms in pharmacovigilance by providing a perspective on their use as one component of a comprehensive suite of signal-detection strategies incorporating clinical and statistical approaches to signal detection. For illustration, data mining exercises involving spontaneous reports submitted to the U.S. Food and Drug Administration (FDA) are used. Several comprehensive reviews were also published (for example, see reference 81).

Despite reported limitations and residual uncertainties associated with the application of computer-based data mining methods in pharmacovigilance, it

can be argued that such methods have expanded the range of credible options available to major healthcare organizations dealing with huge amounts of complex and diverse clinical data to the benefit of patients.

3.6 THE FUTURE

The development and effective use of data mining technologies is considered now a significant competitive advantage in the pharmaceutical industry. Key applications of data mining range from a wide number of advanced chemoinformatics- and bioinformatics-based approaches to employment of toxicogenomic and chemogenomic databases, analysis of clinical data, development of personalized therapies, and so on. Several conceptually different data mining algorithms and software tools have been developed to handle these complex tasks, and we will see further development of these approaches and tools in the future.

In particular, data mining methods will be actively used to rationalize computer-aided drug design to detect specific molecular features that determine pharmacological activity profile, ADME/Tox and physicochemical properties, pharmacokinetic behavior, and so on. This type of analysis used for correlation of molecular properties with specific activities will seriously influence modern strategies of drug design as relatively inexpensive yet comprehensive tools, and therefore will have major importance for the industry. However, it was argued that there is universal agreement that more good experimental ADME/Tox data are needed for use in *in silico* model development, for models are only as good as the data on which they are based.

A plethora of bioinformatics-based applications will play an increasingly significant role in the identification of specific patterns or motifs from genomes and proteomes and thus will provide new insights into human disease and possible therapeutic interventions. This research area will be more and more influenced by post-genome drug discovery strategies integrating chemoinformatics and bioinformatics. An obvious trend in the field is extensive utilization of specific computational approaches to the management, condensation, and utilization of the knowledge obtained from high-throughput screening experiments, and their combination with genome and proteome data.

Another major development for the future is the application of data mining to clinical information databases. The methodology can help reveal patients at higher risk for specific diseases and therefore promises significant preventative potential. In addition, data mining methods for the purposes of pharmacovigilance can help detect ADRs as soon as possible with minimum patient exposure. Wide application of immunoinformatics methods will greatly benefit future immunology and vaccinology, leading to discovery of more effective diagnostics and vaccines.

Data mining methods will also be increasingly applied to the extraction of information not only from chemical, biological, and clinical data, but also from

scientific literature. With the increase in electronic publications, there is an opportunity and a need to develop automated ways of searching and summarizing the literature.

ACKNOWLEDGMENT

The authors gratefully acknowledge Dr. Sean Ekins for *fruitful* collaborations over the past several years and for friendly support. Owing to limited space, it was not possible to cite all *data mining*-related papers; authors' sincere apologies to those omitted.

REFERENCES

1. Hui-Huang Hsu, ed. *Advanced Data Mining Technologies in Bioinformatics*. Hershey, PA, and London, UK: Idea Group Publishing, 2006.
2. Larose DT, ed. *Data Mining Methods and Models*. Weinheim: Wiley Interscience, 2006.
3. Steeg EW, ed. *Principled Data Mining for Drug Discovery: Applications in Biomedicine, Pharmaceutical Science, and Toxicology*. Weinheim: Wiley, 2007.
4. Bajorath J. Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* 2002;1:882–894.
5. Bleicher KH, Bohm HJ, Muller K, Alanine AI. Hit and lead generation: Beyond high-throughput screening. *Nat Rev Drug Discov* 2003;2:369–378.
6. Stahura FL, Bajorath J. Virtual screening methods that complement HTS. *Comb Chem High Throughput Screen* 2004;7:259–269.
7. Yan SF, King FJ, He Y, Caldwell JS, Zhou Y. Learning from the data: Mining of large high-throughput screening databases. *J Chem Inf Model* 2006;46:2381–2395.
8. Parker CN, Schreyer SK. Application of chemoinformatics to high-throughput screening: Practical considerations. *Methods Mol Biol* 2004;275:85–110.
9. Böcker A, Schneider G, Teckentrup A. Status of HTS mining approaches. *QSAR & Comb Sci* 2004;23:207–213.
10. Balakin KV, Savchuk NP. Computational methods for analysis of high-throughput screening data. *Curr Comput Aided Drug Des* 2006;2:1–19.
11. Harper G, Pickett SD. Methods for mining HTS data. *Drug Discov Today* 2006;11(15/16):694–699.
12. Paul EB Jr., Kevin PC, Michael AF, Glenn JM, Joseph SV, Chihae Y. Systematic analysis of large screening sets in drug discovery. *Curr Drug Discov Technol* 2004:1:37–47.
13. Blower P, Fligner M, Verducci J, Bjoraker J. On combining recursive partitioning and simulated annealing to detect groups of biologically active compounds. *J Chem Inf. Comput Sci* 2002;42:393–404.
14. Jones-Hertzog DK, Mukhopadhyay P, Keefer CE, Young SS. Use of recursive partitioning in the sequential screening of G-protein-coupled receptors. *J Pharmacol Toxicol Methods* 1999;42:207–215.

15. Myers PL, Greene JW, Saunders J, Teig SL. Rapid reliable drug discovery. *Today's Chem Work* 1997;6:45–53.
16. Tropsha A, Zheng W. Rational principles of compound selection for combinatorial library design. *Comb Chem High Throughput Screen* 2002;5:111–123.
17. Blower PE, Cross KP, Eichler GS, Myatt GJ, Weinstein JN, Yang C. Comparison of methods for sequential screening of large compound sets. *Comb Chem High Throughput Screen* 2006;9:115–122.
18. Schnur D, Beno BR, Good A, Tebben A. *Approaches to Target Class Combinatorial Library Design. Chemoinformatics—Concepts, Methods, and Tools for Drug Discovery*, pp. 355–378. Totowa, NJ: Humana Press, 2004.
19. Bajorath J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J Chem Inf Comput Sci* 2001;41: 233–245.
20. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: Applications to targets and beyond. *Br J Pharmacol* 2007;152:21–37.
21. Devillers J. *Neural Networks in QSAR and Drug Design*. London: Academic Press, 1996.
22. Vapnik V. *Statistical Learning Theory*. New York: Wiley, 1998.
23. Zernov VV, Balakin KV, Ivashchenko AA, Savchuk NP, Pletnev IV. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness and enzyme inhibition predictions. *J Chem Inf Comput Sci* 2003;43: 2048–2056.
24. Tkachenko SE, Okun I, Balakin KV, Petersen CE, Ivanenkov YA, Savchuk NP, Ivashchenko AA. Efficient optimization strategy for marginal hits active against abl tyrosine kinases. *Curr Drug Discov Technol* 2004;1:201–210.
25. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. *J Pharm Toxicol Methods* 2000;44:235.
26. Walters WP, Murcko MA. Prediction of "drug-likeness". *Adv Drug Deliv Rev* 2002;54:255.
27. Oprea TI. Current trends in lead discovery: Are we looking for the appropriate properties? *J Comput Aided Mol Des* 2002;16:325.
28. Wenlock MC, Austin RP, Barton P, Davis AM, Leeson PD. A comparison of physiochemical property profiles of development and marketed oral drugs. *J Med Chem* 2003;46:1250.
29. Yoshida F, Topliss JG. QSAR model for drug human oral bioavailability. *J Med Chem* 2000;43:2575.
30. Clark DE. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J Pharm Sci* 1999;88:807.
31. Palm K, Stenberg P, Luthman K, Artursson P. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm Res* 1997;14:568.
32. Wessel MD, Jurs PC, Tolan JW, Muskal SM. Prediction of human intestinal absorption of drug compounds from molecular structure. *J Chem Inf Comput Sci* 1998;38:726.

33. Kelder J, Grootenhuis PDJ, Bayada DM, Delbressine LPC, Ploeman J-P. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm Res* 1999;16:1514.
34. Luco JM. Prediction of the brain-blood distribution of a large set of drugs from structurally derived descriptors using partial least-squares (PLS) modeling. *J Chem Inf Comput Sci* 1999;39:396.
35. Lombardo F, Blake JF, Curatolo WJ. Computation of brain-blood partitioning of organic solutes via free energy calculations. *J Med Chem* 1996;39:4750.
36. Crivori P, Cruciani G, Carrupt PA, Testa B. Predicting blood-brain barrier permeation from three-dimensional molecular structure. *J Med Chem* 2000;43: 2204.
37. Dressman JB, Thelen K, Jantratid E. Towards quantitative prediction of oral drug absorption. *Clin Pharmacokinet* 2008;47:655–667.
38. Dearden JC. In silico prediction of ADMET properties: How far have we come? *Expert Opin Drug Metab Toxicol* 2007;3:635–639.
39. Schneider G, Coassolo P, Lave T. Combining in vitro and in vivo pharmacokinetic data for prediction of hepatic drug clearance in humans by artificial neural networks and multivariate statistical techniques. *J Med Chem* 1999;42:5072–5076.
40. Lombardo F, Obach RS, Shalaeva MY, Gao F. Prediction of human volume of distribution values for neutral and basic drugs. 2. Extended data set and leave-class-out statistics. *J Med Chem* 2004;47:1242–1250.
41. Lombardo F, Obach RS, Shalaeva MY, Gao F. Prediction of volume of distribution values in humans for neutral and basic drugs using physicochemical measurements and plasma protein binding. *J Med Chem* 2002;45:2867–2876.
42. Balakin KV, Ivanenkov YA, Savchuk NP, Ivaschenko AA, Ekins S. Comprehensive computational assessment of ADME properties using mapping techniques. *Curr Drug Discov Technol* 2005;2:99–113.
43. Lombardo F, Obach RS, Dicapua FM, Bakken GA, Lu J, Potter DM, Gao F, Miller MD, Zhang Y. A hybrid mixture discriminant analysis-random forest computational model for the prediction of volume of distribution of drugs in human. *J Med Chem* 2006;49:2262–2267.
44. Chohan KK, Paine SW, Waters NJ. Quantitative structure activity relationships in drug metabolism. *Curr Top Med Chem* 2006;6, 1569–1578.
45. Ekins S, Shimada J, Chang C. Application of data mining approaches to drug delivery. *Adv Drug Deliv Rev* 2006;58:1409–1430.
46. Langowski J, Long A. Computer systems for the prediction of xenobiotic metabolism. *Adv Drug Deliv Rev* 2002;54:407–415.
47. Ekins S, Balakin KV, Savchuk N, Ivanenkov Y. Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and Kohonen and Sammon mapping techniques. *J Med Chem* 2006;49:5059–5071.
48. Wilson AG, White AC, Mueller RA. Role of predictive metabolism and toxicity modeling in drug discovery—A summary of some recent advancements. *Curr Opin Drug Discov Devel* 2003;6:123–128.
49. Ekins S, Boulanger B, Swaan PW, Hupcey MAZ. Towards a new age of virtual ADME/TOX and multidimensional drug discovery. *J Comput Aided Mol Des* 2002;16:381.

50. Ekins S, Rose JP. In silico ADME/Tox: The state of the art. *J Mol Graph* 2002;20:305.
51. Van de Waterbeemd H, Gifford E. ADMET in silico modelling: Towards prediction paradise? *Nat Rev Drug Discov* 2003;2:192.
52. Ekins S, Swaan PW. Development of computational models for enzymes, transporters, channels and receptors relevant to ADME/TOX. *Rev Comput Chem* 2004;20:333–415.
53. Hou T, Wang J. Structure-ADME relationship: still a long way to go? *Expert Opin Drug Metab Toxicol* 2008;4:759–770.
54. Ekins S, ed. *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*. Hoboken, NJ: Wiley Interscience, 2007.
55. Muster W, Breidenbach A, Fischer H, Kirchner S, Müller L, Pähler A. Computational toxicology in drug development. *Drug Discov Today* 2008;13:303–310.
56. Eros D, Kövesdi I, Orfi L, Takács-Novák K, Acsády G, Kéri,G. Reliability of logP predictions based on calculated molecular descriptors: A critical review. *Curr Med Chem* 2002;9:1819–1829.
57. Balakin KV, Savchuk NP, Tetko IV. In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: Trends, problems and solutions. *Curr Med Chem* 2006;13:223–241.
58. Faller B, Ertl P. Computational approaches to determine drug solubility. *Adv Drug Deliv Rev* 2007;59:533–545.
59. Balakin KV, Ivanenkov YA, Skorenko AV, Nikolsky YV, Savchuk NP, Ivashchenko AA. In silico estimation of DMSO solubility of organic compounds for bioscreening. *J Biomol Screen* 2004;9:22–31.
60. Jarvis RA, Patrick EA. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans Comput* 1973;C-22:1025–1034.
61. Agrafiotis DK, Lobanov VS. Nonlinear mapping networks. *J Chem Inf Comput Sci* 1997;40:1356.
62. Gillet VJ. Designing combinatorial libraries optimized on multiple objectives. *Methods Mol Biol* 2004;275:335–354.
63. Agrafiotis DK. Multiobjective optimization of combinatorial libraries. *Mol Divers* 2002;5:209–230.
64. Zerikly M, Challis GL. Strategies for the discovery of new natural products by genome mining. *Chembiochem* 2009;10(4):625–633.
65. Yang Y, Adelstein SJ, Kassis AI. Target discovery from data mining approaches. *Drug Discov Today* 2009;14:147–154.
66. Haoudi A, Bensmail H. Bioinformatics and data mining in proteomics. *Expert Rev Proteomics* 2006;3:333–343.
67. Bensmail H, Haoudi A. Data mining in genomics and proteomics. *J Biomed Biotechnol* 2005;2005(2):63–64.
68. Jacoby E, Schuffenhauer A, Floersheim P. Chemogenomics knowledge-based strategies in drug discovery. *Drug News Perspect* 2003;16:93–102.
69. Mestres J. Computational chemogenomics approaches to systematic knowledge-based drug discovery. *Curr Opin Drug Discov Devel* 2004;7:304–313.
70. Schuffenhauer A, Jacoby E. Annotating and mining the ligand-target chemogenomics knowledge space. *Drug Discov Today: BIOSILICO* 2004;2:190–200.

71. Root DE, Flaherty SP, Kelley BP, Stockwell BR. Biological mechanism profiling using an annotated compound library. *Chem Biol* 2003;10:881–892.
72. Horvath D, Jeandenans C. Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces—a novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J Chem Inf Comput Sci* 2003;43:680–690.
73. Rolland C, Gozalbes R, Nicolaï E, Paugam MF, Coussy L, Barbosa F, Horvath D, Revah F. G-protein-coupled receptor affinity prediction based on the use of a profiling dataset: QSAR design, synthesis, and experimental validation. *J Med Chem* 2005;48:6563–6574.
74. Ganter B, Tugendreich S, Pearson CI, Ayanoglu E, Baumhueter S, Bostian KA, Brady L, Browne LJ, Calvin JT, Day GJ, Breckenridge N, Dunlea S, Eynon BP, Furness LM, Ferng J, Fielden MR, Fujimoto SY, Gong L, Hu C, Idury R, Judo MS, Kolaja KL, Lee MD, McSorley C, Minor JM, Nair RV, Natsoulis G, Nguyen P, Nicholson SM, Pham H, Roter AH, Sun D, Tan S, Thode S, Tolley AM, Vladimirova A, Yang J, Zhou Z, Jarnagin K. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J Biotechnol* 2005;119:219–244.
75. Okuno Y, Yang J, Taneishi K, Yabuuchi H, Tsujimoto G. GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res.*, 2006;34:D673–D677.
76. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* 2000;24:236–244.
77. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN, Mesirov JP, Lander ES, Golub TR. Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci USA* 2001;98:10787–10792.
78. Blower PE, Yang C, Fligner MA, Verducci JS, Yu L, Richman S, Weinstein JN. Pharmacogenomic analysis: Correlating molecular substructure classes with microarray gene expression data. *Pharmacogenomics J* 2002;2:259–271.
79. Lindquist M. Use of triage strategies in the WHO signal-detection process. *Drug Saf* 2007:30:635–637.
80. Ye X, Fu Z, Wang H, Du W, Wang R, Sun Y, Gao Q, He J. A computerized system for signal detection in spontaneous reporting system of Shanghai China. *Pharmacoepidemiol Drug Saf* 2009;18:154–158.
81. Wilson AM, Thabane L, Holbrook A. Application of data mining techniques in pharmacovigilance. *Br J Clin Pharmacol* 2004;57:127–134.
82. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Saf* 2002;25:381–392.
83. Mann RD. Prescription-event monitoring–recent progress and future horizons. *Br J Clin Pharmacol* 1998;46:195–201.
84. Forgionne GA, Gangopadhyay A, Adya M. Cancer surveillance using data warehousing, data mining, and decision support systems. *Top Health Inf Manage* 2000;21:21–34.

85. Cerrito P. Application of data mining for examining polypharmacy and adverse effects in cardiology patients. *Cardiovasc Toxicol* 2001;1:177–179.
86. Niu MT, Erwin DE, Braun MM. Data mining in the US Vaccine Adverse Event Reporting System (VAERS): Early detection of intussusception and other events after rotavirus vaccination. *Vaccine* 2001;19:4627–4634.
87. Hauben M, Vegni F, Reich L, Younus M. Postmarketing hepatic adverse event experience with PEGylated/non-PEGylated drugs: A disproportionality analysis. *Eur J Gastroenterol Hepatol* 2007;19:934–941.
88. Knudsen JF, Sokol GH, Flowers CM. Adjunctive topiramate enhances the risk of hypothermia associated with valproic acid therapy. *J Clin Pharm Ther* 2008;33: 513–519.
89. Hauben M, Horn S, Reich L, Younus M. Association between gastric acid suppressants and clostridium difficile colitis and community-acquired pneumonia: Analysis using pharmacovigilance tools. *Int J Infect Dis* 2007;11:417–422.

PART II

CHEMOINFORMATICS-BASED APPLICATIONS

4

DATA MINING APPROACHES FOR COMPOUND SELECTION AND ITERATIVE SCREENING

MARTIN VOGT AND JÜRGEN BAJORATH

Table of Contents

4.1 INTRODUCTION

Advances in genomics, large-scale combinatorial synthesis, and high-throughput biological screening have provided pharmaceutical research with exceedingly large amounts of compounds and biological data. The large body of

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

available data presents the field of data mining with unprecedented opportunities to design and apply computational models, to infer structure–activity relationships, and to prioritize candidate compounds for biological evaluation or library design. Thus, computational analysis and modeling aids in the design of experiments and complements the high-throughput technology-driven approach to drug discovery in a rational manner.

Data mining approaches are an integral part of chemoinformatics and pharmaceutical research. Besides its practical relevance, this field is intellectually stimulating because of the many conceptually diverse methods that have been developed or adapted for chemical and biological data mining. For data mining approaches, a major target area within the chemoinformatics spectrum is virtual compound screening, i.e., the application of computational methods to search large databases for novel molecules having a desired biological activity. The two principal approaches are protein structure-based virtual screening, or docking, and small molecule-based similarity searching. Docking algorithms rely on the knowledge of the three-dimensional (3-D) structure of proteins and their binding sites. A detailed discussion of the multitude of available algorithms and docking techniques is provided, for example, in reviews by Halperin et al. [1] or Klebe [2]. Ligand-based similarity methods are as popular for virtual screening as docking. 3-D approaches such as docking or 3-D ligand similarity searching using pharmacophore representations [3] or shape information [4] have, in principle, higher information content than similarity methods that are based on two-dimensional (2-D) molecular representations. However, a number of studies have shown that docking and other 3-D search techniques are not principally superior to 2-D ligand-based methods [5,6]. For example, in a recent study, McGaughey et al. [6] found that 2-D ligand-based searching performed better than 3-D ligand-based similarity searching or docking on different test cases. Generally, enrichment factors were higher for ligand-based methods than docking, although correctly identified actives were structurally less diverse compared to docking methods. Possible explanations for the often favorable performance of 2-D methods include that the "connection table of a molecule encodes so much implicit information about the 3D structure that using actual 3D coordinates adds little more information," as pointed out by Sheridan and Kearsley [5], and that 2-D methods are not prone to errors associated with modeling of active conformations.

From a data mining point of view, compound classification and filtering techniques are related to similarity analysis. Compound classification methods are often, but not always, used to separate compounds into groups of similar ones or for class label prediction, i.e., a label is assigned to each test compound, which effectively separates a compound collection into two or more distinct classes or sets. Frequently, binary classification models are derived on the basis of learning sets to distinguish, for example, between active and inactive compounds. Furthermore, compound filtering techniques attempt to remove molecules with undesirable properties (e.g., little solubility, toxicity,

synthetic inaccessibility). Such filters are often rule-based and designed to eliminate compounds having known toxic or reactive groups. Filter functions have become especially popular since the introduction of Lipinski's *rule of five* [7], deduced from a statistical survey of known drugs in order to identify compounds having a low probability to be orally available.

Among the long-standing and most widely used data mining methods in chemoinformatics are compound clustering [8–11] and partitioning [12] algorithms that organize database compounds into groups of similar ones with respect to chosen molecular descriptors and chemical reference spaces. Among many different applications, partitioning and clustering are often applied to preselect compounds from screening libraries for biological testing on the basis of already known active molecules. Subsequent iterations of subset selection and biological evaluation often help to substantially reduce the number of compounds that need to be screened in order to identify a sufficient number of novel hits, a process referred to as sequential screening [13]. Moreover, these classification methods are also applied to select compounds for the assembly of target-focused compound libraries [14]. Thus, taken together, there is a broad spectrum of applications for data mining approaches in computer-aided drug discovery and chemoinformatics that makes it attractive to review selected approaches and to highlight their application potential.

This chapter will discuss data mining approaches that are particularly relevant for chemoinformatics applications. Because data mining techniques and their relative performance cannot be separated from the molecular representations that are employed, the chapter will begin with a brief review of popular descriptors. Then exemplary clustering tools and similarity search techniques will be presented. A major focal point will be a discussion of the theoretical foundations of three major data mining approaches that currently experience much attention in chemoinformatics and virtual compound screening: Bayesian modeling, binary kernel discrimination, and support vector machines. We will introduce an approach to predict compound recall rates for Bayesian screening from property distributions of reference and database compounds and, finally, will highlight iterative screening and the assembly of target-focused libraries as attractive application areas for data mining.

4.2 MOLECULAR REPRESENTATIONS AND DESCRIPTORS

The performance of data mining approaches does not only depend on the method itself but also on the chosen molecular representations. Often combinations of numerical chemical descriptors are used to represent a molecule as a vector of descriptor values in descriptor space. Typically, descriptor combinations capture only a part of the chemical information content of a molecule and, although seemingly a triviality, data mining algorithms can only exploit

this information. If it is too limited, data mining will fail. Thus, the choice of molecular representations is indeed a major determinant for the outcome of data mining, regardless of the algorithms that are used. Many different types of descriptors [15] and molecular representations have been introduced [16]. A reason for the continued interest in deriving novel molecular representations might be that conflicting tasks often influence chemical similarity analysis: one aims at the identification of molecules that are similar in activity to known reference compounds, but these molecules should then be as structurally diverse as possible. So, it is desirable for representations to focus on relevant attributes for activity rather than on structural resemblance.

Representations can roughly be separated into three types: one-dimensional (1-D) representations include the chemical composition formula, the simplest molecular view, but also more complex representations such as linear notations including the pioneering SMILES language [17,18] and InChI [19]. SMILES and InChI capture the structure of a molecule in a unique way and are well suited for database searching and compound retrieval. Although not specifically designed for similarity searching, SMILES representations have been used for database mining by building feature vectors from substrings [20–22]. Molecular 2-D representations include connection tables, graph representations, and reduced graphs [23]. Molecular graphs are often employed as queries in similarity searching using algorithms from graph theory for the detection of common substructures. Typically, those algorithms are time-consuming, which limits their applicability for screening large databases. 3-D representations include, for example, molecular surfaces or volumes calculated from molecular conformations. If these representations should be related to biological activity, then binding conformations of test compounds must be known. However, for large compound databases, conformations must usually be predicted, which introduces uncertainties in the use of such representations for compound activity-oriented applications. Pharmacophore models are 3-D representations that reduce molecules to spatial arrangements of atoms, groups, or functions that render them active and are among the most popular tools for 3-D database searching.

Combinations of calculated molecular descriptors are also used to represent molecules and/or to position them in chemical space. Descriptors are in general best understood as numerical mathematical models designed to capture different chemical properties [15]. In many cases, descriptors calculate chemical properties that can be experimentally measured such as dipole moments, molecular refractivity, or log $P(o/w)$, the octanol/water partition coefficient, a measure of hydrophobicity. Descriptors are often organized according to the dimensionality-dependent classification scheme, as discussed above for molecular representations. Thus, dependent on the dimensionality of the molecular representation from which they are calculated, we distinguish 1-D, 2-D, and 3-D descriptors. 1-D descriptors are constitutional descriptors requiring little or no knowledge about the structure of a molecule such as

molecular mass or atom type counts. 2-D descriptors are based upon the graph representation of a molecule. Large numbers of descriptors are calculated from the 2-D structure of chemical compounds. For example, topological descriptors describe properties such as connectivity patterns, molecular complexity (e.g., degree of branching), or approximate shape. Other 2-D descriptors are designed to approximate 3-D molecular features like van der Waals volume or surface area using only the connectivity table of a molecule as input. 3-D descriptors and representations both require knowledge about molecular conformations and geometrical properties of the molecules. Many 2-D and 3-D descriptors vary greatly in their complexity. For example, complex molecular descriptors have been designed to combine multiple descriptor contributions related to biological activity [12] or model surface properties such as the distribution of partial charges on the surface of a molecule [24]. In the following, we will describe graph representations and fingerprint descriptors in more detail.

4.2.1 Graph Representations

In canonical molecular graph representations, nodes represent atoms and edges represent bonds. The use of graph-based algorithms has a long tradition in chemical database searching [25]. The identification of substructures in molecular graphs is hindered by subgraph isomorphism identification, which is a hard problem in computer science and for the treatment of which, in general, no efficient algorithms exist [25]. A special case of compound similarity evaluation on the basis of graph-based representations is to consider the maximum common subgraph (MCS) [26,27], i.e., the largest common substructure. MCS comparison retains most of the structural information of a molecule and consequently detects distinctly similar compounds in a database search. Reduced graphs [23,28] or feature trees [29] simplify graph-based molecular comparisons by combining characteristic chemical features like aromatic rings or functional groups into single nodes and abstract from 2-D structure. This simplifies graph-based comparisons and increases computational efficiency as well as the potential of *scaffold hopping* [30], i.e., the identification of compounds having similar activity but diverse structures.

4.2.2 Fingerprints

Fingerprints are special kinds of descriptors that characterize a molecule and its properties as a binary bit vector. Since many fingerprints have unique designs and are used for similarity searching in combination with selected similarity metrics, they are often also regarded as search methods. In structural fingerprints, each bit represents a specific substructural feature, like an aromatic ring or a functional group of a molecule, and the bit setting accounts either for its presence (i.e., the bit is set on) or absence (off). Fixed-size bit

string representations, where each bit encodes the presence or absence of a predefined structural feature, simplify substructure searching and circumvent the computational complexity associated with the use of graph isomorphism algorithms. Once fingerprints for all compounds in a database have been computed, quantitative fingerprint overlap between query and database compounds is calculated as a measure of molecular similarity. The set of 166 MDL structural keys (MACCS) [31,32] represents a widely used prototype of a fragment-based fingerprint. An early search strategy has been to use fragment-based fingerprints in a fast prescreening step to eliminate large numbers of database compounds lacking encoded fragments present in a query, followed by a subgraph isomorphism search on the remaining molecules [25]. In recent years, increasingly sophisticated fingerprint designs have been introduced that enable database searching beyond prescreening or fragment matching including, for example, pharmacophore fingerprints [33]. These types of fingerprints systematically account for 2-D or 3-D patterns of two to four features such as hydrogen bond donor or acceptor functions, hydrophobic or aromatic moieties, or charged groups, and pairwise distance ranges separating them. For 3-D pharmacophore fingerprinting, test molecules are subjected to systematic conformational search and matches of fingerprint-encoded pharmacophore patterns are monitored. In 2-D pharmacophore fingerprints, bond distances replace spatial distances between feature points, and atom types are often used instead of pharmacophore functions, which is reminiscent of atom pair-type descriptors [34]. Due to the combinatorial nature of pharmacophore patterns, especially 3-D pharmacophore fingerprints can be exceedingly large and often consist of millions of bit positions. Other types of 2-D fingerprints systematically account for connectivity pathways through molecules up to a predefined length. This fingerprint design was pioneered by Daylight [35]. The Daylight fingerprints employ hashing and folding techniques to map the large number of possible pathways to a small number of bits. Furthermore, atom environment fingerprints developed by Glen and coworkers [36] encode the local environment of each atom in a molecule as strings and assemble characteristic strings. Here collections of strings represent the molecular fingerprint, which departs from the classical fixed-length design. Similarly, extended connectivity fingerprints (ECFPs) [37,38] also capture local atom environments. MOLPRINT codes each individual atom environment (either in 2-D or 3-D) up to a certain bond distance range as a fingerprint bit and has been implemented together with Bayesian modeling using multiple template compounds for similarity searching [39,40].

Encoding of numerical property descriptors in a fingerprint format is also possible. For example, the MP-MFP fingerprint [41] assigns 61 property descriptors to individual bits by partitioning their ranges at the median of a compound database (i.e., through binary transformation). Moreover, through equifrequent binning of database descriptor value distributions, the PDR-FP fingerprint encodes a set of 93 molecular property descriptors using only 500 bit positions [42].

4.3 DATA MINING TECHNIQUES

4.3.1 Clustering and Partitioning

Clustering algorithms have been, and continue to be, widely used for compound classification [8–10] and for both diversity- and activity-oriented compound selection. Partitioning algorithms [12] are applied for the same purposes but do not have such a long history in chemoinformatics as clustering methods. Clustering and partitioning methods are often not clearly distinguished in the literature, although they do have a principal difference that is relevant for compound classification and selection: regardless of their algorithmic details, clustering methods involve at some stage pairwise distance or similarity comparisons, whereas partitioning algorithms do not; rather, they generally create coordinate systems in chemical reference spaces into which test compounds fall based on their calculated feature values. As a consequence, partitioning methods can be applied to much larger data sets than conventional clustering techniques, which has become particularly relevant during the age of combinatorial library generation. Both clustering and partitioning methods represent a form of unsupervised learning and thus do not require training sets of known active compounds [43]. Instead, they organize a chemical library into subsets of compounds that are similar according to a chosen metric, given a chemical reference space. Clustering and partitioning are often applied to cover available chemical space by selecting representative compounds from all clusters or partitions. Accordingly, these methods have also been adapted for sequential screening where representative compound subsets are initially selected from a library and are tested to identify novel hits. Then, during iterative rounds, newly identified hits are added to the classification process to select similar compounds from the library for further evaluation [13,44]. Thus, sequential screening integrates diversity- and activity-oriented compound selection schemes.

As already mentioned above, clustering depends on the calculation of intermolecular distances in chemical reference spaces, whereas partitioning is based on establishing a consistent reference frame that allows the independent assignment of coordinates to each database molecule. With the increasing size of data sets, clustering algorithms can become computationally expensive, if not prohibitive. This is especially the case for hierarchical clustering methods where all intermolecular distances need to be considered. For hierarchical-agglomerative clustering methods, clusters are obtained by iteratively combining smaller clusters to form larger ones, beginning with singletons (i.e., single-compound clusters). By contrast, hierarchical-divisive methods start with a single large cluster consisting of all compounds and iteratively split clusters into smaller ones [45]. Besides distinguishing between top–down and bottom–up approaches, hierarchical clustering methods differ in they way by which intercluster distances are measured. Popular methods consider either the minimum, maximum, or average distance of compounds of two clusters.

For example, Ward's clustering algorithm minimizes intracluster variance and maximizes intercluster variance and thus attempts to minimize the increase in information loss when joining clusters [46].

Nonhierarchical methods are generally faster but require to preset the total number of clusters, as in *k*-means clustering [47], or define what constitutes a *neighborhood*, as in Jarvis–Patrick clustering [48]. Cell-based partitioning methods [12,49] and variants like median partitioning [50] are an attractive alternative because of their computational efficiency. These methods assign molecules to cells defined by a combination of descriptor ranges. A prominent and widely applied supervised learning variant for classification problems is the recursive partitioning approach [51,52]. Recursive partitioning divides compound sets along decision trees and attempts to generate homogeneous subsets at the leaves, thereby separating molecules according to activity.

4.3.2 Similarity Searching

Like compound clustering, similarity searching is among the most widely employed approaches in chemoinformatics. The notion of compound similarity and the search for similar molecules are at the core of ligand-based virtual screening concepts. Since the explicit formulation of the *similarity property principle*, which simply states that similar molecules should have similar biological activities [53], a wide variety of concepts of what constitutes molecular similarity have been developed, and a multitude of computational methods for identifying similar molecules in compound databases have been devised. In its most basic form, similarity searching detects common 2-D substructures in the chemical graphs of molecules [25]. As mentioned above, these graph-based approaches are computationally quite expensive, and the need for more efficient alternatives has boosted the popularity of fingerprints to a large extent. Another reason for the popularity of fingerprints is that they can be used to generate search queries if only single bioactive compounds are available as templates, in contrast to other compound classification approaches including machine learning methods.

As discussed above, fingerprints abstract from the chemical structure and make searching of large databases feasible. Importantly, they decouple similarity assessment from direct structural comparisons through the evaluation of bit string similarity. In general, fingerprint-based similarity evaluation depends on two criteria: the type of fingerprint that is used and the chosen similarity measure. Fingerprints can often be easily modified. For example, for structural fingerprints, Durant et al. [32] systematically investigated subsets of the MDL keys for their ability to detect molecules having similar activity in order to optimize sets of structural keys for similarity searching.

In addition to differences in fingerprint design, there also is a variety of similarity measures available for fingerprint comparison [54] including, for example, the Hamming and Euclidean distance. For binary vectors, the

Hamming distance simply counts the number of bit differences between two fingerprints and the Euclidean distance is the square root of the Hamming distance. Most popular in chemical similarity searching is the Tanimoto or Jaccard coefficient, which accounts for the ratio of the number of bits set on in both fingerprints relative to the number of bits set on in either fingerprint. The Hamming and Euclidean distances equally account for the presence or absence of features, i.e., binary complement fingerprints yield the same distance, whereas the Tanimoto coefficient only takes into account bit positions that are set on. For instance, if we consider two fingerprints where 75% of all bits are set on and the two fingerprints overlap in 50% of these bits, then a Tanimoto coefficient of 0.5 is obtained. However, if we take the complement instead, i.e., count the absence of features instead of their presence, a Tanimoto coefficient of 0 is obtained because there is no overlap in missing features (i.e., bit positions set to zero).

Similarity measures enable the ranking of database compounds based on similarity to single or multiple reference compounds and, in successful applications, achieve an enrichment of novel active molecules among the top-ranked compounds. However, the most similar compounds are typically analogues of active reference molecules, and one can therefore not expect to identify diverse structures having similar activity by simply selecting top-ranked database compounds. For the identification of different active chemotypes, similarity value ranges where *scaffold hopping* occurs must be individually determined for combinations of fingerprints and similarity coefficients, which represents a nontrivial task.

In part due to the availability of large databases consisting of different classes of bioactive compounds like the MDDR (a database compiled from patent literature) [55] or WOMBAT [56], similarity searching using multiple reference molecules has become increasingly popular and is typically found to produce higher recall of active molecules compared with calculations using single templates. These findings are intuitive because the availability of multiple compounds increases the chemical information content of the search calculations. For fingerprint searching using multiple reference molecules, different search strategies have been developed that combine compound information either at the level of similarity evaluation or at the level of fingerprint generation. One principal approach is data fusion, which merges the results from different similarity searches [57–60] either by fusing the search results based on the rank of each compound or by using the compound score. This can be achieved, for example, by considering only the highest rank of a database compound relative to each individual template, by calculating the sum of ranks, or by averaging the similarity values of the nearest neighbors. At the level of fingerprints, information from multiple reference molecules can be taken into account by calculation of consensus fingerprints [61], scaling of most frequently occurring bit positions [62], or by determining and using feature value ranges that are most characteristic of template sets [63,64]. Similarity searching is clearly not limited to the use of fingerprint descriptors.

As stated above, reduced graph representations or pharmacophore models are also employed.

Having discussed clustering and similarity search techniques that have a long history in chemical database analysis, in the following, we will focus on three data mining approaches that are based on statistics and machine learning. Because these data mining approaches have in recent years become increasingly popular in chemoinformatics, we discuss their theoretical foundations in some detail.

4.4 BAYESIAN MODELING

Bayesian modeling and Bayesian naïve classifiers are currently widely used for different types of applications in virtual screening [36,65–68] and in compound classification [69–71]. The attractiveness of the Bayesian approach might at least in part be due to its inherent simplicity and the ease of interpretation of the results. Bayesian modeling produces an estimate of the likelihood that compounds exhibit a desired property such as a target-specific activity. Bayesian principles may equally well be applied to continuous, discrete valued descriptors and binary fingerprints. The key aspect of Bayesian modeling is the interpretation of descriptors as random variables following different distributions depending on a certain property or target-specific activity. The Bayesian approach has a sound basis in statistics and relies on some assumptions concerning the representation of compounds. Importantly, it assumes that features are randomly distributed according to a probability distribution that is dependent on the type of compound. Thus, active compounds are expected to show different distributions than inactive ones for descriptors that are relevant for activity. When considering multiple dimensions, i.e., multiple descriptors or different bits in a fingerprint, the assumption that the dimensions have conditionally independent distributions plays a crucial role. The quality of a Bayesian model will largely depend on the knowledge of the distribution of the descriptors and on the validity of the independence assumption. The independence assumption is called the *naïvety* assumption, which will hardly ever be fully met. Nevertheless, Bayesian models have proven to be very successful in practice. Knowledge about the descriptor distributions has to be estimated from the training data. In contrast to similarity search methods where single template searches are feasible, estimates of value distributions cannot be obtained from individual molecules. Discrete descriptors like fingerprint bits can be estimated using frequency counting, which is usually combined with some form of Laplacian correction, because the number of training data points tends to be small. For continuous data, assumptions about the distributions also need to be made. In the absence of further knowledge, generically assuming the presence of Gaussian distributions has been shown to yield promising results in many applications [66,72]. An alternative to assuming specific types of distributions is to discretize con-

tinuously valued descriptors using binning schemes [72,73]. One major drawback of discretization is that a fairly large number of active training compounds are required in order to obtain meaningful histograms for probability estimations. As will be explained below, distance measures in chemical descriptor space can be interpreted in a probabilistic way as likelihoods and yield a theoretical foundation for the appropriateness of metrics like the Euclidean distance. The basic approach is to consider the probability of a compound, c, represented by the (multidimensional) descriptor x_c to show a desired property, A. This probability cannot be estimated directly. Instead, from a set of training compounds known to possess property A, one can estimate the probabilities $P(x_c \mid A)$, i.e., the probability of a compound, c, to show descriptor value x_c given that the compound has property A. Both probabilities are related by the Bayes theorem:

$$P(A \mid x_c) = \frac{P(x_c \mid A)P(A)}{P(x_c)}. \tag{4.1}$$

The probability $P(x_c)$ may be estimated from the training data. However, the probability $P(A)$, i.e., the probability of a compound to possess property A, is generally unknown. When ignoring these terms, one is not able to estimate the probability; rather, one obtains a relative measure termed the likelihood $L(A \mid x_c) \propto P(x_c \mid A)$. Consequently the likelihood ratio

$$R(x_c) = \frac{L(A \mid x_c)}{L(B \mid x_c)} = \frac{P(x_c \mid A)}{P(x_c \mid B)} \tag{4.2}$$

will give a relative likelihood measure of compound c having the desired property A when compared to compounds belonging to a set B not having property A. Figure 4.1 shows Gaussian distributions estimated from samples of a hypothetical descriptor for sets of active and inactive compounds. The height of the curves is dependent on the overall probability for a compound to be active (and is artificially increased for visualization purposes). If compound c is represented by continuous descriptors $x = (x_i)_{i=1\ldots n}$ in an n-dimensional chemical space and the assumptions of descriptor independence and Gaussian distributions are made, then from

$$L(A \mid x) \propto p(x \mid A) = \prod_{i=1}^{n} p(x_i \mid A) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right), \tag{4.3}$$

where μ_i and σ_i are the mean and standard deviation of descriptor i, it follows by considering the negative log likelihood that

$$-\log L(A \mid X = x) \propto \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{2\sigma_i^2} + \text{const.} \tag{4.4}$$

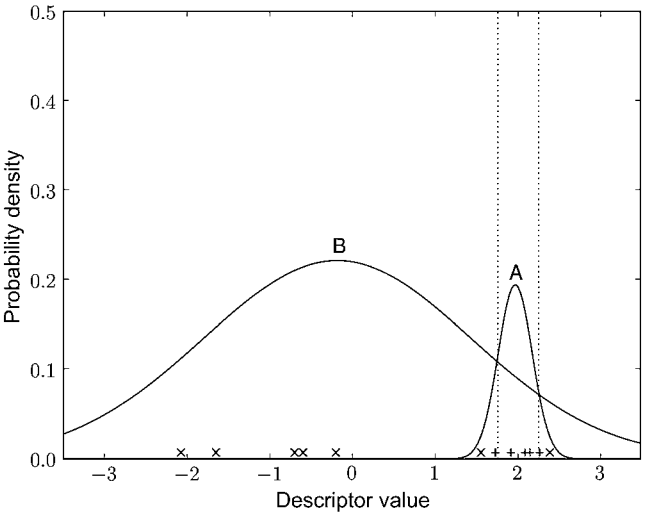


Figure 4.1 Bayesian screening and estimation of probability densities. This schematic representation shows estimates for Gaussian distributions of a hypothetical descriptor based upon a small number of reference samples of inactive (indicated by × curve B) and active (indicated by + curve A) compounds. For illustrative purposes, the curves are scaled assuming that 10% of a compound set is active. The region between the dotted lines indicates the descriptor range for which compounds are more likely to be active than inactive. In practice, the ratio of actives to inactives in a database is unknown, but usually very small, so that still most compounds within the indicated descriptor range will be inactive. However, they are expected to show considerable enrichment in activity compared to a random selection, especially if multiple (uncorrelated) descriptors are combined.

Note that for continuous distributions, the conditional probabilities $P(x \mid A)$ are replaced by the probability density functions $p(x \mid A)$. Thus, normalized Euclidean distance [74] in chemical space is related to the assumed Gaussian distributions of the descriptor values. The ability to relate similarity metrics to descriptor value distributions, given the basic assumption of independence, makes it possible to assess the quality of these measures. It should be noted that the likelihood $L(A \mid x) \propto P(x \mid A)$ is only a relative measure of probability. For instance, if x represents a structural feature that is present in 70% of a class of active compounds A, it might be an indicator of activity. But if this feature is also present in 90% of the compound database, the probability of activity is about 3.8 times higher for the 10% of the molecules that do not possess the structural element.

Bayesian classification takes the likelihoods $L(B \mid x)$ of compounds not possessing property A into account by considering the ratio of these:

$$R(x) = \frac{L(A \mid x)}{L(B \mid x)} = \prod_{i=1}^{n} \frac{P(x_i \mid A)}{P(x_i \mid B)} \tag{4.5}$$

Taking the logarithm yields the *log-odds* score:

$$\log R(x) = \sum_{i=1}^{n} (\log P(x_i \mid A) - \log P(x_i \mid B)). \tag{4.6}$$

By using the negative of the logarithm, minimizing the "distance" log $R(x)$ corresponds to maximizing the odds. Following this approach for (a) an n-dimensional continuous descriptor space and (b) an m-dimensional binary fingerprint representation yields the following similarity measures:

(a)
$$\log R(x) = \frac{1}{2} \sum_{i=1}^{n} \left(\left(\frac{x - \mu_i^B}{\sigma_i^B} \right)^2 - \left(\frac{x - \mu_i^A}{\sigma_i^A} \right)^2 \right) + \text{const.} \tag{4.7}$$

Here μ_i^A and σ_i^A are the sample mean and standard deviation for descriptor i for a set of training compounds A with the desired property like bioactivity, and μ_i^B and σ_i^B are the sample mean and standard deviation of descriptor i of training compounds B not possessing that property.

(b)
$$\log R(v) = \sum_{i=1}^{m} v_i \left(\log \frac{p_i^A}{p_i^B} - \log \frac{1 - p_i^A}{1 - p_i^B} \right) + \text{const.} \tag{4.8}$$

For a fingerprint $v = (v_i)_{i=1\ldots m}$, the Bayesian approach yields a weighting factor of $\log \frac{p_i^A}{p_i^B} - \log \frac{1 - p_i^A}{1 - p_i^B}$ for bit position i, where p_i^A is the relative frequency of bit i being set on for A, and p_i^B is the relative frequency of bit i being set on for B. Similar weighting schemes for binary fingerprints have been introduced [75,76] in the context of substructural analysis methods [77].

When searching for active compounds in a large compound library using a relatively small set of active reference structures, the vast majority of library compounds will be inactive and only relatively few compounds will also be active. In this case, the training set for estimating the probability distributions of active compounds consists of the reference structures and, for all practical purposes, the distributions of the inactive compounds can be well approximated by considering the total compound library, including potential actives, as they only marginally influence the estimates.

The Bayesian approach as described above is not limited to a single type of representation but can also successfully be applied to the combination of different representations like continuous descriptors and binary fingerprints. The MetaBDACCS approach [67] combines descriptors and different fingerprints in a single model and shows a significant increase in performance for a number of biological activity classes [67].

4.4.1 Predicting the Performance of Bayesian Screening

Given the statistical nature of the approach, its success relies on the difference in distribution of descriptors for sets of compounds A and B. In short, the

more the distributions of descriptors differ, the larger the discriminatory power of the descriptors. A suitable quantitative measure for the difference of distributions is the Kullback–Leibler divergence [78]:

$$D[p(x \mid A) \| p(x \mid B)] = \int p(x \mid A) \log \frac{p(x \mid A)}{p(x \mid B)} dx. \tag{4.9}$$

The Kullback–Leibler divergence corresponds to the expected score of the log-likelihood ratio log $R(x)$ for compound class A. Given estimates for the conditional distributions $p(x \mid A)$ and $p(x \mid B)$, the Kullback–Leibler divergence can be calculated analytically. For normally distributed descriptors,

$$D[p(x \mid A) \| p(x \mid B)] = \sum_{i=1}^{n} \left(\log \frac{\sigma_i^A}{\sigma_i^B} + \frac{(\sigma_i^B)^2 - (\sigma_i^A)^2 + (\mu_i^A - \mu_i^B)^2}{2(\sigma_i^A)^2} \right), \tag{4.10}$$

and for fingerprints,

$$D[P(x \mid A) \| P(x \mid B)] = \sum_{i=1}^{n} \left(p_i^A \log \frac{p_i^A}{p_i^B} + (1 - p_i^A) \log \frac{1 - p_i^A}{1 - p_i^B} \right). \tag{4.11}$$

In practice, Equation 4.9 can be used to analyze the fitness of chemical descriptor spaces for virtual screening. The Kullback–Leibler divergence can be used to assess the importance of individual descriptors for the detection of activity for specific biological targets [72]. It is thus possible, by considering only the most discriminating descriptors, to select low-dimensional descriptor representations of molecules individually for virtual screening on specific targets [72].

The relation of the Kullback–Leibler divergence to the log-likelihood ratio can be exploited to establish a quantitative relationship between the Kullback–Leibler divergence and the expected performance of virtual screening calculations [79,80]. The performance of a virtual screening trial can be measured as the ratio of the number of active compounds retrieved in the selected set to the total number of actives in the compound database (i.e., the recall rate). In a first step, virtual screening benchmark trials are performed using a number of different activity classes from a database like the MDDR by calculating the Kullback–Leibler divergence from training sets and by determining the recall rates of actives from a compound database using the Bayesian models based on those training sets. In a second step, a linear regression model relating the logarithm of the Kullback–Leibler divergence to the recall rate is derived to predict recall rates. Figure 4.2 shows two such models based on 40 activity classes from the MDDR using continuous descriptors (A) or MACCS fingerprints (B) to predict the recall rate of the top 100 compounds of a database of 1.4 million molecules [79,80]. The recall rates obtained using seven test classes are seen to correspond well to the rates predicted by the regression model.

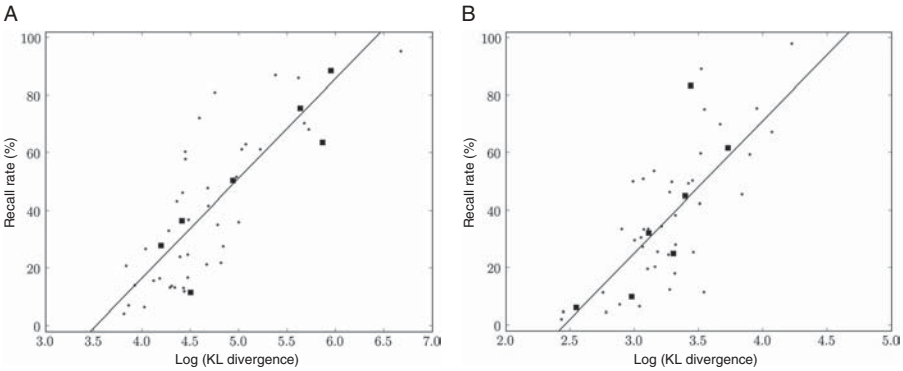


Figure 4.2 Estimation of recall rates based on a set of training classes. The graphs show the relation between Kullback–Leibler (KL) divergence and the recall rate of active compounds from a database. Forty activity classes (small dots) were used in a benchmark approach to establish a linear relationship between the logarithm of KL divergences and the recall rates for the top 100 compounds of a database of about 1.4 million molecules. A linear regression model was trained and seven other classes were used to test the accuracy of the predicted recall. The measured recall rates are represented as squares. (A) shows the result using 142 continuous-valued descriptors and (B) shows the result using the MACCS fingerprint.

4.4.2 Binary Kernel Discrimination

The Bayesian modeling approach described above makes explicit assumptions about feature distributions, specifically assuming independence and the presence of normal distributions for continuous variables. As long as those assumptions are not substantially violated, Bayesian models can be efficiently trained and require only relatively small learning sets because only a limited number of parameters need to be estimated from the data. However, departures from the underlying assumptions might significantly impair the performance of a Bayesian model. If little is known about feature distributions, other nonparametric methods can be used to estimate them. A technique from machine learning applies kernel functions to estimate probability densities, an approach also known as the Parzen window method [47].

The binary kernel discrimination approach [81] introduced Parzen windows applied to binary fingerprints for compound classification and virtual screening. In analogy to Bayesian classification, the likelihood ratio

$$R(x) = \frac{L(A \mid v)}{L(B \mid v)} \propto \frac{p(v \mid A)}{p(v \mid B)} \tag{4.12}$$

is considered and compounds are prioritized accordingly. Suppose training sets A and B containing m_A and m_B compounds, respectively, are given, with A containing compounds having a desired property and B containing others.

Then probability densities $p(v \mid A)$ and $p(v \mid B)$ are estimated using kernel functions:

$$p(v \mid A) = \frac{1}{m_A} \sum_{i=1}^{m_A} K_\lambda (v - v_i^A), \tag{4.13}$$

$$p(v \mid B) = \frac{1}{m_B} \sum_{i=1}^{m_B} K_\lambda (v - v_i^B). \tag{4.14}$$

Here v_i^A and v_i^B are the descriptor values or fingerprints for compounds from A and B, respectively. K_λ is a symmetric multidimensional density function and λ is a smoothing parameter. The estimate $p(v \mid A)$ is a linear combination of probability density functions centered at each point of the training set. The parameter λ controls the *smoothness* of the estimate, i.e., the range of influence around each data point. When continuity is assumed, the estimate converges to the true density function [47]. Figure 4.3A shows the probability density estimate using Gaussian kernel functions with varying standard deviations σ as smoothing parameter for a sample of seven data points. The quality of the estimates will mainly depend on two factors: (1) the number of compounds in the training set and (2) the *nonbias* of the training data. This means that the training data should ideally be a representative subset of test compounds with respect to the descriptor space used for representation. For compound classification, for example, this can only be achieved when learning sets are not merely dominated by analog series, which would skew the data distribution toward a single chemotype.

For fingerprints, i.e., binary vectors v and w of length n, the following kernel function has been suggested:

$$K_\lambda (v, w) = \lambda^{n - \|v - w\|} (1 - \lambda)^{\|v - w\|}, \tag{4.15}$$

where $\|\cdot\|$ is the Hamming distance and λ, the smoothing parameter, ranges from 1 to 0.5. Figure 4.3B illustrates the effect of the smoothing parameter of the binary kernel function on an embedding of data points in a 2-D Euclidean plane. Training must determine an appropriate value for λ. Harper et al. [81] have suggested to increase λ in a stepwise manner from 0.50 to 0.99 and to use leave-one-out cross validation to determine the best parameter setting. Typically, one can expect the parameter λ to increase when more training data become available because each data point needs to cover less range.

Binary kernel discrimination has been shown to be an effective approach to classify and rank test compounds according to their likelihood of activity [82]. An interesting variant of the approach has been introduced by Chen et al. [83], who replaced the Hamming distance with other similarity measures. Their experiments revealed overall preferred performance of the Tanimoto and Dice coefficients over the Hamming distance.

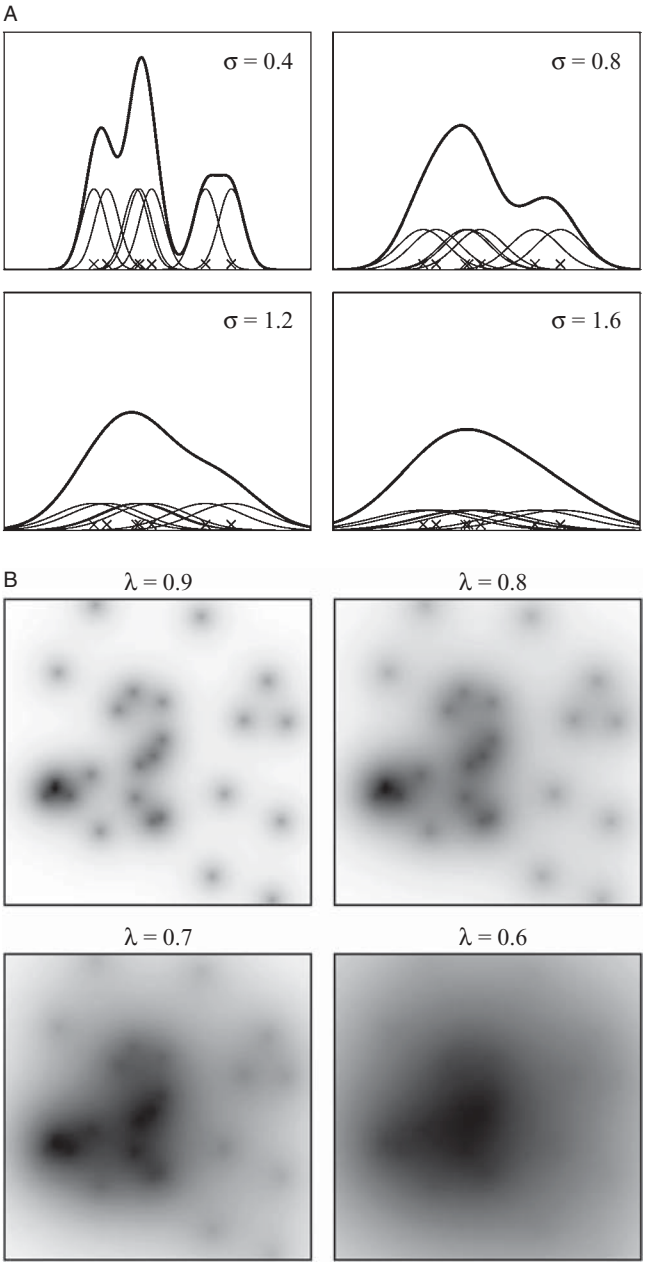


Figure 4.3 Density estimation using Parzen windows. The figures show the nonparametric estimation of probability densities with Parzen windows using different kernel functions. (A) shows the influence of the smoothing parameter σ of a Gaussian kernel function on the estimation of a 1-D probability distribution. (B) schematically illustrates the binary kernel function. Points correspond to compounds embedded in a 2-D plane. The influence of different parameter settings on the smoothness of the resulting probability distribution is shown as gray intensity levels.

4.5 SUPPORT VECTOR MACHINES

In recent years, applications of support vector machines have become very popular in chemoinformatics. Support vector machines are a supervised binary classification approach [84,85]. The basic underlying idea is to linearly separate two classes of data in a suitable high-dimensional space representation such that (1) the classification error is minimized and (2) the margin separating the two classes is maximized. Accordingly, the popularity and success of this method can be attributed to that fact that instead of only trying to minimize the classification error, support vector machines employ structural risk minimization methods to avoid overfitting effects. The structural risk minimization principle implies that the quality of a model does not only depend on minimizing the number of classification errors but also on the inherent complexity of the model. That is, models with increasingly complex structures involve more risk, which means that they do not generalize well, but display significant trends of overfit relative to the training data. Thus, following basic ideas of support vector machines, finding a *maximal* separating hyperplane corresponds to *minimizing* the structural risk.

Overfitting is generally known to be a serious problem in machine learning, which is typically a consequence of using only small training sets but many variables. For classification machines, this would mean using sparse training data, but permitting many degrees of freedom to fit a data-separating boundary. Generally, this situation is referred to as the *curse of dimensionality* and means that with the increase of (feature) dimensionality, the size of training data sets to sample feature space with constant resolution needs to grow exponentially. In principle, a support vector machine implements a linear classifier; however, using the so-called *kernel trick*, i.e., the mapping of data into a high-dimensional space via a kernel function, it also is capable of deriving nonlinear classifiers.

Let us consider a training set of overall size m split into two classes, A and B, of, for instance, active and inactive compounds. Each compound is described by an n-dimensional vector $\mathbf{x}_i$ of numerical features such as descriptor values. Compounds of class A are assigned the value $y_i = +1, i \in A$ and those of class B the value $y_i = -1, i \in B$. If linear separation is possible, the support vector machine is defined by a hyperplane that maximizes the margin, i.e., the closest distance from any point to the separating hyperplane. A hyperplane, H, is defined by a normal vector, $\mathbf{w}$, and a scalar, b, so that

$$H: \langle \mathbf{x}, \mathbf{w} \rangle + b = 0, \tag{4.16}$$

where $\langle \cdot, \cdot \rangle$ defines a scalar product.

For the hyperplane H to separate classes A and B, it is required that all points $\mathbf{x}_i$, $i \in A$ lie on one side of the hyperplane and all points $\mathbf{x}_i$, $i \in B$ on the other. In algebraic terms, this is expressed as

$$\langle \mathbf{x}_i, \mathbf{w} \rangle + b \geq +1 \text{ for } i \in A, \text{ i.e., } y_i = +1 \tag{4.17}$$

$$\langle \mathbf{x}_i, \mathbf{w} \rangle + b \leq -1 \text{ for } i \in B, \text{ i.e., } y_i = -1. \tag{4.18}$$

Combining these inequalities yields

$$y_i \left(\langle \mathbf{x}_i, \mathbf{w} \rangle + b \right) - 1 \geq 0 \text{ for } i = 1 \ldots m \tag{4.19}$$

Points that meet the equality condition and are closest to the separating hyperplane define two hyperplanes,

$$H_{+1} : \langle \mathbf{x}_i, \mathbf{w} \rangle + b = +1 \tag{4.20}$$

and

$$H_{-1} : \langle \mathbf{x}_i, \mathbf{w} \rangle + b = -1, \tag{4.21}$$

parallel to the separating hyperplane H, which determine the margin. Their separating distance is $2/\|\mathbf{w}\|$. So, minimizing $\|\mathbf{w}\|$ with respect to the inequality constraints yields the maximum margin hyperplane, where the inequalities ensure correct classification and the minimization produces the minimal risk, i.e., the best generalization of performance. Those points that lie on the margin are called the *support vectors* because they define the hyperplane H, as can be seen from Figure 4.4. These are the points for which equality holds in Equations 4.17 and 4.18.

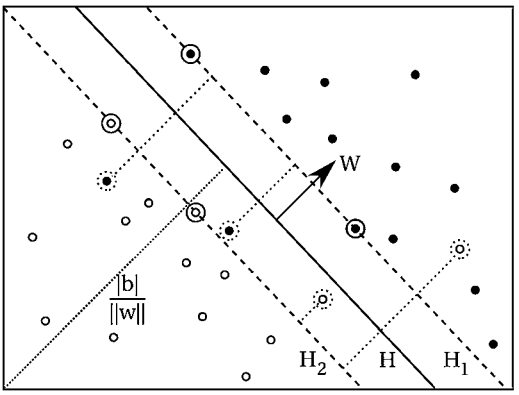


Figure 4.4 Maximal margin hyperplane. The maximal margin hyperplane H is defined by the vector w and the distance $|b|/\|w\|$ from the origin. The support vectors are indicated by solid circles. The classification errors are indicated by the dotted circles. The lines from the margins to the dotted circles indicate the magnitude of the slack variables.

The basic technique for solving optimization problems under constraints is to introduce Lagrange multipliers α_i. The Langrangian

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i y_i \left(\langle \mathbf{x}_i, \mathbf{w}\rangle + b\right) + \sum_{i=1}^{m} \alpha_i \tag{4.22}$$

must be minimized with respect to w and b, and the derivatives of L_P with respect to α_i need to disappear, given the constraints $\alpha_i \geq 0$. Calculating the derivatives with respect to w and b yields the conditions

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \text{ and} \tag{4.23}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0. \tag{4.24}$$

Combining Equations 4.23 and 4.24 with Equation 4.22 yields the dual formulation

$$L_D = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{m,m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \tag{4.25}$$

that must be maximized with respect to α_i under the constraint that $\alpha_i \geq 0$ and $\sum_{i=1}^{m} \alpha_i y_i = 0$. This corresponds to a convex quadratic optimization problem that can be solved using iterative methods to yield a global maximum. If the problem is solved, $\mathbf{w}$ is obtained from Equation 4.23 and b can be obtained from

$$y_i \left(\langle \mathbf{x}_i, \mathbf{w}\rangle + b\right) - 1 = 0 \tag{4.26}$$

for any vector i with $\alpha_i \neq 0$. The vectors i with $\alpha_i \neq 0$ are exactly the support vectors, as the Lagrangian multipliers will be 0 when equality does not hold in Equations 4.17 and 4.18. Once the hyperplane has been determined, compounds can be classified using the decision function

$$f(\mathbf{x}) = \text{sign}\left(\langle \mathbf{x}, \mathbf{w}\rangle + b\right) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i \langle \mathbf{x}, \mathbf{x}_i \rangle + b\right). \tag{4.27}$$

Usually, the condition of linear separability is too restrictive and, therefore, *slack variables* are introduced to the conditions, Equations 4.17 and 4.18, thereby relaxing them to permit limited classification errors:

$$\langle \mathbf{x}_i, \mathbf{w}\rangle + b \geq +1 - \xi_i \text{ for } i \in A \tag{4.28}$$

$$\langle \mathbf{x}_i, \mathbf{w}\rangle + b \leq -1 + \xi_i \text{ for } i \in B \tag{4.29}$$

$$\xi_i \geq 0 \text{ for } i = 1 \ldots m. \tag{4.30}$$

Figure 4.4 illustrates the introduction of slack variables. The dotted lines from the margins represent slack variables with positive values allowing for classification errors of the hyperplane. The objective function to be minimized under those constraints becomes $1/\|\mathbf{w}\| + C\left(\sum_{i=1}^{m} \xi_i\right)^k$, where usually $k = 1$ or $k = 2$ and C is a parameter controlling the penalty of classification errors.

As stated above, support vector machines are not limited to linear boundaries. Nonlinear boundaries can also be achieved by introducing kernel functions. Equation 4.25 only requires the calculation of the scalar product between two vectors and does not require an explicit representation of the vectors. Conceptually, kernel functions correspond to a mapping of the original vectors into a high-dimensional space and calculating the scalar product. Popular kernel functions include, for example, the Gaussian kernel function, polynomial functions, or sigmoid functions:

$$K_\sigma(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2}\right) \tag{4.31}$$

$$K_p(\mathbf{x}_1, \mathbf{x}_2) = (\langle \mathbf{x}_1, \mathbf{x}_2 \rangle + 1)^p \tag{4.32}$$

$$K_{\kappa,\delta}(\mathbf{x}_1, \mathbf{x}_2) = \tanh(\kappa \langle \mathbf{x}_1, \mathbf{x}_2 \rangle - \delta). \tag{4.33}$$

The flexibility of the kernel approach also makes it possible to define kernel functions on a wide variety of molecular representations that need not be numerical in nature. Azencott et al. [22] provide examples of a variety of kernel functions. For 1-D SMILES and 2-D graph representations, a spectral approach is used by building feature vectors recording either the presence or the absence or the number of substrings or substructures. The constructed vectors are essentially fingerprints, and the kernel function is subsequently defined as a similarity measure on the basis of those fingerprints. Using 3-D structures, kernel functions can also be constructed for surface area representations and pharmacophores, or by considering pairwise distances between atoms recorded in histograms. Thus, different types of kernel functions make it possible to tackle diverse classification problems and ensure the general flexibility of the support vector machine approach.

4.6 APPLICATION AREAS

In chemoinformatics and computer-aided drug discovery, support vector machines and binary kernel discrimination have thus far mostly been used for distinguishing between active and inactive compounds in the context of virtual screening [76,81,82]. However, Bayesian models and classifiers have also been used for different applications beyond prediction of active compounds. For

example, Bayesian modeling has been applied to predict compound recall for fingerprint search calculations [79], multidrug resistance [69], or biological targets of test compounds [70]. Nevertheless, for all of these advanced data mining approaches, virtual compound screening is a major application area where the derivation of predictive models from experimental screening data presents a particularly attractive aspect. Models from screening data for activity predictions have also been built using recursive partitioning and hierarchical clustering techniques, but their quality is typically rather sensitive to systematic errors and noise in the data, from which essentially any high-throughput screening (HTS) data set suffers. This is why advanced data mining methods like Bayesian modeling or binary kernel discrimination have become very attractive for these purposes, because these approaches have been shown to be capable of deriving robust models from noisy screening data [73,83].

Typically, models are built from screening data to search other databases for novel active compounds. Thus, HTS data serve as a learning set to derive a support vector machine or a Bayesian or binary kernel discrimination model to classify other database compounds as active or inactive. This makes these data mining approaches also very attractive to aid in iterative experimental and virtual screening campaigns that are often described as sequential screening [86,87]. Iterative cycles of virtual compound preselection from screening libraries and experimental evaluation can substantially reduce the number of compounds that need to be screened in order to identify sufficient numbers of hits for follow-up [86,88]. During these iterations, newly identified hits are usually included in model refinement for subsequent rounds of compound selection. The major aim of these calculations is to continuously enrich small compound sets with active compounds, and this selection scheme can be quite powerful. For example, if only a moderate overall enrichment factor of five is achieved, this means that only 10% of a screening library needs to be tested in order to identify 50% of potentially available hits. Initial approaches to establish sequential screening schemes have predominantly employed recursive partitioning [89,90] or recursion forest analysis [91], but machine learning techniques have recently also been applied [92]. For advanced data mining approaches, sequential screening represents a highly attractive application scenario for several reasons. For example, Bayesian or kernel-based classifiers are much less influenced by screening data noise than standard compound classification methods and, moreover, classifiers can be trained not only to select active compounds but also to deselect efficiently database molecules having a very low probability of activity. Given the fact that the vast majority of database compounds are potential false positives for a given target, efficient compound deselection becomes an important task in screening database analysis and can greatly contribute to achieving favorable enrichment factors during iterative screening campaigns. Thus, we can expect that the interest in machine learning and data mining approaches in virtual and iterative compound screening will further increase.

Another attractive application area for advanced data mining methods is the assembly of target-focused compound libraries. A variety of approaches have been introduced to design target-focused libraries based on ligand or target structure information or a combination of both [14]. In recent years, there has been a clear trend to employ structure design techniques for the generation of focused libraries [93,94], more so than data mining methods. However, conceptually similar to the tasks at hand in iterative screening, major goals of targeted library design include a significant enrichment of molecules having a high probability to display a target-specific activity in compound sets that are much smaller in size than diverse screening libraries. Therefore, data mining also becomes highly attractive for these applications. For example, the ability to predict biological targets for large numbers of database compounds using multiple Bayesian models [70] is expected to substantially aid in prioritizing compounds for the assembly of target-focused libraries. Thus, similar to iterative screening, we can expect that the design of specialized compound libraries will also be a future growth area for data mining applications.

4.7 CONCLUSIONS

In this chapter, we have discussed various data mining approaches and have selected applications in the context of chemoinformatics. Since the performance of data mining methods cannot be separated from the molecular representations that are employed, prominent types of molecular descriptors and representations have also been reviewed. Special emphasis has been put on discussing theoretical foundations of three advanced data mining approaches that are becoming increasingly popular in chemoinformatics and in pharmaceutical research: Bayesian modeling, binary kernel discrimination, and support vector machines. We have particularly highlighted virtual and integrated compound screening schemes and the design of target-focused compound libraries as attractive application areas with future growth potential.

REFERENCES

1. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* 2002;47:409–443.
2. Klebe G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today* 2006;11:580–594.
3. Mason JS, Good AC, Martin EJ. 3-D pharmacophores in drug discovery. *Curr Pharm Des* 2006;7:567–597.
4. Hawkins P, Skillman A, Nicholls A. Comparison of shape-matching and docking as virtual screening tools. *J Med Chem* 2007;50:74–82.
5. Sheridan RP, Kearsley SK. Why do we need so many chemical similarity search methods? *Drug Discov Today* 2002;7:903–911.

6. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoula C, Lindsley S, Maiorov V, Truchon JF, Cornell WD. Comparison of topological, shape, and docking methods in virtual screening. *J Chem Inf Model* 2007;47:1504–1519.
7. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 1997;23:3–25.
8. Willett P, Winterman V, Bawden D. Implementation of nonhierarchic cluster analysis methods in chemical information systems: Selection of compounds for biological testing and clustering of substructure search output. *J Chem Inf Comput Sci* 1986;26:109–118.
9. Barnard JM, Downs GM. Clustering of chemical structures on the basis of two-dimensional similarity measures. *J Chem Inf Comput Sci* 1992;32:644–649.
10. Brown RD, Martin YC. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J Chem Inf Comput Sci* 1996;36:572–584.
11. Brown RD, Martin YC. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J Chem Inf Comput Sci* 1997;37:1–9.
12. Pearlman RS, Smith K. Novel software tools for chemical diversity. *Perspect Drug Discov Des* 1998;9:339–353.
13. Bajorath J. Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* 2002;1:882–894.
14. Schnur D, Beno BR, Good A, Tebben A. *Approaches to Target Class Combinatorial Library Design. Chemoinformatics—Concepts, Methods, and Tools for Drug Discovery*, pp. 355–378. Totowa, NJ: Humana Press, 2004.
15. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH, 2000.
16. Maldonado AG, Doucet JP, Petitjean M, Fan BT. Molecular similarity and diversity in chemoinformatics: From theory to applications. *Mol Divers* 2006;10:39–79.
17. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comp Sci* 1988;28:31–36.
18. Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique smiles notation. *J Chem Inf Comput Sci* 1989;29:97–101.
19. Stein SE, Heller SR, Tchekhovski D. An open standard for chemical structure representation—The IUPAC chemical identifier. In *Nimes International Chemical Information Conference Proceedings*, edited by Collier H, pp. 131–143. Tetbury, UK: Infomatics, 2002. Available at http://www.iupac.org/inchi/ (accessed February 1, 2008).
20. Vidal D, Thormann M, Pons M. LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J Chem Inf Model* 2005;45:386–393.
21. Grant J, Haigh J, Pickup B, Nicholls A, Sayle R. Lingos, finite state machines, and fast similarity searching. *J Chem Inf Model* 2006;46:1912–1918.
22. Azencott CA, Ksikes A, Swamidass SJ, Chen JH, Ralaivola L, Baldi P. One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical, and biological properties. *J Chem Inf Model* 2007;47:965–974.
23. Gillet V, Willett P, Bradshaw J. Similarity searching using reduced graphs. *J Chem Inf Comput Sci* 2003;43, 338–345.

24. Labute P. *Derivation and Applications of Molecular Descriptors Based on Approximate Surface Area. Chemoinformatics—Concepts, Methods, and Tools for Drug Discovery*, pp. 261–278. Totowa, NJ: Humana Press, 2004.
25. Barnard JM. Substructure searching methods: Old and new. *J Chem Inf Comput Sci* 1993;33:532–538.
26. Raymond J, Gardiner E, Willett P. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J Chem Inf Comput Sci* 2002;42:305–316.
27. Willett P. Searching techniques for databases of two- and three-dimensional chemical structures. *J Med Chem* 2005;48:4183–4199.
28. Gardiner E, Gillet V, Willett P, Cosgrove D. Representing clusters using a maximum common edge substructure algorithm applied to reduced graphs and molecular graphs. *J Chem Inf Model* 2007;47:354–366.
29. Hessler G, Zimmermann M, Matter H, Evers A, Naumann T, Lengauer T, Rarey M. Multiple-ligand-based virtual screening: Methods and applications of the MTree approach. *J Med Chem* 2005;48:6575–6584.
30. Barker E, Buttar D, Cosgrove D, Gardiner E, Kitts P, Willett P, Gillet V. Scaffold hopping using clique detection applied to reduced graphs. *J Chem Inf Model* 2006;46:503–511.
31. McGregor M, Pallai P. Clustering of large databases of compounds: Using the MDL "keys" as structural descriptors. *J Chem Inf Comput Sci* 1997;37:443–448.
32. Durant J, Leland B, Henry D, Nourse J. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 2002;42:1273–1280.
33. Mason J, Morize I, Menard P, Cheney D, Hulme C, Labaudiniere R. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J Med Chem* 1999; 42:3251–3264.
34. Carhart RE, Smith DH, Venkataraghavan R. Atom pairs as molecular features in structure-activity studies: Definition and applications. *J Chem Inf Comput Sci* 1985;25:64–73.
35. *Daylight Theory Manual*. Aliso Viejo, CA: Daylight Chemical Information Systems, Inc. 2002. Available at http://www.daylight.com/dayhtml/doc/theory/ (accessed February 1, 2008).
36. Bender A, Mussa HY, Glen RC, Reiling S. Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *J Chem Inf Comput Sci* 2004;44:170–178.
37. Klon A, Glick M, Thoma M, Acklin P, Davies J. Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *J Med Chem* 2004;47:2743–2749.
38. Klon A, Glick M, Davies J. Application of machine learning to improve the results of high-throughput docking against the HIV-1 protease. *J Chem Inf Comput Sci* 2004;44:2216–2224.
39. Bender A, Mussa HY, Glen RC, Reiling S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J Chem Inf Comput Sci* 2004;44:1708–1718.

40. Bender A, Mussa H, Gill G, Glen R. Molecular surface point environments for virtual screening and the elucidation of binding patterns (MOLPRINT 3D). *J Med Chem* 2004;47:6569–6583.
41. Xue L, Godden J, Stahura F, Bajorath J. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J Chem Inf Comput Sci* 2003;43:1151–1157.
42. Eckert H, Bajorath J. Design and evaluation of a novel class-directed 2D fingerprint to search for structurally diverse active compounds. *J Chem Inf Model* 2006;46:2515–2526.
43. Bajorath J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J Chem Inf Comput Sci* 2006;41: 233–245.
44. Engels MFM, Venkatarangan P. Smart screening: Approaches to efficient HTS. *Curr Opin Drug Discov Devel* 2001;4:275–283.
45. Downs GM, Barnard JM. Clustering methods and their uses in computational chemistry. In: *Reviews in Computational Chemistry*, Vol. 18, edited by Lipkowitz KB, Boyd DB, pp. 1–40. Weinheim: Wiley-WCH, 2002.
46. Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 1963;58:236–244.
47. Duda RO, Hart PE, Stork DG. *Pattern Classification*, 2nd edn. New York: Wiley Interscience, 2000.
48. Jarvis R, Patrick E. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans Comput* 1973;C22:1025–1034.
49. Pearlman R, Smith K. Metric validation and the receptor-relevant subspace concept. *J Chem Inf Comput Sci* 1999;39:28–35.
50. Godden J, Xue L, Kitchen D, Stahura F, Schermerhorn E, Bajorath J. Median partitioning: A novel method for the selection of representative subsets from large compound pools. *J Chem Inf Comput Sci* 2002;42:885–893.
51. Chen X, Rusinko A, Young S. Recursive partitioning analysis of a large structure-activity data set using three-dimensional descriptors. *J Chem Inf Comput Sci* 1998;38:1054–1062.
52. Rusinko A, Farmen M, Lambert C, Brown P, Young S. Analysis of a large structure/biological activity data set using recursive partitioning. *J Chem Inf Comput Sci* 1999;39:1017–1026.
53. Johnson M, Maggiora G. *Concepts and Applications of Molecular Similarity*. New York: John Wiley & Sons, 1990.
54. Willett P. Chemical similarity searching. *J Chem Inf Comput Sci* 1998;38: 983–996.
55. *Molecular Drug Data Report (MDDR)*. San Leandro, CA: Elsevier MDL. Available at http://www.mdl.com (accessed February 1, 2008).
56. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M, Oprea TI. WOMBAT: World of molecular bioactivity. In: *Chemoinformatics in Drug Discovery*, edited by Oprea TI, pp. 223–239. New York: Wiley-VCH, 2004.
57. Salim N, Holliday J, Willett P. Combination of fingerprint-based similarity coefficients using data fusion. *J Chem Inf Comput Sci* 2003;43:435–442.

58. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J Chem Inf Comput Sci* 2004;44:1177–1185.
59. Whittle M, Gillet VJ, Willett P. Analysis of data fusion methods in virtual screening: Theoretical model. *J Chem Inf Model* 2006;46:2193–2205.
60. Whittle M, Gillet VJ, Willett P. Analysis of data fusion methods in virtual screening: Similarity and group fusion. *J Chem Inf Model* 2006;46:2206–2219.
61. Xue L, Stahura F, Godden J, Bajorath J. Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J Chem Inf Comput Sci* 2001;41:746–753.
62. Godden J, Furr J, Xue L, Stahura F, Bajorath J. Molecular similarity analysis and virtual screening by mapping of consensus positions in binary-transformed chemical descriptor spaces with variable dimensionality. *J Chem Inf Comput Sci* 2004;44:21–29.
63. Eckert H, Bajorath J. Determination and mapping of activity-specific descriptor value ranges for the identification of active compounds. *J Med Chem* 2006;49:2284–2293.
64. Eckert H, Vogt I, Bajorath J. Mapping algorithms for molecular similarity analysis and ligand-based virtual screening: Design of DynaMAD and comparison with MAD and DMC. *J Chem Inf Model* 2006;46:1623–1634.
65. Xia X, Maliski E, Gallant P, Rogers D. Classification of kinase inhibitors using a Bayesian model. *J Med Chem* 2004;47:4463–4470.
66. Vogt M, Godden J, Bajorath J. Bayesian interpretation of a distance function for navigating high-dimensional descriptor spaces. *J Chem Inf Model* 2007;47:39–46.
67. Vogt M, Bajorath J. Bayesian screening for active compounds in high-dimensional chemical spaces combining property descriptors and molecular fingerprints. *Chem Biol Drug Des* 2008;71:8–14.
68. Watson P. Naïve Bayes classification using 2D pharmacophore feature triplet vectors. *J Chem Inf Model* 2008;48:166–178.
69. Sun H. A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J Med Chem* 2005;48:4031–4039.
70. Nidhi, Glick M, Davies JW, Jenkins JL. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* 2006;46:1124–1133.
71. Klon AE, Lowrie JF, Diller DJ. Improved naïve Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. *J Chem Inf Model* 2006;46:1945–1956.
72. Vogt M, Bajorath J. Bayesian similarity searching in high-dimensional descriptor spaces combined with Kullback-Leibler descriptor divergence analysis. *J Chem Inf Model* 2008;48:247–255.
73. Labute P. Binary QSAR: A new method for the determination of quantitative structure activity relationships. In: *Pacific Symposium on Biocomputing*, Vol. 4, edited by Altman RB, Dunber AK, Hunter L, Klein TE, pp. 444–455. Singapore: World Scientific Publishing, 1999.
74. Godden J, Bajorath, J. A distance function for retrieval of active molecules from complex chemical space representations. *J Chem Inf Model* 2006;46:1094–1097.

75. Ormerod A, Willett P, Bawden D. Comparison of fragment weighting schemes for substructural analysis. *Quant Struct Act Relat* 1989;8:115–129.
76. Wilton DJ, Harrison RF, Willett P, Delaney J, Lawson K, Mullier G. Virtual screening using binary kernel discrimination: Analysis of pesticide data. *J Chem Inf Model* 2006;46:471–477.
77. Cramer R, Redl G, Berkoff C. Substructural analysis. A novel approach to the problem of drug design. *J Med Chem* 1974;17:533–535.
78. Kullback S. *Information Theory and Statistics*. Mineola, MN: Dover Publications, 1997.
79. Vogt M, Bajorath J. Introduction of an information-theoretic method to predict recovery rates of active compounds for Bayesian in silico screening: Theory and screening trials. *J Chem Inf Model* 2007;47:337–341.
80. Vogt M, Bajorath J. Introduction of a generally appicable method to estimate retrieval of active molecules for similarity searching using fingerprints. *ChemMedChem* 2007;2:1311–1320.
81. Harper G, Bradshaw J, Gittins JC, Green DVS, Leach AR. Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J Chem Inf Comput Sci* 2001;41:1295–1300.
82. Wilton D, Willett P, Lawson K, Mullier G. Comparison of ranking methods for virtual screening in lead-discovery programs. *J Chem Inf Comput Sci* 2003;43: 469–474.
83. Chen B, Harrison RF, Pasupa K, Willett P, Wilton DJ, Wood DJ, Lewell XQ. Virtual screening using binary kernel discrimination: Effect of noisy training data and the optimization of performance. *J Chem Inf Model* 2006;46:478–486.
84. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov* 1998;2:121–167.
85. Vapnik VN. *The Nature of Statistical Learning Theory*, 2nd edn. New York: Springer, 2000.
86. Stahura FL, Bajorath J. Virtual screening methods that complement HTS. *Comb Chem High Throughput Screen* 2004;7:259–269.
87. Blower PE, Cross KP, Eichler GS, Myatt GJ, Weinstein JN, Yang C. Comparison of methods for sequential screening of large compound sets. *Comb Chem High Throughput Screen* 2006;9:115–122.
88. Parker CN, Bajorath J. Towards unified compound screening strategies: A critical evaluation of error sources in experimental and virtual high-throughput screening. *QSAR Comb Sci* 2006;25:1153–1161.
89. Jones-Hertzog DK, Mukhopadhyay P, Keefer CE, Young SS. Use of recursive partitioning in the sequential screening of G-protein-coupled receptors. *J Pharmacol Toxicol Methods* 1999;42:207–215.
90. Abt M, Lim Y, Sacks J, Xie M, Young SS. A sequential approach for identifying lead compounds in large chemical databases. *Stat Sci* 2001;16:154–168.
91. van Rhee AM. Use of recursion forests in the sequential screening process: Consensus selection by multiple recursion trees. *J Chem Inf Comput Sci* 2003;43: 941–948.

92. Auer J, Bajorath J. Simulation of sequential screening experiments using emerging chemical patterns. *Med Chem* 2008;4:80–90.
93. Deng Z, Chuaqui C, Singh J. Knowledge-based design of target-focused libraries using protein-ligand interaction constraints. *J Med Chem* 2006;49:490–500.
94. Orry AJ, Abagyan RA, Cavasotto CN. Structure-based development of target-specific compound libraries. *Drug Discov Today* 2006;11:261–266.

5

PREDICTION OF TOXIC EFFECTS OF PHARMACEUTICAL AGENTS

ANDREAS MAUNZ AND CHRISTOPH HELMA

Table of Contents

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

5.1 INTRODUCTION

This chapter describes the prediction of toxic effects with data mining techniques in a stepwise approach. Methods are characterized in terms of principal advantages and shortcomings and interpretability of the results. We seek to present techniques that are effective as well as universally applicable. We also give some software recommendations focusing on open source software, which is not only free but is also transparent and extensible. All packages for the R environment for statistical computing (as well as R itself) are available from CRAN (Comprehensive R Archive Network), the central R repository [1] [(http://www.r-project.org/).

5.1.1 Problem Description

Chemicals influence biological systems in a huge variety of biochemical interactions, mostly on the cellular and molecular level. In toxicology, the aim is to understand the biochemical mechanisms involved and the degree to which chemicals induce toxicological activity in living organisms with respect to a well-defined end point.

In predictive toxicology, we exploit the toxicological knowledge about a set of chemical compounds in order to predict the degree of activity of other compounds. More specifically, we mathematically model the relationship between specific properties of training compounds (i.e., compounds for which the degree of activity is known) and their toxicological activity and apply the model to query compounds (i.e., compounds for which the degree of activity is not known) to obtain predicted activities.

The process of model building is called (quantitative) structure–activity relationship ([Q]SAR). Structure–activity relationships (SARs) are models based on structural features, and quantitative structure–activity relationships (QSARs) rely on quantitative (frequently physicochemical) properties. The most general mathematical form of a (Q)SAR is

$$\text{Activity} = f(\text{physicochemical properties and/or structural features}). \quad (5.1)$$

The training compounds are stored in databases together with their activity values. Formally, we have observed data for n cases $((x_1, y_1), \ldots, (x_n, y_n))$, where each $x_i = (x_{i1}, \ldots, x_{im})$[1] is a feature vector of m input values and each y_i is the associated activity (dependent variable). The observations and corresponding activities can therefore be compactly represented in a data matrix, X (sometimes also referred to as set X), and a corresponding activity vector y. Our primary interest is to predict the unknown activity value y_q for a query compound x_q. A predicted value for x_q is commonly referred to as $f(x_q)$, associated with a confidence value, c, which is derived from certain properties of the model that describe the goodness of the fit. One of these properties is the

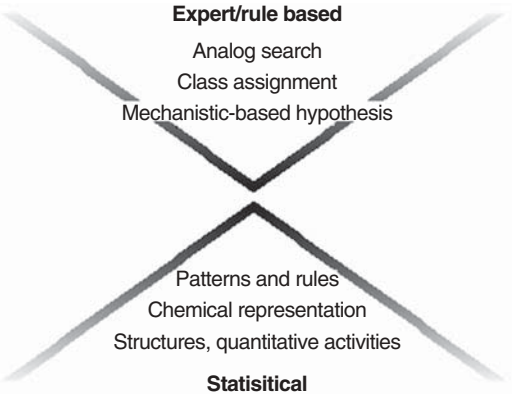


Figure 5.1 Types of (Q)SAR modeling: While expert systems make use of expert knowledge, specifically with feature selection and modeling, statistical approaches derive most things in an automated fashion.

chemical similarity between training compounds and the query compound, denoted as $sim(x_i, x_q)$. For quantitative activities, the prediction process is called regression; for qualitative activities (i.e., a finite set of activity classes), it is called classification.

5.1.2 Predictive Toxicology Approaches

According to Richard [2] and Richard and Williams [3], predictive toxicology models can be classified as statistical and expert/rule-based approaches (see Fig. 5.1). Statistical approaches use general toxic end points and activity values gathered for a wide range of structures and are primarily driven by information inherently present in the data, not from human expert knowledge. Expert/rule-based approaches build (Q)SAR generalizations from individual chemicals to chemical classes based on prior knowledge, heuristics, expert judgment and chemical and biological mechanism considerations.

For the purpose of this chapter, we will focus on statistical (Q)SAR techniques and on the expert system aspects (e.g., categorization, feature selection) that are frequently used in (Q)SAR modeling.

5.1.2.1 Traditional (Q)SAR Models Traditional (Q)SAR methods use linear regression techniques to identify a relationship between chemical features and experimental activities. They rely on the idea that structural properties contribute in a linearly additive way to activity. Usually, the critical molar concentrations C are modeled. The classical approaches are

1. *Hansch analysis.* Physicochemical properties are used as descriptor values (QSAR):

$$\log\left(\frac{1}{c}\right) = a\log P + b\log P^2 + cE + dS + e,$$

where log P is the octanol–water partition coefficient, describing the ability of the agent to reach the target site, and E and S are electronic and steric terms, respectively. Electronic properties relate to binding ability and steric properties describe the bulk and shape of the compound. Descriptor values can be drawn from literature or calculated by computer programs. Relatively few descriptors are needed and they can be interpreted in biochemical terms.

2. *Free–Wilson analysis.* Structural features are used in a group contribution approach (substituents, SAR).

$$\log\left(\frac{1}{c}\right) = \sum_i a_i x_i + \mu,$$

where x_i denotes the presence of group i (0 or 1) and μ the contribution of the unsubstituted compound. Predictions can only be made for substituents already included in the training set. Therefore, a large number of compounds are needed, which yield a large number of features. Hansch analysis and Free–Wilson can also be combined.

The interpretation of linear (Q)SAR models is done rather straightforward by inspecting the most important features (i.e., features with high coefficients). Overfitting is rarely a problem because of the limited expressiveness of the model. For the same reason, the applicability of linear models is restricted to congeneric series with similar modes of action. Another problem with traditional (Q)SAR techniques is the selection of features for end points that are very complex and that incorporate many different and potentially unknown biological mechanisms. In this case, it is very likely to miss important features or to suffer from the "curse of dimensionality" if too many features have been selected.

5.1.2.2 Constraints in Predictive Toxicology Toxicological experiments are frequently expensive, time-consuming, and may require a large number of animal experiments. Therefore, it is usually impossible to create experimental data for congeneric series specifically for (Q)SAR modeling. For this reason, most toxicological (Q)SARs have to rely on existing data sets, which are in many cases very diverse in respect to structure, biological mechanisms, data origin, and quality.

Fortunately, publicly available structural and biological databases (e.g., PubChem [http://pubchem.ncbi.nlm.nih.gov/], Toxnet [http://toxnet.nlm.nih.

gov/], DSSTox [http://www.epa.gov/nheerl/dsstox/]) have grown substantially in recent years. Despite this wealth of information, databases are often characterized by the following properties that make modeling difficult:

- The chemicals are not congeneric; i.e., they do not share a common substructure and act by a common mechanism.
- The activities are noisy with missing values.
- The activity distributions are skewed and/or have other non-normal properties.
- A substantial amount of toxicity data is confidential and is not accessible to the general public.

With data mining techniques from artificial intelligence research, it is possible to use information from diverse databases much more efficiently than traditional (Q)SAR approaches that rely on congeneric compounds. Many of these techniques can be seen as automation of various aspects from the (Q)SAR modeling process. They work similar to a human (Q)SAR expert, who separates the training set into subsets with similar mechanisms, selects and calculates chemical features, and builds (Q)SAR models for the individual subsets. Many of them can be also seen as an attempt to base scientific decisions on sound statistical criteria.

5.1.2.3 Data Mining in (Q)SAR Modeling Data mining can be described as finding nontrivial, previously unknown, and potentially useful information in large amounts of data. In predictive toxicology, data mining techniques can be used for all model building tasks that will be described in the following sections. It is, e.g., possible to create, aggregate, and select relevant features, to group chemicals according to their similarity, or to create complex prediction models. In this context, we see traditional (Q)SAR techniques also as data mining tools that identify linear models in databases with chemical features and experimental toxicity data.

5.1.3 (Q)SAR Model Development

Independent of algorithmic and implementation details, the process of (Q)SAR modeling can be subdivided into five basic steps:

$$\text{Feature generation} \rightarrow \text{feature selection} \rightarrow \text{model learning} \rightarrow \text{model validation} \rightarrow \text{model interpretation}$$

The following sections will be organized according to this sequence, but we can also refine the whole procedure into more detail:

1. definition of the goal of the project and the purpose of the (Q)SAR models;
2. creation or selection of the training set;

3. checking the training set for mistakes and inconsistencies and performing corrections;
4. selection of the features relevant to the project (by expert knowledge or data mining);
5. selection of the modeling technique;
6. exploratory application and optimization of the modeling and feature selection techniques to see if it provides useful results;
7. application of the selected and optimized techniques to the training set;
8. interpretation of the derived model and evaluation of its performance; and
9. application of the derived model, e.g., to predict the activity of untested compounds, or an external test set with known activity values.

It is usually impossible to use all features because they are highly correlated and contain much noise. A high-dimensional feature space is also sparsely populated and hardly interpretable. For this reason, a thorough selection of features is extremely important (step 4). This can be achieved through a combination of objective feature selection and a further refinement step (projection-based or supervised method).

Steps 5–7 employ data mining techniques for distance weighting and distance measures as well as for similarity measurements and regression.

A software package that implements a rather complete (Q)SAR solution using data mining methods is Waikato Environment for Knowledge Analysis (WEKA) [4] (http://www.cs.waikato.ac.nz/ml/weka/). There are also several packages that make chemoinformatics libraries written in other languages available in R [1,5]. A high-level visual workflow approach to data exploration and analysis with interfaces to both WEKA and R is KNIME, the Konstanz information miner (http://www.knime.org/). The OpenTox project (http://www.opentox.org/) aims to build an open source framework for predictive toxicology. It will incorporate many of the tools mentioned in this chapter together with automated validation routines and facilities to build graphical user interfaces.

5.1.3.1 Criteria for the Selection and Evaluation of Data Mining Algorithms The Organisation for Economic Co-operation and Development (OECD) has developed acceptance criteria for (Q)SARs for regulatory purposes [6,7]. Specifically, these are

1. a defined end point;
2. an unambiguous algorithm with a clear description of the mathematical procedure;
3. a defined applicability domain with descriptor and structure space definitions;

4. measures of goodness of fit (r), robustness (q^2) and predictivity (external prediction); and
5. a mechanistic interpretation shall be given, if possible.

These rather broad criteria contain essential aspects of good practice in (Q) SAR modeling. However, for the purpose of data mining applications, these criteria are rather general and do not provide enough algorithmic details for their implementation. Within the following sections, we will propose formal definitions and algorithms for OECD criteria, especially for the assessment of feature space properties, applicability domains, and model validation (items 3 and 4).

The following section will provide more detail about the individual steps that are involved in the development of predictive toxicology models.

5.2 FEATURE GENERATION

The goal of feature generation is the description of chemical structures. There is no set of universal features that describes a compound equally well for all purposes.

The classical (Q)SAR methods (Hansch analysis and Free–Wilson) both employ multiple linear regression to build a model. Hansch analysis was historically used to derive a statistical relationship between measured quantities of chemicals and toxicological activities exhibited by those chemicals. The octanol–water partition coefficient (log P), for example, is closely related to lipophilicity and describes the ability of a chemical to pass membranes in the body. It is therefore correlated with many toxic effects and can be used to statistically model these end points. Hansch analysis uses physicochemical properties and substituent constants, while Free–Wilson uses chemical fragments derived from the 2-D structure. Such descriptors can be (among others)

1. Structural properties (structural alerts/substructures from general feature mining):
   a. structural alerts from experts (substituent constants),
   b. hybrid (refinement of structural alerts by data mining techniques),
   c. substructures derived by graph mining algorithms, and
   d. spectroscopic data;
2. Experimental and calculated physicochemical properties, quantum chemical parameters or graph theoretical indices (electronic, hydrophobic, or steric), e.g., log P, pKa:
   a. measured biological properties, e.g., from short-term assays, high-throughput screening, -omics data.

Structural properties can be obtained directly from chemical compounds and are called primary features, while experimental or calculated quantities are called secondary features.

The selected feature type affects not only the predictive performance but also the biological rationale for the algorithm and the interpretation of individual predictions. The interpretability of models and predictions benefits from features that are well known to chemists and toxicologists and that have a clear mechanistic relevance.

From a statistical point of view, it is important to classify features as qualitative or quantitative. Qualitative features indicate the presence or absence of some feature, while quantitative features give a measured or calculated amount on some numerical scale. Structural features are frequently qualitative, e.g., they indicate the presence or absence of some substructure, and experimental features are frequently quantitative. Historically, both types were used and models are referred to as either SAR or QSAR for qualitative and quantitative features, respectively. Hansch and Free–Wilson analyses use quantitative and qualitative features, respectively (Fig. 5.2).

Open source software projects that provide chemical toolkits and libraries for feature generation and for many other purposes are associated in the blue obelisk group [8], e.g., OpenBabel, CDK, JOELib (http://www.blueobelisk.org/).

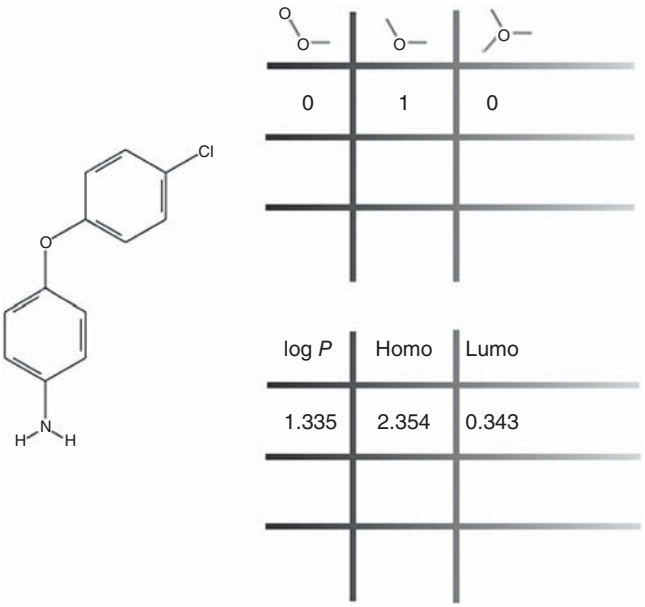


Figure 5.2 Features obtained from a chemical. Upper: qualitative primary features (SAR), below: quantitative secondary features (QSAR).

5.3 FEATURE SELECTION

Traditionally, the (Q)SAR modeler has to use his/her knowledge about toxicological mechanisms to decide which features will be included in a (Q)SAR model. Especially with complex and poorly understood toxic effects, the selection of features is likely to be incomplete and error prone. With data mining, we can use objective criteria to select relevant features automatically from a large set of features in order to filter out noise and to find informative patterns within the data. Using a large feature space together with objective criteria for feature selection reduces the risk of ignoring important features and allows an automated detection of new structural alerts. A basic understanding of statistical tests is vital for the application of feature selection algorithms [9].

5.3.1 Unsupervised Techniques

Methods that do not consider toxic activities (the dependent variable) are called unsupervised techniques. They remove redundant information and/or construct fewer, more informative features. Table 5.1 lists some popular unsupervised techniques for feature selection.

With objective feature selection, each pair of features is compared. This is usually implemented by iteratively adding features to the data matrix X when they pass the tests. In SAR modeling, i.e., with qualitative features, objective feature selection can contain identity, zero and singularity tests, checking for features that occur in the same structures and for features that do not occur or occur only once in the training compounds. In QSAR modeling, i.e., with quantitative features, it is possible to check for standard deviation (a feature carries little information when it has a low standard deviation), singularity (where the values are the same for all compounds except one), and correlation.

Cluster analysis is a procedure for grouping together similar features in clusters, thus enabling the algorithm to pick one representative for each cluster. The problem is to decide *a priori* how many groups should be built as this depends to a large extent on the data. Most popular are techniques that recursively partition the features. A very advanced technique is known as self-organizing maps [10]. Computational complexity varies greatly for these approaches.

TABLE 5.1 Some Popular Unsupervised Techniques for Feature Selection

Name	Theory of Operation	Retains Features?
Objective feature selection	Selects features iteratively	Yes
Cluster analysis	Group correlated features	Yes
Principal component analysis	Projects data to a lower dimension	No

Principal component analysis is a projection of the data to a lower-dimensional vector space, thereby eliminating correlations between features. It works by finding the eigenvalues and eigenvectors of the covariance matrix of X. A rotation matrix is created that projects the data into the vector space made up by the most influential eigenvectors (the principle components), accounting for most of the data's variance. Usually, a decision is made beforehand for a specific percentage of variance and the algorithm uses only the most influential eigenvectors to reach this threshold. By not using all eigenvectors, data compression through dimensionality reduction is achieved. The amount of compression depends on the correlation within the original data. Principal component analysis is a frequently applied technique and well documented [11]. It is available as a function in R [1].

Using principal component analysis harms the interpretability of a model, as the original feature space gets lost. However, the loadings can be inspected to assess the influence of the original features present in the principal components. Objective feature selection and clustering techniques are well behaved in this respect. Unsupervised techniques are not prone to overfitting since only redundant information is removed.

5.3.2 Supervised Techniques

Supervised feature selection tries to select features that correlate well with the dependent variable, i.e., the activities. In the SAR case (i.e., with qualitative features), it is possible to assign significance values to features which can be used as a preprocessing step to selection, and this is discussed first. Similar techniques are available for quantitative features. Significance values for features are also valuable when it comes to model building.

5.3.2.1 Significance Tests Given two different sets of compounds (e.g., compounds with/without a certain substructure), it is interesting to find out whether the two samples differ significantly in respect to their toxicological activities. The activity values form sample distributions and we can use statistical tests to find out if the distributions of both sets differ significantly (see Fig. 5.3). If the difference is significant, it is possible that the investigated substructure contributes to the toxic activity. This association is of course purely statistical, and human expert knowledge is still needed to determine the exact biological mechanisms.

A popular choice for the comparison of qualitative results (e.g., carcinogen/noncarcinogen classifications) is the χ^2 test, whereas the Kolmogorov–Smirnov test can be used for the comparison of qualitative data (e.g., LD_{50} values). The probability (p value) that an observed difference is due to chance can be calculated from the test statistics. A common significance threshold is 0.05, which means that one false positive is accepted in 20 cases.

If multiple tests are performed (e.g., for the evaluation of sets of substructures), the p values have to be corrected. If p is the significance threshold for

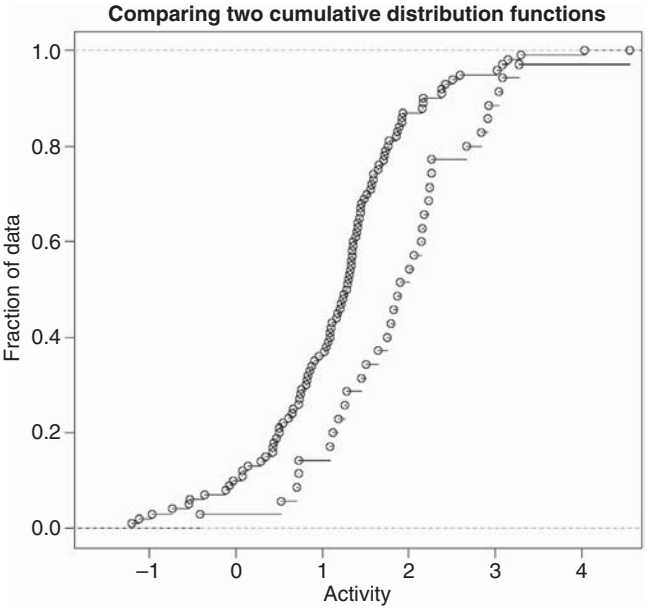


Figure 5.3 A comparison of the cumulative activity distributions of two sets of activity values x and y with sizes 100 and 35, respectively. The mean value of x is 1.0; the mean value of y is 2.0. It is highly unlikely (Kolmogorov–Smirnov test gives $p = 0.0001319$) that x and y have been drawn from the same data source.

a specific test, then $1 - p$ is the probability of drawing a negative feature f_i. For n independent tests, the probability that no single test is positive for f_i is $(1 - p)^n$, which converges to 0 for growing n. This increases the probability of type I errors (false discovery rate). A simple correction is the Bonferroni correction, which divides each p value by n. More sophisticated methods to control the false discovery rate exist [12]. In settings where the absolute values are less important than rankings, corrections can be omitted. There exists an R package for multiple tests (multitest) [13] that features also functions for permutation tests, bootstrapping, and jackknifing procedures that increase the reliability of tests.

The set size is very important for significance tests. A set size below 12 is usually considered "very small," and that below 30 is "small." Mean values differ greatly for very small sets and are still unstable for small sets [9]. In other words, to avoid chance effects, no significance tests should be performed for very small sets. For small sets, permutation tests can be helpful.

5.3.2.2 Supervised Selection In supervised feature selection, a particular selection of features is evaluated and assigned a score (reward signal). This process is iterated many times to identify an optimal feature set. This makes the method computationally expensive and bears the danger of overfitting the

TABLE 5.2 Some Popular Supervised Techniques for Feature Selection

Name	Theory of Operation	Retains Features?
Forward selection/backward elimination	Iterative (de)selection	Yes
Simulated annealing	Probabilistic selection	Yes
Genetic algorithm subset selection	Probabilistic selection	Yes

selection with respect to the training data, reducing the ability to predict external data. Significance tests for features can be used as a preliminary step for supervised feature selection, which is a special case of reinforcement learning [14]. Table 5.2 lists some popular supervised techniques for supervised feature selection.

The naive approach in supervised feature selection is to evaluate all possible subsets of features. However, most of the time, this is computationally too expensive. *Forward selection* starts with an empty set of features and successively adds features that increase the fit, starting with the most significant features. But forward selection has drawbacks, including the fact that each addition of a new feature may render one or more of the already included features nonsignificant. Backward elimination goes the other way round: it starts with all features and removes those that have little contribution to the model. This method also has limitations; sometimes features that would be significant when added to the final reduced model are dropped. Stepwise selection is a compromise between the two methods, allowing moves in either direction.

Simulated annealing switches in each iteration to a different selection of features with a probability that depends on the goodness of fit and a "temperature" parameter, t. The lower the fit and the higher t, the greater the probability for switching. t is decreased with every iteration (therefore the name) until a certain threshold is reached. The idea is to overcome local maxima by "jumping."

Genetic algorithm subset selection successively narrows down the feature space by evolutionary means. It recombines pairs of sets of features by mimicking crossover and mutation to obtain better features. In each "generation," the remaining candidates are evaluated by a "fitness function" and the process repeats itself with the more successful ones.

5.4 MODEL LEARNING

5.4.1 Data Preprocessing

Most predictive toxicology techniques do not work directly on raw experimental measurements but rely on some sort of preprocessing. This can involve statistical calculations (e.g., for the determination of TD_{50} or LD_{50} values) as

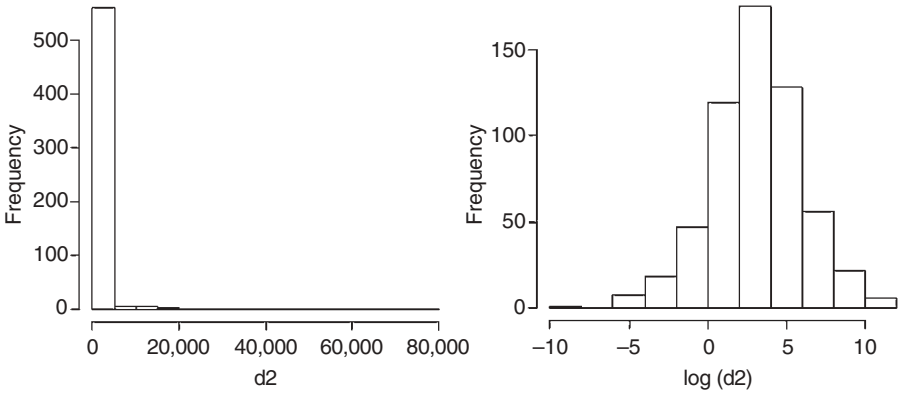


Figure 5.4 Histogram of original (left) and log-transformed database activities (right).

well as expert knowledge (e.g., human carcinogenicity classifications) to aggregate replicates, doses, or multiple experiments into a single value. It is important to understand the properties and limitations of these techniques before attempting to model a derived variable (e.g., Are assumptions behind the procedures verified? Are quantitative values good indicators of toxic potencies? Are the results expressed in molar values?). A description of the data aggregation procedures should be part of the documentation for the first OECD criterium (defined end point).

A common (Q)SAR practice is to log transform quantitative variables to the range of values and to achieve a normally distributed data set. It is still important to check the normality assumption for each data set before parametric methods are applied. If the normality assumption is not met, "binning" the data into discrete values might help. A more generally applicable solution is to use nonparametric methods that make no assumptions about data distributions (Fig. 5.4).

5.4.2 Modeling Techniques

Table 5.3 lists popular modeling techniques for (Q)SAR regression and classification.

Multilinear models have been in use for a long time. As linear equations, they are easy to use and are relatively straightforward to interpret. For n instances, they are defined as the coefficients that minimize the error on a system of n linear equations,

$$y = b_{1x_{i1}} + \ldots + b_{mx_{im}} + d, i \in \{1, \ldots, n\}, \tag{5.2}$$

or, in a more compact notation,

TABLE 5.3 Popular Modeling Techniques

Name	Theory of Operation
Traditional QSARs (Hansch, Free–Wilson) [15]	Multilinear regression on physicochemical properties/structural features
Artificial neural networks [16–18]	Nonlinear multidimensional parameterized model mimicking the function of neurons
Support vector machines [19]	Robust classification algorithm using hyperplanes to split the feature space into class regions
Decision trees and rule learners [20]	Hierarchical rules from recursive partitioning of the training data
k-Nearest neighbor techniques [20]	Derive the prediction from the activities of structurally similar compounds

$$y = \langle X, b \rangle + d, \tag{5.3}$$

where $\langle \cdot . \cdot \rangle$ denotes the normal dot product, and b and d are the coefficients to learn. Multilinear models assume linear relationships between features and activities; therefore, the expressiveness is limited and the model will perform poorly if these conditions are not met. The remaining (nonlinear) models are able to fit diverse data structures (in fact, many can fit arbitrary data), but are frequently too complex for interpretation or have a poor biological rationale. Learning can take very long and overfitting is more likely. For both types of neural networks, several decisions have to be made about architecture, learning rate, and activation functions.

Support vector machines are perhaps the most prominent approach that represents the family of kernel-based techniques (kernel machines). Another member of this family that is quite new to machine learning are Gaussian processes [21]. Successful approaches have demonstrated that kernel machines are more solid than, and can serve as a replacement for, artificial neural networks in a wide variety of fields [22].

To see how support vector machines work, let us consider a classification problem, i.e., $y = \{-1, +1\}^n$. The same techniques are applicable to regression or principal component analysis or any other linear algorithm that relies exclusively on dot products between the inputs. The prediction $f(x_q)$ is obtained by

$$f(x_q) = \text{sign}(\langle x_q, b \rangle + d), \tag{5.4}$$

where sign(.) denotes the sign of the prediction; i.e., f gives a prediction depending on which side of the hyperplane $\langle x, b \rangle + d = 0$ the query structure x_q lies. Finding an optimal separating hyperplane constitutes a quadratic opti-

mization problem. The coefficient b is then a linear combination of some training vector x_i (the support vector):

$$b_j = \Sigma_{i=1\ldots n} a_i y_i x_i, \tag{5.5}$$

which allows to rewrite Equation 5.4 as an integration over the training data

$$f(x_q) = \text{sign}(\Sigma_{i=1\ldots n} a_i y_i \langle x_q, x_i \rangle + d). \tag{5.6}$$

The dot product $\langle x_q, x_i \rangle$ denotes the cosine of the angle between x_q and x_i (assuming unit length of the vectors). It can thus be seen as a similarity measure with geometric interpretation. The dot product is the simplest instance of a kernel function. However, support vector machines usually do not perform learning in the original feature space. The key is to replace x_q and x_i in the right-hand side of Equation 5.6 by higher-dimensional representations $\varphi(x_q)$ and $\varphi(x_i)$, where $\varphi: \mathrm{R}^n \rightarrow \mathrm{R}^m$ with $m > n$ is called a map.

The expression $\varphi(x)$ is not calculated directly in practice due to combinatorial explosion. Support vector machines exploit the fact that it only occurs in dot products in the algorithms. This allows to bypass direct calculation of the map. Instead, a so-called kernel function, $k: \mathrm{R}^m \times \mathrm{R}^m \rightarrow \mathrm{R}$, is used, which calculates $\langle \varphi(x_q), \varphi(x_i) \rangle$ directly in the input space ("kernel trick"). In fact, k can be any positive definite function denoting similarity. The final predictive equation is then given by

$$f(x_q) = \text{sign}(\Sigma_{i=1\ldots n} a_i y_i k \langle x_q, x_i \rangle + d). \tag{5.7}$$

5.4.3 Global Models

If a model is fitted to training data in advance, i.e., without knowing the query structure, then the model is called "global." At query time, global models simply evaluate the model function on the training instance to obtain a prediction. Therefore, global models require low memory and give fast predictions once the training phase is over. However, complex functions in a high-dimensional feature space suffer from data sparseness and are easily overfitted, thereby destroying its predictive ability for new compounds.

Overfitting is the process of fitting a model with many parameters too accurately to the training data. Despite a perfect fit for the training data, the resulting model has poor generalization capabilities and is not predictive for unknown query instances. To avoid overfitting, it is necessary to use additional techniques (e.g., cross validation, Bayesian priors on parameters or model comparison that can indicate when further training does not result in better generalization). The process of overfitting a neural network during training is also known as overtraining.

The effect of data sparseness in high dimensions is due to the so-called curse of dimensionality [23]. Roughly speaking, with increasing dimensions,

subsets of the data span a growing subspace that approaches the whole feature space rapidly. In other words, with a high number of dimensions, the distance between compounds increases and the neighborhoods get sparse.

5.4.4 Instance-Based Techniques (Local Models)

It is frequently possible to identify congeneric subsets within diverse data sets. Such a group of structures can be said to represent a local (Q)SAR. Global (Q)SAR methods may not recognize such local relationships if they do not use very complex (nonlinear) functions and many features.

Local models obtain a prediction for a query structure using its "local neighborhood" rather then considering the whole data set; i.e., they only use training compounds that are similar to the query structure with respect to some distance measure. They can also use fewer features than global models. Local models cannot be built before the query instance is known. Most local models not only defer model learning but also defer clustering the training compounds into neighborhoods until a query instance is to be predicted. Because of that they are also termed "lazy."

With lazy learning, for each distinct query, a new approximation to the target function is created. The approximations are local and differ from one another; therefore, for the whole feature space, many different approximations are used at different locations. The single approximations maybe simple (e.g., linear), but seen as a whole, they can approximate a complex function. They are also robust because they depend only on the data points close to the query instance. In contrast to eager learning, the computational burden for the prediction is higher, since all the training is done at query time.

5.4.4.1 Similarity Measures The idea is to cluster congeneric compounds by chemical similarity and to use only the nearest neighbors as training instances and/or to weight the contribution by distance. The similarities between the query compound and the training compounds are also useful for determining applicability domains and prediction confidences (see Section 5.6).

Traditionally, chemical similarity is determined by expert knowledge to obtain clusters of congeneric chemicals (chemical classes). The assignment of chemical classes is, however, frequently ambiguous and does not necessarily reflect biological mechanisms. For fully automated data mining approaches, a wide variety of similarity indices have been proposed [24].

Willet et al. [24] have reviewed 22 structural similarity indices by searching databases for chemical analogues. They showed that combinations of descriptors perform best, among them the Tanimoto, the Russel–Rao, the simple matching, and the Stiles coefficients. They all work on 2-D fragment bit strings, indicating the presence or absence of structural features in a compound. The Tanimoto index, for example, calculates the ratio of common features between two compounds.

For quantitative features, distance-based indices are also well suited (Euclidean or Mahalanobis distance). A data structure that can be used for

an efficient calculation are kd-trees (libkdtree++ [http://libkdtree.alioth.debian.org/]).

Chemical similarity can also be assessed by supervised techniques (i.e., by taking the training activities into account). The contribution of each feature to the Tanimoto index can be weighted, for example, with the p values of statistical significance tests [25].

5.4.4.2 Prediction from Neighbors and Distance Weighting Having determined the similarities between the query structure and each training structure, these values can be used to select a local neighborhood to the query structure and to train the model on these compounds only. Different methods are available for neighbor selection:

- Counting cutoff: Use the k nearest neighbors, where k is a fixed number.
- Similarity cutoff: Use the neighbors that are more similar than some fixed similarity threshold.
- Soft selection: Use all compounds and weight their contribution to the model by their similarity values, where more similar compounds get higher weights. Doing so is no harm to model precision because distant training points will have little effect on the approximation. The only drawback is that model building takes longer.

Of course, distance weighting can also be applied in the cutoff approaches. In dense populations, a kernel function is frequently used to additionally smooth the similarity. A variety of smoothing functions have been reviewed in [26]. Most widely used are Gaussian kernels of the squared exponential form

$$sim_g(x_i, x_q) = \exp\left(-\frac{1}{2} sim(x_i, x_q)^2\right). \tag{5.8}$$

This kernel creates a progression phase in the neighborhood and generally ameliorates conditions. It can also be stretched by using the general Gaussian probability distribution function with adjustable width.

The actual prediction can then be obtained by rather simple models, e.g., with distance weighted majority votes for classification problems and multilinear regression for regression problems. More complex models can be tried if the simple approaches do not give satisfactory results.

5.5 COMBINATION OF (Q)SAR STEPS

Efficient graph mining techniques are currently a strong research focus of the data mining community. As chemicals can be represented as graphs, many of these techniques can also be used for chemoinformatics and (Q)SAR problems. Most of them focus on the efficient identification of relevant

substructures (combining the feature generation and selection steps) or on using graph structures directly for classification/regression.

5.5.1 Constraint-Based Feature Selection

Complete feature sets can be built by decomposing the structures of the training set into all subgraphs of a certain type (e.g., paths, trees, graphs). As this process is computationally very expensive, various techniques to reduce the search space have been developed. Traditionally, size limits have been used, but this can lead to the loss of large significant fragments. More recently, frequency-based constraints have been introduced (e.g., in MolFea [27], FreeTreeMiner [28], gSpan [29], and Gaston [30]). The idea is to restrict the search space by stating the minimal and/or maximal frequencies in two classes of compounds (e.g., carcinogens/noncarcinogens), and the algorithm finds efficiently all subgraphs that fulfill these constraints.

Although restricting the search for substructures by minimum/maximum constraints is intriguing at first glance, there are several problems associated with this approach:

- The goal of feature selection is to find fragments that are significantly correlated with a toxicological outcome. Most graph mining algorithms support only monotonic constraints (e.g., minimum and maximum frequencies), but test statistics are usually convex. Although extensions for convex functions (e.g., χ^2) exist, they prune the search space rather inefficiently in our experience.
- As frequency-based searches use activity information, it is important to repeat the search whenever the training set changes (e.g., if a query compound has been identified and removed from the database and for each fold during cross validation; see Section 5.7). Having to repeat the fragment search frequently (e.g., for model development or cross-validation runs) may render the initial performance advantage useless. Storing the complete fragment search and repeating only the selection process can be a more efficient alternative.

5.5.2 Graph Kernels

Graph kernels have been developed to incorporate graph structures into support vector predictions (see Section 5.2).The crucial part is to define a kernel that indicates the chemical similarity of two compounds (see also Section 5.4.1). An example that uses substructure fingerprints is the Tanimoto kernel. For two compounds, x_i and x_j, the kernel function is the proportion of feature f that is shared between x_i and x_j:

$$k^{\tau}(x_i, x_j) = \frac{|[f \mid f \subseteq x_i \wedge f \subseteq x_j]|}{|[f \mid f \subseteq x_i \vee f \subseteq x_j]|}. \tag{5.9}$$

Different techniques have been proposed that work on the adjacency matrix of graphs and derive different features (directed or undirected, labeled or unlabeled subgraphs, etc.) as well as marginalized graph kernels that obtain features from Markov random walks. In practice, support vector machines with graph kernels can perform remarkably well (for an extended discussion, see, e.g., Reference 31).

5.6 APPLICABILITY DOMAIN

5.6.1 Definition and Purpose of Applicability Domains

Jaworska et al. define the applicability domain of a (Q)SAR as "the physicochemical, structural or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds" [32]. A critical assessment of the applicability domain is important to distinguish between reliable and unreliable predictions.

The purpose of applicability domains is to tell whether the modeling assumptions are met. With data mining methods, this is a twofold task: (1) Are training compounds similar enough to the query instance? (2) How is the descriptor space populated (e.g., How dense are the training compounds? Is the query compound within the subspace spanned by the training compounds?)?

5.6.2 Determination of Applicability Domains

In traditional (Q)SAR approaches, the applicability domain is determined by the modeled end point and by the selection of compounds and features. In the Hansch analysis, for example, features triggering the end point are selected, and consequently, the applicability domain consists of compounds that contain those features or whose features lie in the respective range, i.e., those that belong to a certain chemical class.

With data mining methods, the practical application of the applicability domain concept requires an operational definition that permits the design of an automatic (computerized), quantitative procedure to determine a model's applicability domain. Although up to now there is no single generally accepted algorithm for determining the applicability domain, there exists a rather systematic approach for defining interpolation regions [33]. The process involves the removal of outliers with the help of a probability density distribution estimation using different distance measures. When using distance metrics, care should be taken to use an orthogonal and significant feature space. This can be achieved by a different means of feature selection and by successive principal component analysis.

For practical applications, a viable approach consists of two steps: (1) transform the training data so that the feature space has acceptable properties (low

dimensionality and orthogonality) and (2) generate a probability density allowing to assess important aspects of the distribution. More specifically, the following steps can serve as a guide toward a reliable confidence index:

- Create a low-dimensional, orthogonal feature space and prune redundant information with objective feature selection followed by principal component analysis with a threshold for variance loss. The optimal threshold can be estimated by cross validation.
- For normally distributed training compounds, create a probability density distribution estimation, taking into account the data's "shape" using Mahalanobis distance, D_M, defined as

$$D_M(x_i) = \sqrt{(x_i - \mu_X)^{\mathrm{T}} R_x^{-1} (x_i - \mu_X)} \tag{5.10}$$

for any data point x_i, where μ_x is the center of the distribution X and R_x is the covariance matrix of the data. Leverage h, which is directly related to Mahalanobis distance, is defined as $h(x_i) = D_M(x_i)/n - 1$ (see Fig. 5.5). For non-normally distributed data, nonparametric methods have to be applied [33]:

- Identify if the query compound is an outlier with the estimated density distribution. As a rule of thumb, a compound, x_i, is an outlier if $h(x_i) > 2p/n$, where p is the number of parameters in the model [9].
- Create a confidence value, c, for every prediction by combining density distribution estimates and a global chemical similarity index (e.g., Tanimoto index). Typically, a well-spread neighborhood in feature

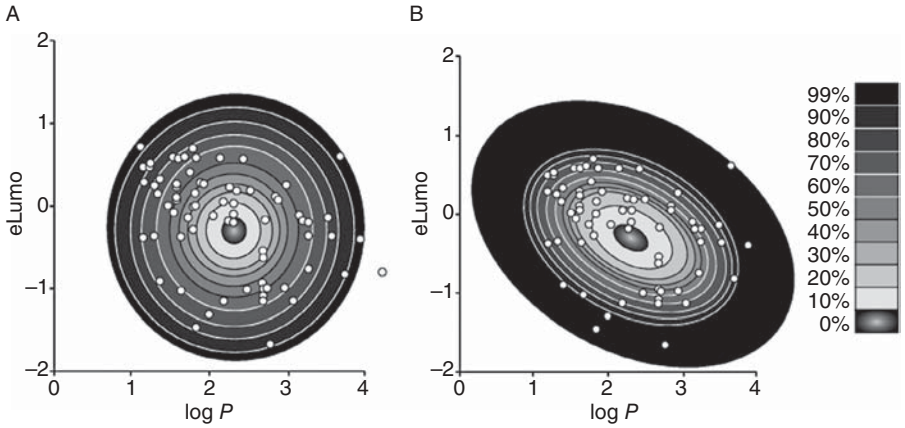


Figure 5.5 Probability density estimation using Euclidean distance (left) and Mahalanobis distance (right) (taken with permission from Reference 33).

space combined with high chemical similarity should give high-quality predictions. Typical implementations use a fixed ratio of chemical similarity and density distribution estimation to determine the confidence. By convention, c ranges between 0 (lowest confidence) and 1 (highest confidence).

A recent approach in this direction, termed "automated lazy learning QSAR," achieved high accuracy using an automatically calculated applicability domain from distributional properties of the training set [34]. Specifically, the applicability domain incorporated the average (Euclidean) distance and standard deviation of distances to the center of the distribution. To account for chemical similarity, the lazy learning approach used similarity weighting based on Gaussian kernels.

Suitable open source software for these purposes is available from the R project [1].

5.7 MODEL VALIDATION

The goal of model validation is to evaluate the performance for untested compounds, i.e., the predictive power of the model. This step is often interleaved with applicability domain estimation: by predicting compounds, it can be assessed how well the applicability domain discriminates between good and bad predictions.

5.7.1 Validation Procedures

5.7.1.1 Retrofitting the Training Set Especially with multilinear (Q)SAR models, predicting the compounds in the training set is still a popular "validation" method, although this technique does not evaluate the performance for unseen instances. The problem is less obvious for multilinear regression because it cannot fit the training data exactly, but many data mining techniques can accommodate any data distribution (e.g., neural networks). If no precautions against overfitting are taken, they achieve 100% accuracy on the training set, but the overfitted function performs poorly for new predictions. For this reason, it is crucial to test every model performance with structures that have not been used for model building.

5.7.1.2 Artificial Validation Sets As it is usually impossible to create experimental data for validation purposes, it is common practice is to split the available data into training and test sets prior to modeling. The model is developed with the training set and the test set is used to validate the model prediction. Although the procedure may seem to be simple and straightforward, there are several possible pitfalls:

- All test set information has to be excluded from the training set. This means that all supervised feature selection methods have to be performed only with training set information.
- The composition of the test set has a huge impact on validation results. If the test set has many compounds within the applicability domain, prediction accuracies will increase; test sets that are very dissimilar to the training set will achieve low accuracies.
- As validation results depend strongly on the test set composition, it would be ideal to validate with a test set that has been drawn randomly from future prediction instances—unfortunately, these are rarely known to the model developer.
- If the training and test set are drawn from the same source, they still share common information, e.g., about activity distributions. This will lead to overly optimistic results for techniques that derive *a priori* probabilities from training set distributions (e.g., naive Bayes).
- If the same test set is used repeatedly for model development and parameter optimization, it is likely that the resulting model is overfitted for a particular test set and will perform poorly for other instances.
- There is a trade-off between training and test set sizes: large training sets improve the model performance, but large test sets improve the accuracy of validation results.

We will argue later in Section 5.7.2.3 that the inclusion of applicability domains in validation results will resolve some of these problems. To enable accurate performance indicators for smaller data sets, cross-validation techniques have been developed. The complete data set is divided into n folds. Each fold serves once as test set for a model based on the remaining $n - 1$ folds. With this procedure, it is possible to obtain unbiased predictions for all compounds of the original data set. It is, however, important to repeat feature selection and parameter optimizations within each cross-validation fold.

5.7.1.3 External Validation Sets The "gold standard" to evaluate model performance is to determine the end point experimentally and to compare the results with predictions. In this case, it is impossible to cheat voluntarily or involuntarily or to use information about the test set distribution for model development. However, external validation sets share two important limitations with other test sets:

- The validation results depend to a large extent on the test set composition and on the fraction of compounds within the applicability domain of the model.
- Validation results with large test set are more reliable than results from small test sets. As a rule of thumb, test sets should contain at least 30 compounds.

5.7.2 Performance Measures

The following discussion of performance measures assumes that a validation set X of size n has been predicted and the goal is to assess the predictive power of the model.

5.7.2.1 Classification We assume a twofold classification; i.e., $f(x_i)$ and y_i can only take two possible values, e.g., active and inactive for all $x_i \in X$. The simplest measure for the potential of the model to differentiate between right and wrong predictions is precision. It is defined as the ratio of correct predictions with respect to a certain confidence threshold, ad, as

$$\text{prec}(\text{ad}) = \frac{|x_i \mid f(x_i) = y_i \wedge c_i > \text{ad}|}{|x_i \mid c_i > \text{ad}|}. \tag{5.11}$$

A counting statistic can be obtained in a contingency table that counts classifications based on the predicted and database activity. When this data is combined with the confidence value c_i obtained from applicability domain estimation (see Section 5.6), a receiver operating characteristic (ROC) curve [35] can be generated. For every confidence threshold, ad, it is possible to calculate the true positive ratio and the false positive ratio as

$$\begin{aligned} \text{tpr}(\text{ad}) &= \frac{|\{x_i \mid f(x_i) = \text{active} \wedge y_i = \text{active} \wedge c_i > \text{ad}\}|}{|\{x_i \mid y_i = \text{active} \wedge c_i > \text{ad}\}|} \quad \text{and} \\ \text{fpr}(\text{ad}) &= \frac{|\{x_i \mid f(x_i) = \text{active} \wedge y_i = \text{inactive} \wedge c_i > \text{ad}\}|}{|\{x_i \mid y_i = \text{inactive} \wedge c_i > \text{ad}\}|}. \end{aligned} \tag{5.12}$$

The true positive rate, tpr, indicates the sensitivity or recall of the model, i.e., how easy the model recognizes actives, and 1 – fpr indicates the specificity of the model, i.e., how robust it is against false alarms at a confidence level of ad. Plotting tpr against fpr for many possible values of ad between 0 and 1 gives the ROC curve (see Fig. 5.6).

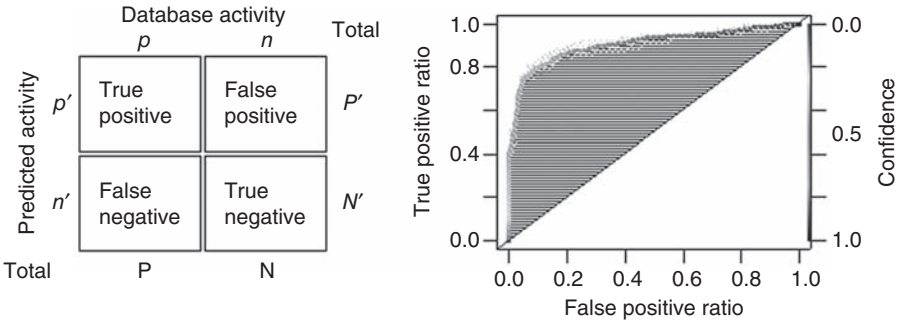


Figure 5.6 An example contingency table and ROC curve.

An ROC curve shows several things. First, it demonstrates that any increase in sensitivity will be accompanied by a decrease in specificity; i.e., there is a trade-off between the two. Second, the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the model, and the closer the curve comes to the 45° diagonal of the ROC space, the less accurate the model. Furthermore, the slope of the tangent line at a specific confidence threshold gives the likelihood ratio for that confidence value of the model. Finally, the area between the curve and the diagonal is a measure of model accuracy. This is a very valuable and usable parameter because it is nonparametric; i.e., it assumes no specific data distribution.

ROCR, a rather powerful library for ROC analysis, which is able to generate a wealth of performance measures for classification, is available for R [36].

5.7.2.2 Regression Choosing a performance measure for regression, i.e., when predicting quantitative values, is not so easy because a counting statistic is not available. A straightforward and nonparametric measure is the mean squared error. It is defined as

$$\text{mse} = \sum_{j=1..m} (x_i - f(x_i))^2.$$

The mean squared error should always be calculated as an unambiguous performance measure. However, this quantity is sensitive to the overall scale of the target values, and it makes sense to normalize by the variance of the training activities to obtain the standardized mean squared error (smse).

If the sample size is not small (i.e., >30) and the data are normally distributed, the degree of correlation between predicted and database activities can be measured with r^2, the squared correlation coefficient. For two normally distributed variables, F and Y, the correlation coefficient is defined as

$$r(F, Y) = \frac{\text{cov}(F, Y)}{\sqrt{\sigma_F^2 \sigma_Y^2}}, \tag{5.13}$$

where σ_F^2 is the variance of F and σ_Y^2 is the variance of Y. More common than r is r^2, the square of r. It can be interpreted as the proportion of the variance explained by the model.

Generally, the higher the r^2, the better the fit of the model, because r^2 describes how well a linear approximation would fit the plot of pairs of Y and F. However, r^2 can only be applied if the two variables are normally distributed [9,37], and this has to be verified in every case unless nonparametric alternatives are used. Acceptable values for (Q)SAR models are $r^2 \geq 0.64$ ($r \geq 0.8$) [38].

Because of the variability of experimental results, it has been argued that the fraction of predictions within one log unit of error (assuming that the data is log transformed) is "acceptable and closer to regulatory needs" than correla-

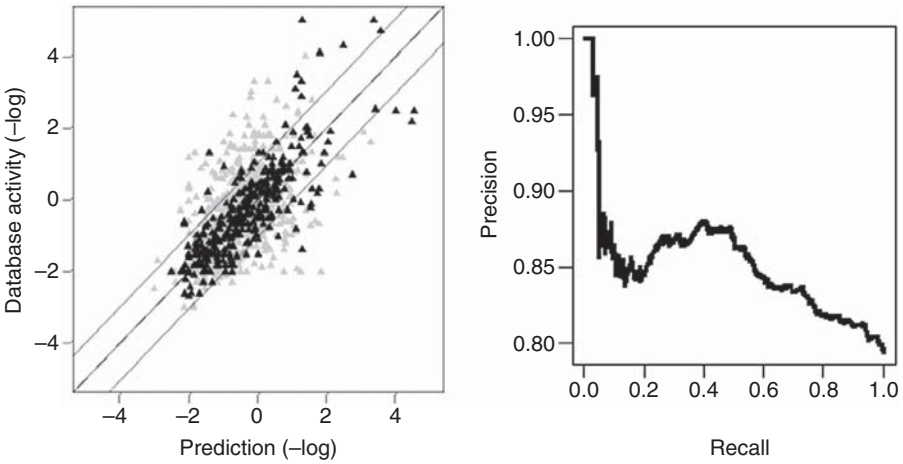


Figure 5.7 Left: predicted versus database activities for the FDAMDD (Federal Drug Administration Maximum Recommended Daily Dose) data set (Matthews EJ, Kruhlak NL, Benz RD, Contrera JF. Assessment of the health effects of chemicals in humans: I. QSAR estimation of the maximum recommended therapeutic dose (MRTD) and no effect level (NOEL) of organic chemicals based on clinical trial data. *Curr Drug Discov Technol* 2004;1:61–76) obtained by leave-one-out cross validation. Compounds within the applicability domain are drawn black, the rest gray. The error limit of one log unit is depicted as parallel to the diagonal. Right: precision versus recall with regard to the error limit (unpublished material by the authors).

tion coefficients [6]. This way, the ease of counting statistics is regained and it is possible to perform ROC analysis. Correlation coefficients maybe also difficult to interpret for nonstatisticians (see Fig. 5.7).

5.7.2.3 Impact of Applicability Domains on Validation Results The purpose of applicability domains is to discriminate reliable from unreliable predictions. For individual predictions, a confidence value can indicate the distance to the applicability domain and the expected quality of the prediction. The actual confidence values depend on the composition of the training set and on the query structure.

For an easier interpretation of results, a cutoff for acceptable confidence indices (and thus an applicability domain with fixed borders) can be introduced.

Validation results depend to a large extent on the test set composition and on the fraction of test compounds within the applicability domain. To compensate for this effect, it is advisable to use only the test set compounds within the applicability domain for model validation, which gives more consistent validation results [39]. Another alternative would be to weight individual predictions with their associated confidence index.

If a counting statistic is available, i.e., a classification of predictions into correct and wrong, a very simple tool related to ROC analysis is

cumulative accuracy (ca). For the k predictions with the highest confidence, calculate

$$\text{ca} = \frac{\sum_{i=1}^{n} c_i * \delta_i}{\sum_{i=1}^{n} c_i}, \tag{5.14}$$

where $\delta_i = 1$ if prediction i is correct and $\delta_i = 0$ else, and c_i is the confidence of prediction i. This calculates the confidence-weighted correct prediction ratio and removes the bias induced by high confidence values from precision (see Section 5.7.2.1).

5.7.3 Mechanistic Interpretation

Many (Q)SAR and data mining techniques can be used to derive a hypothesis about biological mechanisms. However, it is important to remember that most of these techniques have no knowledge about chemical and biological processes. Thus, they cannot reason about mechanisms, but they can provide information that is relevant for a mechanistic assessment (e.g., structural alerts, compounds with similar modes of action). This means that a toxicological researcher has to evaluate only a limited number of possible hypotheses, but expert knowledge is still needed for the identification of mechanisms.

The interpretability of models and individual predictions may depend on several factors:

- *Model complexity.* Interpretability decreases with model complexity and abstraction level, but complex models are frequently needed to accommodate for real world situations. It is, however, not always necessary to interpret complete models. The extraction of specific information (e.g., relevant substructures/properties) and the inspection of rationales for individual predictions may provide more information for toxicologists that complete models.
- *Biological rationale for the algorithm.* Most scientists find it easier to interpret models that have a biological rationale and/or resemble their way of thinking about toxicological phenomena. Techniques based on chemical similarities are very useful in this respect because they support the search for analogs and chemical classes. A mechanistic hypothesis (and a critical evaluation of individual predictions) can be obtained from the inspection of relevant features and from the mechanisms of structurally similar compounds.
- *Visual presentation of the results.* Most data mining programs are hard to use for nondata mining experts and have great shortcomings in the visual presentation of their results. End users with a toxicological background should not be confused with data mining/(Q)SAR terminology and with detailed options for algorithms and parameter settings. The

interface should provide instead an intuitive and traceable presentation of the rationales for a prediction together with links for the access of supporting information (e.g., original data, results in other assays, literature).

5.8 CONCLUSION

The most frequent application of data mining in predictive toxicology is the development of (Q)SAR models. The development of (Q)SAR models requires (1) the generation of features that represent chemical structures, (2) the selection of features for a particular end point, (3) the development of a (Q)SAR model, (4) the validation of the model, and (5) the interpretation of the model and of individual predictions.

For each of these tasks, a large number of data mining techniques are available. Selecting and combining suitable algorithms for the individual steps allows us to develop problem-specific solutions with capabilities that go beyond standardized solutions. However, it is important to understand the properties and limitations of the applied techniques and to communicate them clearly to the model users.

REFERENCES

1. R Development Core Team. R: A Language and Environment for Statistical Computing R. Foundation for Statistical Computing, Vienna, Austria. 2007. Available at http://www.R-project.org. ISBN 3-900051-07-0 (accessed June 6, 2008).
2. Richard AM. Commercial toxicology prediction systems: A regulatory perspective. *Toxicol Lett* 1998;102-103:611–616.
3. Richard AM, Williams CR. Public sources of mutagenicity and carcinogenicity data: Use in structure-activity relationship models. In: *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*, edited by Benigni R. Boca Raton, FL: CRC Press, 2003.
4. Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann. 2005. Available at http://www.cs.waikato.ac.nz/ml/weka/ (accessed August 6, 2008).
5. Guha R. Chemical informatics functionality in R. *J Stat Softw* 2007;18:00–00. Available at http://www.jstatsoft.org/v18/i05 (accessed August 6, 2008).
6. Benigni R, Bossa C, Netzeva T, Worth A. Collection and evaluation of (Q)SAR models for mutagenicity and carcinogenicity. 2007. http://ecb.jrc.ec.europa.eu/documents/QSAR/EUR_22772_EN.pdf (accessed June 25, 2008).
7. Pavan M, Netzeva T, Worth A. Validation of a QSAR model for acute toxicity. *SAR QSAR Environ Res* 2006;17:147–171.
8. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner JK, Willighagen EL. The blue obelisk–interoperability in chemical informatics. *J Chem Inf Model* 2006;46:991–998.
9. Crawley MJ. *Statistics: An Introduction Using R*. Chichester, UK: Wiley, 2005.

10. Guha R, Serra JR, Jurs PC. Generation of QSAR sets with a self-organizing map. *J Mol Graph Model* 2004;23:1–14.
11. Jolliffe IT. *Principal Components Analysis*. New York: Springer, 2002.
12. Yoav B, Yekutieli D. The Control of the False Discovery Rate in Multiple Testing Under Uncertainty. *Ann Stat* 2001;29:1165–1188.
13. Pollard KSDudoit Svan der Laan MJ. Multiple testing procedures: R multitest package and applications to genomics. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pp. 251–272 (edited by Gentleman R, et al.). New York: Springer Science + Business Media, 2005.
14. Russell SJ, Norvig P. *Artificial Intelligence: A Modern Approach*, 2nd edn. Upper Saddle River, NJ: Prentice Hall, 2002.
15. Franke R, Gruska A. General introduction to QSAR. In: *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*, edited by Benigni R. Boca Raton, FL: CRC Press, 2003.
16. Papa E, Villa F, Gramatica P. Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in *Pimephales promelas* (fathead minnow). *J Chem Inf Model* 2005;45:1256–1266.
17. Eldred DV, Weikel CL, Jurs PC, Kaiser KL. Prediction of fathead minnow acute toxicity of organic compounds from molecular structure. *Chem Res Toxicol* 1999;12:670–678.
18. Serra JR, Jurs PC, Kaiser KL. Linear regression and computational neural network prediction of tetrahymena acute toxicity for aromatic compounds from molecular structure. *Chem Res Toxicol* 2001;14:1535–1545.
19. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed July 6, 2008).
20. Mitchell TM. *Machine Learning*. Columbus, OH: The McGraw-Hill Companies, Inc., 1997.
21. Rasmussen CE, Williams CKI. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. London: MIT Press, 2005.
22. Schölkopf B, Smola AJ. *Learning with Kernels*. London: MIT Press, 2002.
23. Bellman R. *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.
24. Holliday J, Hu C, Willett P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb Chem High Throughput Screen* 2002;5:155–166.
25. Helma C. Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and salmonella mutagenicity. *Mol Divers* 2006;10:147–158.
26. Atkeson CG, Moore AW, Schaal S. Locally weighted learning. *Artif Intell Rev* 1997;11:11–73.
27. Helma C, Kramer S, De Raedt L. The molecular feature miner MolFea. Proceedings of the Beilstein-Institut, Workshop, Bozen, Italy, May 13–16, 2002.
28. Rückert U, Kramer S. Frequent free tree discovery in graph data. In: *Proceedings of the ACM Symposium on Applied Computing (SAC 2004)*, 2005, pp. 564–570.

29. Yan X, Han J. gSpan: Graph-based substructure pattern mining. In: *ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining*. Washington, DC: IEEE Computer Society, 2002.
30. Nijssen S, Kok JN. The Gaston tool for frequent subgraph mining. Electronic notes in theoretical computer science. In: *Proceedings of the International Workshop on Graph-Based Tools (GraBaTs 2004)*, vol. 127, issue 1. Elsevier, 2005.
31. Ralaivola L, Swamidass SJ, Saigo H, Baldi P. Graph kernels for chemical informatics. *Neural Netw* 2005;18:1093–1110.
32. Jaworska JS, Comber M, Auer C, Van Leeuwen CJ. Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environ Health Perspect* 2003;111:1358–1360.
33. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR applicability domain estimation by projection of the training set descriptor space: A review. *Altern Lab Anim* 2005;33(5):445–459.
34. Zhang S, Golbraikh A, Oloff S, Kohn H, Tropsha A. A novel automated lazy learning QSAR (ALLQSAR) approach: Method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *J Chem Inf Model* 2006;46:1984–1995.
35. Egan JP. *Signal Detection Theory and ROC Analysis*. New York: Academic Press, 1975.
36. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: Visualizing classifier performance in R. *Bioinformatics* 2005;21:3940–3941.
37. Anscombe FJ. Graphs in statistical analysis. *Am Stat* 1973;27:17–21.
38. Cronin MT, Livingstone DJ. *Predicting Chemical Toxicity and Fate*. Boca Raton, FL: CRC Press, 2004.
39. Benigni R, Netzeva TI, Benfenati E, Bossa C, Franke R, Helma C, Hulzebos E, Marchant C, Richard A, Woo YT, Yang C. The expanding role of predictive toxicology: An update on the (Q)SAR models for mutagens and carcinogens. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* 2007;25:53–97.

6

CHEMOGENOMICS-BASED DESIGN OF GPCR-TARGETED LIBRARIES USING DATA MINING TECHNIQUES

KONSTANTIN V. BALAKIN AND ELENA V. BOVINA

Table of Contents

6.1 INTRODUCTION

Modern chemogenomics is a special discipline studying the biological effect of chemical compounds on a wide spectrum of biological targets. Currently, insights from chemogenomics are increasingly used for the rational compilation of screening sets and for the rational design and synthesis of directed chemical

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

libraries to accelerate drug discovery. However, considering huge amounts of existing chemical and biological data (compounds, targets, and assays), analysis and effective exploration of the data represent a very complex problem. This chapter discusses specific issues associated with the chemogenomics-based data mining strategies including chemogenomics databases, annotated libraries, homology-based ligand design, and design of target-specific libraries, in the context of G protein-coupled receptor (GPCR)-targeted drug design.

GPCRs are the largest family of membrane-bound receptors and are also the targets of an estimated 30% of marketed drugs. About 400 GPCRs identified in the genome are considered to be good therapeutic targets [1,2]. At the same time, only 30 receptors are currently addressed by marketed drugs suggesting great potential for the development of novel chemical entities that modulate the activity of GPCRs.

Over the past few years, numerous computational algorithms have been introduced to build a robust basis for the rational design of chemical libraries, including GPCR-focused sets. The observed trend is that a molecular diversity alone cannot be considered to be a sufficient component of a library design. High-throughput screening (HTS) of large diversity-based libraries is still a common strategy within many pharmaceutical companies for the discovery of GPCR leads. However, as noted by many researchers in the field, there is no evidence that high-throughput technologies, including parallel synthesis/combinatorial chemistry and HTS, provided the expected impedance to the lead discovery process [3,4]. A number of computational approaches have been implemented for the design of GPCR-focused libraries. These include pharmacophore and target structure-based design strategies, approaches based on specific structural recognition motifs, and specific methods of data mining.

Despite successful application in several lead discovery programs [5,6], practical utility of the target structure-based approach in the screening of GPCR-biased chemical libraries is still limited. This is presumably due to the lack of structural data, detailed knowledge of the ligand binding mode, and inherent issues concerning scoring functions. In addition, there are specific concerns associated with particular classes of GPCRs. For example, assembly of meaningful compound sets targeting peptidergic G protein-coupled receptors (*p*GPCRs) still poses a considerable challenge. In this situation, rapid, reliable, and conceptually simple ligand-based strategies are of importance, especially for cases when the structural information is scarce.

Knowledge-based data mining methods discussed in the following section successfully complement modern strategies in GPCR-directed drug discovery.

6.2 DATA MINING TECHNIQUES IN THE DESIGN OF GPCR-TARGETED CHEMICAL LIBRARIES

Dimensionality reduction techniques belong to a powerful cluster of methods in data mining, which can identify nonobvious relevant information for

further exploitation. In the case of compounds represented either by fingerprints or by a number of descriptors, a set of available data is multidimensional. To explore such information, it is necessary to map the data points into two-dimensional (2-D) or three-dimensional (3-D) space. The aim of the mapping procedure is to preserve the topology of the multidimensional matrix, so that data points that are close together in the multidimensional environment are accurately represented by the space with a reduced number of variants.

There are several methods for projection computation based on neural and statistical approaches. Specifically, topology- and distance-preserving mappings, e.g., self-organizing feature map of Kohonen [7] or distance-preserving nonlinear mapping of Sammon [8] (discussed in Chapter 16 of this book), are well suited for data visualization and data mining purposes.

As an illustration, Savchuk et al. used self-organizing maps (SOMs) for analysis and visualization of different groups of GPCR ligands based on seven calculated molecular descriptors [9]. In this experiment, tachykinin NK_1 antagonists (1400 compounds), muscarinic M_1 agonists (563 compounds), and β_3-adrenoceptor agonists (433 compounds) appeared to be clustered at distinctly different areas of the map. Such maps for particular groups of ligands can be used for predicting potential subtype-specific activity.

A virtual screening procedure based on a topological pharmacophore similarity and SOMs was applied to optimizing a library of P_1 purinergic human A_{2A} receptor antagonists [10]. Initially, a SOM was developed using a set of biologically tested molecules to establish a preliminary structure–activity relationship (SAR). A combinatorial library design was performed by projecting virtually assembled new molecules onto the SOM. A small focused library of 17 selected combinatorial products was synthesized and tested. On average, the designed structures yielded a threefold smaller binding constant (~33 versus ~100 nM) and 3.5-fold higher selectivity (50 versus 14) than the initial library. A most selective compound revealed a 121-fold relative selectivity for A_{2A} versus A_1. This result demonstrated that it was possible to design a small, activity-enriched focused library with an improved property profile using the SOM virtual screening approach. The strategy might be particularly useful in projects where structure-based design cannot be applied because of a lack of receptor structure information, for example, in the many projects aiming at finding orphan G protein-coupled receptor (oGPCR) modulators.

By contrast to SOM, nonlinear maps (NLMs) represent all relative distances between all pairs of compounds in the 2-D version of a descriptor space. The distance between two points on the map directly reflects similarity of compounds. NLMs have been initially used for the visualization of protein sequence relationships in 2-D and for comparisons between large compound collections, represented by a set of molecular descriptors [11].

Differences between several receptor-specific groups of GPCR ligands were investigated by using Kohonen SOMs [9]. Because of some problems in the analysis of multidimensional property spaces inherent to Kohonen SOM

methodology, the same team of researchers performed a complementary study using the algorithm of nonlinear mapping (N.P. Savchuk, S.E. Tkachenko, and K.V. Balakin, unpublished data). The specific aim of this work was to discriminate synthetic small-molecule ligands to *p*GPCRs and nonpeptidergic G protein-coupled receptors (non-*p*GPCRs). As input variables, electrotopological state (E-state) indices [12] were used, which encode information about the both topological environment of an atom and the electronic interactions resulting from all other atoms in the molecule. Unlike other types of molecular descriptors, E-state indices are easily and unambiguously calculated, and, at the same time, they encode some essential molecular features characterizing the topology, polarity, and hydrogen-bonding capabilities of a molecule. A 593-compound database of known GPCR modulators was collected including both launched drugs as well as compounds that entered preclinical and clinical trials. Two nonoverlapping data sets were included in the database: the first data set consisted of 186 *p*GPCR ligands; the second data set included 407 agents active against non-*p*GPCRs. A total of 24 E-state indices were calculated for each molecule.

Figure 6.1 shows distributions of compound categories on the Sammon map. Small-molecule *p*GPCR ligands are shown as black circles and non-*p*GPCR-active drugs are indicated as white circles. There are clear differences

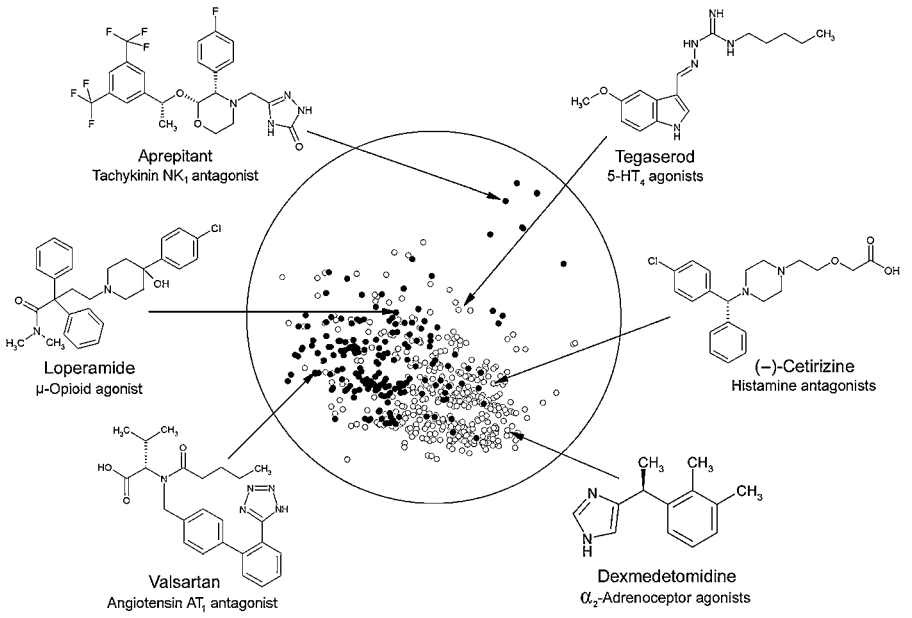


Figure 6.1 Nonlinear map illustrating the differences between small-molecule synthetic ligands to peptidergic (black circles) and nonpeptidergic (white circles) GPCRs expressed in terms of atomic electrotopological state.

in their location. Shown structures and locations of three launched *p*GPCR-active drugs (on the left) and three non-pGPCR-active drugs (on the right) provide some visual clues for their discrimination. In general, *p*GPCR ligands are topologically more complex and have an increased number of polar functional groups. This method provides a reasonable basis for the assessment of the *p*GPCR activity potential. For example, it can be used as a computationally inexpensive virtual filtering procedure in the design of *p*GPCR-targeted libraries. The described exercise also illustrates the increased complexity of synthetic *p*GPCR ligands, expressed in terms of atomic E-state, as compared to non-*p*GPCR ligands.

Recently, Vogt and Bajorath reported the design of target-selective chemical spaces using CA-DynaMAD, a mapping algorithm that generates and navigates flexible space representations for the identification of active or selective compounds [13]. The algorithm iteratively increases the dimensionality of reference spaces in a controlled manner by evaluating a single descriptor per iteration. For seven sets of closely related biogenic amine GPCR antagonists with different selectivity, target-selective reference spaces were designed and used to identify selective compounds by screening a biologically annotated database.

The modern computational tools provide interactive, fast, and flexible data visualizations that help analyze complex structure–activity dependencies. However, visualization alone is often inadequate when large numbers of data points need to be considered. Powerful data mining methods that are to search for meaningful intervariable relationships in large multidimensional databases are now being used for the design of GPCR-targeted libraries.

Recursive partitioning (RP) is simple yet powerful statistical method of choice for the analysis of SAR in large complex data sets [14,15]. In the field of GPCR-targeted library design, RP algorithm was used for the analysis of a large number of μ-opioid receptor ligands [16]. It was shown that the optimized RP model "discovered" the existence of the two main ("morphine-like" and "meperidine-like") μ ligand subtypes, represented as the two main "active nodes" of the receptor modulator tree.

The RP technique can also be applied to a sequential screening [17] of compound libraries for a particular GPCR activity. The sequential approach screens compounds iteratively for activity, analyzes the results, and selects a new set of compounds for the next screening round based on a previous data set. The purpose of this iterative process is to maximize information about ligand–receptor interactions and to minimize early-stage discovery costs. Jones-Hertzog et al. employed the sequential approach to the analysis of data obtained from 14 GPCR assays [18]. Several cycles of screening appeared to be more efficient than screening all the compounds in a large collection.

Another study [19] was focused on GPCRs that are activated by positively charged peptide (GPCR-PA^+) ligands. Using special partitioning algorithm based on five calculated molecular descriptors, a region of chemical property space enriched in GPCR ligands was identified. This information was used to

build a "test" library of 2025 individual compounds to probe space associated with the endogenous GPCR-PA^+ ligands. The library was evaluated by HTS against three different *p*GPCRs, namely, rMCH, hMC4, and hGnRH. It yielded considerably more active ligands (4.5–61.0-fold) compared with a control set of 2024 randomly selected compounds.

In the past 10–15 years, methods based on artificial neural networks (ANNs) have been shown to effectively aid in the design of target-specific libraries. Several papers have appeared in the literature that describe successful use of different neural network approaches to distinguish different categories of GPCR-active compounds in large data sets [20–22].

Researchers from Nippon Shinyaku studied the probabilistic neural network (PNN), a variant of normalized radial basis function (RBF) neural networks, as a predictive tool for a set of 799 compounds having activities against seven biological targets including histamine H_3 and serotonin 5-HT_{2A} GPCRs [23]. The compounds were taken from the MDL Drug Data Report (MDDR) database to represent both distinct biological activities and diverse structures. Structural characteristics of compounds were represented by a set of 24 atom-type descriptors defined by 2-D topological chemical structures. The modeling was done in two ways: (1) compounds having one certain activity were discriminated from those not having that activity, and (2) all compounds were classified into seven classes corresponding to their biological target. In both cases, around 90% of the compounds were correctly classified in the internal test sets. For example, in the binary classification task, the percentages of correctly classified histamine H_3 and serotonin 5-HT_{2A} ligands were 81.4% and 90.8%, respectively.

The data mining algorithms mentioned here were successfully employed to identify regions of "chemical space" occupied by GPCR ligands. These relatively inexpensive and comprehensive algorithms correlating molecular properties with specific activities play an increasingly significant role in chemical library design. The ability to identify compounds with the desired target-specific activity and to optimize a large number of other molecular parameters (such as absorption, distribution, metabolism, and excretion/toxicity [ADME/Tox] related properties, lead- and drug-likeness) in a parallel fashion is a characteristic feature of these methods. In the latter case, library design can be considered a multiobjective optimization problem, which has become a topic of growing interest over the last decade in the pharmaceutical industry. An overview of the general methodology for designing combinatorial and HTS experiments rooted in the principles of multiobjective optimization has been presented by Agrafiotis [24].

As further enhancement of data mining strategies, chemogenomics approaches provide novel opportunities in the design of targeted libraries through better understanding of the relationships between GPCR sequence and compounds that interact at particular receptors. In the following section, we describe key technologies that have been developed to date in the field of chemogenomics-based drug discovery.

6.3 MINING THE CHEMOGENOMICS SPACE

6.3.1 Annotated Libraries

A key specific tool usually used in chemogenomics-based applications is annotated libraries, which can be defined as chemical databases of compounds tested in a variety of biological assays such as an extensive series of gene expression analyses, *in vitro* and *in vivo* target-specific activity assays, pharmacokinetics and toxicity experiments, and other similar data collected from scientific literature to yield an integrated chemoinformatics system. Such annotated libraries allow for the annotation-based exploration of the information-rich space. To address this need, most of the developed technologies based on annotated libraries integrate chemical database management platforms with data mining and modeling tools making it possible to establish the relationships between chemical structures and their multiple biological activity profiles.

Historically, one of the first attempts to analyze a complex annotated ligand–target space was the work published in 1997 by Weinstein et al. [25]. The authors proposed a general concept of "information-intensive approach" for the analysis of biological activity patterns in the National Cancer Institute's (NCI's) screening panel. Recent years witnessed rapid development and maturation of technologies based on annotated libraries [26–29]. Notably, this progress was paralleled by the technological advances in combinatorial chemistry, HTS, cheminformatics, protein crystallography, and data mining. As a drug discovery tool, the annotated libraries can be used in a wide variety of applications.

For example, NCI's annotated screening database has been extensively analyzed to identify novel antitumor compounds with better potential for successful clinical trials and for market approval [30–33]. Thus, an extensive study of ca. 20,000 compounds tested against 80 of NCI's tumor cell lines was performed using Kohonen SOMs [31]. Figure 6.2 shows a fragment of this map, which defines locations of several groups of antitumor drugs; this map served for generation of hypotheses for the rational discovery of antitumor agents. Annotated library was applied for probing a variety of biological mechanisms and related biological targets in cellular assays [34] and for identifying a biological target or mechanism of action of a chemical agent [35].

A classification scheme based on annotated library was used to identify cysteine protease targets in complex proteomes and predicts their small-molecule inhibitors [36]. Publicly available selectivity data were employed to create a chemogenomic kinase dendrogram for 43 kinases [37]. An annotated library representing a map of small molecule–protein interaction for 20 clinical compounds across 119 related protein kinases has been created using KinomeScan technology [38]. Using this map, many novel interactions were identified, including tight binding of the p38 inhibitor BIRB-796 to an imatinib-resistant variant of the Abl kinase and binding of imatinib to the Src

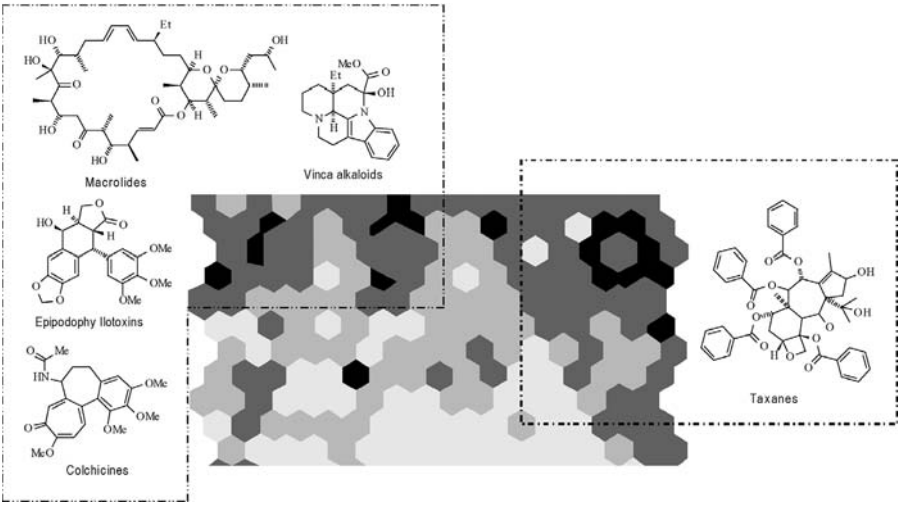


Figure 6.2 Fragment of Kohonen SOM, which defines the location of several tumor-specific groups of drugs [31].

family kinase Lck. Analysis of annotated library directed toward nuclear receptors has yielded scaffolds with highly promiscuous nuclear receptor profiles and nuclear receptor groups with similar scaffold promiscuity patterns [39]. This information is useful in the design of probing libraries for deorphanization of activities as well as for devising screening batteries to address the selectivity issues. Researchers at Amphora Discovery Corp. reported the design of an annotated knowledge database currently consisting of >30 million data points based on the screening of 130,000 compounds versus 88 targets [40].

Annotated library analyzed with a powerful data mining algorithm can be a useful cheminformatics tool to address ADME/Tox issues [41–43]. For example, a nonlinear map demonstrated differences in the sites of preferable localization of compounds with good human intestinal absorption (HIA) and agents well permeable trough blood–brain barrier (BBB) [41]. These observations, which are consistent with the fact that the phenomena of BBB and HIA permeability are different in their nature, are valuable in the design of orally active drugs, which are not intended to cross the BBB (for example, to avoid their central nervous system [CNS] toxicity).

6.3.2 Technologies Based on Annotated Databases

The mentioned researches demonstrating the value of annotated libraries in drug discovery triggered rapid growth of supporting industrial technological solutions (Table 6.1).

TABLE 6.1 Technologies Based on Annotated Compound Databases

Technology	Company	Type of Technology
BioPrint®	Cerep, http://www.cerep.fr/	Totally annotated multipurpose database
DrugMatrix®	Iconix Biosciences, Inc., http://www.iconixbiosciences.com	Totally annotated multipurpose database
WOMBAT	Sunset Molecular Discovery LLC, http://www.sunsetmolecular.com/	General-purpose databases with experimental data obtained from literature
StARlite™	Inpharmatica Ltd., http://www.inpharmatica.co.uk/	General-purpose databases with experimental data obtained from literature
WOMBAT-PK	Sunset Molecular Discovery LLC, http://www.sunsetmolecular.com	Annotated databases with a focus on ADME/Tox data
AurSCOPE ADME/Drug–Drug Interactions, AurSCOPE hERG Channel	Aureus Pharma, http://www.aureus-pharma.com/	Annotated databases with a focus on ADME/Tox data
Kinase Knowledgebase™	Eidogen-Sertanty, http://www.eidogen-sertanty.com/	Annotated databases with a focus on specific protein targets
AurSCOPE GPCR, AurSCOPE Ion Channels, AurSCOPE Kinase	Aureus Pharma, http://www.aureus-pharma.com/	Annotated databases with a focus on specific protein targets
ChemBioBase™	Jubilant Biosys Ltd., http://www.jubilantbiosys.com/	Annotated databases with a focus on specific protein targets

The BioPrint database was constructed by scientists at Cerep using a systematic profiling of over 2500 marketed drugs, failed drugs, and reference compounds, in a panel of 159 well-characterized *in vitro* assays including 105 binding assays (nonpeptide, peptide and nuclear receptors, ion channels, amine transporters), 34 enzyme assays (10 kinases, 10 proteases, 5 phosphodieterases), and 20 ADME/Tox assays (solubility, absorption, Cytochrome P450 [CYP] mediated drug–drug interaction). BioPrint also includes a large *in vivo* data set based on clinical information (therapeutic uses, adverse drug reactions, pharmacokinetics, and drug–drug interactions) for nearly all active agents. In addition, BioPrint presents a system for understanding the relationships between all *in vitro* pharmacology assays in the database and reported adverse drug reactions. All *in vitro* data are produced in Cerep's laboratories

with controlled overall data quality including test compound identity, purity, and stability. The BioPrint platform integrates proprietary software as well as data mining and modeling tools, which make it possible to determine *in vitro* pharmacology and/or ADME patterns that correlate with specific biological activities or clinical effects (for example, see References 44–46).

DrugMatrix is a flagship technology at Iconix (http://www.iconixbiosciences.com/), one of the first comprehensive research tools in the new field of toxicogenomics [43]. DrugMatrix consists of data from profiles on drugs and toxic substances using more than 10,000 large-scale gene expression microarrays, *in vivo* histopathology data, molecular pharmacology assays, and literature curation studies [47]. The latest release of DrugMatrix contains the profiles derived from administering 638 different compounds to rats. Compounds include Food and Drug Administration (FDA)-approved drugs, drugs approved in Europe and Japan, withdrawn drugs, drugs in preclinical and clinical studies, biochemical standards, and industrial and environmental toxicants. Analysis of toxicogenomic data from DrugMatrix allowed to accurately predict 28-day pathology from the gene expression readout at day 5, a time point when classic tools (clinical chemistry and histopathology) showed no evidence of toxicity in low-dose protocols [47]. This work demonstrated that genomic biomarkers can be more sensitive than traditional measurements of drug-induced biological effects.

BioPrint and DrugMatrix are totally annotated databases based on in-house bioactivity assays. Uniform and comprehensive annotations using advanced technological platforms create many valuable opportunities for pharmaceutical developers. However, a relatively small number of compounds in these databases still limit their utility in the design and analysis of novel lead chemotypes.

Other databases listed in Table 6.1 generally include information collected by manual or automatic literature screening, which remains the most popular approach for assembling annotated compound libraries. A massive amount of information is available in scientific literature about the active compounds and their biological properties, and thorough analysis of the literature is an essential yet resource-intensive knowledge generation method. Information from the diverse data sources is extracted using expert curators and is then placed into a database in a uniform format. The automatic buildup of compound annotation can be based on Medline literature reports [34] or web-based resources screened using special search engines (for example, Aureus Pharma's AurQUEST technology). The curated literature information on known biochemistry, pharmacology, toxicology, and other aspects of a drug's activity allows users to effectively mine and interpret experimental information related to their candidate molecules.

World of Molecular Bioactivity (WOMBAT) is a technology developed at Sunset Molecular Discovery (http://www.sunsetmolecular.com/) [48]. The latest version of WOMBAT (2008.1) contains 220,733 entries (192,529 unique SMILES) representing 1979 unique targets, captured from 10,205 papers

published in medicinal chemistry journals between 1975 and 2007. Annotated WOMBAT-PK database integrates knowledge from target-driven medicinal chemistry and clinical pharmacokinetics data. WOMBAT-PK is integrated within the WOMBAT database; its current version contains 1125 entries (1125 unique SMILES), totaling over 6509 clinical pharmacokinetic measurements. WOMBAT database was used for prediction of biological targets for chemical compounds [49,50].

StARlite annotated database (http://www.admensa.com/StARlite/) launched in 2005 by Inpharmatica comprises information on ca. 300,000 bioactive compounds including related pharmacology and target information abstracted from two journals, *Journal of Medicinal Chemistry* (from 1980 to the present) and *Bioorganic and Medicinal Chemistry Letters* (from 1991 to the present). Chemical structures in StARlite are available in 2-D and 3-D forms enabling 2-D and 3-D searching. There are over 1.3 million activity data points, which cover functional, binding, and ADME/Tox assays as well as some calculated molecular parameters. There are over 5000 unique molecular targets searchable by sequence and by various accession codes such as Swiss-Prot, TREMBL, and GenBank. This database can be used for navigating through compound, assay, activity, and target relationships, for obtaining target family chemotype portfolios, and for elucidating SAR, selectivity, and potency profiles.

AurSCOPE® developed at Aureus Pharma is a series of annotated chemical databases containing biological and chemical information related to a given pharmacological effect. Current databases available include AurSCOPE Global Pharmacology Space (GPS), AurSCOPE GPCR, AurSCOPE Kinase, AurSCOPE Ion Channels, AurSCOPE Nuclear Receptor, AurSCOPE Protease, AurSCOPE ADME/Drug–Drug Interactions, and AurSCOPE hERG Channel. Using AurSCOPE databases, specific *in silico* predictive models have been developed, useful for the design of focused libraries with decreased hERG-related side effects [51], as well as increased kinase [52] and ion channel [53] potency.

Kinase Knowledgebase developed at Eidogen-Sertanty is a database of biological activity data, SARs, and chemical synthesis data focused on protein kinases. It is based on Eidogen-Sertanty's proprietary web-based technology for capture, curation, and display of biological activity and chemical synthesis data from scientific literature and patents. This database currently covers more than 4600 journal articles and patents with over 370,000 SAR data points. The overall number of unique small-molecule structures for kinase modulators is greater than 469,000. The number of kinase targets with assay data is more than 390, and the number of annotated assay protocols is more than 16,700. Structural data available in Kinase Knowledgebase allow researchers to group known kinase inhibitors in scaffold groups and to lay out a project plan around patentable chemotypes [54].

ChemBioBase is a set of annotated target-specific ligand databases developed at Jubilant Biosys Ltd. The ChemBioBase products integrate assay data

and target information with chemical structure data collected from published patents and articles. Currently available databases are targeted toward protein kinases, GPCR, nuclear hormone receptors, ion channels, proteases, and phosphodiesterases.

It should be noted that the huge amount of literature data is highly heterogeneous in terms of organization of publication, scientific quality and protocol description. The systematic procedure of data collection and analysis should ensure that the context is always described as exhaustively as possible, and without variation due to publication analysts. A principal problem for many technologies from Table 6.1 is a nonuniform annotation of compounds. The gaps can create serious difficulties in data mining and thus reduce practical utility of these technologies in the analysis of complex interrelated biological activity phenomena.

In summary, current industrial technologies based on annotated libraries provide valuable possibilities in the design of novel chemistry to improve overall lead and library quality. The total activity profile of a compound in an annotated library comprises multiple signatures representing its structure, on- and off-target mechanistic effects, ADME/Tox data, and so on. The key problem is how to make use of these profiles in making decisions that improve the quality of drug discovery and development.

6.3.3 Chemogenomics-Based Design of GPCR Ligands

Effective identification of high-quality hits and leads across diverse classes of GPCR targets can also be based on a systematic analysis of structural genomics data [55]. Several approaches to explore the chemogenomics knowledge space of GPCR ligands and their receptors were recently reported, combined with their use in generating GPCR-directed libraries.

Researchers at Cerep reported an analysis of properties for more than 500 drugs screened against 42 targets. Based on this target–ligand database, they derived similarity metrics for both targets and ligands based on fuzzy bipolar pharmacophore fingerprints [45]. It was observed that ligands to subtypes of one target or closely related target families usually have similar ligand binding profiles, whereas homologically distant targets (such as 5-HT-binding GPCRs and ion channels), despite the common endogenous ligand, have a different ligand binding profile.

BioPrint database developed at Cerep can support medicinal chemists in the prioritization of hit series from HTS and in lead optimization stages. It can also be used as a computational tool for the development and improvement of SAR models, for *in silico* screening, and for the rational design of compound libraries. As an illustration, a QSAR model accounting for "average" GPCR binding was built from a large set of experimental standardized binding data extracted from the BioPrint database (1939 compounds systematically tested over 40 different GPCRs) [46]. The model was applied

to the design of a library of "GPCR-predicted" compounds. To validate the model, a 10% sample of the proposed GPCR-targeted library (240 randomly selected products) was experimentally assayed for binding on a panel of 21 diverse GPCRs. A 5.5-fold enrichment in positives was observed when comparing the "GPCR-predicted" compounds with 600 randomly selected compounds predicted as "non-GPCR" from a general collection. The obtained results suggest the usefulness of the model for the design of ligands of newly identified GPCRs, including orphan receptors.

Annotated ligand–target libraries provide reference sets for homology-based focused library design. Based on the structure–activity relationship homology (SARAH) principle formulated by Frye in 1999 [56], the knowledge obtained in the screening experiments for one target could be directly applied to lead discovery for its homologues and isoforms. Based on the concept of homology-based similarity searching, the researchers at Novartis developed an annotation scheme for the ligands of four major target classes, enzymes, GPCRs, nuclear receptors, and ligand-gated ion channels for *in silico* screening and combinatorial design of targeted libraries [57]. According to their approach, the homology-based library design consists of several principal steps (Fig. 6.3A). Initially, gene sequences for targets that have been identified by genomics approaches are cloned and expressed as target proteins that are suitable for screening. Using the annotated ligand–target database, at least one target with known ligands is selected that is homologous to this new target. Then the known ligands of the selected target are combined to a reference set. Finally, the potential ligands for the new target are searched based on their similarity to the reference set.

Figure 6.3B shows a part of this annotation scheme. It contains information about the amine-binding class A GPCRs and their ligands. As an example, a retrospective *in silico* experiment was described, in which 270 dopamine D_2 receptor ligands were used as a reference set. All compounds in the candidate set were ranked by their similarity to a reference compound set, and then compounds with Tanimoto similarity indices from 1 to 0.6 were analyzed. Authors noted that this homology-based similarity search is suitable for the identification of ligands binding to receptors closely related to a reference system. As an illustration, Figure 6.4 shows structures of dopamine D_2 (left column) and serotonin 5-HT_{1A} (right column) receptor ligands as examples for structurally similar ligands of two relatively distant amine-binding class A GPCRs. The same group of scientists reported modified [58] homology-based similarity searching based on special molecular representations, so-called Similog keys.

Another chemogenomics approach to the design of GPCR-targeted libraries has been developed by scientists at BioFocus [59]. Thematic analysis defines a common consensus binding site for all GPCRs in the upper half of the transmembrane (TM) region. Commonly occurring small sets of amino acids (themes) are identified from primary sequence overlays and are associated with ligand fragments (motifs) using SAR information. Multiple themes

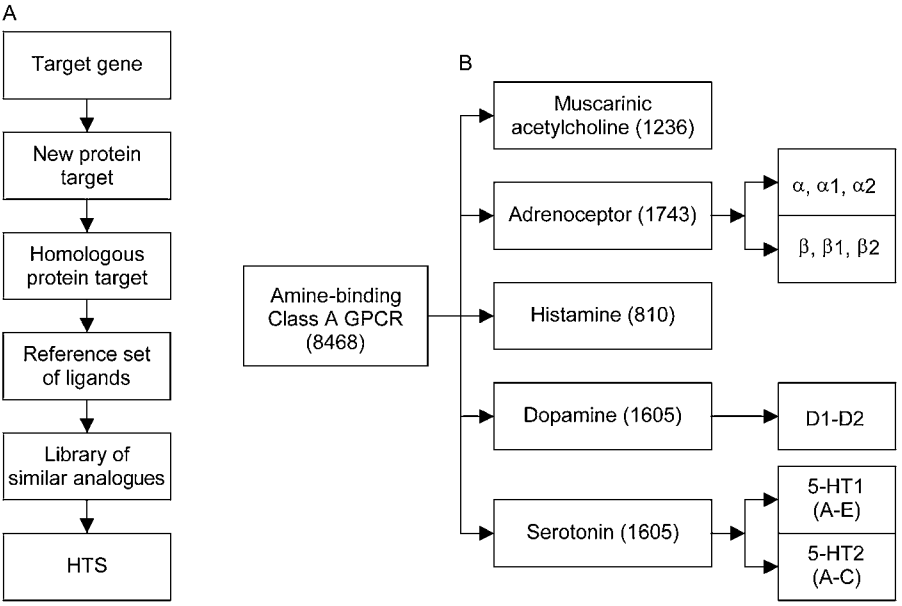


Figure 6.3 (A) Homology-based targeted library design. (B) Part of ligand–target classification scheme (amine-binding class A GPCR) used for homology-based similarity search [57].

have been identified across the GPCR family and these have been associated with motifs to create a design tool for combinatorial libraries and lead optimization.

A combination of chemogenomics and target structure-based approaches is expected to lead to an increase in the efficacy of HTS. Mutagenesis and SAR studies furnished the notion of a central binding site consisting of three subsites within the TM region of GPCRs [60]. Using this hypothesis, researchers at Biovitrum classified GPCRs via a chemogenomics approach, which defines the three subsites of the binding sites by manual docking of 5-HT, propranolol, and 8-OH-DPAT into a homology model of the 5-HT_{1A} receptor with experimentally verified interactions as constraints [61]. The chemogenomic classification was followed by a collection of bioactive molecular fragments and virtual library generation. By applying the strategy to the serotonin 5-HT_7 receptor, a focused virtual library with 500 members was created. To evaluate the library, compounds active at the 5-HT_7 receptors were collected from the literature. Furthermore, a virtual library was created from all commercially available building blocks of a similar composition to assess the benefit of the design process. Principal component analysis of molecular descriptors suggested library focus to the region in the chemical space defined by the reported actives. An enrichment factor in the range of

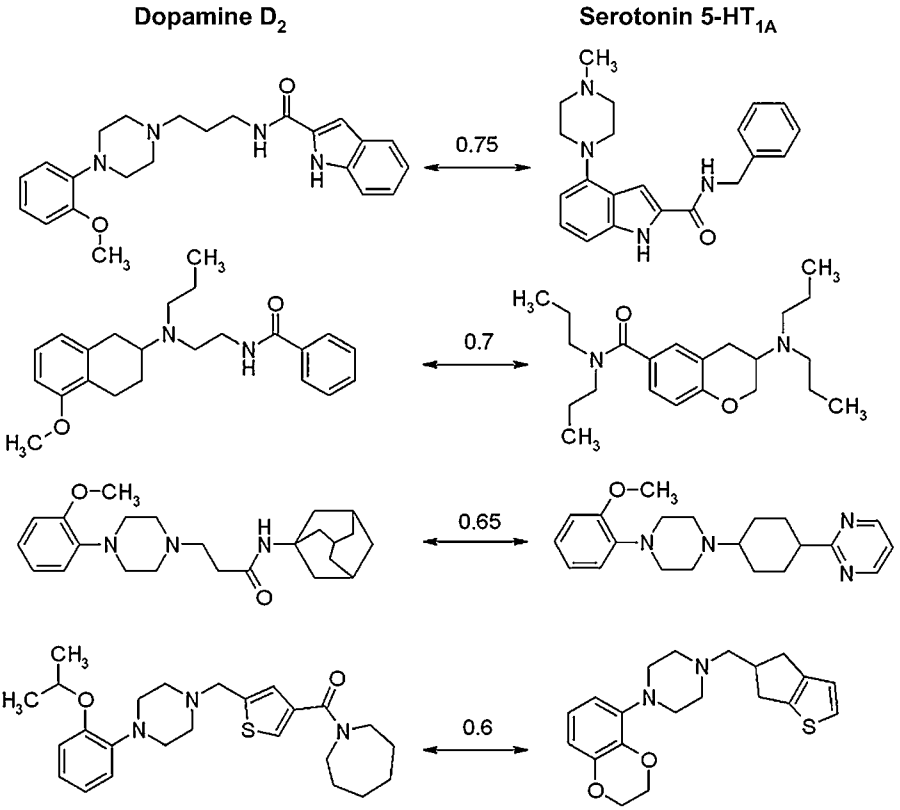


Figure 6.4 Serotonin 5-HT_{1A} receptor binding compounds can be identified using similarity to the D_2 reference set [57]. Tanimoto similarity coefficients are shown for each pair of ligands.

2–4 was reported for the 5-HT_7-targeted library when compared to the reference library.

Using an annotated set based on screening of 130,000 compounds versus 88 different targets [40], multiple chemical scaffolds were identified, allowing for prioritization and expanding SAR through medicinal chemistry efforts. Target selection in the annotated library encompassed both target families (kinases, proteases, phosphatases, GPCRs, ion channels, and lipid-modifying enzymes) and pathways.

Bock and Gough developed a virtual screening methodology [62] that generated a ranked list of high-binding small-molecule ligands for oGPCRs. They merged descriptors of ligands and targets to describe putative ligand–receptor complexes and used support vector machine (SVM) algorithm to discriminate real complexes from ligand–receptor pairs that do not form complexes. A similar approach was used to correlate ligand–receptor descriptions to the

corresponding binding affinities [63]. The authors of this work modeled the interaction of psychoactive organic amines with the five known families of amine GPCRs. The model exploited data for the binding of 22 compounds to 31 amine GPCRs, correlating chemical descriptions and cross descriptions of compounds and receptors to binding affinity.

Jacob et al. presented an *in silico* chemogenomics approach specifically tailored for the screening of GPCRs [64], which allowed to systematically test a variety of descriptors for both the molecules and the GPCRs. The authors tested 2-D and 3-D descriptors to describe molecules and five ways to describe GPCRs, including a description of their relative positions in current hierarchical classifications of the superfamily and information about key residues likely to be in contact with the ligand. The performance of all combinations of these descriptions was evaluated on the data of the GLIDA database [65], which contains 34,686 reported interactions between human GPCRs and small molecules. It was observed that the choice of the descriptors has a significant impact on the accuracy of the models. The developed method based on SVM algorithm was able to predict ligands of oGPCRs with an estimated accuracy of 78.1%.

In summary, a systematic exploration of the annotated ligand–target matrix for selected target families appears to be a promising way to speed up the GPCR-directed drug discovery. The principal challenge for technological platforms and computational approaches mentioned in previous sections of this chapter is to develop computational methods capable of deciphering information contained in annotated libraries and effectively displaying the results for more effective "next-step" decisions in drug candidate selection and development.

6.4 CHEMOGENOMICS-BASED ANALYSIS OF CHEMOKINE RECEPTOR LIGANDS

The following exercise focuses specifically on the application of a specific multidimensionality reduction technique in the context of the chemogenomics approach to derive information from simultaneous biological evaluation of multiple compounds on a set of coherent biological targets belonging to an actual class of GPCRs, chemokine receptors.

6.4.1 Mapping the Chemogenomics Space of GPCR Ligands

Kohonen SOMs, a compound classification method used in this work for correlation of molecular properties with specific activities, play a significant role in modern virtual screening strategies. Applications of this algorithm ranges from the identification of compounds with desired target-specific activity, which constitutes an essential part of the virtual screening ideology, to the

prediction of a wide spectrum of key pharmacologically relevant features including biological activity, pharmacokinetic and ADME/Tox profiles, and various physicochemical properties.

6.4.1.1 Annotated Knowledge Database As a first step in the analysis of the chemogenomics space, we collected a knowledge database. It included a set of drug compounds with experimentally defined activity against the biological targets of interest. The information was extracted from several commercially available pharmaceutical databases, such as Prous Ensemble (http://www.prous.com/) and WOMBAT [48] databases, as well as proprietary knowledge databases that can be easily used as the separate or joint source of information about structures and their specific activities. In addition to approved therapeutic drugs, the database also included lead compounds entered in advanced clinical/preclinical trials. Structures were extracted according to the assigned activity class, where the class indicates a common target-specific group such as GPCRs, kinases and proteases, nuclear receptors and ion channels as well as more than 150 subclasses (specific GPCRs, kinase and protease enzymes, etc.). Prior to the statistical experiments, the molecular structures were filtered and normalized in order to fulfill certain criteria (by analogy to Reference 21). The final database used in the modeling experiments included ca. 16,500 structures of drug compounds. The overall objective was to investigate differences between various groups of GPCR-specific ligands based on their physicochemical properties.

6.4.1.2 Software In the described experiments, we used SmartMining (http://www.chemdiv.com/) and InformaGenesis (http://www.informagenesis.com/) programs, which are originally developed and scientifically validated as specific computational tools for chemical database management, descriptor calculation, and data mining. Thus, InformaGenesis is a powerful software with an integrated module for Kohonen SOM generation and a wide number of advanced modifications and complex-specific modalities. The program has been designed to work under Windows operating system. In addition to basic Kohonen settings and learning parameters, InformaGenesis includes significant algorithmic-based improvements, such as "Neural Gas," "Duane Desieno," "Noise Technique," "Two Learning Stages and 3D Architecture" as well as several unique algorithms and specific methods, for instance, "Corners," "Gradient," and "Automatic Descriptor Selection Algorithm" (ADSA). InformaGenesis is completely adapted for the analysis of large sets of data of different types and dimensionality. For descriptor calculation, we used SmartMining program (http://www.chemdiv.com/). It calculates more than a hundred fundamental molecular descriptors that are generally divided into several logical and functional categories, including the basic specific physicochemical features such as log *P*, number of H-bond donors, H-bond acceptors, and rotatable bonds; topological and electrotopological descrip-

tors, such as Zagreb, Wiener, and E-state indexes; as well as *quasi*-3-D descriptors, such as van der Waals volume and surface. All descriptors are directly calculated by using well-known common models and approximations borrowed from the scientific literature [66]. In addition, several algorithms have been progressively modified to obtain more exact feature prediction or/and calculation; for example, van der Waals parameters are calculated fairly accurately considering the overlapped volumes and/or surfaces. The theoretical aspects of Kohonen SOMs are described in several comprehensive papers including Chapter 16 of this book.

6.4.1.3 Generation of the Map Then more than 100 different molecular descriptors including physicochemical properties (for example, log *P*, $VDW_{vol/surf}$ and MW), topological descriptors, such as Zagreb, Winner, and E-state indexes, as well as various structural descriptors, such as number of H-bond acceptors/donors and rotatable bond number (RBN), were calculated. For reduction of the number of input variables, we used a special algorithm, automatic descriptor selection (ADS), implemented in InformaGenesis software. The ADS method is based generally on preorganization of Kohonen neurons and assigned weight coefficients according to several common principles. Conceptually, the method resembles a sensitivity analysis widely used in computational modeling. Gradually adding the next descriptor, it painstakingly attempts to find the optimal positions of input objects with a maximum degree of dissimilarity between each other following the corresponding metric distances. Starting from any corner of the Kohonen map, each subsequent vector of descriptor values passing straight through the map walks step by step across the perimeter until the best separation among the input objects is achieved. During each cycle, it can also be amplified by a minor learning procedure to estimate the total sensitivity of a temporary fixed vector net. As a rule, the selection procedure is continued until the predefined number of descriptors is achieved. As a result of the ADS procedure, the final descriptor set included seven molecular descriptors: Zagreb index; E-state indexes for structural fragments >C–, $–CH_2–$, and $–CH_3$; the number of H-bond donors; HB2 (a structural descriptor that encodes the strength of H-bond acceptors following an empirical rule); log *P*.

After all the preparatory procedures were complete, the reference database with selected molecular descriptors was used for the development of a SOM-based *in silico* model. The whole self-organizing Kohonen map of ca. 16,500 pharamaceutical leads and drugs (not shown) demonstrates that the studied compounds occupy a wide area on the map, which can be characterized as the area of drug-likeness.

Distribution of various target-specific groups of ligands on the Kohonen map demonstrates that most of these groups have distinct locations in specific regions of the map. As an illustration, Figure 6.5 shows the population of four large target-specific groups, which occupy distinct regions of the map. A possible explanation of differences in their location is in the fact that

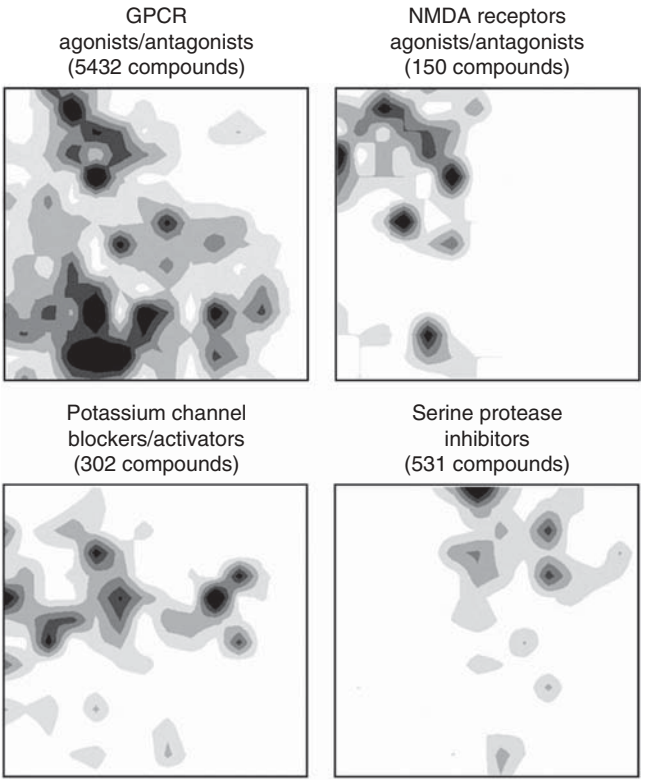


Figure 6.5 Distribution of four large target-specific groups of ligands on the Kohonen map. NMDA = N-methyl-D-aspartate.

closely related biotargets often share a structurally conserved ligand binding site. The structure of this site determines molecular properties that a receptor-selective ligand should possess to effectively bind to the site. These properties include specific spatial, lipophilic, and H-binding parameters, as well as other features influencing the pharmacodynamic characteristics. Therefore, every group of active ligand molecules can be characterized by a specific combination of physicochemical parameters statistically differentiating it from other target-specific groups of ligands. This observation is consistent with the basic principle of chemogenomics originally formulated by Klabunde [67]: "similar receptors bind similar ligands." Another possible explanation of the observed phenomenon is different pharmacokinetic requirements to drugs acting on different biotargets. For example, organ- and tissue-specific distribution of biotargets can influence physicochemical properties of their ligands.

The described algorithm represents an effective and relatively simple computational procedure for the selection of target-biased compound subsets

compatible with high-throughput *in silico* evaluation of large virtual chemical libraries. Whenever a large enough training set of active ligands is available for a particular receptor, the Kohonen map can be generated to identify specific sites of location of target activity groups of interest. Then the developed map can be used for testing any available chemical databases with the same calculated descriptors. The Kohonen mapping procedure is computationally inexpensive and permits real-time calculations with moderate hardware requirements. Thus, for a training database consisting of 16,500 molecules with seven descriptors using standard settings, approximately 1 hour is required for a standard PC (Pentium 3-GHz processor) on a Windows XP platform to train the network; the time increases almost linearly with the size of the database.

Our own experience and literature data demonstrate that Kohonen SOMs are an efficient data mining and visualization tool very useful in the design of chemical libraries, including the design of focused compound sets in the context of the chemogenomics approach. The approach, however, has some limitations. First of all, SOM algorithm is designed to preserve the topology between the input and grid spaces; in other words, two closely related input objects will be projected on the same or on close nodes. At the same time, the SOM algorithm does not preserve distances: there is no relation between the distance between two points in the input space and the distance between the corresponding nodes. The latter fact sometimes makes the training procedure unstable, when the minor changes in the input parameters lead to serious perturbation in the output picture. As a result, it is often difficult to find the optimal training conditions for better classification. Another potential problem is associated with the quantization of the output space. As a result, the resolution of low-sized maps can be insufficient for effective visualization of subtle differences between the studied compound categories.

6.4.2 GPCR Target Classes

Using the constructed map, it is possible to explore the area of GPCR ligands. Thus, compounds acting specifically on different GPCR subclasses including α/β-adrenoceptors, dopamine D1–D4 receptors, tachykinin NK1/NK2, and serotonin and chemokine receptors can also be successfully separated within the same map. For illustration, Figure 6.6 shows the distribution of four GPCR ligand subclasses, which are located separately in different areas within the Kohonen map with insignificant overlap.

Since a key objective of our research is to analyze the chemokine receptor superfamily, generally, in the context of the chemogenomics approach adopted specifically for compound library design, we have also studied the distribution of compounds within the Kohonen map active against different chemokine subclasses.

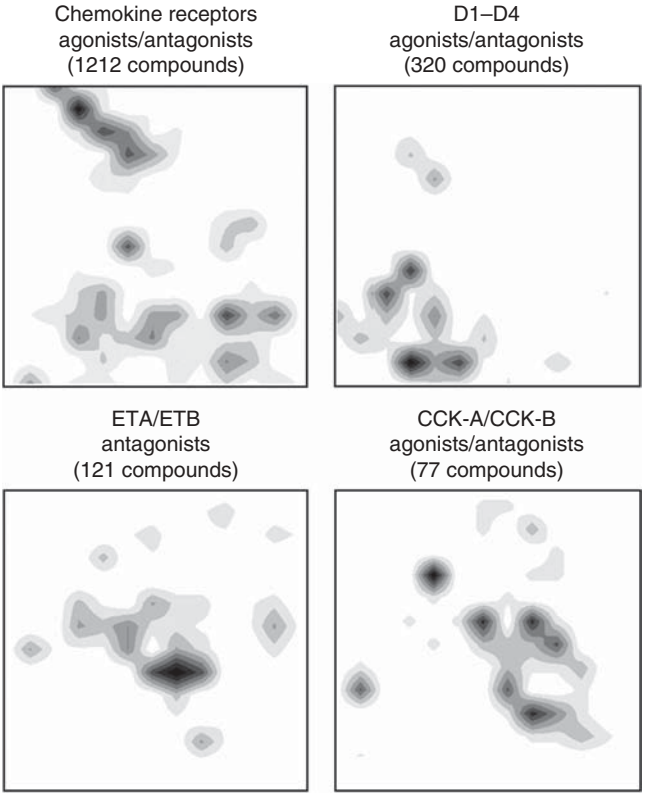


Figure 6.6 Distribution of four large GPCR-specific groups of ligands on the Kohonen map. ETA/ETB = Endotheline receptors A and B; CCK-A/CCK-B = cholecystokinin receptors A and B.

6.4.3 Similarity across the Chemokine Receptor Superfamily

Following the fundamental principle of chemogenomics, receptors are no longer viewed as single entities but are grouped into sets of related proteins or receptor families that are explored in a systematic manner. This interdisciplinary approach aimed primarily to find the links between the chemical structures of bioactive molecules and the receptors with which these molecules interact.

According to basic principles of chemogenomics, for a drug target of interest, known drugs and ligands of similar receptors, as well as compounds similar to these ligands, can serve as a convenient starting point for drug discovery. The obvious question here is "How can a receptor similarity be defined?" A review by Rognan [55] provides a comprehensive classification and overview of chemogenomics approaches, defines principles of receptor and/or ligand similarity, and presents case studies on how this knowledge has been applied

to rational drug design. In particular, Rognan formulates the following levels of receptor similarity:

- receptor class (e.g., GPCRs),
- receptor subclass (e.g. chemokine receptors),
- overall sequence homology (phylogenetic tree), and
- similarity of active binding site (3-D structure or one-dimensional [1-D] sequence motifs).

It is important to note that the reported chemogenomics approaches usually apply the classification of target families (such as ion channels, kinases, proteases, nuclear receptors, and GPCRs) or protein subfamilies (such as tyrosine kinases, chemokine receptors, and serine proteases) without taking into account similarities of the determined or assumed ligand binding sites. However, there are strong evidences that only a complex analysis of receptors, which includes formal receptor classification, sequence homology, 3-D similarity, and active binding site construction, provides a relevant and adequate strategy toward modern chemogenomics concepts.

Chemokines (chemotactic/chemoattractant cytokines) are highly basic, small, secreted proteins consisting on average 70–125 amino acids with molecular masses ranging from 6 to 14 kDa, which mediate their effects through binding to seven transmembrane domains (7TMs) of the specific family of GPCRs located on target cell membranes. The chemokine superfamily includes a large number of ligands that bind to a smaller number of receptors [68,69]. It is a well-known fact that multiple chemokine ligands can bind to the same receptor and vice versa, and such a complexity and promiscuity of receptor binding introduce an additional challenge in understanding the common mechanism of chemokine ligand binding. At the same time, with respect to chemogenomics, this feature of chemokine ligand–receptor recognition provides a valuable starting point to investigate key interrelationships across the chemokine receptor subfamily.

Since chemokine receptors are members of the common GPCR family, the two first similarity criteria formulated by Rognan are being fulfilled successfully. Currently, there are more than 20 functionally signaling chemokine receptors and more than 45 corresponding chemokine ligands in humans [70]. The chemokine ligands and receptors have been divided into several major groups based on their expression patterns and functions. In addition, their genomic organization also provides an alternative chemokine classification based on their phylogenetic trees.

The chemokine receptor CXCR4 possesses multiple fundamental functions in both normal and pathologic physiology. CXCR4 is a GPCR receptor that transduces signals of its endogenous ligand, the chemokine CXCL12 (stromal cell-derived factor-1 [SDF-1], previously SDF1-α). The interaction between CXCL12 and CXCR4 plays a critical role in the migration of progenitors

during embryologic development of the cardiovascular system, the hematopoietic system, the CNS, and so on. This interaction is also known to be involved in several intractable disease processes, including HIV infection, cancer cell metastasis, leukemia cell progression, rheumatoid arthritis (RA), asthma, and pulmonary fibrosis. Unlike other chemokine receptors, CXCR4 is expressed in many normal tissues, including those of the CNS, while it is also commonly expressed by over 25 different tumor cells including cancers of epithelial, mesenchymal, and hematopoietic origins. [71]. Since CXCR4 is the most actively studied chemokine receptor, and a number of small-molecule compounds are currently known to modulate its basic functions, it is quite reasonable to investigate similarity links between this and other chemokine receptors based on Rognan's criteria. According to phylogenetic dendograms [72], CXCR4 is located closely to CXCR1, CXCR2, as well as CXCR3. This means primarily that these receptors possess a similar genotype and, based on this observation, they can be logically grouped into the common CXCR family differed genetically from the CCR subclass but not significantly. It seems to be perfectly reasonable to investigate small-molecule space around the whole CXCR subclass; however, the last similarity criterion is still not considered. A binding site composition and a corresponding space cavity jointly play a key role in the ligand binding process. Furthermore, the majority of ligand–receptor complexes are not static structures; they can change dynamically upon ligand binding. In addition, enforced conformational changes across the active binding site can also be achieved by ligand partial binding followed by internal cavity formation fitted appropriately for deep embedding. There are several scientific reports highlighting the partial sequence homology (25–30%) and high-binding site similarity between CCR5 and CXCR4 [73]. For instance, using MembStruk methods to develop 3-D protein structures for CXCR4 and CCR5 and the HierDock protocol to define the binding site for both of these receptors, it was clearly shown that the two binding sites, even though being on different sides of their receptors, have similar characteristics [74]. In both cases, CCR5 and CXCR4 MembStruk structures are also used to correctly identify the binding site regions according to mutational studies. In addition, a high degree of similarity was also determined for CCR5 and CCR3 [75]. Therefore, from the chemogenomics point of view, it is of practical relevance to test the agents acting against CXCR4 also on activity toward CXCR1-3 and CCR3/5. Thus, compounds are profiled against a set of receptors and are not tested against single targets.

Figure 6.7 shows a fragment of the phylogenetic tree for chemokine receptors and sites of location of CXCR4 and CCR5 ligands on the Kohonen map described above. The figure demonstrates that ligands that bind to CCR5 are located closely to CXCR4 on the Kohonen map with significant overlapping. Therefore, such a map can be used for chemogenomics-based discovery of either CXCR4 or CCR5 ligands.

Combining the results of our computational modeling and theoretical analysis, it can reasonably be concluded that the applied mapping technique rep-

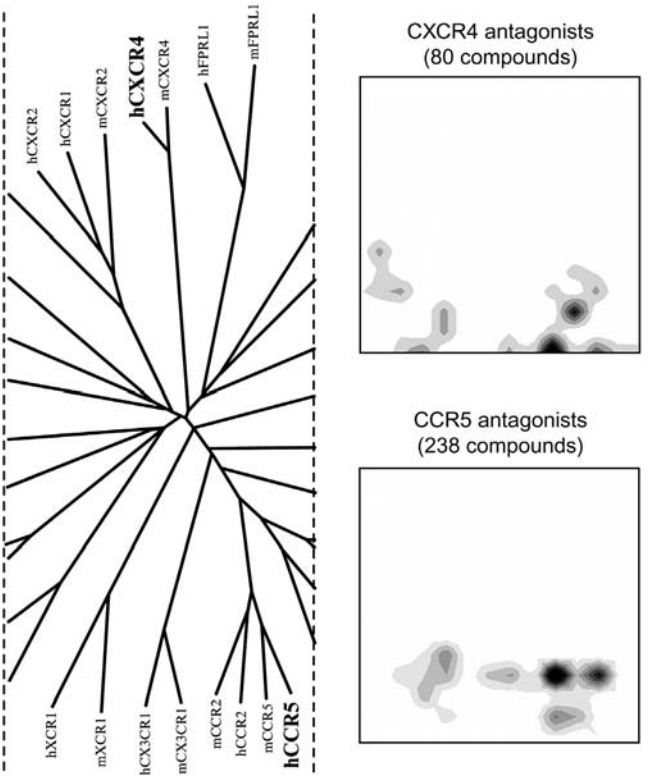


Figure 6.7 Homology-based design of chemokine receptor ligands.

resents a useful approach to filtering combinatorial libraries for selection of target-specific subsets including those acting against the chemokine receptor superfamily. It permits to significantly reduce the size of initial multidimensional chemistry space and can be recommended as a classification and visualization tool for practical combinatorial design. It is important that this method is complementary to other target and ligand structure-based approaches to virtual screening. In addition, Kohonen-based SOMs are fully compatible with both the high-throughput virtual screening protocols and the analysis of small- to medium-sized combinatorial libraries.

6.5 CONCLUSION

At present, drug discovery technologies are undergoing radical changes due to both amazing progress in genomic research as well as the massive advent of combinatorial synthesis and high-throughput biological screening. Moreover, it can be argued that, during the past decade, the main paradigm

in medicinal chemistry has been turning gradually from traditional receptor-specific studies and biological assays to a novel cross-receptor vision. Currently, such approach becomes increasingly applied within the whole pharmaceutical research to enhance the efficiency of modern drug discovery. Chemogenomics, as an alternative route to innovative drug discovery, provides novel insights into receptor–ligand interaction and molecular recognition by the analysis of large biological activity data sets. Rational drug design strategies that are based primarily on the chemogenomics approach often complement HTS for finding chemical starting points for novel drug discovery projects.

The greatest impact of the chemogenomics approaches can be expected for targets with sparse or without ligand information as well as for targets lacking structural 3-D data. For these targets, classical drug design strategies like ligand- and structure-based virtual screening and/or de novo design cannot be applied. Key methodologies underlying the chemogenomics approaches are annotated knowledge databases and specific data mining tools.

In the described experiment illustrating one possible approach to the chemogenomics-based design of chemokine receptor-targeted libraries, we have applied the algorithm of self-organizing Kohonen maps for the analysis of clinically validated therapeutic agents and approved drug compounds. The developed models can be used for the selection of screening candidates from chemical databases. The applied virtual screening technology is focused on a small molecular level, as opposed to target structure-based design or docking methodology. A leitmotif of this method is a ligand-based strategy realized in the context of a chemogenomics concept. The described method represents a consistent and valuable approach toward both the rational drug design and gathering information from the simultaneous biological evaluation of many compounds on multiple biological targets.

REFERENCES

1. Wise A, Gearing K, Rees S. Target validation of G-protein coupled receptors. *Drug Discov Today* 2002;7:235–246.
2. Lagerström MC, Schiöth HB. Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov* 2008;7:339–357.
3. Oprea TI. Chemical space navigation in lead discovery. *Curr Opin Chem Biol* 2002;6:384–389.
4. Proudfoot JR, John R. Drugs, leads, and drug-likeness: An analysis of some recently launched drugs. *Bioorg Med Chem Lett* 2002;12:1647–1650.
5. Evers A, Klabunde T. Structure-based drug discovery using GPCR homology modeling: Successful virtual screening for antagonists of the alpha1A adrenergic receptor. *J Med Chem* 2005;48:1088–1097.

6. Cavasotto CN, Orry AJW, Abagyan RA. Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein coupled receptors. *Proteins* 2003;51:423–433.
7. Kohonen T. *Self-Organizing Maps*, 3rd edn. New York: Springer Verlag, 2000.
8. Sammon, JE. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, 1969;C-18, 401-409.
9. Savchuk NP, Tkachenko SE, Balakin KV. *Cheminformatics in Drug Discovery*, pp. 287–313. Weinheim: Wiley-VCH, 2004.
10. Schneider G, Nettekoven M. Ligand-based combinatorial design of selective purinergic receptor (A2A) antagonists using self-organizing maps. *J Comb Chem* 2003;5:233–237.
11. Agrafiotis DK, Myslik JC, Salemme FR. Advances in diversity profiling and combinatorial series design. *Mol Divers* 1999;4:1–22.
12. Kier LB, Hall LH. *Molecular Structure Description: The Electrotopological State*. San Diego, CA: Academic Press, 1999.
13. Vogt I, Bajorath J. Design and exploration of target-selective chemical space representations. *J Chem Inf Model* 2008;48:1389–1395.
14. Rusinko A III, Farmen MW, Lambert CG, Brown PL, Young SS. Analysis of a large structure/biological activity data set using recursive partitioning. *J Chem Inf Comput Sci* 1999;39:1017–1026.
15. Chen X, Rusinko A, Young SS. Recursive partitioning analysis of a large structure-activity data set using three-dimensional descriptors. *J Chem Inf Comput Sci* 1998;38:1054–1062.
16. Horvath D. Recursive partitioning analysis of μ-opiate receptor high throughput screening results. *SAR QSAR Environ Res* 2001;12:181-212.
17. Weber L, Wallbaum S, Broger C, Gubernator K. Optimization of the biological activity of combinatorial compound libraries by a genetic algorithm. *Angew Chem Int Ed Engl* 1995;34:2280–2282.
18. Jones-Hertzog DK, Mukhopadhyay P, Keefer CE, Young SS. Use of recursive partitioning in the sequential screening of G-protein-coupled receptors. *J Pharmacol Toxicol Methods* 1999;42:207–215.
19. Lavrador K, Murphy B, Saunders J, Struthers S, Wang X, Williams J. A screening library for peptide activated G-protein coupled receptors. 1. The test set. *J Med Chem* 2004;47:6864–6874.
20. Balakin KV, Tkachenko SE, Lang SA, Okun I, Ivashchenko AA, Savchuk NP. Property-based design of GPCR-targeted library. *J Chem Inf Comput Sci* 2002;42: 1332–1342.
21. Balakin KV, Lang SA, Skorenko AV, Tkachenko SE, Ivashchenko AA, Savchuk NP. Structure-based versus property-based approaches in the design of G-protein-coupled receptor-targeted libraries. *J Chem Inf Comput Sci* 2003;43: 1553–1562.
22. Manallack DT, Pitt WR, Gancia E, Montana JG, Livingstone DJ, Ford MG, Whitley DC. Selecting screening candidates for kinase and G-protein coupled receptor targets using neural networks. *J Chem Inf Comp Sci* 2002;42:1256–1262.
23. Niwa T. Prediction of biological targets using probabilistic neural networks and atom-type descriptors. *J Med Chem* 2004;47:2645–2650.

24. Agrafiotis DK. Multiobjective optimization of combinatorial libraries. *Mol Divers* 2002;5, 209–230.
25. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ Jr., Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL, Buolamwini JK, van Osdol WW, Monks AP, Scudiero DA, Sausville EA, Zaharevitz DW, Bunow B, Viswanadhan VN, Johnson GS, Wittes RE, Paull KD. An information-intensive approach to the molecular pharmacology of cancer. *Science* 1997;275:343–349.
26. Mestres J. Computational chemogenomics approaches to systematic knowledge-based drug discovery. *Curr Opin Drug Discov Devel* 2004;7:304–313.
27. Bredel M, Jacoby E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat Rev Genet* 2004;5:262–275.
28. Savchuk NP, Balakin KV, Tkachenko SE. Exploring the chemogenomic knowledge space with annotated chemical libraries. *Curr Opin Chem Biol* 2004;8:412–417.
29. Balakin KV, Tkachenko SE, Kiselyov AS, Savchuk NP. Focused chemistry from annotated libraries. *Drug Discov Today* 2006;3:397–403.
30. Blower PE, Yang C, Fligner MA, Verducci JS, Yu L, Richman S, Weinstein JN. Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data. *Pharmacogenomics J* 2002;2:259–271.
31. Rabow AA, Shoemaker RH, Sausville EA, Covell DG. Mining the National Cancer Institute's tumor-screening database: Identification of compounds with similar cellular activities. *J Med Chem* 2002;45:818–840.
32. Wallqvist A, Rabow AA, Shoemaker RH, Sausville EA, Covell DG. Establishing connections between microarray expression data and chemotherapeutic cancer pharmacology. *Mol Cancer Ther* 2002;1:311–320.
33. Rickardson L, Fryknäs M, Haglund C, Lövborg H, Nygren P, Gustafsson MG, Isaksson A, Larsson R. Screening of an annotated compound library for drug activity in a resistant myeloma cell line. *Cancer Chemother Pharmacol* 2006;58: 749–758.
34. Root DE, Flaherty SP, Kelley BP, Stockwell BR. Biological mechanism profiling using an annotated compound library. *Chem Biol* 2003;10:881–892.
35. Klekota J, Brauner E, Roth FP, Schreiber SL. Using high-throughput screening data to discriminate compounds with single-target effects from those with side effects. *J Chem Inf Model* 2006;46:1549–1562.
36. Greenbaum DC, Arnold WD, Lu F, Hayrapetian L, Baruch A, Krumrine J, Toba S, Chehade K, Brömme D, Kuntz ID, Bogyo M. Small molecule affinity fingerprinting. A tool for enzyme family subclassification, target identification, and inhibitor design. *Chem Biol* 2002;9:1085–1094.
37. Vieth M, Higgs RE, Robertson DH, Shapiro M, Gragg EA, Hemmerle H. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim Biophys Acta* 2004;1697:243–257.
38. Fabian MA, Biggs WH III, Treiber DK, Atteridge CE, Azimioara MD, Benedetti MG, Carter TA, Ciceri P, Edeen PT, Floyd M, Ford JM, Galvin M, Gerlach JL, Grotzfeld RM, Herrgard S, Insko DE, Insko MA, Lai AG, Lélias JM, Mehta SA, Milanov ZV, Velasco AM, Wodicka LM, Patel HK, Zarrinkar PP, Lockhart DJ. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat Biotechnol* 2005;23:329–236.

39. Cases M, García-Serna R, Hettne K, Weeber M, van der Lei J, Boyer S, Mestres J. Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family. *Curr Top Med Chem* 2005;5:763–772.

40. Janzen WP, Hodge CN. A chemogenomic approach to discovering target-selective drugs. *Chem Biol Drug Des* 2006;76:85–86.

41. Balakin KV, Ivanenkov YA, Savchuk NP, Ivashchenko AA, Ekins S. Comprehensive computational assessment of ADME properties using mapping techniques. *Curr Drug Discov Technol* 2005;2:99–113.

42. Krejsa CM, Horvath D, Rogalski SL, Penzotti JE, Mao B, Barbosa F, Migeon JC. Predicting ADME properties and side effects: The BioPrint approach. *Curr Opin Drug Discov Devel* 2003;6:470–480.

43. Ganter B, Tugendreich S, Pearson CI, Ayanoglu E, Baumhueter S, Bostian KA, Brady L, Browne LJ, Calvin JT, Day GJ, Breckenridge N, Dunlea S, Eynon BP, Furness LM, Ferng J, Fielden MR, Fujimoto SY, Gong L, Hu C, Idury R, Judo MS, Kolaja KL, Lee MD, McSorley C, Minor JM, Nair RV, Natsoulis G, Nguyen P, Nicholson SM, Pham H, Roter AH, Sun D, Tan S, Thode S, Tolley AM, Vladimirova A, Yang J, Zhou Z, Jarnagin K. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J Biotechnol* 2005;119:219–244.

44. Mao B, Gozalbes R, Barbosa F, Migeon J, Merrick S, Kamm K, Wong E, Costales C, Shi W, Wu C, Froloff N. QSAR modeling of in vitro inhibition of cytochrome P450 3A4. *J Chem Inf Model* 2006;46:2125–2134.

45. Horvath D, Jeandenans C. Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces—A novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J Chem Inf Comput Sci* 2003;43:680–690.

46. Rolland C, Gozalbes R, Nicolaï E, Paugam MF, Coussy L, Barbosa F, Horvath D, Revah F. G-protein-coupled receptor affinity prediction based on the use of a profiling dataset: QSAR design, synthesis, and experimental validation. *J Med Chem* 2005;48:6563–6574.

47. Engelberg A. Iconix Pharmaceuticals, Inc.—Removing barriers to efficient drug discovery through chemogenomics. *Pharmacogenomics* 2004;5:741–744.

48. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M, Oprea TI. *Chemoinformatics in Drug Discovery*, pp. 223–239. New York: Wiley-VCH, 2004.

49. Nidhi, Glick M, Davies JW, Jenkins JL. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* 2006;46:1124–1133.

50. Nigsch F, Bender A, Jenkins JL, Mitchell JB. Ligand-target prediction using winnow and naive Bayesian algorithms and the implications of overall performance statistics. *J Chem Inf Model* 2008;48:P2313–P2325.

51. Dubus E, Ijjaali I, Petitet F, Michel A. In silico classification of hERG channel blockers: A knowledge-based strategy. *Chem Med Chem* 2006;1:622–630.

52. Ijjaali I, Petitet F, Dubus E, Barberan O, Michel A. Assessing potency of c-Jun N-terminal kinase 3 (JNK3) inhibitors using 2D molecular descriptors and binary QSAR methodology. *Bioorg Med Chem* 2007;15:4256–4264.

53. Ijjaali I, Barrere C, Nargeot J, Petitet F, Bourinet E. Ligand-based virtual screening to identify new T-type calcium channel blockers. *Channels (Austin)* 2007;1: 300–304.
54. Schürer SC, Tyagi P, Muskal SM. Prospective exploration of synthetically feasible, medicinal relevant chemical space. *J Chem Inf Model* 2005;45:239–248.
55. Rognan D. Chemogenomic approaches to rational drug design. *Br J Pharmacol* 2007;152:38–52.
56. Frye SV. Structure-activity relationship homology (SARAH): A conceptual framework for drug discovery in the genomic era. *Chem Biol* 1999;6:R3–R7.
57. Schuffenhauer A, Zimmermann J, Stoop R, van der Vyver JJ, Lecchini S, Jacoby E. An ontology for pharmaceutical ligands and its application for *in silic*o screening and library design. *J Chem Inf Comput Sci* 2002;42:947–955.
58. Schuffenhauer A, Floersheim P, Acklin P, Jacoby E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J Chem Inf Comp Sci* 2003;43: 391–405.
59. Crossley R. The design of screening libraries targeted at G-protein coupled receptors. *Curr Top Med Chem* 2004;4:581–588.
60. Jacoby E, Fauchere JL, Raimbaud E, Ollivier E, Michel A, Spedding M. A three binding site hypothesis for the interaction of ligands with monoamine G protein-coupled receptors: Implications for combinatorial ligand design. *Quant Struct Activ Relat* 1999;18:561–572.
61. Nordling E, Homan E. Generalization of a targeted library design protocol: Application to 5-HT7 receptor ligands. *J Chem Inf Comput Sci* 2004;44:2207–2215.
62. Bock JR, Gough DA. Virtual screen for ligands of orphan G protein-coupled receptors. *J Chem Inf Model* 2005;45:1402–1414.
63. Lapinsh M, Prusis P, Uhlén S, Wikberg JES. Improved approach for proteochemometrics modeling: application to organic compound-amine G protein-coupled receptor interactions. *Bioinformatics* 2005;21:4289–4296.
64. Jacob L, Hoffmann B, Stoven V, Vert JP. Virtual screening of GPCRs: an in silico chemogenomics approach. *BMC Bioinformatics* 2008;9:363.
65. Okuno Y, Yang J, Taneishi K, Yabuuchi H, Tsujimoto G. GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res* 2006;34:D673–D677.
66. Todeschini R, Consonni V, Mannhold R, Kubinyi H, Timmerman H. *Handbook of Molecular Descriptors*. New York: Wiley, 2000.
67. Klabunde T. Chemogenomic approaches to ligand design. In: *Ligand Design for G-Protein-Coupled Receptors*, edited by D. Rognan, pp. 115–135. Weinheim: Wiley-VCH, 2006.
68. Zlotnik A, Yoshie O. Chemokines: A new classification system and their role in immunity. *Immunity* 2000;12:121–127.
69. Yoshie O, Imai T, Nomiyama H. Chemokines in immunity. *Adv Immunol* 2001;78: 57–110.
70. Balakin KV, Ivanenkov YA, Tkachenko SE, Kiselyov AS, Ivachtchenko AV. Regulators of chemokine receptor activity as promising anticancer therapeutics. *Curr Cancer Drug Targets* 2008;8:299–340.

71. Allavena P, Marchesi F, Mantovani A. The role of chemokines and their receptors in tumor progression and invasion: potential new targets of biological therapy. *Curr Cancer Ther Rev* 2005;1:81–92.

72. Zlotnik A, Yoshie O, Nomiyama H. The chemokine and chemokine receptor superfamilies and their molecular evolution. *Genome Biol* 2006;7:243.

73. Pérez-Nueno VI, Ritchie DW, Rabal O, Pascual R, Borrell JI, Teixidó J. Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand-receptor docking. *J Chem Inf Model* 2008;48:509–533.

74. Spencer EH. *Development of a structure prediction method for G protein-coupled receptors*. Thesis. California Institute of Technology, 2005.

75. Efremov R, Truong MJ, Darcissac EC, Zeng J, Grau O, Vergoten G, Debard C, Capron A, Bahr GM. Human chemokine receptors CCR5, CCR3 and CCR2B share common polarity motif in the first extracellular loop with other human G-protein coupled receptors. *Eur J Biochem* 2001;263:746–756.

7

MINING HIGH-THROUGHPUT SCREENING DATA BY NOVEL KNOWLEDGE-BASED OPTIMIZATION ANALYSIS

S. Frank Yan, Frederick J. King, Sumit K. Chanda, Jeremy S. Caldwell, Elizabeth A. Winzeler, and Yingyao Zhou

Table of Contents

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

7.1 INTRODUCTION

Small-molecule drug discovery has evolved dramatically over the last two decades. First, most of the therapeutics that have been commercialized for the past 50 years were developed for approximately 500 known biological targets. It is now estimated, however, that the number of potential drug targets is at least an order of magnitude larger; this implies that a plethora of currently undiscovered therapeutic opportunities exists [1]. In order to identify novel drug targets and to accelerate the development of innovative treatments, a myriad of technologies have been developed or revolutionized. These include various high-throughput (HT) and "-omics" technologies, which provide an opportunity to systematically address the function of each member of the genome, transcriptome, or proteome. Second, advancements in automation equipment, assay formats, and combinatorial chemical libraries have transformed the small-molecule compound screening process. A common feature for all of these techniques is the generation of large data sets; it is not uncommon to generate 1,000,000 plus data points from a single high-throughput screening (HTS) campaign. For example, from the public domain, the National Institutes of Health (NIH) Roadmap initiatives have made available the data from over 1,700 assays via the National Center for Biotechnology Information (NCBI) PubChem database. The data analysis approaches described in this chapter primarily are based on the HTS database extant at the Genomics Institute of the Novartis Research Foundation (GNF, San Diego, CA, USA), where over 1.6 million compounds have been evaluated in more than 200 HTS campaigns to date.

A representation of a typical workflow for an early-stage drug discovery program illustrates how informatics techniques can impact multiple aspects of the process. As shown in Figure 7.1, the capacity in terms of the number of compounds being studied decreases rapidly as the workflow progresses; i.e., only a very limited number of compounds can be followed up at each following stage. However, the upsurge in primary data points that are now generated by a typical HTS has led to a concomitant increase in the number of candidate lead compounds for postscreening follow-up studies. To alleviate the potential bottleneck resulting from the increase in HTS capabilities, two general approaches have been undertaken: (1) accelerating the throughput of the downstream processes, e.g., *in vitro* pharmacokinetic (PK) assays, and (2) applying informatics technologies in order to eliminate compounds that would likely be triaged at a later stage. If successful compound prioritization occurs at an early lead discovery phase, it has a far-reaching impact toward the overall success of the project.

Informatics approaches that can be employed readily and reliably in order to effectively analyze extensive and complex data sets are crucial components of the current drug discovery process. For example, it has been found that a significant portion of confirmed active HTS hits is triaged at a later stage due to undesirable selectivity and/or toxicity profiles. Those compounds are ini-

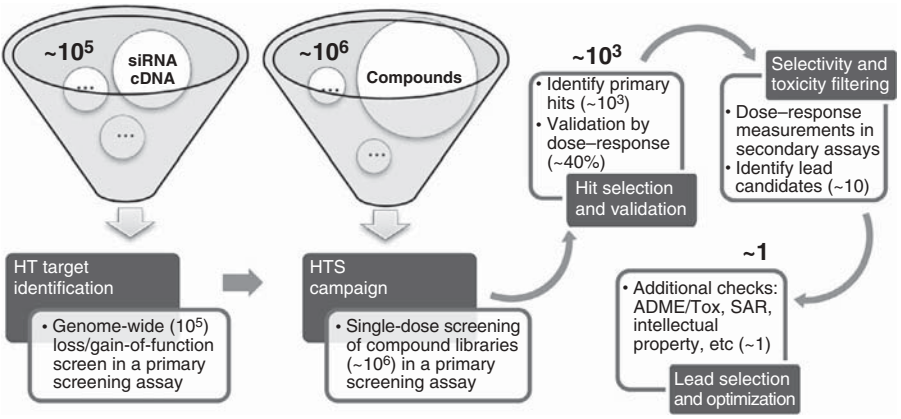


Figure 7.1 A typical high-throughput (HT) early lead discovery workflow. Genome-wide function screens identify potential novel drug targets for HTS campaign. A full library (1,000,000 plus) HTS typically produces primary hits in the order of several thousands. In many cases, an aliquot of these compounds are plated, serially diluted, and then rescreened using the original primary assay for confirmation of their activity along with a determination of their potency and efficacy. Once confirmed, the active hits are subsequently studied in a battery of secondary assays in order to assess their selectivity and toxicity profiles; this process normally leads to a small number of "lead candidates." These molecules are then prioritized by several criteria, including their *in vitro* pharmacological profile such as absorption, distribution, metabolism, and excretion (ADME), general cytotoxicity (TOX), scaffold structure–activity relationship (SAR) strength, and potential intellectual property (IP) space. The desired outcome is the identification of several lead scaffolds/templates suitable for further medicinal chemistry optimization.

tially determined as true positives since their unfavorable characteristics cannot be recognized from the analysis of individual data sets. With the rapidly growing HTS databases, it is possible to identify those frequent hitters by data mining existing database instead of carrying out laborious experimental profiling. The knowledge obtained from numerous HTS campaigns provides an opportunistic resource that can impact the prioritization of lead compounds for any future HTS. However, only a very small number of studies have been published that systematically mine large HTS databases [2–4]. This is probably due to the fact that most existing chemoinformatics approaches were designed for analyzing individual HTS data sets rather than data matrices containing data from multiple assays, as well as the limited availability of large-scale HTS data sets that are usually proprietary for pharmaceutical companies.

Challenges listed above can be addressed by a novel algorithm dubbed knowledge-based optimization analysis (KOA). KOA was first invented for gene expression-based gene function prediction based on the guilt-by-association (GBA) principle [5]. The algorithm was initially named as ontology-based pattern identification (OPI) according to this first bioinfor-

matics application; however, we have decided to use the term KOA in this chapter as this algorithm can take advantage of virtually any knowledge base, gene ontology (GO) information being only one of them. The KOA algorithm identifies optimal data analysis protocols and yields reliable mining results through maximizing rediscovery of existing knowledge. The idea of KOA is indeed very general and therefore can be adapted to solve a wide range of informatics problems, particularly in relevance to drug discovery (Table 7.1). It is of no surprise that the original KOA algorithm was later modified to take

TABLE 7.1 Published KOA Applications

Domain	Problem	Data Set	Knowledge	Reference
Bioinformatics	Predict gene function based on mRNA coexpression patterns	A gene by sample matrix, one genotype vector per gene	Gene ontology (GO)	[5–7]
Chemoinformatics	Improve primary hit confirmation rate in an HTS campaign	An activity vector, one number per compound	Clusters of structurally similar compounds	[8]
Chemoinformatics	Identify compounds that are both chemically and biologically similar; identify promiscuous HTS hitters	A compound by assay matrix, one activity profile per compound	Clusters of structurally similar compounds	[2,9]
Chemoinformatics	Identify HTS fingerprints for a given mechanism of action (MOA)	A compound by assay matrix, one activity profile per compound	MOA database (e.g., Medical Subject Headings [MeSH] database)	[10]
Bioinformatics	Improve siRNA confirmation rate in a functional genomic screen	An activity vector, one number per siRNA	siRNA–gene many-to-one mapping	[11]
Bioinformatics	Discovery cis-regulatory motif elements	Promoter sequences	Clusters of coexpressed genes	[12]
Bioinformatics	Gene function prediction in high-content imaging screen	A gene by phenotype matrix, one phenotype vector per gene	GO	[13]

advantage of the structure–activity relationship (SAR) principle and was aptly applied to better mine HTS data sets in order to address some of the chemoinformatics challenges mentioned above. In particular, KOA was applied to individual HTS campaigns for primary hit selection and showed a significant improvement in hit confirmation rate, i.e., as high as 80% compared to the typical 40% using the cutoff method [8]. In addition, it was applied to mine large corporate HTS databases consisting of ~80 different HTS campaigns and to identify compound families that demonstrate strong neighborhood behavior in terms of their selectivity profiles [9]. By combining both applications and without requiring additional experimental resource, single-dose HTS data sets can now be mined to provide high-quality HTS hits that have a high chance of not only being confirmed but also of passing the later toxicity filter.

In this chapter, the KOA algorithm is explained in nonmathematical terms by using a single-assay small interfering RNA (siRNA) hit selection problem as an example. By explaining how one can identify more confirmable targets in a single loss-of-function siRNA screen, we outline the ideas behind KOA. The resultant KOA solution was recently blindly tested in several whole-genome siRNA screens; experimental results provided solid validations for the knowledge-based optimization strategy. We then discuss the expression-based gene function prediction problem and show how KOA mines large data matrices. Analogous problems in the area of compound HTS are then discussed in detail. In particular, we demonstrate how KOA can effectively mine HTS data sets for better hit prioritization and identification of promiscuous hits. Comparisons between KOA and other related algorithms found in the literature are also discussed.

7.2 KOA ALGORITHM—CONCEPT, VALIDATION, AND ITS APPLICATIONS IN TARGET IDENTIFICATION

7.2.1 KOA Analysis for HT siRNA Function Screening

The advent of RNA interference (RNAi) technology has provided scientists with important screening tools to associate the reduction of expression of a particular mRNA with a cellular phenotype. siRNAs are a class of 19–29 base pair double-stranded RNA molecules that are designed to reduce the mRNA levels of a specific gene target. Knockdown of a specific gene candidate can be associated with a particular cellular phenotype, which may provide insights into a biological mechanism that can lead to a therapeutic target. The completion of the sequencing of the human genome enabled scientists with the ability to design siRNAs against each mRNA species and to allow siRNA screens on a "genome-wide" scale. In theory, the number of individual siRNAs required to assess the phenotypic effects of mRNA depletion for each individual member of an entire genome is equivalent to the number of genes expressed for the particular organism. In practice, however, the number of

entities typically used for a genome-wide siRNA screening is much larger. This is because individual siRNAs typically have a range of specificity and efficacy against their intended target, which cannot be ascertained readily. In order to increase the probability that the library encompasses active siRNA for each gene and to reduce the likelihood that a screening hit is due to an off-target effect (OTE), two or more unique siRNAs targeted against the same gene are typically screened. One consequence of this approach is that the number of data points produced from each genome-wide siRNA screen can be extensive. Also, the decision to pursue further analysis of a gene target if only a single siRNA scored as a hit in the primary screen often is made arbitrarily.

Specifically, the data obtained from a large siRNA screen are traditionally ranked according to their individual activity scores, and the "top X" number of wells is "hit picked" for reconfirmation and validation studies. In most cases, the cutoff assigned to top X is set by logistical constraints (e.g., given the available resources, how many siRNAs can be analyzed) and not all siRNAs with notable activity are hit picked. Our experience is that the hit confirmation rates using the cutoff method are typically low (about <20%), which has been always considered as an important challenge in the interpretation of data sets obtained from several large-scale siRNA screens. However, the confirmation rate can be significantly improved by the KOA algorithm detailed in Scheme 7.1.

To explain how KOA works, in this example, we assume a hypothetical small collection of 40 independently designed siRNAs that target 14 genes (one to four siRNAs per gene). A range of siRNAs per gene was chosen since this reflects our personal experience; the number of siRNAs per target can fluctuate over time due to the merger of multiple siRNA libraries, change in gene structures, elimination of nonspecific siRNAs, availability of reagent, and so on. The 40 siRNAs are shown in Figure 7.2, ranked according to their activities so that the most potent wells are placed at the top. siRNAs are pattern-filled according to their intended target gene identities; i.e., each pattern represents a unique gene. For the purpose of discussion, we assume our follow-up capacity to be eight wells. Hits identified by the cutoff method and KOA method are highlighted in the "cutoff hits" and "KOA hits" columns, respectively. Cutoff is the most popular hit-picking method, which goes for the top eight most potent wells.

For a given gene, the accumulative hypergeometric p values (Scheme 7.1, Eq. 7.1) are calculated for each siRNA, and the curve dips at each siRNA targeting the gene itself (large filled circle in Fig. 7.2B). The global minimum (indicated by arrow) is then identified, which separates siRNAs into two groups: hits and outliers. KOA outliers are also marked as "X" in the "KOA outlier" column (Fig. 7.2A). For gene C in black, its three siRNAs correspond to hypergeometric p values of 0.08, 0.01, and 0.66, respectively. The minimum p value 0.01 is obtained when its first two siRNAs are considered as hits and the last as outlier (black circles in Fig. 7.2B). Assuming gene C is a true nega-

Data set: an activity vector for a total of N_T siRNAs, one activity number per siRNA
Knowledge: the design of the siRNA library, where S_{ij} is the jth siRNA designed to target the ith gene, $j = 1, \ldots, n_i$
Hypothesis: siRNAs of the same gene tend to be coactive or coinactive
Output: a list of siRNAs that are considered to be true hits
Algorithm:

1. Rank all siRNAs based on their activities in descending order (most potent on top).
2. For each gene i and
3. for each siRNA$_{ij}$ ($j = 1, \ldots, n_i$),
4. calculate enrichment factor $f_{ij} = p(N_T, n_i, R_{ij}, j)$;
5. $j^* = \arg\min_j f_{ij}$ and $f_i^* = f_{ij^*}$; assign f_i^* to all siRNA$_{ij}$ ($j = 1, \ldots, n_i$);
6. siRNA$_{ij}$ with $j \le j^*$ are marked as positives; siRNA$_{ij}$ with $j > j^*$ are removed as outliers.
7. Rank all positive siRNAs based on f_i^* in ascending order, then by R_{ij} in ascending order.

R_{ij} = the rank number of siRNA S_{ij} in the sorted list; p = accumulated hypergeometric distribution function (Eq. 7.1). Assume a box contains n white balls, $N - n$ black balls, and N in total. If one randomly takes R balls from the box, the probability of obtaining r or more white balls in the selection is

$$p(N, n, R, r) = \sum_{k=r}^{\min(R,n)} \frac{\binom{n}{k}\binom{N-n}{R-k}}{\binom{N}{R}}. \tag{7.1}$$

Scheme 7.1 Outline of the KOA siRNA hit selection algorithm.

tive and all its three siRNA readings are noise, the chance of finding at least two out of three siRNAs in the top three activity slots is only 1%. Therefore, gene C is unlikely to be a negative and should be ranked high. All three siRNAs are assigned a KOA score of 0.01 (steps 3–6 in Scheme 7.1). Similarly, for gene D, its four siRNAs correspond to p values of 0.2, 1.0, 1.0, and 0.9, respectively (gray circles in Fig. 7.2B). Only the first siRNA is considered as a hit and the remaining three are removed as outliers. The hypergeometric model estimates the chance of finding at least one out of four siRNAs in the top two activity slots by chance is 20%, which means gene D is likely to be a negative and should be ranked low. This type of analysis is repeated for all the 14 genes, and positive siRNAs are initially ranked by their gene p values (ascending) and then by individual activities (potent to weak); the best eight siRNAs are highlighted as KOA hits and mostly inactive siRNAs are automatically identified as KOA outliers (marked as "X" in Fig. 7.2A). Five out of the top eight hits between cutoff and KOA algorithms are in common (marked as

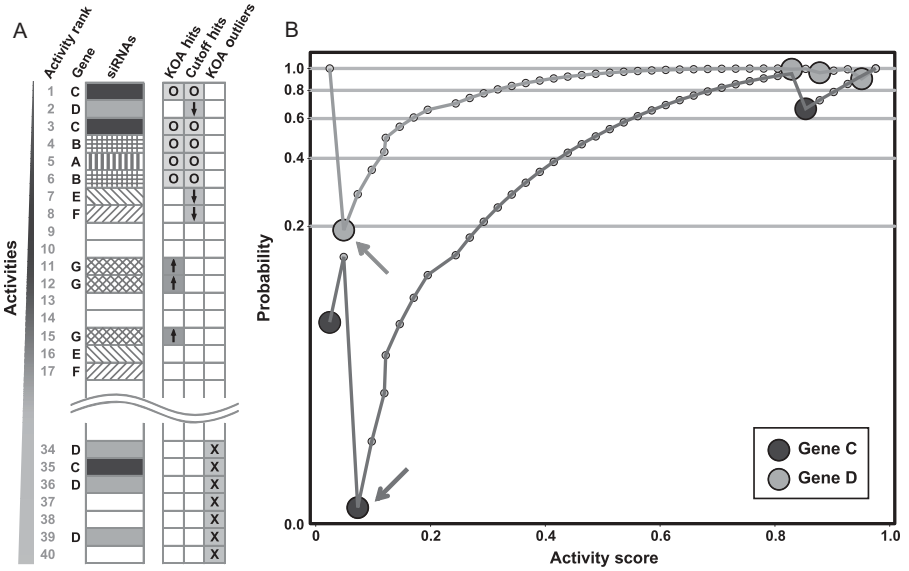


Figure 7.2 Illustration of KOA algorithm in a hypothetical siRNA screen. (A) Hits selected by either KOA/OPI and cutoff methods. (B) Probability scoring curves for genes C and D. The activity threshold for each gene is determined by global minima as indicated by an arrow. siRNAs with activities weaker than the corresponding thresholds (to the right) are considered as outliers and are eliminated.

"O" in Fig. 7.2A), although their hit ranks may differ. KOA-only hits are marked as "↑" and cutoff-only hits as "↓."

It is intuitive that a gene having four relatively active siRNAs is more likely to be confirmed, relative to a gene with one very active siRNA and two inactive siRNAs. In the scenario above, the latter gene would be selected by the cutoff method due to the high ranking of the single very active siRNA. The KOA algorithm, however, incorporates the underlying siRNA library design and considers the behavior of all siRNAs of the same gene in its scoring function. In an activity-sorted list, multiple siRNAs for a true-positive gene would tend to be positioned toward the top. Such an upward bias in signal distribution would not occur if the gene were a true negative; the KOA scoring function essentially statistically characterizes such a bias in signal distribution.

The two methods result in six unique hits. For gene D, only one out of four siRNAs scored as a hit. Therefore, the active well was assigned as a false positive and thus was deprioritized by KOA (↓). Genes E and F both have two siRNAs each, all relatively active. Therefore, both were designated as hits by the cutoff method. However, gene G has three siRNAs and all are relatively active as well. Since strong activity associated with three out of three siRNAs provides strong evidence that gene G is an authentic hit, KOA ranks

(↑) gene G ahead of genes E and F. By using individual siRNA signal alone, the cutoff method selects gene D, which is likely to be a false positive and fails to identify gene G, which is likely to be a false negative. The KOA algorithm, on the other hand, takes into account the predicted redundant activities of several siRNAs that target the same gene and scores the gene probabilistically. As the number of genes being studied increases and statistical fluctuation decreases, the statistical advantage of KOA will be magnified and the improvement of its hit list is expected to be more pronounced.

The KOA algorithm may be biased toward genes where more targeted siRNAs are present in the screening collection. This can be clarified by gene D (four siRNAs, discussed previously) being scored poorly and gene B (two siRNAs) being ranked as the second best gene for having both siRNAs determined to be highly potent. Even gene A with a single siRNA can still be ranked favorably under KOA for its unusually high potency. Nevertheless, the assessment of genes with single or fewer siRNAs being less compelling compared to genes with multiple active siRNAs agrees well with statistical sampling theory. Given a limited validation capacity, recruitment of genes with a single active siRNA (e.g., gene D) is often made at the expense of missing genes of multiple active siRNAs (e.g., gene G). This is why the cutoff method results in a much lower confirmation rate at the end.

7.2.2 Experimental Validation of KOA by Genome-Wide siRNA Screening

The fundamental biological assumption we made in the above KOA algorithm is that multiple siRNAs that target a "true-positive" gene will tend to be coactive probabilistically and those of a "true negative" gene will tend to be coinactive. However, exceptions to this conjecture could occur. The KOA algorithm aims to assess this possible scenario by using library design knowledge, i.e., the library mapping between genes and their siRNA wells. Using the null hypothesis that each gene is inactive, KOA uses an iterative process to find the minimum p value in order for each gene to be ranked as favorably as possible. We have used a small library of 40 siRNAs to explain how KOA models the amenability of a gene being validated and have shown that conceptually it has many advantages over the cutoff method. However, it would be more compelling if the hits picked by both KOA and cutoff algorithms could be compared experimentally. A recent siRNA study by König et al. did exactly this [11], where the KOA algorithm was referred to as the "redundant siRNA activity (RSA)" analysis. Using a genome-wide siRNA library targeting approximately 19,628 human genes, containing on average three wells per gene and with two siRNAs per well (~6 siRNAs per gene, total of 53,860 wells), three independent inhibition biological assays were used to screen the complete collection. For simplicity, a summary of only the findings for assay B mentioned in that study is reported. Both RSA and cutoff algorithms initially were applied in order to identify the top 55 wells. Interestingly, the two

analyses shared an overlap of only 11 wells. Each siRNA that was identified as a hit using either approach was rescreened independently using the original assay. A well was deemed reconfirmed only if at least one of the siRNAs gave an assay signal less than 50% of the mean value for the respective assay plate. A gene was considered validated only when two or more siRNAs were confirmed.

Figure 7.3 plots the reconfirmation rate as a function of hit ranks in the initial screen analysis by either the KOA or the cutoff method. siRNAs identified using the KOA methodology clearly show higher rates of reconfirmation, namely, 100–40% compared to 0–16% in the cutoff method. Careful inspection shows that the discrepancy between the two approaches is likely due to the reconfirmation rates of the most active siRNAs predicted by each method (Fig. 7.3). For instance, the confirmation rate for the most active siRNAs determined by the cutoff analysis is initially around 0% and then gradually increases. The cutoff method relies on the popular presumption that the wells with a higher activity have a higher reconfirmation rate, which is unlikely to hold up based upon the experimental results. In contrast, our experience has been that the most active wells are often caused by experimental artifacts and, therefore, the primary result is difficult to reproduce. This phenomenon causes an "abnormal" confirmation curve and has been previously documented for small-molecule HTS campaigns [14,15]. If a gene is represented by a single well and scores as a positive, there is little choice but to treat the well as a true

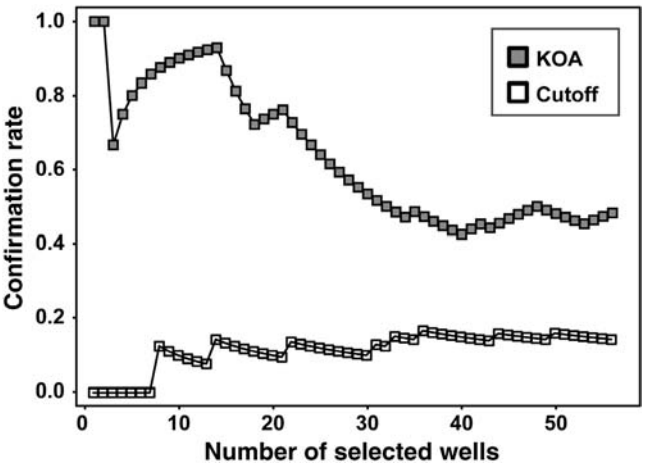


Figure 7.3 Confirmation curves of hits selected by KOA and cutoff methods in a genome-wide siRNA screen. KOA/RSA hits show a "normal" curve, where the most confirmable siRNAs were picked first and the confirmation rate gradually decreases from 100% and plateaus at around 42%. The cutoff hits show an "abnormal" curve, where the low-quality hits are picked first. As more hits are made available, the confirmation rate gradually increases to a level of approximately 18%. Figure courtesy of Nature Publishing Group.

hit. However, if a gene has several siRNAs across multiple wells and only a single well shows strong activity, one would reason the active well might be an artifact. It is such knowledge-based analysis strategy that helps the KOA/RSA algorithm to avoid those superficial active wells picked by the cutoff method and instead improves the confirmation rate through a more extensive sampling of the large-scale siRNA data sets. As there currently exists no parallel methodology for the analysis of large-scale RNAi data sets [16], application of KOA should significantly enhance the interpretation of large-scale RNAi data through the exclusion of false-positive activities derived from both experimental artifacts and off-target activities.

7.2.3 KOA for *In Silico* Gene Function Prediction

Quantifying a phenotype using a loss-of-function screen such as the one described above is an important HT target identification approach that does not rely on any prior knowledge to the relevant cellular target. It can, in principle, identify a new signal transduction pathway or a potential drug target. However, many screens are initiated with at least some prior knowledge of a relevant signal transduction pathway for what is being studied. Consequently, the project goal is often to distinguish new pathway members or additional biological activities of the known ones. Matured HT technologies, such as expression profiling, have produced large data sets consisting of transcript abundance measurements for many different tissues, treatments, disease states, and pathological stages. A documented approach to correlate the functions of known genes with those that have similar gene expression profiles is based on the GBA principle. This enables expeditious *in silico* gene function prediction for a significant portion of the transcriptome. The KOA algorithm was in fact originally designed to take advantage of the existing GO knowledge base for identifying statistically significant gene expression patterns and was originally named as OPI. The gene function prediction problem bears a great deal of similarity to our latter discussion on the identification of promiscuous HTS hits, showing the synergy of data mining tool development between bioinformatics and chemoinformatics.

Given a large gene expression matrix, where each row represents a gene and each column represents an array experiment, the goal of gene function prediction is the identification of gene clusters that not only share a similar expression pattern but also share statistical enrichment in a certain functional category. Then all the cluster members are predicted to play a role in the corresponding function based on GBA. Traditionally, unsupervised clustering algorithms such as *k*-means clustering [17,18] and hierarchical clustering [19] have been applied to group genes based on their similarities in expression profiles first, and then resultant clusters are examined for potential functional enrichments. This two-step approach is suboptimal for two reasons. First, unsupervised clustering algorithms rely on subjective parameters, such as the number of partitions in the *k*-means clustering or the similarity threshold in

Data set: an expression matrix for a total of N_T genes, one expression profile per gene
Knowledge: gene ontology (GO), where S_{ij} represents a piece of knowledge that gene j is a known member of the ith function category in GO, $j = 1, \ldots, n_i$
Hypothesis: genes sharing the same function tend to share similar expression profiles
Output: a list of genes that are considered to be members of function category i
Algorithm:

1. Construct a metaprofile Q_i for function i from all known gene members S_{ij}, $j = 1, \ldots, n_i$.
2. Rank all N_T genes based on their profile similarity against Q_i in descending order (most similar on top).
3. For each gene S_{ij}, $j = 1, \ldots, n_i$,
   4. calculate enrichment factor by $f_{ij} = p(N_T, n_i, R_{ij}, j)$;
5. $j^* = \arg\min_j f_{ij}$ and $f_i^* = f_{ij^*}$; assign f_i^* to gene S_{ij} $(j = 1, \ldots, n_i)$;
6. S_{ij} with $j \le j^*$ are marked as positives; S_{ij} with $j > j^*$ are removed as outliers.

R_{ij} = the rank number of S_{ij} in the sorted list; p = accumulated hypergeometric distribution function (Eq. 7.1).

Scheme 7.2 Outline of the KOA gene function prediction algorithm.

isolating subtrees in hierarchical clustering. Second, there are no readily available statistical controls, i.e., the odds of obtaining the resultant clusters as a matter of chance is difficult to measure. KOA outlined in Scheme 7.2 aims to overcome these shortcomings.

Using a previously published *Plasmodium falciparum* (malaria) cell cycle gene expression matrix (2234 genes across 16 parasite life stages) as an example, we attempted to identify new targets related to the "cell–cell adhesion" (GO:0016337) GO term. Prior studies discovered 10 gene members (labels shown in Fig. 7.4), and we used their profiles as "baits" to recruit new GO members by detecting "unusual" similarity in their expression patterns. The quantification of "unusualness" was an important factor in the evaluation. With a too stringent threshold most of the 10 known genes described above fell below the threshold and became false negatives. On the other hand, a too generous threshold would have led to low-quality predictions by including too many false positives. Other data analysis approaches simply applied a subjective cutoff, such as 0.8 in the Pearson correlation coefficient, regardless of which particular GO was being studied [20]. However, the goal of KOA is to find a balance between the two extremes of "unusualness without the application of any subjective parameters."

The first step entails the construction of a metaprofile that best represents the GO:0016337 family by combining the expression patterns for the 10 known gene members. More specifically, one may either select one of the 10 profiles,

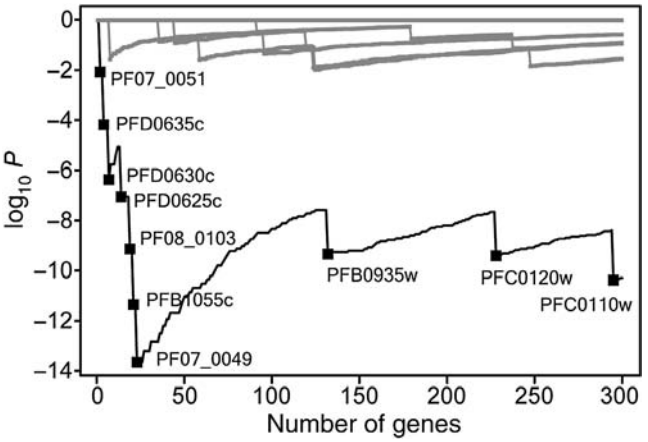


Figure 7.4 Illustration of the KOA algorithm in predicting new "cell–cell adhesion" genes using *Plasmodium falciparum* cell cycle microarray data. The gray lines represent probability curves obtained by shuffling gene labels. The black line stands for the real probability curve, of which the low p values observed are neither due to chance nor due to the multiple iterations. Figure courtesy of Oxford Journals.

where the greatest number of remaining GO members share a minimum degree of similarity, or may simply take a weighted average of all 10 profiles (or any other reasonable approach). Our typical method is to obtain multiple metaprofiles and to interrogate the KOA algorithm in order to identify the profile that results in the best knowledge optimization. Against the given metaprofile, all 2234 genes are ranked according to their profile similarities, and the 10 known genes are positioned between ranks 2nd to 300th. Genes ranked near the top of the list share a greater similarity to the metaprofile. Therefore, these genes should have a greater likelihood of having the desired activity than genes near the bottom of the list. The next step is to determine a nonsubjective similarity threshold, so that the function label GO:0016337 can be reassigned to as many of the 10 known genes as possible (knowledge rediscovery), while being assigned to as few unknown genes as possible (assuming that most of them are negatives). This is done by an iterative knowledge optimization routine that is described below.

If one accepts the top two genes as the final candidates, we rediscover the function for the second gene, PF07_0051, and predict the top gene to be a new function member (Fig. 7.4). Statistically, if one randomly selects two out of the 2234 genes and at least one of the 10 known genes fall into the selection, the probability of this occurring by chance is about 1%. In the next iteration, if one accepts the top four genes as the final candidates, i.e., we rediscover both PF07_0051 and PFD0635c (ranked fourth) and the analysis predicts two new function members, the false discovery rate (FDR) remains at 50% as before, but the true-positive rate (TPR) increases from 1/10 to 1/5,

since we now improved our knowledge set by correctly labeling PFD0635c. Statistically, if one randomly selects 4 out of 2234 genes and at least 2 of the 10 known genes fall into the selection, the p value is 0.01%. Therefore, the quality of our prediction improves significantly. In the rth iteration, if one accepts the top R genes containing r known genes, the chance of randomly selecting R genes from N that also include at least r out of the n known genes by chance is $p(N, n, R, r)$ (see Eq. 7.1).

Clearly, the smaller the probability, the more likely the resultant gene candidates have unusual associations with the given GO term. All of the iterations described above render a probability score curve (Fig. 7.4), where the global minimum is associated with the gene ranked 23rd, PF07_0049 (p value as low as 10^{-14}). Therefore, the first 23 genes are considered the best candidates with an FDR of 30% (only counting genes with some GO annotations) and a TPR of 70% (7 out of 10 are rediscovered). Figure 7.4 also shows an attempt to recruit additional known genes from the remaining three outliers (PFB0935w, PFC0120w, and PFC0110w) in order to increase the TPR at the expense of significantly increasing the FDR. The knowledge of the 10 known genes is therefore rediscovered to the best extent (7 out of 10) as quantitatively characterized by the probability scoring function (Eq. 7.1). The resultant 23 genes not only share the similar expression profile (correlation coefficients >0.48) but also are highly enriched in GO:0016337 because they contain seven known members (p value = 10^{-14}). The KOA algorithm has been successfully applied to several gene function prediction studies, which has led to many robust predictions that have been cross validated by protein network data, by cross-species coexpression patterns, and by other experimental evidence (Table 7.1) [7].

7.3 APPLICATIONS OF THE KOA APPROACH IN SMALL-MOLECULE HTS DATA MINING

We have shown how KOA can be applied to improve the confirmation rate in siRNA screening by making use of both the gene-well mapping knowledge and the assumption that wells for the same gene tend to be coactive or coinactive. We have also shown how KOA can identify a group of gene candidates that share unusual coexpression patterns by making use of both the prior ontology knowledge base available for a small collection of known genes and the GBA principle. Next, we will illustrate how KOA can be modified and aptly applied to data mining single-assay and multiassay small-molecule HTS data and to detect relationships not easily obtainable by other methods. Considering the similarities between the expected goals for analyzing siRNA functional genomics and small-molecule HTS screening, it is intuitive that a parallel hit-picking strategy would work efficiently for both types of screens. In this way, KOA provides an opportunity to synergize data mining efforts between bioinformatics and chemoinformatics.

The top X method is still widely used in the HTS primary hit selection process, which often causes a rather low confirmation rate largely due to the error-prone, noisy nature of single-dose HTS. Many data analysis methods used today [21] were originally designed for modeling data obtained from an old screening paradigm called sequential screening or smart screening [22–24], when the screening throughput was still relatively low. Although these methods are helpful for understanding the HTS data, they are not directly applicable to addressing many of the major challenges that are faced in the current HTS-based lead discovery process. For example, parameters such as compound structure analysis usually are not included as part of the decision making for the selection of primary hits; such consideration is introduced only after hit confirmation. This is suboptimal since medicinal chemists often are willing to trade a potent scaffold with limited SAR opportunities for a scaffold with the opposite properties. This is because the SAR landscape of a scaffold is usually considered to be a critically important aspect of a lead candidate for its suitability for further optimization.

7.3.1 Scaffold-Based HTS Compound Triage and Prioritization for Improved Lead Discovery

Similar to HT siRNA screening, a typical confirmation rate using the top X method for primary HTS hit selection is in the range of low 40%, mainly because of the noisy and error-prone nature of single-dose HTS currently employed in most screens. Due to the limited capacity of the follow-up hit validation, which usually involves determining compound dose-dependent responses, and the associated higher cost, it is important to improve the primary hit confirmation rate, which can help provide increased numbers of compounds of good quality for the ensuing lead discovery process. In addition, since the great majority of compounds tested in an HTS is triaged at this hit identification phase, it obviously has an important far-reaching effect to the entire multistep drug discovery process. Thus, extra care needs to be taken in this primary hit selection step.

Unlike siRNA screening libraries, small-molecule libraries typically are not designed to contain multiple samples of identical structures. Consequently, a single evaluation of a compound's assay activity typically is the sole parameter that is employed for hit designation, and it may appear that the screening activity is the only parameter one can rely on and that the top X method is the only choice for selecting hits. Fortunately, in reality, large HTS compound libraries used in pharmaceutical companies usually carry a certain redundancy in terms of compound chemical structures [25–28]; i.e., structurally similar compounds are often screened together in one HTS campaign in a single-dose format. This is not surprising considering the source of compounds in the screening library, which includes (1) libraries purchased from vendors whose catalogs often overlap; (2) libraries from combinatorial synthesis often lead to intensive sampling of a small chemical space; (3) compounds synthesized

by previous lead optimization efforts and put back into the screening deck, in which case those advanced medicinal compounds typically share common scaffolds; (4) even if a company intentionally tries to construct a diversified screening collection, they often have to purchase compounds in the unit of a plate instead of picked wells for cost consideration. The widely used SAR principle implies that structurally similar compounds may share similar activity, which further indicates that if a compound scaffold is truly active, activities from the scaffold family members tend to bias toward the high-activity region. This bias in distribution is unlikely to happen by chance when the scaffold is not actually active. Therefore, by clustering compounds into scaffold families, we create a knowledge set where members of the same compound family tend to be active and inactive coherently. By replacing siRNAs with compounds and genes with scaffolds, the abovementioned KOA algorithm applied in siRNA function screens can be modified only slightly and can be readily applicable to HTS hit triage (Scheme 7.3).

Yan et al. applied this knowledge-based hit-picking approach to a cell-based HTS campaign carried out in-house using the internal corporate compound library [8]. Following quality control and normalization to eliminate obvious artifacts and outliers, a total of ~1.1 million compounds with single-

Data set: an activity score vector for a total of N_T compounds, one activity score per compound

Knowledge: design of the compound library, where S_{ij} is the jth compound in the ith scaffold family, $j = 1, \ldots, n_i$

Hypothesis: SAR; structurally similar compounds tend to be coactive or coinactive

Output: a list of compounds that are considered to be true hits

Algorithm:

1. Cluster all compounds, C, based on their structure similarity into scaffold families.
2. Rank all compounds based on their activity scores in descending order (potent on top).
3. For each scaffold family i and
    4. for each family member, compound C_{ij} ($j = 1, \ldots, n_i$),
        5. calculate enrichment factor $f_{ij} = p(N_T, n_i, R_{ij}, j)$;
    6. $j^* = \text{argmin}_j f_{ij}$ and $f_i^* = f_{ij^*}$; assign f_i^* to all C_{ij} ($j = 1, \ldots, n_i$).
    7. C_{ij} with $j > j^*$ are removed as outliers from the data set.
8. Rank all remaining compounds based on f_i^* in ascending order, then by R_{ij} in ascending order.

R_{ij} = the rank number of C_{ij} in the sorted list; p = accumulated hypergeometric distribution function (Eq. 7.1).

Scheme 7.3 Outline of KOA hit selection algorithm for compound HTS.

dose activity data were obtained, among which the 50,000 most active compounds were analyzed by the KOA approach. Using Daylight fingerprints and the Tanimoto coefficient as chemical structure similarity measures [29], the 50,000 compounds were clustered into scaffold families based on a threshold value of 0.85 [30]. Then, each compound scaffold family was scored and prioritized according to Scheme 7.3. Both KOA and the top X methods selected the 2000 best hits, respectively, and Figure 7.5A shows the accumulative confirmation rate plots (i.e., the ratio between the number of confirmed actives over the number of selected compounds) for both approaches. It is noteworthy that the two methods selected distinctive sets of compounds.

As shown in Figure 7.5A, in the top X method, the confirmation rate is very low when only a small number of compounds are selected (~200). This is mainly because the significant number of compounds with erroneously high activities were potentially caused by experimental artifacts and were not identified by the post-HTS quality control procedure. The confirmation rate gradually improved as more compounds were selected, reaching a maximum of about 55%, at which point about 1000 compounds were picked. In contrast,

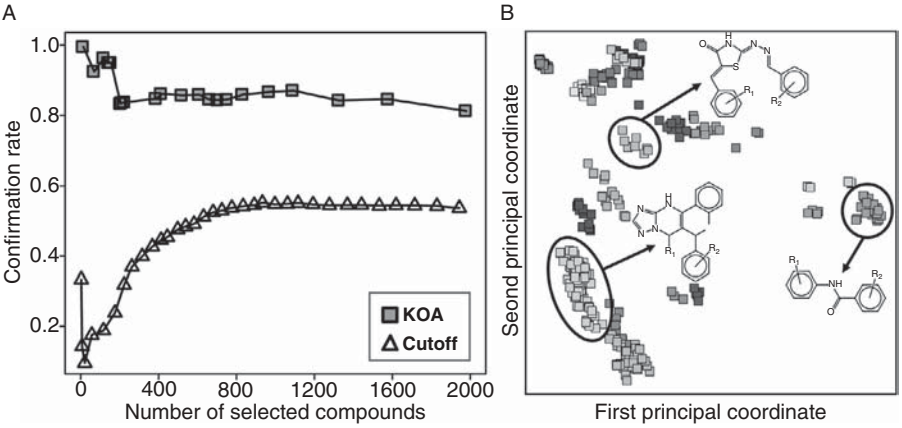


Figure 7.5 Primary hit selection in an HTS campaign. (A) Both KOA and cutoff methods were applied to pick the ~2000 best compounds and were assayed in a dose–response format for confirmation. KOA hits show a "normal" curve, where the most confirmable compounds were picked first, and the confirmation rate gradually decreases and plateaus at around 80%. The cutoff hits show an "abnormal" curve, where the low-quality hits were picked first. As more hits were made available, the confirmation rate gradually increased to a level around 50%. (B) The hits from KOA methods are projected into two-dimensional space by principal component analysis, where both the chemical diversity and the SAR strength of each scaffold family are readily visualized. This is because the KOA algorithm already incorporates the SAR principle in the process of hit ranking. Figure courtesy of American Chemical Society.

the KOA approach generated substantially better results. A high confirmation rate of over 95% was achieved when only ~150 compounds were picked. Furthermore, the false positives seen in the top X approach were effectively eliminated by this KOA approach. It was also able to maintain a high confirmation rate (~85%) with an increased number of selected compounds, and it consistently performed better than the top X method (~55%). As the KOA and the top X approaches selected distinctive sets of compounds, additional experiments were carried out to retest those compounds picked by the KOA approach but not the top X method in order to estimate the potential false negative rate by the traditional cutoff-based method. A total of 825 compounds out of the first 1108 compounds picked by the KOA approach were considered as inactive based on the top X method. Furthermore, 202 of these "inactive" compounds were retested due to compound availability, and 144 compounds were shown to be actually active. This resulted in a confirmation rate of ~71% for the "inactives," even higher than the "active" compound confirmation rate of ~55% by the top X method. This clearly demonstrates the ability of the KOA approach to effectively rescue false negatives determined by the cutoff-based method. By eliminating highly active false positives and retrieving false negatives, the KOA approach is able to substantially improve the quality of primary HTS hits.

The KOA-based HTS hit selection approach is in essence driven by the hypothesis of SAR; that is, chemically similar compounds within a scaffold shall demonstrate a certain level of similarity in assay activity. The KOA approach is capable of picking promising scaffolds with good activity and/or SAR, instead of individual, unrelated compounds like the top X method does. In addition, the SAR rule is just a probabilistic rule, which means that given two chemically similar compounds, there is a probability that they may have similar activity, i.e., for a compound family, only a fraction of its members may show similar activities. The KOA approach is able to provide an individualized activity cutoff value R_{ij^*} for each compound scaffold i based on a rigorous statistical test [5] (step 6 in Scheme 7.3), which in turn determines the fraction of the compound family that actually meets the SAR rule. This is in sharp contrast to the one-cutoff-fits-all approach employed by the top X method. Furthermore, the hits selected by this novel approach contain considerably more information than those from the top X method. For example, as shown in Figure 7.5B, it includes statistical significance, scaffold information, and SAR profiles. In addition, the KOA approach can also be applied to the secondary screening results when dose–response data are available, e.g., IC_{50} or EC_{50} data. Iteratively applying the knowledge-based KOA approach in the HTS compound selection process can significantly improve the quality of HTS hits from the very beginning of the drug discovery process and may help facilitate discovering lead series with high information content [31], as information such as scaffolds and SAR derived in this early HTS hit selection step are considered as favorable characteristics for promising lead series [31–33].

7.3.2 Identify Promiscuous and Toxic Scaffolds by Mining Multiassay HTS Database

It has been observed that many confirmed hits from a cell-based HTS study are later eliminated due to general toxicity. Indeed, toxic compounds identified in a cell-based "antagonist" screen are technically "authentic" hits but likely will be categorized as a "false positive" from a pharmacological point of view later in the project. On one hand, general cytotoxicity cannot be reliably predicted by computational approaches and it requires using a significant number of screens to profile a series of confirmed HTS hits in a panel of secondary assays that include an assessment of cellular viability. The repeated attrition of compounds with no selectivity in various HTS campaigns presents another bottleneck for lead discovery and a questionable use of limited available resources. On the other hand, since those generally cytotoxic compounds are likely to be found active across numerous cellular assays that are used to identify antagonists, it is possible to identify these compounds and/or compound scaffolds by mining the HTS database. This would provide a "filter" that could be used to eliminate these compounds and would shift the attention toward more promising chemical starting points in the lead discovery process.

Very few existing chemoinformatics algorithms have been described specifically for a corporate-wide HTS database, probably due to intellectual property issues. Horvath and Jeandenans demonstrated the concept of generalized neighborhood behavior; i.e., structurally similar compounds may have similar biological profiles in a small-scale multiassay (42 targets $\times$ 584 compounds) HTS data proof-of-concept study [34]. In this study, such structure–profile relationship (SPR) is characterized by the overall optimality criterion and consistency criterion, together with various descriptor-based structural similarity measures [34,35]. Similar successful applications of such analysis of compound HTS profiles across multiple assays have also been reported [36,37]. However, in these studies, the probabilistic nature of the SPR is not sufficiently considered, and each compound family is treated in the same way, much like the "one-size-fits-all" rule used in the top X approach for hit selection [8]. Other HTS data mining methods, which attempt to delineate relationships between compound scaffolds and target families, have also been reported [38–45]. It should be noted that most previous studies focus on the relationships between compounds and the "druggable" protein target families, such as G protein-coupled receptors (GPCRs), kinases, and proteases [46,47]. These methods are not straightforwardly applicable to study promiscuous and/or general toxic HTS hitters, where the targets may not be known and the HTS data can be intrinsically noisy [48].

Here we applied the KOA algorithm to study the correlations between compounds and assay formats on a scaffold level, which can be used to identify artifactual results and promiscuous hitters. In order to do so, first we need to identify compound families (and the core members) with strong SPR. Two major challenges remain. First, the biological HTS profile correlation among

the member compounds can be simply caused by chance. In an extreme case, if the HTS profile only contains two assays, any two compounds show either perfect Pearson correlation or anticorrelation, which, however, is only a statistical artifact. Second, as mentioned above, SAR and SPR are merely probabilistic rules; i.e., given a family of compounds with similar structures, only a fraction of it may share the similar profile. Therefore, it is important to identify those core members that do satisfy the SPR rule while excluding the outliers. Most existing data mining approaches employ some clustering analysis based on either compound structural similarity or biological profile, but not both, e.g., the widely used cluster image map (CIM) method [49]. However, when the biological profiles are used to cluster the compounds, it is found that compounds within the same scaffold are often scattered around on the clustered map, which causes great difficulty to visually identify reliable, meaningful correlations between scaffolds and their biological effects. Visual inspection has also been applied to locate common tree components when both structural and biological profile similarities are used [50]. How to simultaneously take advantage of both chemical and biological data in data mining still remains a great challenge. The KOA algorithm described here (Scheme 7.4) can be

Data set: an activity matrix of size $N_T \times Q$, i.e., a total of N_T compounds across Q HTS assays

Knowledge: design of the compound library, where S_{ij} is the jth compound in the ith scaffold family, $j = 1, \ldots, n_i$

Hypothesis: SPR; structurally similar compounds tend to share a similar activity profile

Output: lists of compounds that share both a similar scaffold and a similar selectivity profile across Q assays

Algorithm:

1. Cluster all compounds, C, based on their structure similarity into scaffold families.
2. For each scaffold i,
   3. construct a representative biological profile $\boldsymbol{Q}_C$;
   4. score compound i based on the similarity of its profile, $\boldsymbol{Q}_i$, against $\boldsymbol{Q}_C$;
   5. rank all compounds based on their similarity score in descending order.
   6. For each family member C_{ij} ($j = 1, \ldots, n_i$),
      7. calculate enrichment factor $f_{ij} = p(N_T, n_i, R_{ij}, j)$;
   8. $j^* = \text{argmin}_j f_{ij}$ and $f_i^* = f_{ij^*}$; assign f_i^* to all C_{ij} ($j = 1, \ldots, n_i$).
   9. C_{ij} with $j \le j^*$ are marked as core members; C_{ij} with $j > j^*$ are removed as outliers.

R_{ij} = the rank number of C_{ij} in the sorted list; p = accumulated hypergeometric distribution function (Eq. 7.1).

Scheme 7.4 Outline of the KOA algorithm to identify the core members of each compound cluster.

TABLE 7.2 List of the High-Throughput Screens Used in This Study

Classification	Category	Inhibition	Induction
Assay format	Enzyme activity	12	1
	Proliferation (cellular)	19	0
	Reporter gene (cell-based assay)	19	23
Readout	Fluorescence	10	1
	Alamar blue (fluorescence)	17	0
	Luciferase	21	25
Target type	GPCR	5	3
	Kinase	10	1
	Nuclear receptor	3	6
	Protease	3	0

adapted to address these challenges and to extract meaningful knowledge from the large-scale, noisy HTS database by utilizing both biological and chemical data. Table courtesy of American Chemical Society.

We have compiled a data matrix from our internal HTS database, containing 33,107 compounds across 74 assays. Annotation of the assays is shown in Table 7.2. Figure 7.6A illustrates the KOA approach to mining multiassay HTS data. The heat map shows the biological profiles of a family of five compounds (compounds **1–5**) across 74 HTS assays. If the average profile of all five compounds is used as the representative query pattern $\boldsymbol{Q}_C$ and all 33,107 compounds are ranked based on the profile similarity compared to $\boldsymbol{Q}_C$, these five compounds of interest are ranked at the 2nd, 4th, 6th, 5162nd, and 6606th positions, respectively. Figure 7.6A also shows how the logarithmic hypergeometric p value varies as the similarity cutoff value is lowered from 1.0 to 0.0, and it clearly shows that compounds **1–3** are the selected core members of this scaffold family and that their average profile is considered as the true representative profile for the scaffold.

This result is also in line with the structural difference of the compounds within this family. As shown in Figure 7.6A, the core members (compounds **1–3**) all have a propylpiperidine substitution attached to the nitrogen of the central pyrrolidine ring, while the substitution is isopropane for the two outliers (compounds **4** and **5**). It is noteworthy that based on chemical fingerprints alone, these two sets of compounds cannot be separated using a clustering program, while by including additional biological profile data, the KOA algorithm is able to effectively distinguish them. It is shown here that for compounds with more than 85% chemical similarity in a compound scaffold family, only three out of five compounds share the same HTS activity profile, clearly demonstrating the probabilistic nature of the SPR. Furthermore, the KOA algorithm is able to identify the core subset of the compound family, which shares reliable SPR, indicated by a statistically significant p value of 10^{-10}. This is highly beneficial for developing reliable quantitative structure–activity relationship (QSAR) models, as the SAR principal is the underpinning of such

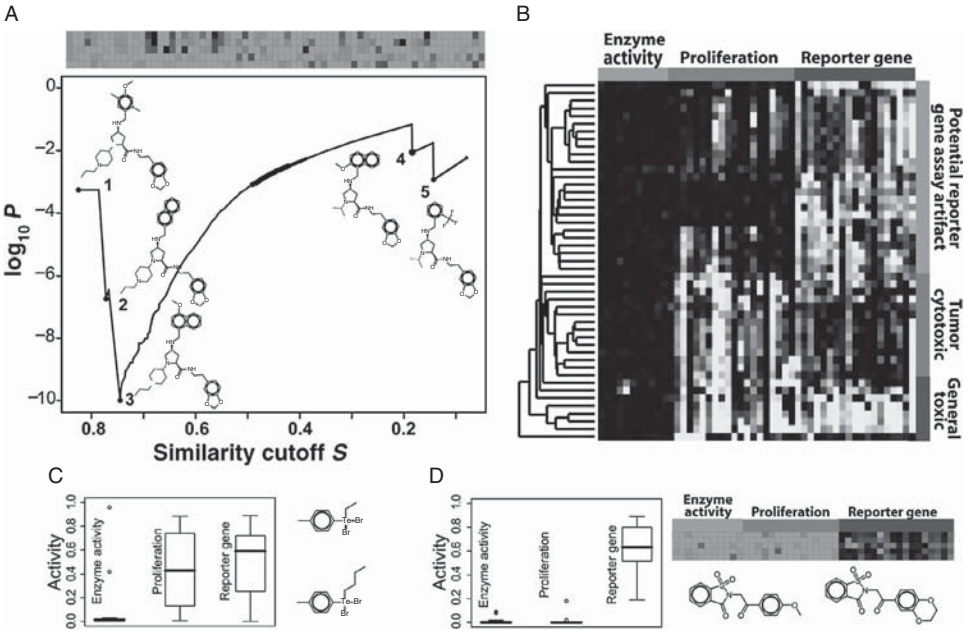


Figure 7.6 (A) Illustration of the KOA algorithm. The compound family contains five members, among which three core members share a similar activity profile and a similar chemical scaffold. The other two outlier compounds, albeit structurally similar, are effectively identified and excluded from subsequent analysis. Black in the heat map corresponds to a normalized activity score close to 1, and gray is for a score close to 0. (B) Overview of representative biological profiles of selected compound scaffold families identified in the data mining exercise. Each row in the heat map represents the median assay activity profile for a differentially behaved scaffold family. The assays are sorted according to their formats; strong inhibition is represented by white and weak inhibition is in black. The hierarchical clustering revealed three generic patterns for a compound scaffold based on assay format—general toxic, tumor cytotoxic, and potential reporter gene assay artifact. (C) Tellurium-containing general toxic scaffold family. (D) Benzisothiazolone scaffold as a potential reporter gene assay artifact. Courtesy of American Chemical Society.

studies, and it has been shown that more data do not necessarily help develop a more accurate QSAR model [51].

The KOA algorithm was applied to all the scaffold families obtained from clustering the 33,107 compounds, and the representative HTS profile from each of the family was determined. Approximately 50 out of the 74 assays in the HTS biological profile are inhibition assays, where the assay conditions were optimized in order to identify antagonists. For each representative HTS profile, the activity scores from the inhibition assays were grouped based on the categories in every assay classification (Table 7.2). For instance, based on assay format, all the inhibition activity scores from a representative HTS profile are grouped into three categories, namely, enzyme activity, prolifera-

tion, and reporter gene (Table 7.2). Statistical tests such as analysis of variance (ANOVA) and Kruskal–Wallis tests, which are the multigroup version of the familiar parametric *t*-test and the nonparametric Wilcoxon test, respectively [52], were then applied. Specifically, the null hypothesis here is that the assay activities from an HTS activity profile shall not show differences among various categories, and a low probability value (usually defined as <0.01) rejects the null hypothesis, which implies statistically significant differential activities among different categories. In this study, we assigned both ANOVA test and Kruskal–Wallis test probability scores to each compound scaffold to minimize false correlations, i.e., insignificant correlations due to randomness between compounds and biological profiles. The standard box plot, which shows median, lower, and upper quartile information in a succinct manner, is used for visualization (Fig. 7.6C and 7.6D). It offers an effective visual tool to further examine the behaviors of the compound families in each assay category of interest (Fig. 7.6).

It is important for drug discovery programs if one can identify compound scaffolds that may give technology-related screening artifacts, such as promiscuous hitters, by mining the corporate HTS database using rigorous statistical methods. For example, generally cytotoxic compounds may show consistently high activities in many cell-based assays, while compounds that are known to form aggregates may also display misleadingly high activities in enzyme inhibition assays [53]. An overview of the KOA results from data mining our HTS database is shown in Figure 7.6B.

Indeed, compound scaffolds that appeared to have a screening profile consistent with a general cytotoxicity mechanism of action were clearly identified (the bottom of the heat map in Fig. 7.6B). Scaffolds that consistently showed inhibition in reporter gene assays are located in the middle of the heat map (also Fig. 7.6D). The underlying pharmacological mechanism of a compound being a frequent hitter was recently studied [4]. It was hypothesized that the mechanisms that cause promiscuity can be due to general cytotoxicity, modulation of gene expression efficiency, luciferase reporter gene artifacts, color shifting, and so on. Further mechanism-based analysis shows that frequent hitters are often related to apoptosis and cell differentiation, including kinases, topoisomerases, and protein phosphatases. When evaluating a particular compound over a wide range of HTS, a typical hit rate as low as 2% is expected, whereas frequent hitters often are scored as hits in at least 50% of the assays performed. These numbers suggest the significant value of deprioritizing such undesirable compounds at the early HTS hit selection stage. An *in silico* approach such as the KOA, mining can effectively address these practical lead discovery issues without additional laborious laboratory work.

Furthermore, Yan et al. has also demonstrated that the same KOA method can lead to the identification of target classes for specific compound scaffolds by statistically testing target-type categories (Table 7.2) of the representative HTS profile [2]. This type of characterization of the compound families would be extremely helpful for designing more specific screening libraries, especially

for companies that wish to cover wide chemical space with limited screening efforts [54,55] or for those screens that are cost prohibitive or are not amenable to HT formats. Compounds in a diversified collection should represent different lead islands in the chemical space. As discussed above, the core members of a scaffold family identified by the KOA algorithm belong to a subset of compounds that contain statistically reliable SAR/SPR. Those representative compounds from each scaffold family might be suitable candidates for constructing such diversity-oriented libraries because of their capacity to best capture the SAR information with minimum structural redundancy.

7.4 OTHER RELATED APPROACHES FOR BIOLOGICAL DATA MINING

Although HTS compound activity data across a large number of screens are hard to obtain, gene expression data matrices consisting of thousands of genes across dozens of microarrays are fairly common. Such data are often mined by various unsupervised clustering algorithms, where clustering results tend to depend on several subjective parameters and there are no readily available statistical p values to indicate their biological significance or insignificance. The KOA algorithm is a knowledge-guided clustering algorithm, in the sense that it relies on existing knowledge to automatically determine the cluster boundaries. By optimally reproducing prior knowledge, KOA intrinsically contains a self-validation component for estimating both FDRs and p values of resultant clusters. Here, we summarize a few key differences between KOA and some related algorithms published by previous studies.

7.4.1 *k*-Means Clustering Algorithm

The k-means clustering algorithm is a partitioning method that separates underlying objects into k groups according to their similarities. This algorithm has been widely applied in summarizing unique patterns from many biological systems. The same malaria cell cycle data set mentioned above was also analyzed by a robust k-means algorithm, resulting in 15 groups [17]. However, there are several intrinsic disadvantages associated with this algorithm. First, the desirable number of clusters, k, is not only difficult to determine but is often conceptually nonexistent. Second, genes are often involved in multiple biological processes; arbitrarily forcing each of them into one cluster distorts our biological understanding. Young et al. therefore compared the resultant clusters obtained by both k-means and KOA methods [6]. It was found that for almost all of the GO functional categories described by the k-means clusters, KOA-generated clusters had comparable or greater statistical significance. For example, by switching to KOA algorithm, "antigenic variation" (GO:0020033) cluster went from a p value of 4×10^{-8} to a p value of 9×10^{-40}. In addition, by relying on each GO category as a piece of knowledge to seed

a cluster, KOA offered much higher resolution in functional space. For example, the *k*-means cluster #15 was found by KOA to further consist of subpatterns of a cell invasion cluster of 53 genes, an apical complex cluster of 82 genes, and a rhoptry cluster of 6 genes.

7.4.2 Iterative Group Analysis Algorithm

Hierarchical clustering offers several advantages compared to the *k*-means algorithm, e.g., it does not require inputting a *k* value and, therefore, permits "fuzzy" cluster boundaries. However, the resultant tree can be difficult to interpret, and there are no objective ways to identify a local subtree for the purpose of function assignment. Breitling et al. introduced an iterative group analysis algorithm [56] and demonstrated a different application of the similar knowledge-based analysis approach in identifying differentially expressed gene classes. Toronen applied another knowledge-based analysis method to identify best-scoring clusters (subtrees) on top of an expression-based hierarchical gene tree [57]. These algorithms share the similar idea as KOA that it is essential to use different similarity thresholds for different gene classes and that thresholds should be determined based on the GO knowledge base.

7.4.3 Gene Set Enrichment Analysis (GSEA)

Mootha et al. also designed a knowledge-based optimization algorithm called GSEA, which relies only on using annotated genes in the GO database to enrich weak differentially expressed signals [58]. Using this approach, they were able to successfully determine the proliferator-activated receptor-gamma coactivator 1 (PGC-1) responsive pathway to be involved in type 2 diabetes mellitus. The GSEA algorithm, however, entirely relies on prior biological annotations, which makes it inapplicable to the functional annotation of uncharacterized genes in its original proposed form. Despite the limitations, GSEA has also found applications in many bioinformatics problems [59,60].

7.5 CONCLUSION

HT technologies are widely being used in the pharmaceutical industry, which has generated a vast amount, but often noisy, biological, chemical, and pharmacological data. These large data sets clearly hold many potential discoveries, which can only be unearthed with robust data mining tools. Through the many successful applications of KOA algorithm in both target identification and lead discovery processes, it has been demonstrated that KOA is one of such ideal tools that has been validated by solid experimental results. By exploiting the existing knowledge, such as library design or GO, and by making use of cornerstone principles or hypotheses such as guilt by association and SAR/SPR, KOA can address many bioinformatics and chemoinformatics

challenges. With an increasing number of applications in various drug discovery phases, we expect KOA will play a more significant role in contributing to a better lead discovery work flow, resulting in higher-quality lead series.

REFERENCES

1. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? *Nat Rev Drug Discov* 2006;5:993–996.
2. Yan SF, King FJ, He Y, Caldwell JS, Zhou Y. Learning from the data: Mining of large high-throughput screening databases. *J Chem Inf Model* 2006;46:2381–2395.
3. Bender A, Young DW, Jenkins JL, Serrano M, Mikhailov D, Clemons PA, Davies JW. Chemogenomic data analysis: Prediction of small-molecule targets and the advent of biological fingerprint. *Comb Chem High Throughput Screen* 2007;10:719–731.
4. Crisman TJ, Parker CN, Jenkins JL, Scheiber J, Thoma M, Kang ZB, Kim R, Bender A, Nettles JH, Davies JW, Glick M. Understanding false positives in reporter gene assays: In silico chemogenomics approaches to prioritize cell-based HTS data. *J Chem Inf Model* 2007;47:1319–1327.
5. Zhou Y, Young JA, Santrosyan A, Chen K, Yan SF, Winzeler EA. *In silico* gene function prediction using ontology-based pattern identification. *Bioinformatics* 2005;21:1237–1245.
6. Young JA, Fivelman QL, Blair PL, de la Vega P, Le Roch KG, Zhou Y, Carucci DJ, Baker DA, Winzeler EA. The *Plasmodium falciparum* sexual development transcriptome: A microarray analysis using ontology-based pattern identification. *Mol Biochem Parasitol* 2005;143:67–79.
7. Zhou Y, Ramachandran V, Kumar KA, Westenberger S, Refour P, Zhou B, Li F, Young JA, Chen K, Plouffe D, Henson K, Nussenzweig V, Carlton J, Vinetz JM, Duraisingh MT, Winzeler EA. Evidence-based annotation of the malaria parasite's genome using comparative expression profiling. *PLoS One*, 2008, 3:e1570.
8. Yan SF, Asatryan H, Li J, Zhou Y. Novel statistical approach for primary high-throughput screening hit selection. *J Chem Inf Model* 2005;45:1784–1790.
9. Yan SF, King FJ, Zhou Y, Warmuth M, Xia G. Profiling the kinome for drug discovery. *Drug Discov Today* 2006;3:269–276.
10. Zhou Y, Zhou B, Chen K, Yan SF, King FJ, Jiang S, Winzeler EA. Large-scale annotation of small-molecule libraries using public databases. *J Chem Inf Model* 2007;47:1386–1394.
11. König R, Chiang CY, Tu BP, Yan SF, DeJesus PD, Romero A, Bergauer T, Orth A, Krueger U, Zhou Y, Chanda SK. A probability-based approach for the analysis of large-scale RNAi screens. *Nat Methods* 2007;4:847–849.
12. Young JA, Johnson JR, Benner C, Yan SF, Chen K, Le Roch KG, Zhou Y, Winzeler EA. *In silico* discovery of transcription regulatory elements in *Plasmodium falciparum*. *BMC Genomics* 2008;9:70.
13. Rines DR, Gomez M, Zhou Y, DeJesus P, Grob S, Batalov S, Labow M, Huesken D, Mickanin C, Hall J, Reinhardt M, Natt F, Lange J, Sharp DJ, Chanda SK,

Caldwell JS. Whole genome functional analysis identifies novel components required for mitotic spindle integrity in human cells. *Genome Biol* 2008;9:R44.

14. Fay N, Ullmann D. Leveraging process integration in early drug discovery. *Drug Discov Today* 2002;7:S181–S186.
15. Bajorath J. Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* 2002;1:882–894.
16. Shedden K, Chen W, Kuick R, Ghosh D, Macdonald J, Cho KR, Giordano TJ, Gruber SB, Fearon ER, Taylor JM, Hanash S. Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data. *BMC Bioinformatics* 2005;6:26.
17. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder AA, Batalov S, Carucci DJ, Winzeler EA. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* 2003;301: 1503–1508.
18. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet* 1999;22:281–285.
19. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998;95:14863–14868.
20. Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* 2004;5:18.
21. Parker CN, Schreyer SK. Application of chemoinformatics to high-throughput screening: Practical considerations. *Methods Mol Biol* 2004;275:85–110.
22. Engels MFM, Venkatarangan P. Smart screening: Approaches to efficient HTS. *Curr Opin Drug Discov Devel* 2001;4:275–283.
23. Young SS, Lam RL, Welch WJ. Initial compound selection for sequential screening. *Curr Opin Drug Discov Devel* 2002;5:422–427.
24. Young SS, Hawkins DM. Using recursive partitioning analysis to evaluate compound selection methods. *Methods Mol Biol* 2004;275:317–334.
25. Golebiowski A, Klopfenstein SR, Portlock DE. Lead compounds discovered from libraries. *Curr Opin Chem Biol* 2001;5:273–284.
26. Golebiowski A, Klopfenstein SR, Portlock DE. Lead compounds discovered from libraries: Part 2. *Curr Opin Chem Biol* 2003;7:308–325.
27. Oprea TI. Chemical space navigation in lead discovery. *Curr Opin Chem Biol* 2002;6:384–389.
28. Rose S, Stevens A. Computational design strategies for combinatorial libraries. *Curr Opin Chem Biol* 2003;7:331–339.
29. Csizmadia F. JChem: Java applets and modules supporting chemical database handling from web browsers. *J Chem Inf Comput Sci* 2000;40:323–324.
30. Wilton D, Willett P, Lawson K, Mullier G. Comparison of ranking methods for virtual screening in lead-discovery programs. *J Chem Inf Comput Sci* 2003;43:469–474.
31. Alanine A, Nettekoven M, Roberts E, Thomas AW. Lead generation–Enhancing the success of drug discovery by investing in the hit to lead process. *Comb Chem High Throughput Screen* 2003;6:51–66.

32. Hann MM, Oprea TI. Pursuing the leadlikeness concept in pharmaceutical research. *Curr Opin Chem Biol* 2004;8:255–263.
33. Bleicher KH, Bohm HJ, Muller K, Alanine AI. Hit and lead generation: Beyond high-throughput screening. *Nat Rev Drug Discov* 2003;2:369–378.
34. Horvath D, Jeandenans C. Neighborhood behavior of *in silico* structural spaces with respect to *in vitro* activity spaces—A novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J Chem Inf Comput Sci* 2003;43:680–690.
35. Horvath D, Jeandenans C. Neighborhood behavior of *in silico* structural spaces with respect to *in vitro* activity spaces—A benchmark for neighborhood behavior assessment of different *in silico* similarity metrics. *J Chem Inf Comput Sci* 2003;43: 691–698.
36. Froloff N. Probing drug action using in vitro pharmacological profiles. *Trends Biotechnol* 2005;23:488–490.
37. Fliri AF, Loging WT, Thadeio PF, Volkmann RA. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc Natl Acad Sci USA* 2005;102:261–266.
38. Bredel M, Jacoby E. Chemogenomics: An emerging strategy for rapid target and drug discovery. *Nat Rev Genet* 2004;5:262–275.
39. Jacoby E, Schuffenhauer A, Popov M, Azzaoui K, Havill B, Schopfer U, Engeloch C, Stanek J, Acklin P, Rigollier P, Stoll F, Koch G, Meier P, Orain D, Giger R, Hinrichs J, Malagu K, Zimmermann J, Roth HJ. Key aspects of the Novartis compound collection enhancement project for the compilation of a comprehensive chemogenomics drug discovery screening collection. *Curr Top Med Chem* 2005;5: 397–411.
40. Fischer HP, Heyse S. From targets to leads: the importance of advanced data analysis for decision support in drug discovery. *Curr Opin Drug Discov Devel* 2005;8:334–346.
41. Vieth M, Sutherland JJ, Robertson DH, Campbell RM. Kinomics: Characterizing the therapeutically validated kinase space. *Drug Discov Today* 2005;10:839–846.
42. Bocker A, Schneider G, Techentrup A. Status of HTS data mining approaches. *QSAR Comb Sci* 2004;23:207–213.
43. Root DE, Flaherty SP, Kelley BP, Stockwell BR. Biological mechanism profiling using an annotated compound library. *Chem Biol* 2003;10:881–892.
44. Covell DG, Wallqvist A, Huang R, Thanki N, Rabow AA, Lu XJ. Linking tumor cell cytotoxicity to mechanism of drug action: An integrated analysis of gene expression, small-molecule screening and structural databases. *Proteins* 2005;59:403–433.
45. Kubinyi H, Muller G. *Chemogenomics in Drug Discovery: A Medicinal Chemistry Perspective*. Weinheim: Wiley-VCH, 2004.
46. Lowrie JF, Delisle RK, Hobbs DW, Diller DJ. The different strategies for designing GPCR and kinase targeted libraries. *Comb Chem High Throughput Screen* 2004;7:495–510.
47. Whittaker M. Discovery of protease inhibitors using targeted libraries. *Curr Opin Chem Biol* 1998;2:386–396.

48. Fishman MC, Porter JA. Pharmaceuticals: A new grammar for drug discovery. *Nature* 2005;437:491–493.
49. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ Jr., Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL, Buolamwini JK, van Osdol WW, Monks AP, Scudiero DA, Sausville EA, Zaharevitz DW, Bunow B, Viswanadhan VN, Johnson GS, Wittes RE, Paull KD. An information-intensive approach to the molecular pharmacology of cancer. *Science* 1997;275:343–349.
50. Kibbey C, Calvet A. Molecular Property eXplorer: A novel approach to visualizing SAR using tree-maps and heatmaps. *J Chem Inf Model* 2005;45:523–532.
51. Martin YC. Challenges and prospects for computational aids to molecular diversity. *Perspect Drug Discov Des* 1997;7–8:159–172.
52. Zar JH. *Biostatistical Analysis*. Upper Saddle River, NJ: Prentice Hall, 1999.
53. Seidler J, McGovern SL, Doman TN, Shoichet BK. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J Med Chem* 2003;46: 4477–4486.
54. Goodnow RA Jr., Guba W, Haap W. Library design practices for success in lead generation with small molecule libraries. *Comb Chem High Throughput Screen* 2003;6:649–660.
55. Webb TR. Current directions in the evolution of compound libraries. *Curr Opin Drug Discov Devel* 2005;8:303–308.
56. Breitling R, Amtmann A, Herzyk P. Iterative Group Analysis (iGA): A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics* 2004;5:34.
57. Toronen P. Selection of informative clusters from hierarchical cluster tree with gene classes. *BMC Bioinformatics* 2004;5:32.
58. Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, Villeneuve A, Sladek R, Xu F, Mitchell GA, Morin C, Mann M, Hudson TJ, Robinson B, Rioux JD, Lander ES. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci USA* 2003;100:605–610.
59. Sweet-Cordero A, Mukherjee S, Subramanian A, You H, Roix JJ, Ladd-Acosta C, Mesirov J, Golub TR, Jacks T. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat Genet* 2005;37:48–55.
60. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;313:1929–1935.

PART III

BIOINFORMATICS-BASED APPLICATIONS

8

MINING DNA MICROARRAY GENE EXPRESSION DATA

Paolo Magni

Table of Contents

8.1 INTRODUCTION

Investigating the effects of a compound at the cellular level is certainly an appealing challenge in the drug discovery and development process. That is especially true in the early phase of the discovery process during target identification and validation. The identification of intracellular pathways that are

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

perturbed by a chemical compound contributes to a better understanding of the mechanism of action of a drug and its possible side effects and potentially leads to the identification of a gene signature correlated with efficacy or safety [1–5]. Moreover, the comparison of the effects at the cellular level of a lead compound on several pathological cell lines (e.g., several tumor cell lines) may allow to early highlight the class of pathology for which the compound is promising. Several techniques are nowadays available for monitoring the effects at both the protein and transcriptional levels. In particular, thanks to the recent advances in high-throughput technology and in molecular biological knowledge mainly by means of the Human Genome Project [6,7] and its correlated projects, it is now possible to detect and to monitor simultaneously the expression levels of thousands of genes (ideally even the whole genome) in only one experiment. There are many tools to measure gene expression, such as northern blotting, (quantitative) real-time polymerase chain reaction (RT-PCR) or serial analysis of gene expression (SAGE), but certainly the most appropriate tools for a parallel analysis of multiple genes are DNA microarrays [8]. Since their appearance in the late 1990s [9,10], they have become standard tools for genome-scale gene expression analysis with well-established biological protocols applied in research laboratories all over the world. They are currently used in several application fields ranging from cancer research to cell cycle investigation, from clinical diagnosis to drug discovery, from pattern discovery of coordinating genes to gene function discovery. However, the typology and the huge volume of collected data create some peculiar difficulties that a plethora of published methods tries to overcome. Adopting a suitable data analysis procedure selected in accordance with the conducted experiment and with the goal of the study is a nontrivial task. The purpose of this chapter is to present, after a brief introduction about microarray technology, a general view of data mining techniques currently used in gene expression analysis and to classify them, providing in such a way a sort of user's guideline.

8.2 MICROARRAY TECHNOLOGY

Microarrays are small devices suitable for the parallel (or simultaneous) investigation of a hundred or thousand conditions of interest. The basic idea is to put on a small solid substrate with a surface of a few square centimeters a great number of probes, exploiting the advance of micrometer technology and image processing capabilities. The main conceptual ingredients at the basis of the microarray technology are (1) a device (called array) on which a great number of probes are orderly deposited along a prefixed grid, (2) a "biological mechanism" that allows to "switch on" a detectable signal in a subset of probes on the basis of the conditions of the investigated sample, and (3) the acquisition of an image of the array and the subsequent "quantitative" evaluation of the intensity of its different spots corresponding to the different switched-on

probes. Microarrays differ from one another for the analyzed biological sample, for the exploited biological mechanism, for the chosen detectable signal, and for the technology adopted to build the array itself. In any case, they allow to investigate previously intractable problems and to find novel potential results. Researchers are using microarrays to try to identify crucial aspects of growth and development of several organisms, as well as to explore the genetic causes of many human diseases.

8.2.1 Types of Microarrays

Several types of microarrays have been developed for different purposes. A first rough subdivision can be made on the basis of the adopted probes, i.e., tissues, antibodies, genomic DNA, cDNA, and mRNA. In the following, the most important devices are briefly mentioned.

Tissue microarrays have been recently developed to facilitate tissue-based research [11]. Core tissue biopsies are arrayed into a recipient paraffin block by using a tissue arrayer, which generates a "fine" regular matrix of cores. DNA, RNA, or proteins are then targeted through *in situ* investigations as for conventional histological tissue. Typically, a block contains up to 600 tissue biopsies (sample diameter ranging from 0.6 to 2.0 mm). They are suitable for the analysis of multiple tissue samples [12,13].

Protein microarrays use antibodies as probes. They are suitable to simultaneously evaluate the expression profile of multiple proteins for which antibodies are available [14].

Comparative genomic hybridization (CGH) and single nucleotide polymorphism (SNP) arrays are suitable for DNA analysis. They use genomic DNA probes. As DNA microarrays (see below), they represent the convergence of DNA hybridization, fluorescence microscopy, and solid surface DNA capture. In particular, CGH arrays detect genomic DNA gains and losses due to insertion/deletion events of large DNA regions (5–10 kb) also noncoding, or changing in the number of copies of particular genes. Several thousands of probes, derived from most of the known genes and interesting noncoding regions of the genome, are printed on a glass slide. DNA from two samples are differently labeled using different fluorophores and are hybridized to probes. The ratio of the fluorescence intensity of the samples in a probe is a measure of the copy number changes of the respective locus in the genome [15–17].

SNP arrays are used to detect mutations or single nucleotide polymorphisms within a population. As SNPs are highly conserved throughout evolution and within a population, the map of SNPs serves as an excellent genotypic marker for research. SNP arrays are a useful tool to study the whole genome [18]. They are suitable for the investigation of individual disease susceptibility, disease evolution, or drug effects and therapy efficacy.

DNA microarrays (sometimes also indicated as mRNA microarrays or expression arrays) instead are useful to monitor the transcriptional activity of cells and therefore to perform gene expression analysis. They are particularly

indicated to evaluate the effect of a disease, of a drug, or of the development and differentiation processes in cellular activity.

Although all the microarrays, as briefly highlighted, exploit the same basic ideas and show common peculiarities also in data mining (at least for what concerns the first steps of the analysis), they are quite different with respect to their goal, to the experimental procedure, and to the most advanced data mining techniques. This chapter focuses on DNA microarrays.

8.2.2 DNA Microarrays

DNA microarrays can be divided in two main slightly different classes: cDNA spotted arrays [9] and oligonucleotide arrays [10]. cDNA arrays exploit the competitive hybridization (like CGH arrays) of two labeled samples on several probes. Probes are "long" expressed sequence tags (ESTs) (1–2 kb long) obtained from cDNA libraries amplified by PCR in separate physical containers and printed (ideally in the same quantity) to a glass slide by a robot following a regular grid (Fig. 8.1).

Ideally, one spot corresponds to one transcript. Typically, spots are 120–250 μm separated, so that tens of thousands of transcripts can be spotted together in one slide. Obviously, in each spot, several copies of the same probe are spotted to allow a quantitative evaluation of the transcripts.

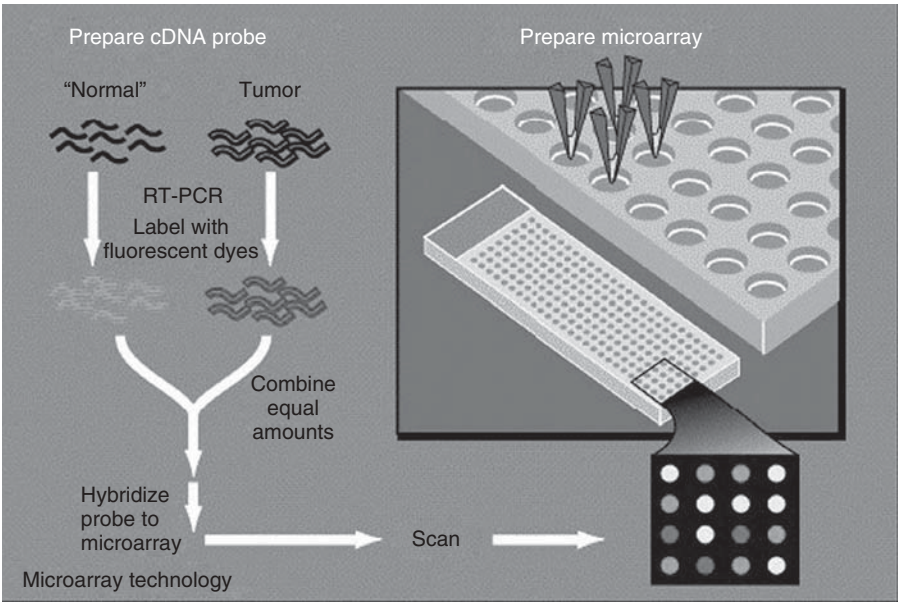


Figure 8.1 Robotically spotted cDNA microarrays hybridized to two samples: normal and tumor. Arrays are built by spotting PCR-amplified cDNA. Two samples are labeled and hybridized.

On the other hand, oligonucleotide arrays exploit the photolithographic technological abilities coming from microelectronic industry to synthesize *in situ* short oligonucleotides (25 bases long and thus these arrays are also called 25 mers) (Fig. 8.2).

This automatic and highly standardized technique for the production of the arrays allows to obtain a high probe density and a high precision and reproducibility of the building process. Differently from cDNA, one probe does not correspond to a transcript, but several (10–20) probes are required to identify uniquely a transcript. Moreover, with this platform, the hybridization is made on a single sample basis, so that no simultaneous comparison is possible between two conditions. This means that typically, different arrays have to be compared applying a suitable normalization transformation across separate microarray data sets in order to make meaningful comparisons. Affymetrix GeneChip technology is nowadays the most popular one for what concerns oligonuclotide arrays. This technology allows to produce high-density arrays (about 600,000 probes in the same array) containing in only one chip the whole genome also of the most complex organisms, such as *Homo sapiens*.

Other solutions have been recently proposed such as 60 mers by Agilent and Applied Biosystem. Agilent microarrays are spotted chips rely on the *in situ* synthesis of probes at or near the surface of the microarray slide by ink

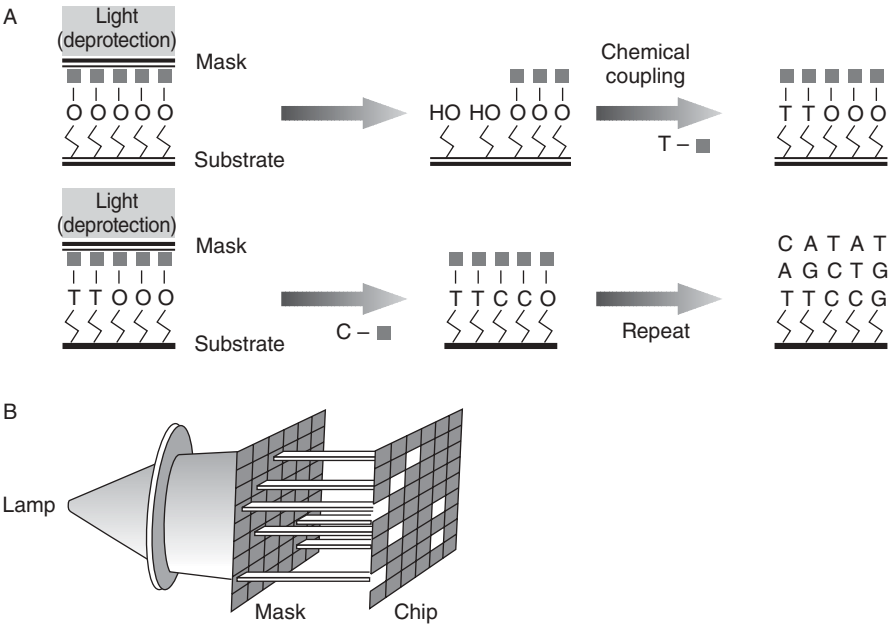


Figure 8.2 The photolithographic construction of Affymetrix microarrays. Using selecting masks, photolabile protecting groups are light activated for DNA synthesis and photoprotected DNA bases are added and coupled to the intended coordinates.

jet printing using phosphoramidite chemistry [19]. Longer nucleotides (60 mers instead of the 25 mers of Affymetrix) allow a more specific hybridization and thus one probe per gene is included on the array differently from Affymetrix and similarly to cDNA. The standard experimental paradigm of this chip compares mRNA abundance in two different biological samples on the same microarray (as cDNA). Agilent oligonucleotide microarrays like conventional spotted microarrays employ a two-color scheme (but also one-color chips are available). The reference is generated either from "normal" cells or tissues, or from a standardized mRNA mix, sometimes termed as "universal control," collected from the transcriptome of a variety of cells or tissues. The universal RNA provides a sort of reference signal for the majority of investigating conditions.

Applied Biosystems microarrays are 60-mer oligonucleotide spotted chips too. They are a single-color channel platform; therefore, one sample for an array is analyzed as for the Affymetrix chip. They use chemiluminescence to measure gene expression levels and fluorescence to grid, normalize, and identify microarray probes. Recently, an interesting work has compared this platform with Affymetrix chip, investigating in both platforms the effect of a cell cycle inhibitor compound, previously characterized for mechanism of action, in tumor cells [5].

8.2.3 Sample Preparation, Labeling, and Hybridization

Every platform requires a specific protocol to be followed in sample preparation, labeling, and hybridization. Each producer delivers together with the chip a detailed experimental protocol, besides all the kits necessary to perform an experiment. As an example, in the following, the cDNA and Affymetrix protocols are summarized. The two procedures are quite different even if the conceptual steps are the same.

The typical workflow for the Affymetrix chip is depicted in Figure 8.3.

Initially, sample preparation starts by isolating total RNA, from which mRNA is extracted and subsequently converted in cDNA using a reverse transcriptase enzyme and an oligo-dT primer containing a T7 polymerase site for 5′ to 3′ to start the retrotranscription. The resulting cDNA is purified and, if necessary, stored in a freezer. This is the sample preparation part. The second step is labeling. cDNA is transcribed *in vitro* in cRNA (this step also allows amplification) and is labeled with biotin. cRNA is purified and fragmented in the presence of metal ions to allow the hybridization overnight of the prepared sample with the short oligonucleotide probes. The chip is then washed and an image is acquired by a high-resolution scanner.

The sample preparation procedure for cDNA experiments is the same as the one illustrated for the Affymetrix chip. The only difference is that mRNA has to be extracted from two samples (see Fig. 8.1). Both samples are labeled separately by a reverse transcription with different fluorescent dyes (Cy3-green and Cy5-red). If required, labeled cDNA is amplified by PCR and then

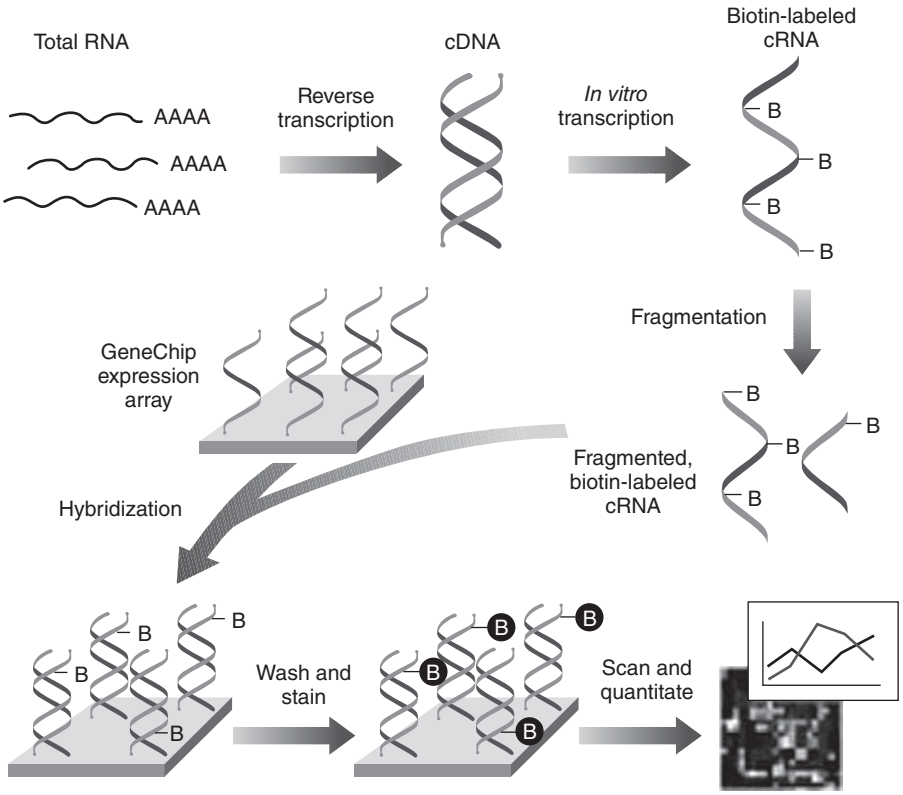


Figure 8.3 Workflow of a typical experiment with Affymetrix arrays.

is hybridized to the cDNA clones spotted on the array. Finally, the array is scanned for both colors (corresponding to the two samples) separately and two images are acquired. Typically, the two images are fused into a singular one, building the well-known images with red, green, and yellow spots.

8.2.4 From Arrays to Numbers: Acquisition and Preprocessing

As already said, after sample preparation, labeling, and hybridization, all the platforms based on fluorescence signals require a (one or two channels) scanning of the array, a quantification of the signal, and a so-called preprocessing of the obtained data in order to make comparable measurements of different arrays. At the end of this step, a matrix of gene expression levels with genes (or transcripts) on the rows and arrays on the columns on which data mining techniques have to be applied is obtained (Fig. 8.4).

In this section, the essential information about these steps will be provided, also because they are necessary to fully understand the nature of the data to be analyzed.

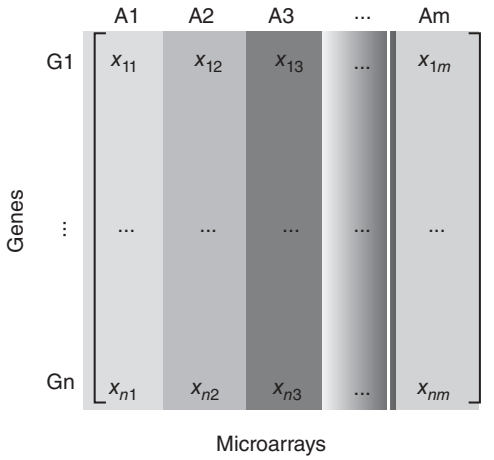


Figure 8.4 DNA microarray data matrix. Gene expression levels coming from a DNA microarray experiment. On the rows are the monitored genes; on the columns are the investigated experimental conditions corresponding to the different arrays.

In the first step, a digital image is obtained by "reading" the array with high-resolution scanners. Probes are recognized by imaging software tools and are labeled by superimposing on the acquired image the "theoretical" grid of probes used to build the arrays. This phase is called gridding. Then, pixels are grouped around the center of each spot and are divided in the foreground and background (segmentation phase). The intensity extraction phase in which, for each probe, the intensities of the foreground and background pixels are summarized in the two respective signals follows. Both commercial and freeware softwares are available to perform gridding, segmentation, and signal extraction steps, e.g., ScanAlyze, GenePix, and QuantArray, which are the most popular ones.

The goal of these phases is to minimize the noise sources. Subsequently, the probes with low-quality signal, i.e., the ones in which the foreground is lower than the background, are removed. For the remaining ones, the foreground signals of the probes related to the same transcript are further summarized to obtain a single value for each transcript, after a background correction (summarizing phase). This step is particularly important for Affymetrix arrays in which several probes correspond to different portions of the same transcript (or gene). Even if this step is very important, it is not well established, and several algorithm have been proposed [20–23] and implemented in software tools such as MAS5.0, dChip, or RMA-Bioconductor.

The last phase of preprocessing is normalization. Normalization is necessary to make comparable different measurements and to remove systematic errors. Normalization can be viewed as a sort of calibration procedure [24]. It can occur both within the array (to remove the intra-array variability) and

between arrays (to remove the interarray variability). The causes of the difference between two measurements can be numerous, e.g., a different efficiency in fluorescent incorporation during the sample preparation, a different efficiency in scanning the arrays (within and between arrays), and a different efficiency in the hybridization (within and between arrays), in addition to a different level of the gene expressions. Several normalization algorithms have been proposed. They are generally based on the hypothesis that the majority of the gene expression levels does not change over the different conditions (between arrays) or in each subregion of the array. Therefore, for example, the simplest methods assume that the median (or mean) is constant between arrays and in subregions of the same array. These methods, called global or linear normalization schemes, assume that all probes have to be scaled by the same normalization factor. More complex normalization procedures based on quartiles and percentiles have been proposed [25,26]. Other methods are based on local linear corrections of the intensity (e.g., lowess normalization [27]).

After normalization, data are generally log transformed because of the large range of expression values and their high asymmetric distribution. Moreover, for the Affymetrix array, a linear dependence between gene expression level and log intensity has been shown too [28].

It is important to note that the complex measuring procedure of the gene expression is affected not only by systematic errors partially removed by the normalization, but also by a technical variability due to the overall experimental procedure, the array manufacturing process, the signal acquisition, or the image processing (Fig. 8.5).

For example, a recent study on tumor cells treated with anticancer compounds has shown that also the day of the hybridization can induce an even greater variability than the effect of compounds [29].

To the technical variability, it is necessary to add the biological variability, owing to the genetic differences or the different environmental conditions. This variability is also present in synchronized cells coming from the same cellular line and grown in the same conditions. The greater the global experimental variability, the lower the ability to find genes differentially expressed in the different investigated conditions. Therefore, it is important to choose experimental designs that allow to evaluate, through technical replicates and biological pooling, the experimental variability and subsequently to reduce its effects on the results of the study by adopting suitable statistical techniques [29]. Erroneous data can be identified if a sufficient number of repeated experiments are performed. However, economic constraints and biological sample availability have to be considered also for a good experimental design.

To conclude this discussion, some comments about the experimental reproducibility are reported. A series of studies have been made for evaluating the comparability of the results across various commercial and homemade microarray platforms, with contradictory results. A number of groups have reported limited concordance of results across platforms [30–32], raising the crucial

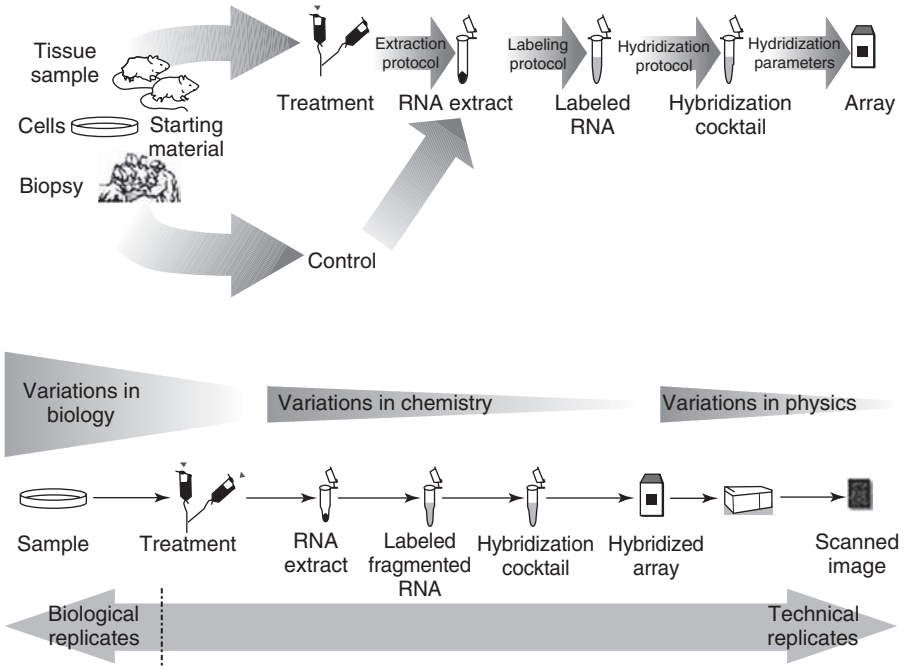


Figure 8.5 Source of variability in DNA microarray experiments.

question of the reliability of the DNA microarray techniques and their results. More recent studies have reached more positive conclusions about the possibility of comparing data coming from different centers or platforms, reinforcing the emerging concept that data treatment and choice of the comparison metric play a fundamental role in the analysis of gene expression data [5,33–36].

8.3 DATA MINING TECHNIQUES

To analyze the huge amount of data collected by microarray technologies, it is fundamental to select the most appropriate data mining instruments from statistics, artificial intelligence, signal analysis, patter recognition, and so on. Typically, data of microarray experiments coming from several arrays are related to different investigated phenotypes (e.g., normal and pathological subjects) or to the same subject under different conditions (e.g., after different drug treatments), or to different tissues. As already said, at the end of preprocessing, microarray data can be represented by a matrix (Fig. 8.4) in which rows represent genes or transcripts and columns represent the different arrays. The two matrix dimensions are very different, with the matrix having several thousands of rows (n) and tens/hundreds of columns (m). It is important to note that in microarray analysis, the distinction between variables and obser-

vations is not trivial: microarray data are an example of the so-called transposable data, that is, data in which variables depend on the question around which the experiment is built. Therefore, the data matrix can be analyzed both on the rows and on the columns in accordance with the main question formulated by the researcher.

8.3.1 Kinds of Experiments

In the literature, there is a large amount of microarray experiments that differ from one another for the goal of the investigation. To better understand which are the most suitable data mining techniques that have to be used in different situations, it is useful to subdivide the typical experiments in homogeneous groups on the basis of simple characteristics. First, experiments can be divided in static and dynamic ones. The first group includes experiments in which the time evolution is not explicitly investigated. The second one, instead, considers explicitly the evolution of the subject under study over time, monitoring its transcriptome over a time span. Classical experiments of the first class concern the study of two or more groups of subjects that differ from one another for one characteristic that is the object of the study. Classical examples are the comparison of samples coming from normal or pathological subjects, or cells subjected to different treatments or behavioral conditions (e.g., cancer cells subjected to an anticancer drug). In these cases, each column of the data matrix (corresponding to a different array) can be associated with one of the investigated conditions. Two different questions can be formulated in these studies: one reading the matrix on the rows and the other reading the matrix on the columns. In the first case, the question is gene oriented; that is, the investigation aims at finding genes that are differently expressed in the two or more investigated conditions. These studies correspond to the classical statistical studies of the biomedical research for group comparison (e.g., treated/untreated, case/control) conducted at cellular level. The data mining techniques used in these cases will be discussed in Section 8.3.2. These gene selection techniques present some interesting peculiarities and pose some specific problems mainly due to the large amount of genes and the few observations. In the second case, the question is oriented to the subjects; that is, the investigation aims at discovering the so-called molecular fingerprint of each group. The problem can be approached as a feature selection/classification problem by means of supervised techniques (Section 8.3.3). The main problem is again that the number of features (genes) is much larger than the number of observations in each class (arrays). To obtain a signature and to verify the goodness of the phenotypic class subdivision or the presence of possible genotypical subclasses, an unsupervised analysis can be also performed, neglecting the class information associated to each column. Arrays can be grouped by means of clustering techniques (Section 8.3.4) on the basis of the gene expression values. Therefore, an *a posteriori* evaluation and characterization of each cluster has to be made considering phenotypic class labels. Also in this situa-

tion, a molecular fingerprint can be derived, even if it has only a descriptive purpose and is not appropriate for classification because of overfitting problems.

About dynamic investigations, classical experiments are related to the temporal evolution study of the transcriptome of one subject in one or more different conditions. Classical examples are the study of the cell cycle, the differentiation and the development of an embryo, or the temporal effect of a specific substance as a drug on the cell activity. In these cases, each column of the data matrix (corresponding to a different array) can be associated with one of the time points of the investigated time span. Therefore, for each gene, an expression (temporal) profile can be immediately obtained by reading the matrix on the rows and by sorting the columns in ascending order. Also for dynamic studies, two different questions can be formulated: one reading the matrix on the rows and another one reading the matrix on the columns, even if the column-oriented analysis is in general less interesting and frequent. In the first case, the question is gene oriented; that is, the aim of the investigation is to find genes that move over the time span. Data mining techniques suitable to cope with this problem will be discussed in Section 8.3.2. They present some peculiarities from a statistical point of view, because, in time series, measurements cannot be considered as independent samples. It can also be of interest to group genes that show similar temporal profiles. Temporal clustering techniques have to be used and they are described in Section 8.3.4. In the second case, the question is instead oriented to the time points; that is, the aim of the investigation is to group time points that show similar gene expressions, clustering columns, and then, if it is of interest, to discover the molecular fingerprint characterizing each cluster. This problem can be approached as a feature selection/classification problem by means of supervised techniques (Section 8.3.3).

8.3.2 Gene Selection

The main problem in the analysis of microarray data is the high number of genes and the low number of arrays. Therefore, the first step of the analysis of gene expression data is gene selection. The main idea is to remove genes that are not significant for the analysis. For this purpose, several techniques have been proposed and are often combined together.

Low expressed genes can be eliminated because the measured signal is particularly noisy. Although the idea is simple, the choice of the threshold is not trivial and it is platform dependent. Each technological platform has adopted a method to make this selection. Affymetrix, for example, defines probe signals as absent/marginal/present on the basis of a complex procedure that requires, also but not only, that the detected signal is over a certain threshold. Applied Biosystem defines another criterion based on the combined analysis of the fluorescent and chemoiluminescent signals that both have to be over fixed thresholds. The selection criteria based on thresholds on the

detected probe values are based on a single array. However, the analysis we are interested in involves several arrays. In this way, additional gene selection criteria have been defined. For example, genes showing a limited variation across arrays are generally removed from the analysis, applying variational filtering methods, to avoid that variations due only to noise are attributed to the different investigated conditions.

In static experiments, the method usually adopted in the literature in the early papers on microarray data analysis is the fold change approach. It is still widely used nowadays when no replicated measurements are available. A gene is not filtered out and is considered as differentially expressed over the investigated conditions if its expression level changes of a factor overcoming a threshold, Th. In other words, a gene is differentially expressed if it happens that at least in two different arrays, $|\log(x_i) - \log(x_j)| > \text{Th}$. Note that this method fixes a threshold that is intensity independent. However, it is known that measurement errors are intensity dependent and that low values are noisier (see Fig. 8.6 as an example).

Therefore, in the presence of replicates, it is possible to adopt an advanced formulation of the fold change method, normalizing the difference of logs through the intensity-dependent standard deviation (SD) estimated from replicates [37]. Then we have $|\log(x_i) - \log(x_j)| > 2 \times \text{Th} \times \text{SD}_{(\log(x_i)+\log(x_j))/2}$.

In the presence of replicates, it was also proposed to use statistical tests to select differentially expressed genes, distinguishing them from differences that

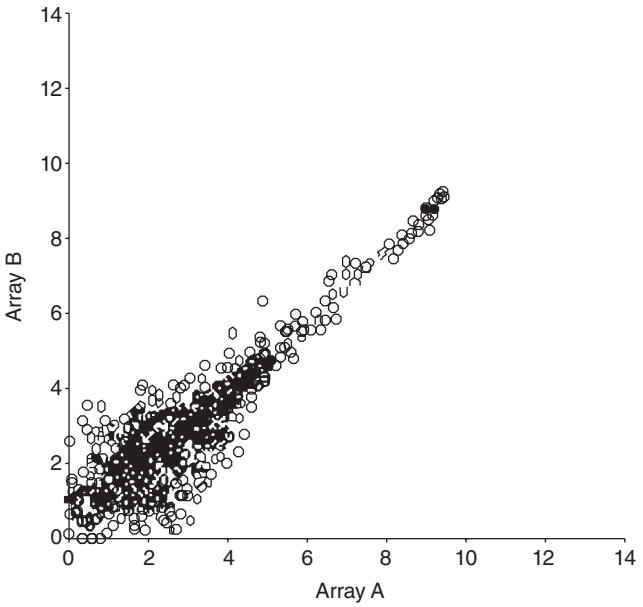


Figure 8.6 Expression levels in two technical replicates with Affymetrix microarrays (log scale on both axes).

occur only by chance. In particular, *t*-test (to compare two conditions) and analysis of variance (ANOVA) (to compare more than two groups) were used. Both tests assume that data are normally distributed and that the variance is the same in the different groups. Normal assumption has never been demonstrated in real data sets and seems to be particularly critical especially for Affymatrix data [28,37]. For these reasons also, nonparametric statistical tests as the Wilcoxon–Mann–Whitney rank-sum test or the Kruskal–Wallis test have been proposed. Another problem in using statistical tests is that the few replicates available in each experiment do not generally allow to estimate accurately the null hypothesis. For example, it is difficult to estimate the experimental variability and in particular its dependence from the signal intensity. Finally, the simultaneous application of statistical tests to thousands of genes requires a correction for multiple tests to control the type I error (i.e., the probability of considering as differentially expressed genes that are different only by chance). A widely used correction (even if particularly conservative) is the Bonferroni correction, which suggests to perform multiple tests dividing the desired significance level by the number of genes on which the test has to be applied. Being gene expression levels highly correlated or at least certainly not independent, this correction is very conservative, and the resulting test has a low potency. Note that the correction for multiple tests is particularly useful when genes have to be selected as biomarkers and then when the significance level is the main focus, whereas the potency is not an important requirement. On the other hand, if the goal is only a reduction in gene numbers to be considered in further analysis, potency is an important parameter and significance is less important. In other words, in the last situation, it is more important not excluding genes that are actually differently expressed [38].

Alternative to the Bonferroni correction, a plethora of modified *t*-tests, accounting for multiple comparisons, has been proposed, even if they are not very popular. An alternative approach to the significance level correction is the control of the false-positive rate or the false discovery rate, i.e., the number of the wrongly selected genes. Briefly, in this approach, the significance level of the statistical tests is fixed in order to obtain a desired false discovery rate. In other cases, permutation strategies were instead adopted to derive the null hypothesis. In particular, the null hypothesis is obtained by simulation, distributing randomly the arrays in the investigated groups several times without considering their real membership.

Dynamic experiments require a supplementary discussion. In fact, in these experiments, gene temporal profiles are collected without or at least with very few replicates. The selection is usually performed to obtain the list of genes that show an expression level in one or more time points significantly different from the baseline. The main problem of these experiments is that measurements at the different time points are not independent samples, being generated from the same dynamic process. Therefore, statistical tests are no longer valid. However, to investigate only the effect of a treatment (e.g., a drug) and

to remove other effects from the analysis, the problem is often reformulated as the comparison of two dynamics experiments, one of which is a reference situation (e.g., treated and control cells). In this study, genes that show a different dynamics between the two experiments have to be selected. Frequently, in these cases, the time series obtained through the point-to-point differences between the two investigated situations is analyzed, as it is the time series to be studied (similarly to classical matched pairs study).

The fold change method has been used also for dynamic study. In fact, it does not require any assumption on the independence of the measurements, being a nonstatistical method. Alternatively, the interquartile range (IQR) is used as a measurement of the time series variability. A low IQR value is synonymous with limited variability.

An ad hoc method has been developed to select genes in a dynamic experiment in the presence of case/control experiments and of a certain number of replicates [39]. More specifically, it is a statistical test based on the evaluation of the area between the two temporal profiles. The experimental area is then compared to that of the null hypothesis generated by assuming that the two temporal profiles are different only by chance. The null distribution is built randomly, generating a temporal profile starting from the analysis of the replicates.

8.3.3 Classification

Supervised classification methods adopt different strategies to derive rules (the classifier) able to establish the membership class of each example, minimizing classification error. The main techniques are linear discriminant analysis, support vector machines, naive Bayes classifiers, nearest neighbors, decision trees, induction of rules, and so on. Independent from the adopted specific classification technique, the principal rule is that the performance of a classifier has to be evaluated on a different data set from the one used to learn the classification rule. Therefore, in general, the experimental data are divided into two groups called learning and test sets. They are not necessarily of the same dimension, but they preferably contain each class in the same proportion. Moreover, to make the evaluation of the classifier independent from the specific test set and training set, several training and test sets are generated from the original data set and performances are averaged. Several methods are proposed to generate training and test sets. The 10-fold cross validation, in which the full data set is divided in 10 parts (9 used as training set and 1 as test set by turns), is one of the best known among these methods. However, in the microarray case, which suffers from the small number of cases (or examples), a more suitable technique is the leave-one-out method in which by turns, one case is considered as a test set and all the others are used as a training set. Bootstrap techniques can also be adopted to generate training sets and test sets. The classification performance expressed in terms of sensitivity and specificity is often represented by the receiver operating characteristic (ROC) curve.

Supervised classification methods have been widely applied for mining DNA microarray data to derive new prognostic and diagnostic models, in particular, in cancer research and in pharmacology and in functional genomics. Some examples are reported here.

In clinical application mainly related to the oncology, DNA microarrays have been widely applied to perform molecular classification. In this case, the classes are a certain number of mutually exclusive diseases. The classification problem is thus to find a decision rule that should correctly diagnose the patient's disease on the basis of the DNA microarray data. From a methodological viewpoint, this problem suffers from the "small *m*–large *n*" situation, i.e., a small number of cases (few patients, tens/hundreds) and a large number of classification attributes (many genes, tens of thousands). Thus, it is generally suitable to apply dimensionality reduction algorithms, such as principal component analysis or independent component analysis and/or gene selection methods, including statistical tests, as discussed in the previous section [40]. However, it is important that gene selection is made by using only the training set and not the whole data set to avoid overfitting. Note that, if gene selection is made as a part of a classification procedure that involves an iterative process on different training sets (like in cross-validation methods), for each run, the feature selection has to be remade on the current training set and thus the selected features can differ from one run to another. Alternatively, to overcome the problem of the large *n* and the small *m*, genes can be selected on the basis of the *a priori* knowledge, for example, focusing the attention on those genes that are involved in some pathways of interest [41].

After gene selection and dimensionality reduction, many algorithms have been proposed to perform molecular diagnosis. Support vector machines and random forests are nowadays considered as the state-of-the-art approach to deal with this class of problems.

Supervised classification algorithms are also applied to derive prognostic models from DNA microarrays, i.e., a prognosis on the outcomes of a certain disease on the basis of the molecular information coming from a certain patient. Many papers have been published in cancer research, although, due to the dimensionality problems previously mentioned, the model proposed has poor generalization properties and cannot be easily applied in clinic routine [42].

Another area of great interest from an application viewpoint is pharmacology, with particular reference to the oncology field. For example, the lymphoma leukemia project [43] has developed a method to predict survival after chemotherapy for diffuse large β-cell lymphoma. In this study, the gene expressions of 160 patients treated with antracycline chemotherapy were used to build a Cox survival model. The model was then tested on 80 patients, showing good performances in predicting 5-year survival and in providing interesting hypotheses on the patients who are good therapy responders.

Finally, in functional genomics, it is possible to build a training set with a number of gene expression profiles with known biological or molecular func-

tions. The training set is used to learn a set of decision rules that allow to classify genes with unknown function on the basis of their expression values. For example, Brown et al. [44] successfully applied support vector machines to the analysis of yeast gene expression data.

Recently, a classification method able to cope with replicates has been proposed to manage heterogeneity and uncertainty in a probabilistic framework [13]. It was originally applied to tissue microarrays, but it should be useful also in the case of DNA microarrays in the presence of replicate measurements.

8.3.4 Clustering

Clustering techniques are part of the standard bioinformatics pipeline in the analysis of DNA microarray data to group lines or rows of the matrix of data. Therefore, the main goals of cluster analysis are (1) finding groups of genes with similar expression profiles over different experimental conditions, including different time points or different patients; (2) finding groups of experimental conditions (patients, toxic agents) that are similar in terms of their genome-wide expression profiles. In the first case (functional genomics), the main hypothesis underlying the application of clustering methods is that genes with similar expression patterns, i.e., coexpressed genes, are involved in the same cellular processes [8]. In this case, in general, experiments collect samples over time. Sometimes, both genes and arrays are clustered in a two-step procedure.

All clustering approaches aim at finding a partition of a set of examples (genes) on the basis of a number of measurements (gene expression values); the partition corresponds to natural groups in the data or clusters. Clustering algorithms search partitions that satisfy two main criteria: (1) the internal cohesion; i.e., the examples of a cluster should be similar the others in the same cluster; and (2) the external separation; i.e., the examples of one cluster should be very different from the examples of every other cluster.

Among the different computational strategies proposed in the literature, we can distinguish three main classes of algorithms: (1) distance-based methods, (2) model-based methods, and (3) template-based methods. Below, a survey of these approaches is reported. In the following, we denote the set of expression measurements of the ith gene as $x_i = \{x_{i1}, \ldots, x_{im}\}$, where x_{ij} is the jth measurement, with $j = 1, \ldots, m$ and $i = 1, \ldots, n$. x_i will also be called expression profile of the ith gene.

8.3.4.1 Distance-Based Methods Clustering methods based on similarity are the most used approaches in the bioinformatics pipeline. These methods rely on the definition of a distance measure between gene expression profiles and group together genes with a low distance (or high similarity) between each other. The distance is computed in the m-dimensional space of the available measurements. The methods differ from one another for the adopted distance

measurement and for the strategy adopted to build up clusters. The most used distance measures are the Euclidean distance and the Pearson correlation coefficient. For what concerns the strategy applied to build up clusters, the two most popular families of methods are partitional clustering and hierarchical clustering.

In partitional clustering, the m-dimensional measurement space is divided in k regions, corresponding to the k clusters; the number of regions is often defined in advance by the data analyst. The different partitional clustering methods, including k-means, k-medoids, and self-organizing maps [40], have been largely applied to the analysis of gene expression data. The k-means algorithm and its variants in general start by (randomly) selecting different k cluster centers, assigning each gene to the nearest center. Then, the mean of each cluster is computed and centers are updated. The genes are then reassigned at the new centers and the algorithm is iterated until it converges to a stable subdivision. Self-organizing maps are a technique starting from building a low-dimensional map (in general, two-dimensional maps are considered) in which each point represents a cluster. Genes are then associated with one or more points of the map in such a way that the clusters that are close in the map are also similar in the original m-dimensional space.

On the other hand, hierarchical clustering algorithms are divided into agglomerative and divisive ones. The former starts with n groups of one element, corresponding to the n examples, and then through $n - 1$ consecutive steps, it progressively clusters the data into groups with a larger number of examples until a single cluster with n examples is obtained [40]. The latter starts with one group of n genes and progressively splits the data in smaller clusters until n clusters of one example are obtained. Agglomerative hierarchical clustering is the most used method in functional genomics [45]. When the collected data are time series, the Pearson correlation coefficient is used as a similarity metric. As a matter of fact, the (standardized) correlation between two gene profiles well describes the biological notion of coexpressed, and maybe coregulated, genes [45]: two genes are similar if their temporal profiles have the same "shape," even if the absolute values are very different. Moreover, the correlation similarity allows to cluster counterregulated genes.

The result of agglomerative clustering is depicted with a binary tree known as a dendrogram: the leaves of the dendrogram are the initial clusters with one example, while internal nodes represent the clusters obtained by grouping the examples that correspond to the child nodes. One of the main reasons of clustering technique success stands in the joint visualization of the dendrogram and of a color map (known as a heat map) of the gene expression levels in the different experimental conditions, as shown in Figure 8.7.

This type of visualization clearly shows the homogeneity regions of the gene expression profiles, highlighting the natural clusters in the data and giving the user the possibility to assess the quality of the clusters obtained from the

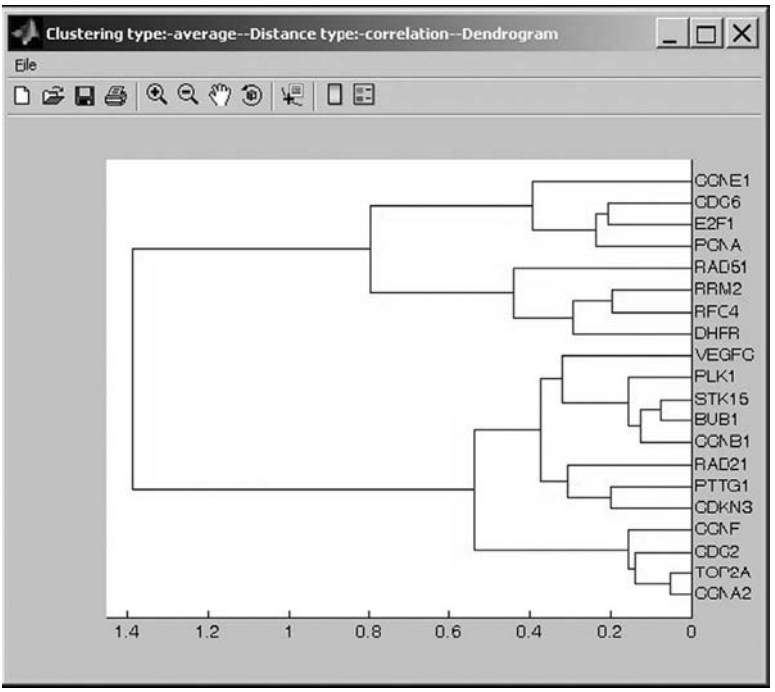


Figure 8.7 Results obtained by applying the hierarchical clustering algorithm to 20 genes involved in the human cell cycle (for details, see Reference 57). Genes are grouped with a decreasing level of similarity from the right to the left. The length of each branch of the tree (dendrogram) is directly proportional to the distance between the clusters that are grouped together: the more similar the clusters are, the smaller the branch is and vice versa.

algorithm. Hierarchical clustering and its tree visualization, although quite intuitive, are criticizable for the arbitrariness in determining the number of clusters.

8.3.4.2 Model-Based Clustering The use of distance-based methods for clustering gene expression time series may suffer from the failure of one of the assumptions underlying distance computation: the applied distance measures are usually invariant with respect to the order of measurements, assuming them to be independent from one another. This assumption is clearly not valid in the case of time series. Several alternative approaches have been proposed to deal with this problem, ranging from a transformation of the original time series to an alternative definition of the distance function [46]. However, given the nature of data, characterized by a small number of points (from some units to tens of measurements) and a small signal-to-noise ratio, an interesting solution is represented by a different class of clustering algo-

rithms, called model-based methods. The main assumption of model-based clustering is that the data are randomly extracted from a population made of a number of subpopulations, correspondent to the clusters, each one characterized by a different probability density function [47]. The subpopulations and their density functions are the model generating the data. Therefore, each gene profile x_i is assumed to be drawn from the probability distribution $f(x_i, \theta)$ given by

$$f(x_i, \theta) = \sum_{k=1}^{c} p_k f_k(x_i, \theta)$$

The clustering problem is thus transformed into a model-selection problem, which can be solved relying on probability and statistical modeling techniques. In time series analysis, each cluster is modeled by a different stochastic process, which is supposed to generate the data. Usually, it is assumed that all the time series can be described by the same class of stochastic process, and that the clusters differ from one another only because of different parameter values. Denoting with Y as the set of available examples (in our case, the $n \times m$ matrix of DNA microarray data), with M as the clustering model of the data, and with θ as the parameter of the stochastic model generating the data, there are two main approaches for model selection that have been proposed in the literature: (1) the maximum likelihood approach, which searches the model that maximizes the likelihood function $p(Y \mid \theta, M)$, i.e., the probability of the data given the model M and the parameter θ; and (2) the Bayesian approach, which searches the model that maximizes $p(M \mid Y)$, i.e., the posterior probability of the model M given the data Y.

In the maximum likelihood approach, the likelihood in general cannot be directly maximized and the expectation maximization strategy can be usefully adopted. The problem is iteratively solved through a two-step procedure: in the first one called E step, the probability that x_i belongs to the kth cluster is computed fixing the model parameters; in the second one, called M step, the parameter estimates are properly updated [48]. The number of cluster can be fixed in advance or chosen by cross validation.

In the Bayesian approach, instead, the posterior probability of a model is computed by the Bayes theorem. The posterior distribution $p(M \mid Y)$ is proportional to the product of the marginal likelihood $p(Y \mid M)$ and of the prior distribution $p(M)$. The prior distribution $p(M)$ is the estimate of the probability of each model M, before having observed the data. The marginal likelihood $p(Y \mid M)$ is a function of θ and M as follows:

$$p(Y \mid M) = \int p(Y \mid M, \theta) p(M) \, d\theta \tag{8.1}$$

Although the Bayesian approach allows to exploit prior information on the clusters, all the models are usually considered *a priori* equally likely and, thus, the marginal likelihood is maximized. If the number of available data $N = n \times m$ is high, maximizing the marginal likelihood is equivalent to find a compromise

between the likelihood of a model and the number of its parameters. In fact, there exists a theorem showing that, if $N \to \infty$,

$$\log(P(Y \mid M)) = \log(P(Y \mid M, \theta_M)) - \frac{1}{2}\log(N)\nu(M) + O(1),$$

where $\nu(M)$ is the number of the degrees of freedom of the model M and θ_M is the estimate of the model parameters [49]. If we choose θ_M as the maximum likelihood estimate, the Bayesian approach looks for models with high likelihood and low dimensionality.

Model-based clustering has been successfully applied to cluster gene expression time series. The CAGED [50] software is one of the most interesting tools for clustering time series in functional genomics. CAGED assumes that the time series are generated by an unknown number of autoregressive stochastic processes. This assumption, together with a number of hypotheses on the probability distribution of the autoregressive model parameters and of the measurement error, allows to compute in close form the integral of Equation 8.1, i.e., the marginal likelihood, for each model M and thus for each possible clustering of the data. Since it is computationally unfeasible to generate and to compare all possible models, it is necessary to couple marginal likelihood computation with an efficient search strategy in the cluster space. For this purpose, CAGED exploits an agglomerative procedure, similar to the one used in hierarchical clustering. The time series are iteratively clustered, selecting at each step the aggregation that maximizes the marginal likelihood. In this way, CAGED is able to select the optimal number of clusters by ranking the marginal likelihood of each level of the hierarchy. Finally, the results can be shown in the same way of hierarchical clustering, with a dendrogram coupled with a heat map.

Recently, different methods have been proposed to improve the CAGED approach by relaxing some of its hypotheses, such as the stationary of the process generating the data or the regular sampling time grid. In particular, more general stochastic processes have been applied, such as random walks [51] or hidden Markov models [52].

In conclusion, the main advantages of the model-based clustering algorithms when used to analyze dynamic DNA microarray experiments are (1) an explicit description of the autocorrelation model of the data, (2) the ability of working also with very short time series, (3) the possibility of managing missing data in a sound probabilistic framework, and (4) the opportunity to determine automatically the optimal number of clusters on the basis of data without being forced to fix it in advance. The main problems are related to the necessity of assuming a reasonable model generating the data and the poor computational efficiency.

8.3.4.3 Template-Based Clustering Gene expression time series are usually characterized by a small number of time points. A recent review has

shown that more than 80% of the time series available in the Stanford Microarray Database has a number of points that are smaller or are equal to 8. The main reasons are related to the high cost and high complexity of those experiments. Since the data are also noisy, alternative clustering strategies have been investigated. One of those strategies is to group the time series on the basis of the matching of the series with a pattern or a template, which may have qualitative characteristics, such as the presence of an increasing or decreasing trend, of an up and down behavior. If the templates are already available, the clustering problem becomes a pattern-matching one, which can be also carried on with qualitative templates [53]. In most of the cases, the templates are not available and the template-based clustering approaches must automatically find the qualitative templates in the data. For example, the method proposed in Reference 54 and implemented in the software STEM [55] starts by enumerating all possible qualitative patterns of a gene profile of m time points, given the parameter c, which represents the possible unit changes of each gene from a time point to the next one. For example, if $c = 2$, each gene may increase or decrease of one or two qualitative units from one point to the next (or to remain steady). This allows to generate $(2c + 1)(m - 1)$ qualitative templates. The second step of the algorithm reduces the number of such templates to a number k predefined by the user. The reduction is performed by clustering the qualitative profiles on the basis of their mutual distance. After this step, the original time series are assigned to the k clusters with a nearest neighbor strategy. The Pearson correlation is used as a similarity function. Finally, the number of clusters is further reduced by (1) computing the statistical significance of each group, through a permutation-based test, and by (2) eventually aggregating the remaining clusters that are closer than a predefined threshold.

Another template-based approach was proposed in Reference 53, where the time series data are modeled as a set of consecutive trend temporal abstractions, i.e., intervals in which one of the increasing, decreasing, steady templates is verified. Clustering is then performed in an efficient way at three different levels of aggregation of the qualitative labels. At the first level, gene expression time series with the same sequence of increasing or decreasing patterns are clustered together. At the second level, time series with the same sequence of increasing, steady, or decreasing patterns are grouped, while at the third level, the time series sharing the same labels on the same time intervals are clustered together. The results of this method, known as temporal abstraction (TA) clustering, can be visualized as a three-level hierarchical tree and thus it is easy to be interpreted. Finally, an interesting knowledge-based template clustering has been presented by Hvidsten et al. [56]. In their work, whose main goal was to find descriptive rules about the behavior of functional classes, they grouped and summarized the available gene expression time series by resorting to template-based clustering. They first enumerated all possible subintervals in the time series and labeled them as increasing, decreasing, and steady with a temporal abstraction-like procedure. Then, they

clustered together genes matching the same templates over the same subintervals. In this way, a single gene may be present in more than one cluster.

Rather interestingly, the different clustering approaches can be now applied in an integrated way thanks to TimeClust, a new software tool freely downloadable, which allows the clustering of time series of gene expression data with distance-based, model-based, and template-based methods [57].

8.4 SUMMARY

The high-throughput technologies for the measurements of the gene expression represent an interesting opportunity but also propose a new challenge for drug discovery, evaluation, and administration processes. The main opportunity is the possibility to overcome the classical reductionist paradigm that studies few genes at the same time, allowing the investigation of the whole transcriptome. On the other hand, these techniques create the problem of managing, analyzing, modeling and interpreting huge volumes of data. The complexity of this challenge is not only computational but is also due to the necessity to put microarray data in the context of the post-genomics information. In this chapter, an overview of the main data mining techniques, useful to analyze microarray data, has been presented. It ranges from image acquisition and preprocessing to supervised and unsupervised classification, through gene selection procedures. Although these techniques are very useful to select a molecular fingerprint of cellular life, it is necessary to remember that the functional interpretation of that molecular fingerprint is still a manual process and represents one of the most important obstacles to the full efficacy of the microarray technology. In fact, the interpretation of the results and the formulation of the hypothesis about biological mechanisms at the basis of the cellular behavior, monitored by microarrays, require the integration of information about gene annotations and descriptions, about metabolic and cellular pathways, in which genes are involved, about their physical position on the genome, and so on. [41].

Despite their widespread use, DNA microarrays have limitations that researchers must consider. First of all, the analysis of the transcriptome is based on three hard assumptions: (1) there is a close correspondence between mRNA transcription and its associated protein translation; (2) all mRNA transcripts have an identical life span; and (3) all cellular activities and responses are entirely programmed by transcriptional events. Actually, mRNA activity and induced levels of proteins are not always well aligned. Translational and posttranslational regulatory mechanisms that affect the activity of various cellular proteins are not examined by the analysis of the transcriptome and thus by DNA microarrays. The promising field of proteomics is starting to address these issues, for example, by using proteomic microarray. Moreover, differential gene expression analysis is not a stand-alone technique; results must be confirmed through direct examination of selected genes. These analy-

ses are typically done at the level of RNA blot or quantitative RT-PCR, to examine transcripts of specific genes, and/or at the protein level, analyzing protein concentration using immunoblots or enzyme-linked immunosorbent assay (ELISA).

However, DNA microarrays are expected to become a routine, and they are widely used for disease diagnosis and classification, which anticipates the future availability of home testing kits, for example, in cancer. Eventually, microarrays could be used as a routine diagnostic tool with which treatments could be tailored for an individual patient [4].

Moreover, the use of microarrays for target identification and validation is currently being explored. The potential discovery of a gene, which, when knocked down, destroys only cancer cells, could indicate an approach for a new cancer therapy. The combination of DNA microarray analysis with the RNAi technology is a very powerful tool for drug discovery [58]. The use of cell microarrays for large-scale RNA interference studies should improve research in this field [59].

REFERENCES

1. Debouck C, Goodfellow PN. DNA microarrays in drug discovery and development. *Nat Genet* 1999;21:48–50.
2. Gardner TS, di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 2003;301: 102–105.
3. Nakatsu N, Yoshida Y, Yamazaki K, Nakamura T, Dan S, Fukui Y, Yamori T. Chemosensitive profile of cancer cell lines and identification of genes determining chemosensitivity by an integrated bioinformatical approach using cDNA arrays. *Mol Cancer Ther* 2005;4:399–412.
4. Jayapal M, Melendez AJ. DNA microarray technology for target identification and validation. *Clin Exp Pharmacol Physiol* 2006;33:496–503.
5. Bosotti R, Locatelli G, Healy S, Scacheri E, Sartori L, Mercurio C, Calogero R, Isacchi A. Cross platform microarray analysis for robust identification of differentially expressed genes. *BMC Bioinformatics* 2007;8(Suppl 1):S5.
6. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B,

Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. *Science* 2001;291:1304–1351.

7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki

K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ; International Human Genome Sequencing Consortium. Initial sequencing and analysis of human genome. *Nature* 2001;409:860–921.

8. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. *Nat Genet* 1999;21(Suppl 1):33–37.
9. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;270: 467–470.
10. Lockhart DJ, Dong H, Byrne MC, Folletie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14: 1675–1680.
11. Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 1998;4:844–847.
12. Braunschweig T, Chung JY, Hewitt SM. Tissue microarrays: Bridging the gap between research and the clinic. *Expert Rev Proteomics* 2005;2:325–336.
13. Demichelis F, Magni P, Piergiorgi P, Rubin MA, Bellazzi R. A hierarchical Naive Bayes model for handling the uncertainty in classification problems: An application to tissue microarrays. *BMC Bioinformatics* 2006;7:514.1–514.12.
14. MacBeath G. Protein microarrays and proteomics. *Nat Genet* 2002;32:526–532.
15. Solinas-Toldo S, Lampel S, Nickolenko S, Stilgenbauer J, Benner A, Lichter H, Dohner T, Cremer P. Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* 1997;20:399–407.
16. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* 1998;20:207–211.
17. Ulger C, Toruner GA, Alkan M, Mohammed M, Damani S, Kang J, Galante A, Aviv H, Soteropoulos P, Tolias PP, Schwalb MN, Dermody JJ. Comprehensive genome-wide comparison of DNA and RNA level scan using microarray technology for identification of candidate cancer-related genes in the HL-60 cell line. *Cancer Genet Cytogenet* 2003;147:28–35.
18. Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, Liu W, Yang G, Di X, Ryder T, He Z, Surti U, Phillips MS, Boyce-

Jacino MT, Fodor SP, Jones KW. Large-scale genotyping of complex DNA. *Nat Biotechnol* 2003;21:1233–1237.

19. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 2001;19:342–347.
20. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biol* 2001;2:1–11.
21. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003;31:1–8.
22. Naef F, Socci ND, Magnasc M. A study of accuracy and precision in oligonucleotide arrays: Extracting more signal at large concentrations. *Bioinformatics* 2003; 19:178–184.
23. Affymetrix. *Genechip Expression Analysis Technical Manual*. Santa Clara, CA: Affymetrix Inc., 2004.
24. Hoffmann R, Seidl T, Dugas M. Profound effect of normalization of detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol* 2002;3:1–11.
25. Amaratunga D, Cabrera J. Analysis of data from viral DNA microchips. *J Am Stat Assoc* 2001;96:1161–1170.
26. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003;19:185–193.
27. Yang Y, Dudoit S, Luu P, Speed TP. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 2002;30:e16.
28. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003;4:249–264.
29. Magni P, Simeone A, Healy S, Isacchi A, Bosotti R. Summarizing probe levels of Affymetrix arrays taking into account day-to-day variability. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (forthcoming).
30. Tan PK, Downey TJ, Spitznagel EL, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* 2003;31:5676–5684.
31. Jarvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi O, Monni O. Are data from different gene expression microarray platform comparable? *Genomics* 2004;83:1164–1168.
32. Kothapalli R, Yoder SJ, Mane S, Loughran TP. Microarrays results: How accurate are they? *BMC Bioinformatics* 2002;3:22.
33. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JGN, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martinez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W. Multiple-laboratory comparison of microarrays platforms. *Nat Methods* 2005;2:1–5.

34. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbuch J. Independence and reproducibility across microarray platforms. *Nat Methods* 2005;2:337–344.
35. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res* 2005;33:5914–5923.
36. Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, Bumgarner RE, Bushel PR, Chaturvedi K, Choi D, Cunningham ML, Deng S, Dressman HK, Fannin RD, Farin FM, Freedman JH, Fry RC, Harper A, Humble MC, Hurban P, Kavanagh TJ, Kaufmann WK, Kerr KF, Jing L, Lapidus JA, Lasarev MR, Li J, Li YJ, Lobenhofer EK, Lu X, Malek RL, Milton S, Nagalla SR, O'malley JP, Palmer VS, Pattee P, Paules RS, Perou CM, Phillips K, Qin LX, Qiu Y, Quigley SD, Rodland M, Rusyn I, Samson LD, Schwartz DA, Shi Y, Shin JL, Sieber SO, Slifer S, Speer MC, Spencer PS, Sproles DI, Swenberg JA, Suk WA, Sullivan RC, Tian R, Tennant RW, Todd SA, Tucker CJ, Van Houten B, Weis BK, Xuan S, Zarbl H; Members of the Toxicogenomics Research Consortium. Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* 2005;2:351–356.
37. Tu Y, Stolovitzky G, Klein U. Quantitative noise analysis for gene expression microarray experiment. *Proc Natl Acad Sci USA* 2002;99:14031–14036.
38. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Statistical Science* 2003;18:71–103.
39. Di Camillo B, Toffolo G, Nair SK, Greenlund LJ, Cobelli C. Significance analysis of microarray transcript levels in time series experiments. *BMC Bioinformatics* 2007;8(Suppl 1):S10.
40. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer, 2001.
41. Bellazzi R, Zupan B. Towards knowledge-based gene expression data mining. *J Biomed Inform* 2007;40:787–802.
42. 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;415:530–536.
43. Rosenwald A, Wright G, Chan W, Connors J, Campo E, Fisher R, Gascoyne R, Muller-Hermelink H, Smeland E, Staudt L. The use of molecular profiling to predict survival after chemotherapy for diffuse large β–cell lymphoma. *N Engl J Med* 2002;346:1937–1947.
44. Brown M, Grundy W, Lin D, Cristianini N, Sugnet C, Furey T, Ares M, Haussler D. Knowledge-based analysis of microarray gene-expression data by using support vector machines. *Proc Natl Acad Sci USA* 2000;97:262–267.
45. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998;95:863–868.
46. Aach J, Church G. Aligning gene expression time series with time warping algorithms. *Bioinformatics* 2001;17:495–508.
47. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. *Bioinformatics* 2001;17:977–987.

48. Qin ZS. Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics* 2006;22:1988–1997.
49. Schwartz G. Estimating the dimension of a model. *Ann Stat* 1978;6:461–464.
50. Ramoni M, Sebastiani P, Kohane IS. Cluster analysis of gene expression dynamics. *Proc Natl Acad Sci U S A* 2002;99:9121–9126.
51. Ferrazzi F, Magni P, Bellazzi R. Random walk models for Bayesian clustering of gene expression profiles. *Appl Bioinformatics* 2005;4:263–276.
52. Schliep A, Schunhuth A, Steinhoff C. Using hidden Markov models to analyze gene expression time course data. *Bioinformatics* 2003;20:i255–i263.
53. Sacchi L, Bellazzi R, Larizza C, Magni P, Curk T, Petrovic U, Zupan B. TA-Clustering: cluster analysis of gene expression profiles through temporal abstractions. *Int J Med Inform* 2005;74:505–517.
54. Ernst J, Nau GJ, Bar-Joseph Z. Clustering short time series gene expression data. *Bioinformatics* 2005;21:i159–i168.
55. Ernst J, Bar-Joseph Z. STEM: A tool for the analysis of short time series gene expression data. *BMC Bioinformatics* 2006;7:191.
56. Hvidsten TR, Laegreid A, Komorowski J. Learning rule-based models of biological process from gene expression time profiles using gene ontology. *Bioinformatics* 2003;19:1116–1123.
57. Magni P, Ferrazzi F, Sacchi L, Bellazzi R. TimeClust: A clustering tool for gene expression time series. *Bioinformatics* 2008;24:430-432.
58. Wheeler DB, Bailey SN, Guertin DA, Carpenter AE, Higgins CO, Sabatini DM. RNAi living-cell microarrays for loss-of-function screens in *Drosophila melanogaster* cells. *Nat Methods* 2004;1:127–132.
59. White J. The future of microarray readers. *Pharm Discov* 2004;30–34.

9

BIOINFORMATICS APPROACHES FOR ANALYSIS OF PROTEIN–LIGAND INTERACTIONS

MUNAZAH ANDRABI, CHIOKO NAGAO, KENJI MIZUGUCHI, AND SHANDAR AHMAD

Table of Contents

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

9.1 INTRODUCTION

Protein–ligand interactions are essential to all aspects of eukaryotic functions. A thorough understanding of such interactions is likely to provide us with a technology to better understand the mechanism of disease, to design novel drugs, and to control biological functions. Bioinformatics has played a significant role in compiling information on known protein–ligand interactions and has made it possible to share and query that information, to predict interaction regions and nature of interactions, and, finally, to design new molecules with desired interaction properties. In this chapter, we provide an overview of bioinformatics approaches to studying protein–ligand interactions and discuss some of the problems facing this enormously important subject of research.

9.2 LIGANDS IN BIOINFORMATICS

9.2.1 Definition of a Ligand

In the simplest sense, the term ligand in the current context refers to a (typically soluble) molecule that binds to another biological molecule to perform or to inhibit a specific (or nonspecific) function. The corresponding molecule

to which a ligand binds is called a receptor. The term ligand in chemistry is used to describe an atom (or a group of atoms) that is bound to a central (typically a metallic) atom in another molecule [1,2]. Its biochemical definition is more permissive and states as follows [2]:

> If it is possible *or convenient* to regard *part of* a polyatomic molecular entity as central, then the atoms, groups *or molecules* bound to that part are called ligands.

Centrality here does not necessarily reflect the geometric nature of a molecule but may refer to the active site or a functionally important region in a protein or DNA. Thus, a ligand may be most generally thought of as an atom or a molecule, attached to some specific location in a protein or DNA. For the purpose of elucidating protein–ligand interaction, the term may imply a metallic ion such as zinc or iron, a small molecule such as carbohydrate, another single protein such as a hormone or a neurotransmitter, or even a protein–protein complex of several peptide chains. In this widest sense of the term, ligands include DNA, RNA, and proteins. The process of attachment or binding of a ligand atom or molecule is correspondingly called ligation, and a ligand is said to be ligating its receptor during its activity. The arbitrariness or context dependence of the term ligand is obvious from the fact that the same atom or molecule may be differently called ligand or receptor in a sequence of biological events, depending upon the stage of the event being referred to. As an example taken from Reference 2,

> four calcium ions are ligands for calmodulin, when the protein is regarded as central; four carboxylate groups of calmodulin ligate (are ligands of) each calcium ion when this ion is regarded as central. It is the ligand that is said to ligate the central entity, which is said to be ligated.

It is clear that a ligand simply needs to be attached to a receptor, and the nature of attachment is not the main concern in designating it as a ligand. Indeed ligands are bound to their receptors or ligate other molecules utilizing all kinds of interactions ranging from ionic (as in the case of metallic ions), covalent (e.g., Gly-Tyr-Phe domains attached to proline-rich peptides) to hydrophobic (as in most protein–protein interactions) and hydrogen bonded (as in many protein–sugar and protein–DNA pairs).

Likewise, as is obvious from the above discussion, there is no limit on the size of an atom or a molecule to be called a ligand, as protein–protein complexes composed of thousands of atoms qualify as ligands just as single-atom ions like zinc and copper do. In the next sections, we take a look at some examples of protein–ligand interactions, illustrating the diversity of molecular structures and biological functions in which protein–ligand interactions are observed.

9.2.2 Inhibition of Enzyme Activity

Enzymes catalyze biochemical reactions by specifically acting on their substrate molecules. Most enzymes have a specific smaller region that works as an active site and participates in the principal (enzymatic) activity. Another small molecule attached to these active sites may inhibit or activate this process. The Kyoto Encyclopedia of Genes and Genomes (KEGG) contains a database of ligands, which is linked to enzymes, providing information about enzyme activity and inhibition by ligands. [3,4]. For example, an enzyme produced by malignant cells, cancer procoagulant, acts on peptide bonds and cleaves the Arg–Ile bond in factor X to form factor Xa [5]. The activity of this enzyme can be inhibited by peptidyl diazomethyl ketones and peptidyl sulfonium salts [6]. In fact, peptidyl diazomethyl ketones are a class of ligands that specifically bind to proteinase enzymes, and the specificity in these compounds is provided by two or three amino acids in its molecule, generally represented as Z-R1-R2-R3, where Z is the ketone functional group and R1, R2, and R3 are amino acid residues attached by a peptide bond (R3 is optional) [7,8].

9.2.3 Metal Ligands

In some proteins, single metallic ions such as zinc and copper are essential to the function of a protein to which they are attached as a ligand (see, for example, Harris [9] and Karlin et al. [10]). Significant conformational changes may occur in proteins due to the metallic ligand binding (e.g., Qin et al. [11]). These conformational changes throw a huge challenge for the comparative modeling of protein structures from sequence and are therefore a subject of intensive research. As a specific example, significant conformational changes were observed when a zinc ion was replaced by copper in a zinc finger DNA-binding protein [12]. Another class of DNA-binding proteins, called histones, is often reported to have a metal-binding domain [13]. Another example of metal binding to proteins is that of a calcium-binding protein, calmodulin [14,15]. Figure 9.1 shows the geometric arrangement of such interaction. Interaction of metals with proteins is also central in the studies on toxicity by mercury and other heavy metals [16–18].

9.2.4 Carbohydrate Ligands

Carbohydrate ligands interact with proteins in a wide range of biological processes such as infection by invading microorganisms and the subsequent immune response, leukocyte trafficking and infiltration, and tumor metastasis [19–27]. Bioinformatics approaches to studying these interactions are still in the early stages, particularly because few protein–carbohydrate complex structures are available and there is still a long way to predict the exact nature of these interactions. Although there are bioinformatics solutions, some of which predict binding sites and others can attach sugars to structure models, many more issues remain to be addressed [28,29].

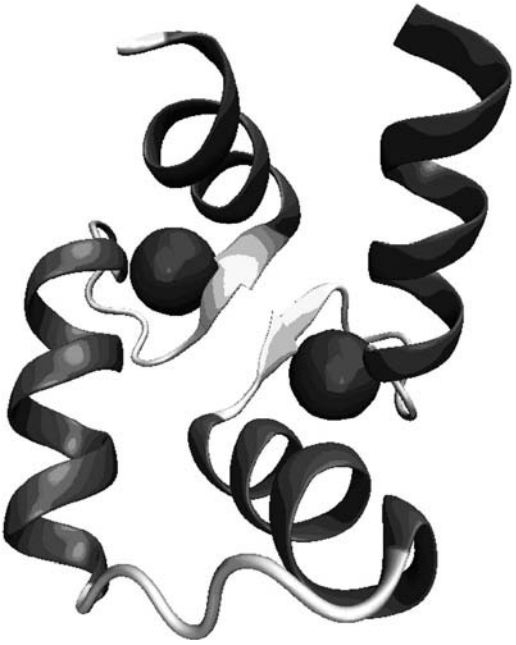


Figure 9.1 C-terminal of calcium-bound calmodulin protein (PDBID 1J7P). See color insert.

9.2.5 Other Small Molecules as Ligands

There are a large number of other ligands that are known to interact with proteins in various functions like charge transport, energy storage, and controlling an enzymatic action. Examples include ATP, phosphate ions, sulfate ions, nitrate, oxygen, and carbon monoxide [30–35]. Some statistics on their available structure complexes are presented in a later section. Here, it is enough to bear in mind that a ligand is not a homogeneous or similarly acting group of molecules but refers to almost an entire range of organic and inorganic substances that interact with biological molecules in general but proteins in particular.

9.2.6 Protein Ligands

Many proteins interact with other proteins forming a ligand–receptor pair, which is one of the most common types of protein–protein interactions (higher-order oligomerization may be an example of another type of protein–protein interaction, not included in the ligand–receptor category). Protein–protein interactions in the signal transduction pathways form a typical example. A specific example of proteins involved in signaling could be that of transmembrane proteins, G protein-coupled receptors (GPCRs) that receive signals

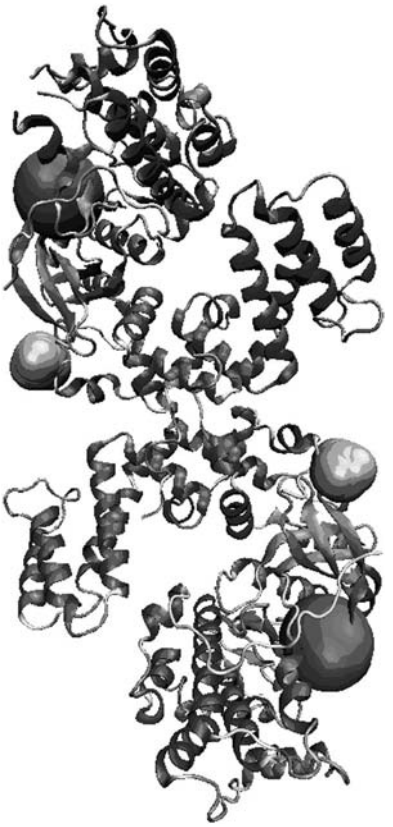


Figure 9.2 G protein-coupled receptor kinase 6 bound to ligands Mg (red) and PO4 (green) (PDBID 2ACX). See color insert.

from hormone proteins (ligands) [36]. GPCRs and their interactions with protein ligands are of enormous pharmaceutical interest as they form one of the most common targets of modern drugs [37]. Many times hormone–receptor interactions in GPCRs are accompanied by interactions with other (small) ligands, making the study of protein–ligand interactions much more complex and diverse (see example in Fig. 9.2).

9.2.7 DNA and RNA Ligands

DNA and RNA are well-known molecules carrying all the hereditary information in living organisms. Formation of proteins based on information encoded in DNA and with support from relevant RNA molecules is a task central to the origin, development, and regeneration of living systems. However, gene expression is tightly controlled by proteins such as transcription factors, enzymes, and histones. These proteins interact with nucleic acids in a specific

or nonspecific manner and the study of such interactions is naturally of crucial importance. Structures of thousands of proteins, complexed with DNA and RNA, are now available in the Protein Data Bank (PDB), which allows a thorough inspection of their interaction geometries [38–42]. Just as the most protein–nucleic acid interactions that occur on a specific location in DNA, the interaction site on the protein is also localized and binding residues on these proteins may also be clearly identified [39,43]. In this way, calling the protein or the nucleic acid as ligand and the other as target or ligated molecule seems to be largely a matter of syntax, and for the purpose of the current discussion, nucleic acid is treated as a ligand.

Based on proteins' ability to interact, several synthetic ligands have been designed, which specifically interact with proteins. Short fragments of nucleic acids designed for interaction with target proteins are termed as DNA- or RNA aptamers (we call them simply aptamers in the current discussion, although we are aware that there are also peptide aptamers, which we shall specifically mention, if discussed). The typical target for interactions with DNA aptamers are proteins such as thrombin, PKCδ, and platelet-derived growth factors [44–46] (Fig. 9.3).

Synthetic DNA aptamers have also been shown to specifically recognize and bind to other molecules such as carbohydrates (e.g., Yang et al. [47]). However, current review has been limited to study DNA ligands interacting with proteins only.

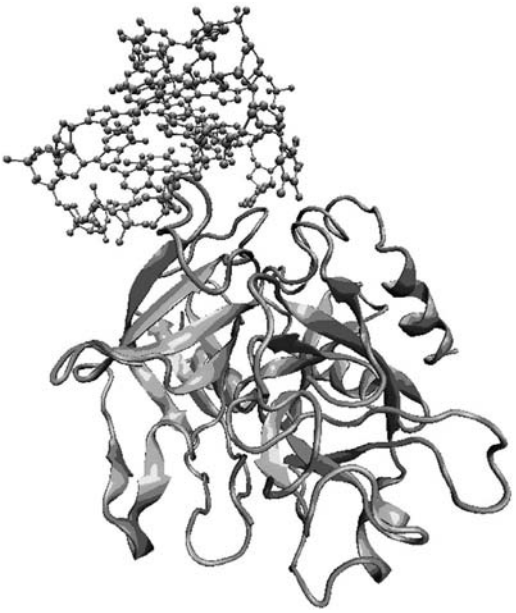


Figure 9.3 Thrombin-binding DNA aptamer (PDBID: 1HAP). See color insert.

9.3 REPRESENTATION AND VISUALIZATION OF LIGANDS

With such a large number of ligands in play, it is a huge task to even name them in a systematic way. Just like naming any other biological entity such as an organism or a gene, different names for the same ligands may also be found. Thus, it is utterly important to have a uniform system that not only identifies ligands by names but may also provide basic information about atomic arrangements within them. Larger ligand molecules such as DNA and proteins follow the naming convention of their own; smaller ones may not be so easy to characterize. The situation may become more complex for referring to a particular atom or a functional group within the ligand as the names and numbering assigned to them may vary in different systems of representation. More importantly, information about chemical bonds and branching of molecules should be readable by computers, for any large-scale processing such as filtering and screening, which form a very important area of application for studying ligands for their potential application as drugs. It is therefore necessary to take a brief look at the representation schemes and conventions that are most reliably used to identify and represent ligands.

9.3.1 Linear Text-Based Representations

The simplest way to present details of atomic arrangements within a molecule is to use one-dimensional character strings. For the most efficient methods of writing, these methods should follow some basic principles:

1. One molecule should be represented by one string.
2. Representation should carry information about the atomic arrangements and branching of chains and functional groups.
3. There should be a way to represent most common types of bonds such as single, double, and triple bonds.
4. Representations should be compact and should use minimum possible characters.
5. Representations should (preferably) be unique, so that two-dimensional (2-D) drawings, which are shown in various orientations, should have only one unique string corresponding to them. This imposes some standards on where to start writing the string and which branches of a chain should get preference.

Realizing the importance of these methods of representations, many linear input systems have been developed, both by individual groups as well as by large international committees such as the International Union of Pure and Applied Chemistry (IUPAC). Some of them are listed here.

9.3.2 Simplified Molecular Input Line Entry System (SMILES)

This method of representing 2-D information of molecules on a one-dimensional string was first proposed by Weininger in 1988 [48] and was further elaborated by Weininger et al. in 1989 [49].

Despite a significant contribution of SMILES in standardizing alphanumeric notations for representing molecules, there remain some issues in the finer aspects of details in their arrangements, hybridization state, chirality, and so on. Thus, ever since the first standards were proposed by Weininger, modifications were made, both to resolve ambiguities as well as to provide additional information. For a long time, a commercial company, Daylight Chemical Information Systems, Inc. (http://www.daylight.com/), has led the standards and specifications of complete molecular notations in SMILES format. A recent effort by a community project called OpenSmiles (http://www.opensmiles.org/) has now been started to make a noncommercial and publicly available set of standards for SMILES. Elaborate details of proposed conventions and standards are available on their respective web sites, and brief summary of general principles in SMILES notation is given below.

SMILES notation recognizes that a molecule can be described by means of (1) the names of atoms, (2) linkages or chemical bond types between atoms, (3) spatial atomic arrangements such as branching of a molecule, and (4) aromatic character of connectivity, which requires additional information about chain closing. In addition to this, more elaborate ideas such as stereochemistry (trans/cis nature of atomic arrangements), chirality, and isotope information may be added to give further details about the molecule. To achieve this, standard naming conventions are developed. For example, atoms are largely represented by their standard symbols (e.g., Na for sodium). To represent bonds, single bonds are assumed default and a notation is omitted; double bonds are shown by the "=" sign and a triple bond by "#." Hydrogen atoms on carbon are not shown and are inferred from the bond notations. Thus, for example, SMILES notations for methane and ethane are "C" and "CC," respectively, only; ethene is shown as "C=C" and ethyne as "C#C." Table 9.1 shows some more examples of molecules and their corresponding SMILES notation. Sometimes, group of atoms as a functional group are

TABLE 9.1 SMILES Representations of Some Simple Molecules

Common Name	Chemical Formula	SMILES
Ethane	CH_3–CH_3	CC
2-Methyl pentane	CH_3–$CH(CH_3)$–CH_2–CH_2–CH_2	CC(C)CCC
Ethanoic acid	CH_3–CH_2–COOH	CCC(=O)O
Benzene	C_6H_6	cccccc
Oxygen molecules	O_2	OO

needed to describe molecules and square brackets are used to do so, as also to denote the isotopes, e.g., [OH] shows alcohol and [13C] shows C13 isotope of carbon. Branching is shown by standard parentheses; e.g., CC (#N)(C)C is a notation for a (#N), (C), and C as three branches attached to CC. Further details may be seen in the user manuals from Daylight or OpenSMILES.

9.3.3 SMILES Arbitrary Target Specification (SMARTS)

Denoting molecules serves many purposes well but does not help if one wants to search through a large database of molecules based on small structural element within them, because small substructure information, most useful for molecular screening, may have been lost in the overall SMILES notation of the molecule. Moreover, an unambiguous standard to describe what one is looking for in a list of molecules needs to be defined. This task is achieved by an extension of SMILES, called SMARTS. As the name suggests an arbitrary target within a molecule can be searched by describing its substructure in terms of SMILES. Remember that SMILES represents a complete molecule and cannot be used to describe atoms without describing their complete bonding environment, which is not so for SMARTS. This notation is also maintained by Daylight Chemical Information Systems, Inc., and the detailed set of standards can be found on their website (http://www.daylight.com/). There exists a variant of SMARTS, developed by OpenEye Scientific Software, Inc. (http://www.eyesopen.com/). One important point of debate is how to describe complex aromatic elements. In daylight SMARTS, an aromatic part of molecules is annotated by first identifying the "smallest set of smallest rings (SSSR)," whereas Openeye SMARTS prefers to count the number of aromatic rings attached to each atom and to design queries based on this count. A critical review of some of these issues was undertaken by Downs et al. [50], which can be referred to for further reading.

9.3.4 SYBYL Line Notation (SLN)

This is another text-based method for representing molecules developed by the commercial company Tripos, Inc. [51]. This notation tries to integrate information about the reactions, data search queries, and molecular information in a single notation.

9.3.5 Formats for Writing 2-D Coordinates

While SMILES, SMARTS and SLN are useful to unambiguously describe atomic arrangements within molecules, the detailed structures cannot be incorporated within them. Bond angle, actual bond lengths, and so on, require precise description of atomic positions of all atoms within the molecule. Many of these molecules may be projected on a plane, and therefore a simplified structure can be written by describing x and y coordinates of molecules just

like any other geometric object in a plane. More precise and complete three-dimensional (3-D) structures can be shown by giving the Cartesian coordinates of all atoms in molecules. However, coordinates of a large number of molecules again need to follow a standard way to write the atom names, their spatial positions, and their connectivities. There are a number of formats to write these coordinates, and these are reviewed elsewhere [52]. Among them, PDB format, Tripos's Mol2 format, Accelrys/MSI BGF, and Chem3D of Chemdraw are some of the most commonly used formats for writing atomic coordinates of small and large molecules. It may, however, be noted that the molecules written in one format can easily be converted to any other format by readily available software (e.g., OpenBabel, http://openbabel.sourceforge.net/wiki/Main_Page).

9.3.6 Molecular Editors

There are a number of tools available for editing a molecular structure by drawing predefined rings and branches and by combining various substructures. They also provide facilities to convert formats of molecular information. Some of these tools are available to directly work on the web. A nonexhaustive list of these tools is provided in Table 9.2.

9.4 IDENTIFYING INTERACTIONS FROM STRUCTURE

9.4.1 Protein–Ligand Complexes and Their Databases

3-D structures of ligand as well as protein molecules play a significant role in the so-called lock-and-key mechanism of molecular recognition [53]. 3-D structures of proteins complexed with their ligands are obtained by X-ray, nuclear magnetic resonance (NMR) or other methods, and PDB is the primary source for such complex structures [54]. Protein–ligand complex structures give detailed information about all atomic positions in a protein as well as its ligand. Most ligands (except nucleic acids and proteins) are identified by the key word HETATM. For our purpose, all HETATM records except water molecules may be treated as ligands. Based on this criterion, there are more than 7000 ligand molecules in PDB at the moment (December 2007). Among the most abundant ligands in PDB are some amino acids (e.g., alanine), sulfate (SO4), metallic ions (sodium, zinc, and calcium), and sugar derivatives such as glycerol. Interestingly, the number of protein–ligand complexes has grown more rapidly than proteins without a ligand during recent years. In fact, about 70% of all PDB entries today have at least one ligand molecule in their structure. A simple count carried out at the time of finalizing this text (January 2008) depicted that the number of ligands per PDB entry has grown from 1.3 to 1.9 since 1993. This could, however, simply imply the availability of higher-resolution structures during recent years.

TABLE 9.2 List of Molecular Editors and Commercial, Free, and Web-Based Applications

Editor	Citation	License
ACD/ChemSketch	http://www.acdlabs.com/download/chemsk.html	Commercial/freeware
ChemDraw	http://www.adeptscience.co.uk/products/lab/chemoffice/chemdraw.html	Propriety/commercial
XDrawChem	http://xdrawchem.sourceforge.net/	Free software
Smormoed	http://www.hungry.com/~alves/smormoed/	Free software/BSD license
JChemPaint	http://almost.cubic.unihyphen;koeln.de/cdk/jcp	Free software
BKchem	http://bkchem.zirael.org/	Free software/GPL license
OpenChem	http://openchemwb.sourceforge.net/	Free software/GPL license
ChemTool	http://ruby.chemie.unihyphen;freiburg.de/~martin/chemtool/	Free software
molsKetch	http://molsketch.sourceforge.net/	Free software/GPL
EasyChem	http://easychem.sourceforge.net/	Free software/GPL
Instant JChem Standard	http://www.chemaxon.com/product/ijc.html	Commercial
Instant JChem personal	http://www.chemaxon.com/product/ijc.html	Free software
Ghemical	http://www.uku.fi/~hassine/projects/ghemical/	Free software/GPL
Avogadro0.0.3	http://avogadro.sourceforge.net/wiki/Get_Avogadro	Free software
Online Editors		
PubChem	http://pubchem.ncbi.nlm.nih.gov/search/search.cgi	Free software
Molinspiration WebME	http://www.molinspiration.com/docu/webme/index.html	Free software
O=CHem JME Molecular Editor	http://www.usm.maine.edu/~newton/jme/index.htm	Free software
Marvin Molecule Editor and Viewer	http://www.chemaxon.com/demosite/marvin/index.html	Free for academic research and teaching

GPL = GNU General Public License; BSD=Berkeley Software Distribution.

A huge number of protein–ligand complexes pose a data management and analysis challenge. This has led to a number of subsets of PDB as secondary databases, with ligand–protein complex information derived from PDB. Most prominent of these databases are Protein–Ligand Database (PLD), LIGAND,

and PDB-LIGAND databases [3,55,56]. In addition PDBSite, EzCatDB, GLIDA, and Relibase are examples of specialized databases that provide useful information about protein–ligand complexes [57–61].

9.4.2 Binding Site From Complexes

In order to carry out any kind of analysis of protein–ligand interactions based on their complex structure, one often starts with the identification of a binding site on protein. The purpose of this identification can be either to determine amino acid preferences to be in the interface, machine readable notation to develop a prediction method, or to elaborately characterize a binding area to discover targets and design inhibitors [62–64]. The definition of binding site may therefore vary in meaning and scope. For example, a whole patch on a protein surface may be considered a single binding region for a ligand even though some of the residues in protein form no physical contact with ligand. On the other hand labeling of each residue to be interacting or noninteracting may be required in some cases such as in the application of machine learning methods [39,40,65]. Whether it is a hydrophobic or electrostatic patch, or it is a single amino acid residue in protein, the criterion to label it as interacting is neither unique nor obvious. Some of the most common methods to label binding sites or regions are thus listed in the following.

9.4.3 Definition Based on Change in Accessible Surface Area (ASA)

Regions of protein and ligand that come in contact with each other or participate in interactions are variously named as binding site, interface area, or, in a more specific context, active site. These terms are largely qualitative in nature and are therefore not very suitable for treatment by computers without manual assignment for each protein–ligand interaction. One of the more objective methods to annotate each residue or atom to be in the interface or not is to calculate the solvent accessibility or the ASA of protein (and ligand, if necessary) first in their complexed state and then by removing protein and ligand and calculating the ASA again in the isolated environments [66–68]. In other words, any atom or residue that becomes fully or partially inaccessible to solvent probes (typically water) in complex form and is accessible in its isolated components (free protein or free ligand) may be treated as the interface or binding region. Using this method, each residue in the protein sequence can be labeled as binding or nonbinding, provided that its complex structures are available. There remain a few parameters that need to be fixed for such a definition. First of all, how much change in ASA is large enough to label a residue to be interface residue? Many works use a permissive definition based on ΔASA of as small as 1 Å^2 (ΔASA is defined as the difference between the solvent ASA of a residue in its complex and isolated or free structure). However, some works have used a different criterion for ΔASA cutoff to define interface residue [69]. In some cases, proteins undergo significant

conformational changes upon binding with ligands, and mapping of binding sites obtained from protein–ligand complexes to free proteins is possible by first obtaining binding sites from complexed structures and then by finding exact alignments between the sequences of the complexed and free proteins [66]. The probe radius of the solvent may be another concern, but most ASA calculation programs such as DSSP use 1.4-Å default size for water probe and are assumed to be satisfactory [70].

The definition of ASA change seems to be plausible and carries additional information about the strength of hydrophobic interactions. However, this definition does not carry much information about the nature of contact between ligand and proteins and assigns more significance to larger residues compared with smaller ones, in which change in ASA may be too small to measure. Also, some atomic pairs may interact at a distance too large to be captured by a small probe radius of 1.4 Å. Another problem with ASA-based definition is that the structures of some complexes may be solved at poor resolution and ASA is calculated with much error, leaving a change in ASA with even greater statistical error. In published literature, binding or interface residues are often obtained based on definitions of direct geometric contacts between protein and ligand atoms [39,40,65].

9.4.4 Definition Based on Geometric Contacts

A definition of binding site or interface residue (or atom) based on geometric contacts is also derived from the structure of protein–ligand complexes, just as the definition based on ΔASA. Formally speaking, a residue or atom of the protein is considered to be in contact with that of the ligand if the distance between them is less than a predefined cutoff. In the case of atomic contact, it is straightforward, but in the case of residues, there is no single point that can naturally represent its position without loss of information; hence, instead of directly determining residue–residue contacts, information about atomic contacts between them is usually taken into consideration. In many cases, a residue may be considered in contact with a ligand if any atom of the residue is in geometric contact with the ligand. In some other cases, distance only from the backbone or side chain atoms of the protein is considered. In small data sets, each geometric contact may be physically examined for its likelihood of interaction, but in a large-scale bioinformatics approach, contacts must be automatically assigned. Some atomic pairs may be assigned meaningful physical contact such as hydrogen bonds with high confidence, but that is not the case with most pairs of atoms in which we do not even know what kind of contact they might make or how much they may contribute to the interaction energy. This leaves the choice of cutoff distances somewhat uncertain. In a broader sense, any-to-any atomic pair contacts are most widely used and typical cutoff distances range from 3 to 6 Å. Although it is unlikely that atomic pairs at 6-Å distance will contribute significantly to the overall energy, some authors have used these values apparently to artificially increase the number

of binding sites for dealing with prediction problems (trying to make the binding and nonbinding data more balanced). However, we have shown that increasing the cutoff distance for marking binding sites does not add any value to prediction, at least in the case of DNA-binding proteins [28,71]. In our studies on predicting DNA-binding sites, we find that a 3.5 Å cutoff for any-to-any atomic contact is more efficient than other prevailing definitions [71]. Manual examination of binding sites is likely to do well for the purpose of developing elaborate empirical rules and may give additional insights into the nature of protein–ligand interactions [29,72–74]. Usual tools for visualizing protein structure (e.g., VMD, RasMol, PyMol) and contact maps provide the most fundamental information about protein–ligand contacts [75–77]. Some specialized tools have been developed for visual or quantitative analysis of protein–ligand contacts [78]. LIGPLOT is one such powerful tool that provides a schematic 2-D view of interactions between protein and ligand atoms. Figure 9.4 is a typical LIGPLOT representation of interaction between phosphopeptide A (Tyr-Val-Pro-Met-Leu, phosphorylated Tyr) and the SH2 domain of the V-src tyrosine kinase transforming protein. The ligand (residues 201–205 of chain B) has its phosphorylated tyrosine shown toward the bottom of the picture. The interactions shown are those mediated by hydrogen bonds and by hydrophobic contacts. Hydrogen bonds are indicated by dashed (green) lines between the atoms involved, while hydrophobic contacts are represented by an arc with spokes radiating toward the ligand atoms they contact. The contacted atoms are shown with spokes radiating back [78].

9.4.5 Solvent Accessibility and Binding Sites

Residues that are accessible to water are obviously also accessible at least partially to ligands, whereas buried residues are not. Thus, it is quite useful to know the solvent accessibility of each residue in order to do a first level of filtering of candidate sites for ligand binding. Solvent accessibility of residues can be obtained from readily available software, databases, and web servers [70,79–81]. Some servers provide additional information about cavities and pockets [82]. Yet others give a graphical view of the arrangement of residues in various levels of solvent accessibility [78,81]. All these methods require the knowledge of protein structures. However, solvent accessibility can also be predicted with reasonable confidence from sequence information only [83,84]. Thus, it may be possible to make a preliminary estimate of potential binding sites by looking at the amino acid sequence and ASA predictions, although at a very rough scale.

9.5 IDENTIFYING INTERACTIONS FROM *IN VITRO* THERMODYNAMIC DATA

Although protein–ligand complexes provide a huge insight into the spatial arrangement and nature of physical contact between atomic pairs, their exact

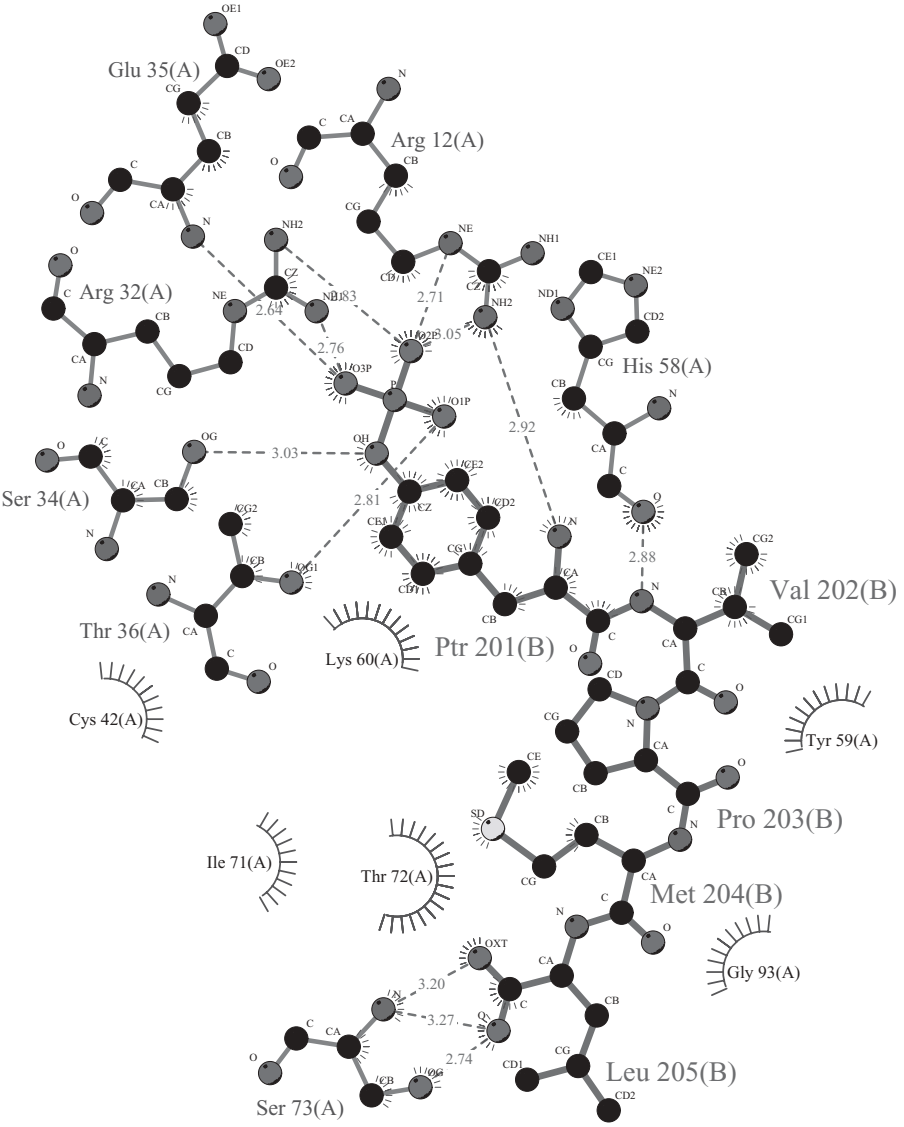


Figure 9.4 SH2 domain complexed with a peptide containing phosphotyrosine (PDBID: 1SHA). See color insert.

role remains uncertain. More decisive information about protein–ligand interaction may be obtained by measuring binding free energy and related parameters. Several experimental methods are available to do so, but here our concern is the computationally useful outcome of those experiments. A variety of information is sought in these methods. First, the strength of a known

protein–ligand interaction is estimated by performing binding experiments and by determining the amount of free energy released when a free protein–ligand pair is brought together to bind with each other under controlled condition [85]. Second, the role of individual residues in the interaction is assessed by site-directed mutagenesis, where one residue is replaced by another and the binding free energy in the two cases is measured under identical conditions—the difference being the effect of mutation [86]. Also, binding of candidate inhibitors, potentially useful for drug design, is also estimated from free energy measurements. Finally, the competitive and reversible nature of protein–ligand interactions is studied to design novel inhibitors for targeted protein–ligand interactions. In the following, we give a brief account of most widely used parameters to measure the strength of protein–ligand interaction.

9.5.1 Measurement Units

9.5.1.1 Free Energy Free energy in the context of protein–ligand interactions typically refers to Gibbs's energy, implying that the entropic contributions should be implicitly taken care of. Standard units for free energy and changes therein are kJ/mol and kcal/mol. Free energy of interaction between ligand and protein therefore would actually refer to the change in Gibbs's energy values in the bound (complexed) and unbound (free) states of the proteins and ligands. Thus, it is customary to refer to ΔG values by the term free energy, wherein the differential nature of measurement is understood. There is, however, another level of free energy change, especially in the context of stabilization or destabilization of protein–ligand interaction due to external factors such as temperature, pH, and buffers and evolutionary factors such as mutations in protein. Free energy changes in the stability of protein–ligand interactions in such cases are measured by the same units but refer to $\Delta\Delta G$ instead of ΔG and do not require explicit measurement or computation of ΔG values.

9.5.2 Association and Dissociation Constants (k_d and k_a) and IC_{50}

The number of protein–ligand molecular pairs in the bound and unbound state at equilibrium depends on the concentration of the ligand. The ratio of the molecular concentrations in the bound and unbound states determines their association constants (the reciprocal is the dissociation constant). Free energy change upon binding is related to association constant by the simple expression

$$\Delta G = -RT \log(K_a)$$

where $K_a = [AB]/[A][B]$, where $[A]$ and $[B]$ are the concentrations of protein and ligand, respectively. Many times, the activity of an enzyme is inhibited by

a ligand, and the quantity of interest is the ligand concentration that could reduce the enzyme (protein) activity to 50% of its maximum value. This quantity is called IC_{50} and is measured in concentration units.

9.6 THERMODYNAMIC DATABASES OF PROTEIN–LIGAND INTERACTIONS

Large numbers of experiments reporting the strength of protein–ligand interactions under different thermodynamic conditions have been carried out. Bioinformaticians have tried to compile them in the form of searchable databases. The most prominent databases reporting the thermodynamic protein–ligand interactions are reviewed below.

9.6.1 Protein–Ligand Interaction Database (ProLINT)

Sarai Lab (including one of the authors of this article [S. Ahmad]) has been compiling the thermodynamic data of protein–ligand interactions since 1998. Although a full public release has not been made, the database has often been previewed on several occasions [87]. Each protein–ligand interaction is made up of several sets of information, viz, ligand information, protein information, thermodynamic information, clinical information, and literature information. Proteins and ligands are identified by their PDB Code, Swiss-Prot or Protein Information Resource (PIR) codes, SMILES notation, and enzyme classification number. Thermodynamic information is in the form of association constants and free energy changes. Each entry is drawn from published literature and hence literature information modules provide necessary citation information. Each entry is also associated with any disease or clinical information and therefore the database has a huge promise of use in medically related bioinformatics research.

9.6.2 AffinDB

Developed at the University of Marburg, Germany, Affinity database is actually a thermodynamic database linking all binding information to their PDB entries [88].Currently, AffinDB consists of 748 affinity data in the form of dissociation constants, IC50, or related binding units. This data correspond to 474 entries in the PDB.

9.6.3 BindingDB Database

BindingDB, developed at the Center for Advanced Research in Biotechnology, University of Maryland, is a database of experimentally determined binding affinities for protein–ligand complexes [89].The main focus of this database is the proteins, which are drug targets or potential drug targets and for which structural information is available in the PDB. BindingDB currently

holds~20,000 binding data for ~11,000 different small-molecule ligands and 110 different drug targets.

There are other databases of structural and thermodynamic aspects of interactions (e.g., comprising protein–protein interactions [90]), but only those dealing with small ligands are included above.

9.7 DATA ANALYSIS AND KNOWLEDGE GENERATION

One of the principal goals of bioinformatics in the recent years has been the development and analysis of databases, more fashionably called knowledge bases, highlighting the fact that mere compilation of information is not enough, unless accompanied by relevant knowledge [91,92]. Thermodynamic and structural data of protein–ligand interactions have been repeatedly analyzed, sometimes across a global set of interactions and other times in a particular family of proteins, type of ligands, or a group of interactions [93–95]. Some of the databases of protein–ligand interactions, derived from their chemical structures (as against the thermodynamics) are discussed below:

9.7.1 Relibase+ and Its Retiring Precursor Relibase

Relibase is one of the early and most important databases of protein–ligand complexes and related information [61]. Relibase+ is a more advanced version, available for commercial users only. Relibase is a database developed for the analysis of protein–ligand complex structures and allows additional features for the development of databases on structures drawn from these complexes.

Basic and advanced features of Relibase and/or Relibase+ are summarized as follows.

9.7.1.1 Web-Based Access Relibase can be accessed through a web interface working on a client–server mode. This allows for extensive and easy sharing of information, without any portability concerns.

9.7.1.2 Search Engine On the server side, there is a search engine that can scan a large number of precompiled entries based on a variety of query terms such as ligand SMILES or SMARTS ligand name, protein name, and other information. Relibase provides for 2-D and 3-D substructure searches. Relibase allows visualizing protein–ligand interactions in three dimensions. One very powerful feature of Relibase+ is the automatic superposition of related binding sites to compare ligand binding modes, water positions, ligand-induced conformational changes, and so on. Relibase+ includes a crystal packing module for detailed investigation of crystallographic packing effects around ligand binding sites. It also provides functionality for detection of unexpected similarities among protein cavities (e.g., active sites) that share little or no sequence

homology. The two most important aspects of any modern database system are their integration with other database and the ability of users to integrate or query them through their own applications; Relibase+ takes care of both of these requirements.

9.7.2 ZINC Database

This database was developed by Irwin and Shoichet at the University of California, San Francisco, CA, USA [96]. Although ZINC does not explicitly deal with interactions, it is valuable in the analysis of interactions as it provides a comprehensive list of commercially available ligand molecules. Like Relibase, it has a powerful search engine, by which molecules satisfying particular conditions may be searched. Queries to the database can be made among others by the chemical properties of molecules (e.g., net charge, log *P*, rotatable bonds, and polar surface area), using full SMILES or SMARTS. This is extremely useful for the design of inhibitors and for finding out candidates for competitive binding to proteins.

9.7.3 PROCOGNATE

Developed at European Bioinformatics Institute (EBI), PROCOGNATE is a database of cognate ligands for the domains of enzyme structures in CATH, SCOP, and Pfam [97]. PROCOGNATE assigns PDB ligands to the domains of structures based on structure classifications provided by CATH, SCOP, and Pfam databases [98–100]. Cognate ligands have been identified using data from the ENZYME and KEGG databases and compared to the PDB ligand using graph matching to assess chemical similarity [4,101].Cognate ligands from the known reactions in ENZYME and KEGG for a particular enzyme are then assigned to enzyme structures that have Enzyme Commission (EC) numbers [102].

9.8 ANALYSIS OF DATABASES

As mentioned above, several primary and secondary sources of information on thermodynamic and structural aspects of protein–ligand interactions have been compiled. These databases result in useful knowledge, which is obtained by a thorough analysis of these databases and related information. Most widely analyzed aspects of protein–ligand interactions may be grouped under several categories, some of which are discussed below.

9.8.1 Binding Propensities

One of the first analyses possible about protein–ligand interactions is that of residue preferences for binding to particular ligands. In essence, it is a measure

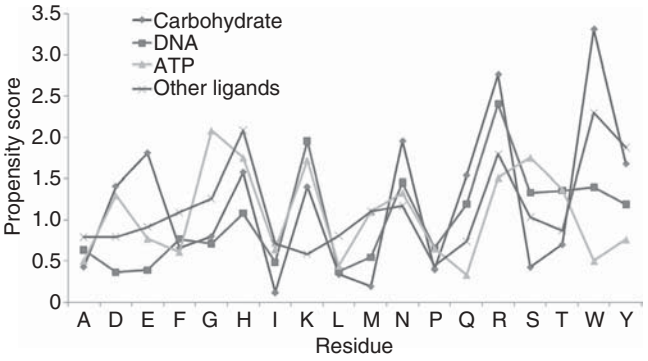


Figure 9.5 Propensity scores of residues in ATP, DNA, carbohydrate, and other ligand binding sites. See color insert.

of the ratio of residue populations in the protein–ligand interface to the rest of the protein or in the overall protein including interface [28]. In some cases, propensities are calculated within the surface residues, taking into account the fact that binding occurs only in the solvent-exposed residues [103].

Different residues have different propensities for different ligands, depending upon the nature of their interactions. For example, DNA and other ligands with negative charge on the surface prefer to interact with basic residues such as Lys and Arg. On the other hand, sugars and similar ligands preferentially interact with Trp due to their structural compatibility. As a case in point, propensity of 20 amino acids to interact with carbohydrates, DNA, and another well-known ligand ATP has been shown in Figure 9.5.

Residue preference for a ligand may be derived from protein–ligand complexes if and only if we know which residues take part in binding or interaction. However simple it might look, it is a difficult task to identify each residue as binding or not, especially if one is dealing with a large data set, which requires automatic assignment or labeling of residues in terms of their binding. This issue has been discussed above in the section on definition of binding sites.

9.8.2 Neighbor Effects and Machine Learning Methods

If certain residues are preferred over others to bind to a given ligand, why not all residues of that kind do so? The answer lies in the environment in which a residue is found. This environment basically puts additional constraints on the viability of amino acid–ligand interaction. Second, a single residue may not be enough to complete the binding process. Thus, the sequence and structural neighbors and spatial arrangement of atoms play the role of ultimate selectivity of interaction. A big bioinformatics challenge is to automatically extract the knowledge of these neighbor and environmental effects and to

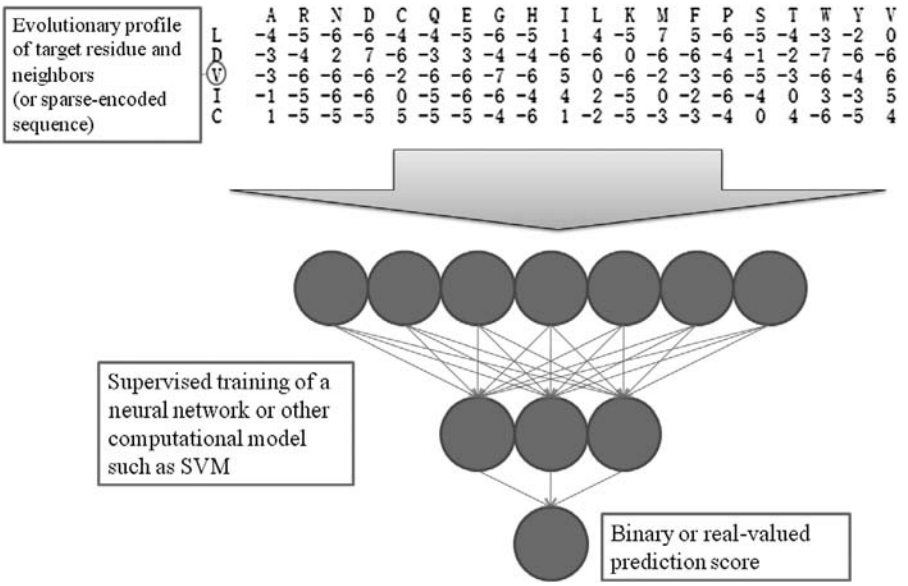


Figure 9.6 Neural network model for binding site prediction using evolutionary information. SVM = support vector machines.

definitively tell (predict) which residues will interact to the ligand in question. For a large-scale prediction based on these considerations, machine learning methods have been shown to be particularly successful [28,29,38,39,65,104–106]. A general scheme to predict binding sites using a machine learning method may be schematically represented as in Figure 9.6.

In brief, the whole idea may be summarized into four steps:

1. Scan a protein's amino acid sequence for each of its residues and characterize the sequence or structural environment in a finite-dimensional vector (input vector).
2. Define binding state or label for each residue (target vector, which is actually a binary value scalar).
3. Find a relationship between input vector and the target vector using a machine learning algorithm such as neural network, support vector machine, or any other model.
4. Cross validate by fitting several models on different data sets and by testing them on independent samples.

These methods are successful to a certain degree and may be powerful for a high-throughput analysis at a genomic scale. However, in most real cases, these models cannot utilize all the information that may be useful for making

predictions. More accurate results are likely to come from docking experiments (see later sections). Nonetheless, these models provide a powerful and fast platform to quickly scan thousands of sequences for potential binding sites and may even be used for constructing initial poses for a docking experiment. Work in that direction is rather rare and we are trying to take some early steps toward this course.

9.9 SIMULATIONS AND MOLECULAR DOCKING

Molecular docking is one of the most widely used techniques to predict the binding mode of protein–ligand interactions. It may be noted that docking of larger molecules such as proteins and DNA is not yet handy and hence most of the discussion in the following section refers to the docking of small ligands with proteins. The technique of docking arguably comes closest to the experimentally verifiable nature of interactions and has been widely used for screening a small number of ligands for their potential use as drugs in the form of inhibitors, which selectively interact with targeted sites on the protein structure [107–112]. Recent advancements have even allowed the use of docking at a large scale, extending the reach of this method to screen a large collection of potentially useful ligands [113,114]. This latter method is called high-throughput docking (HTD). The basic principle of all docking methods, small or large scale, is the same, and they differ in their details and manner of implementation. In general, in a docking experiment, one attempts to find the 3-D structure of a protein–ligand complex from the known structures of proteins and ligands. It is obvious that the fundamental requirement to initiate docking is the availability of 3-D structures of the interacting partners. Large numbers of proteins have been sequenced for which no such structure is available, and hence this method cannot deal with such proteins. However, advances in comparative modeling and structural genomics projects have made structures of many proteins available as structures of small ligands are relatively easy to model [115]. These molecules are the principal target of application for docking studies. The problem of modeling the structure of the protein–ligand complex from the structure of its constituent protein and ligand—*docking*—may be broken into following stages:

1. *Generation of possible poses as an ensemble or a trajectory in time.* One of the major problems in docking, similar to any ab initio method of structure prediction, is the possibility that there is a prohibitively large number of geometric positions that need to be scanned for a possible mode of interaction. A single snapshot of such geometric arrangement is called a pose [116,117]. Any docking method starts with some pose and an evaluation of its energetics. If the starting pose is too far from a real situation, the system is most likely to attain a local minimum energy configuration and no useful information will emerge. Thus, it is very

important to start with a reasonable pose close to the experimentally viable geometric arrangement.

In the drug-discovery scenario, binding sites are generally known and constructing a reasonable initial pose is not so difficult. However, in cases where there is no information about the sites of interaction or the availability of similar complexes, the problem becomes more challenging. This has led to the development of a number of methods to provide an estimate of interaction sites using which protein–ligand complex poses may be generated. In larger molecules, generating exact poses is a difficult task and many times, less ambitious approaches of finding potential interacting surfaces are used [118].

2. *Energy calculations or scoring functions to determine the suitability of each pose for interaction.* Once a candidate protein–ligand interaction pose has been generated, the next problem is to rank different poses on the bases of their energetics. In principle, we should be able to determine the most stable pose by the application of quantum mechanics and solve the problem exactly. However, the problem of protein–ligand interaction has rarely ever been solved from purely quantum-mechanical considerations because of the high complexity of such interactions and the strong effects of solvents, which are difficult for such handling [119,120]. In a more practical sense, the problem has been divided into two parts: one is the development of force fields, which try to define interaction energies in terms of pairs of atomic groups and distances between them, and the other is to score energies and to find the best candidate interaction [121–123]. Interaction energy between atomic groups largely comes from the structures of known complexes or their thermodynamic parameters and therefore often takes care of actual physiological perturbations rather implicitly. Scoring interactions is another important problem of the energy calculation process [106,124]. The problem of scoring docked pairs of protein–ligands is often encountered in drug-discovery technology [124–126]. The problem arises from several reasons. First of all, some force fields such as those based on statistical potentials do not calculate energies in absolute units. Second, there are a number of energy terms coming from protein and ligand conformational changes on one hand and protein–ligand direct interaction on the other, further complicated by solvents and other environments. Third, different ligands may not necessarily bind exactly in the same way or in the same site on the protein, complicating their comparison. Thus, special scoring functions are required to rank a number of hits while searching for potential candidate ligands likely to interact with the protein. Many scoring functions derived from classical, statistical, and quantum mechanical considerations are available [121,122,127,128]. However, it has been argued that there is no single scoring function that can be universally used for all protein–ligand interactions, and most scoring functions perform better for one or the other class of protein–ligand interactions [124,126].

9.10 HIGH THROUGHPUT DOCKING (HTD)

As stated above, one important stage in *in silico* drug design is to scan a library of potential drugs (ligands) that can possibly bind to a selected target on protein. Since structures of proteins and ligands are known in most such cases, a reasonable set of candidate ligands may be obtained by systematically docking them on the target protein sites [129]. Many tools are now available to achieve this task [117,130–135]. It has been shown that the comparison of docking software is not easy as there is no universal principle to evaluate their performance. Some guiding principles have been reviewed in published literature [124,136].

Indiscriminate scanning of ligand databases for their interactions with proteins has been replaced by additional filtering techniques [137]. These techniques allow input of additional information of protein–ligand interactions in order to carry out a screening of ligand databases and to create a smaller ensemble of ligands, which can then be used for HTD.

9.11 CONCLUSION

Almost all interaction of proteins may be regarded as protein–ligand interactions, which occur in structurally and functionally diverse environments. The bioinformatics challenge is to understand the features of proteins important for interactions with known ligands and the ability to predict ligand binding sites in proteins, and to select ligands that could bind to a previously known binding site. Machine learning, docking, and other bioinformatics approaches have played important roles in the advancement of this subject, which provided not only a better understanding of the interaction but also technology to design novel ligands and proteins, with desired properties.

REFERENCES

1. IUPAC. *Compendium of Chemical Terminology*. Oxford: Blackwell Scientific Publications, 1987.
2. IUPAC. Nomenclature announcement. *Arch Biochem Biophys* 1992;294: 322–325.
3. Goto S, Nishioka T, Kanehisa M. LIGAND: Chemical database for enzyme reactions. *Bioinformatics* 1998;14:591–599.
4. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 1999;27:29–34.
5. Falanga A, Gordon SG. Isolation and characterization of cancer procoagulant: A cysteine proteinase from malignant tissue. *Biochemistry* 1985;24:5558–5567.
6. Falanga A, Shaw E, Donati MB, Consonni R, Barbui T, Gordon S. Inhibition of cancer procoagulant by peptidyl diazomethyl ketones and peptidyl sulfonium salts. *Thromb Res* 1989;54:389–398.

7. Green GD, Shaw E. Peptidyl diazomethyl ketones are specific inactivators of thiol proteinases. *J Biol Chem* 1981;256:1923–1928.
8. Brocklehurst K, Malthouse JP. Mechanism of the reaction of papain with substrate-derived diazomethyl ketones. Implications for the difference in site specificity of halomethyl ketones for serine proteinases and cysteine proteinases and for stereoelectronic requirements in the papain catalytic mechanism. *Biochem J* 1978;175:761–764.
9. Harris ED. Cellular copper transport and metabolism. *Annu Rev Nutr* 2000;20: 291–310.
10. Karlin S, Zhu ZY, Karlin KD. The extended environment of mononuclear metal centers in protein structures. *Proc Natl Acad Sci USA* 1997;94:4225–4230.
11. Qin PZ, Feigon J, Hubbell WL. Site-directed spin labeling studies reveal solution conformational changes in a GAAA tetraloop receptor upon Mg2+-dependent docking of a GAAA tetraloop. *J Mol Biol* 2007;35:1–8.
12. Hutchens TW, Allen MH. Differences in the conformational state of a zinc-finger DNA-binding protein domain occupied by zinc and copper revealed by electrospray ionization mass spectrometry. *Rapid Commun Mass Spectrom* 1992;6:469–473.
13. Saavedra RA. Histones and metal-binding domains. *Science* 1986;234:1589.
14. Wang CL, Aquaron RR, Leavis PC, Gergely J. Metal-binding properties of Calmodulin. *Eur J Biochem* 1982;124:7–12.
15. Foffani MT, Battistutta R, Calderan A, Ruzza P, Borin G, Peggion E. Conformational and binding studies on peptides related to domains I and III of calmodulin. *Biopolymers* 1991;31:671–681.
16. Li Y, Yan XP, Chen C, Xia YL, Jiang Y. Human serum albumin-mercurial species interactions. *J Proteome Res* 2007;6:2277–2286.
17. Wei Y, Fu D. Binding and transport of metal ions at the dimer interface of the Escherichia coli metal transporter YiiP. *J Biol Chem* 2006;281:23492–23502.
18. Sirois JE, Atchison WD. Effects of mercurials on ligand- and voltage-gated ion channels: A review. *Neurotoxicology* 1996;7:63–84.
19. Sharon N, Lis H. Lectins as cell recognition molecules. *Science* 1989;246: 227–234.
20. Lasky LA. Selectins: Interpreters of cell-specific carbohydrate information during inflammation. *Science* 1992;258:964–969.
21. Sastry K, Ezekowitz RA. Collectins: Pattern recognition molecules involved in first line host defense. *Curr Opin Immunol* 1993;5:59–66.
22. Barondes SH, Cooper DN, Gitt MA, Leffler H. Galectins: Structure and function of a large family of animal lectins. *J Biol Chem* 1994;269:20807–20810.
23. Hoppe HJ, Reid KB. Collectins—Soluble proteins containing collagenous regions and lectin domains—And their roles in innate immunity. *Protein Sci* 1994;3: 1143–1158.
24. Rosen SD, Bertozzi CR. The selectins and their ligands. *Curr Opin Cell Biol* 1994;6:663–673.
25. Sharon N, Lis H. Lectins-proteins with a sweet tooth: Function in cell recognition. *Essays Biochem* 1995;30:59–75.

26. Sharon NH, Lis H. Lectins: Carbohydrate-specific proteins that mediate cellular recognition. *Chem. Rev* 1998;98:637–674.
27. Karlsson KA. Meaning and therapeutic potential of microbial recognition of host glycoconjugates. *Mol Microbiol* 1998;29:1–11.
28. Malik A, Ahmad S. Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. *BMC Struct Biol* 2007;7:1.
29. Shionyu-Mitsuyama C, Shirai T, Ishida H, Yamane T. An empirical approach for structure-based prediction of carbohydrate binding sites on proteins. *Protein Eng* 2003;16:467–478.
30. Sarno S, Salvi M, Battistutta R, Zanotti G, Pinna LA. Features and potentials of ATP-site directed CK2 inhibitors. *Biochim Biophys Acta* 2005;1754:263–270.
31. Hirsch AK, Fischer FR, Diederich F. Phosphate recognition in structural biology. *Angew Chem Int Ed Engl* 2007;46:338–352.
32. Lindahl U. Heparan sulfate-protein interactions–A concept for drug design? *Thromb Haemost* 2007;98:109–115.
33. Fischer K, Barbier GG, Hecht HJ, Mendel RR, Campbell WH, Schwarz G. Structural basis of eukaryotic nitrate reduction: Crystal structures of the nitrate reductase active site. *Plant Cell* 2005;17:1167–1179.
34. Gong W, Hao B, Chan MK. New mechanistic insights from structural studies of the oxygen-sensing domain of Bradyrhizobium japonicum FixL. *Biochemistry* 2000;39:3955–3962.
35. Yoshikawa S, Shinzawa-Itoh K, Nakashima R, Yaono R, Yamashita E, Inoue N, Yao M, Fei MJ, Libeu CP, Mizushima T, Yamaguchi H, Tomizaki T, Tsukihara T. Redox-coupled crystal structural changes in bovine heart cytochrome c oxidase. *Science* 1998;280:1723–1729.
36. Kolakowski LF Jr. GCRDb: A G-protein-coupled receptor database. *Receptors Channels* 1994;2:1–7.
37. Filmore D. It's a GPCR world. *Modern Drug Discovery (American Chemical Society)* 2004;7(11):24–28.
38. Ahmad S, Sarai A. Moment-based prediction of DNA-binding proteins. *J Mol Biol* 2004;341:65–71.
39. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* 2004;20:477–486.
40. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics* 2005;6:3.
41. Prabakaran P, Siebers JG, Ahmad S, Gromiha MM, Singarayan MG, Sarai A. Classification of protein-DNA complexes based on structural descriptors. *Structure* 2006;14:1355–1367.
42. Luscombe NM, Austin SE, Berman HM, Thornton JM. An overview of the structures of protein-DNA complexes. *Genome Biol* 2000;1:001.1–001.37.
43. Ahmad S, Kono H, Araúzo-Bravo MJ, Sarai A. ReadOut: Structure-based calculation of direct and indirect readout energies and specificities for protein-DNA recognition. *Nucleic Acids Res* 2006;1:124–127.

44. Bock LC, Griffin LC, Latham JA, Vermaas EH, Toole JJ. Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *Nature* 1992; 355:564–566.

45. Green LS, Jellinek D, Jenison R, Östman A, Heldin CH, Janjic N. Inhibitory DNA ligands to platelet-derived growth factor B-chain. *Biochemistry* 1996;35: 14413–14424.

46. Sullenger BA, Gilboa E. Emerging clinical applications of RNA. *Nature* 2002; 418:252–258.

47. Yang Q, Goldstein IJ, Mei HY, Engelke DR. DNA ligands that bind tightly and selectively to cellobios (systemic evolution of ligands by exponential enrichment / aptamer / cellulose / saccharide). *PNAS* 1998;95:5462–5467.

48. Weininger D. SMILES: A chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;28:31–36.

49. Weininger D, Weininger A, Weininger JL. SMILES: 2. Algorithm for generation of unique SMILES notation. *J Chem Inf Comput Sci* 1989;29:97–101.

50. Downs GM, Gillet VJ, Holliday JD, Lynch MF. A review of ring perception algorithms for chemical graphs. *J Chem Inf Comput Sci* 1989;29:172–187.

51. Ash S, Malcolm A, Cline R, Homer W, Hurst T, Smith GB. SYBYL Line Notation (SLN): A versatile language for chemical structure representation. *J Chem Inf Comput Sci* 1997;37:71–79.

52. Araúzo-Bravo MJ, Fujii S, Kono H, Ahmad S, Sarai A. Sequence-dependent conformational energy of DNA derived from molecular dynamics simulations: toward understanding the indirect readout mechanism in protein-DNA recognition. *J Am Chem Soc* 2005;127:16074–16089.

53. Koshland DE. The key-lock theory and the induced fit theory. *Angew Chem Int Ed Engl* 1994;33:2375–2378.

54. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–242.

55. Puvanendrampillai D, Mitchell JBO. Protein ligand database (PLD): Additional understanding of the nature and specificity of protein–ligand complexes. *Bioinformatics* 2003;19:1856–1857.

56. Shin JM, Cho DH. PDB-ligand: A ligand database based on PDB for the automated and customized classification of ligand-binding structures *Nucleic Acids Res* 2005;33:D238–D241.

57. Ivanisenko VA, Grigorovich DA, Kolchanov NA. PDBSite: A database on biologically active sites and their spatial surroundings in proteins with known tertiary structure. In: *The Second International Conference on Bioinformatics of Genome Regulation and Structure (BGRS 2000)*, Vol. 2. http://www.bionet.nsc.ru/meeting/bgrs2000/thesis/bgrs2000_2.pdf (accessed September 1, 2009).

58. Grigorovich DA, Ivanisenko VA, Kolchanov NA. Structure and format of the EnPDB database accumulating spatial structures of DNA, RNA and proteins. In: *The Second International Conference on Bioinformatics of Genome Regulation and Structure (BGRS 2000)*, Vol. 2. http://www.bionet.nsc.ru/meeting/bgrs2000/thesis/bgrs2000_2.pdf (accessed September 1, 2009).

59. Nagano N. EzCatDB: The Enzyme Catalytic-mechanism Database. *Nucleic Acids Res* 2005;33:D407–D412.

60. Okuno Y, Tamon A, Yabuuchi H, Niijima S, Minowa Y, Tonomura K, Kunimoto R, Feng C. GLIDA: GPCR—ligand database for chemical genomics drug discovery—Database and tools update. *Nucleic Acids Res* 2008;36:D907–D912.
61. Hendlich M, Bergner A, Günther J, Klebe G. Relibase: Design and development of a database for comprehensive analysis of protein-ligand interactions. *J Mol Biol* 2003;326:607–620.
62. Keil M, Exner TE, Brickmann J. Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network. *J Comput Chem* 2004;25:779–789.
63. Jones S, Shanahan HP, Berman HM, Thornton JM. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res* 2003;31:7189–7198.
64. Tsuchiya Y, Kinoshita K, Nakamura H. PreDs: A server for predicting dsDNA-binding site on protein molecular surfaces. *Bioinformatics* 2005;21:1721–1723.
65. Gou IB, Li Z, Hwang RS. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins* 2006;64:19–27.
66. Porollo A, Meller J. Prediction-based fingerprints of protein-protein interactions. *Proteins* 2007;66:630–645.
67. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem* 2002;269: 1356–1361.
68. Keskin O, Ma B, Nussinov R. Hot regions in protein-protein interactions: The organization and contribution of structurally conserved hot spot residues. *J Mol Biol* 2005;345:1281–1294.
69. Jones S, Thornton JM. Analysis of protein–protein interaction sites using surface patches. *J Mol Biol* 1997;272:121–132.
70. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;12:2577–2637.
71. Andrabi M, Mizuguchi K, Sarai A, Ahmad S. Benchmarking and analysis of DNA-binding site prediction using machine learning methods, *Proc. IEEE Int. Joint Conf. Neural Networks*, June 1–6, 2008, Hong Kong NN0554, pp. 1746–1750.
72. Laskowski RA, Thornton JM, Humblet C, Singh J. X-SITE: Use of empirically derived atomic packing preferences to identify favorable interaction regions in the binding sites of proteins. *J Mol Biol* 1996;259:175–201.
73. Raha K, van der Vaart AJ, Riley KE, Peters MB, Westerhoff LM, Kim H, Merz KM Jr. Pairwise decomposition of residue interaction energies using semi empirical quantum mechanical methods in studies of protein-ligand interaction. *J Am Chem Soc* 2005;127:6583–6594.
74. Böhm HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des* 1994;8:243–256.
75. Humphrey W, Dalke A, Schulten K. VMD—Visual molecular dynamics. *J Mol Graph* 1996;14:33–38.
76. Sayle RA, Milner-White EJ. RasMol: Biomolecular graphics for all. *Trends Biochem Sci* 1995;20:374.

77. DeLano WL. *The PyMOL Molecular Graphics System*. San Carlos, CA: DeLano Scientific. 2002. Avaiable at http://www.pymol.org/ (accessed September 1, 2009).
78. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions. *Protein Eng* 1995;8:127–134.
79. Eisenhaber F, Argos P. Improved strategy in analytical surface calculation for molecular system—Handling of singularities and computational efficiency. *J Comput Chem* 1993;14:1272–1280.
80. Hubbard SJ, Thornton JM. NACCESS, Version 2.1.1, Department of Biochemistry and Molecular Biology, University College, London NACCESS. 1996. Available at http://wolf.bms.umist.ac.uk/naccess/ (accessed September 1, 2009).
81. Ahmad S, Gromiha MM, Fawareh H, Saraim A. ASA-View: Graphical representation of solvent accessibilty for the entire Protein Data Bank BMC. *Bioinformatics* 2004;5:51.
82. Laskowski RA, Chistyakov VV, Thornton JM. PDBsum more: New summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res* 2005;33:D266–D268.
83. Ahmad S, Gromiha MM, Sarai A. RVP-net: Online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics* 2003;19: 1849–1851.
84. Wagner M, Adamczak R, Porollo A, Meller J. Linear regression models for solvent accessibility prediction in proteins. *Journal of Computational Biol* 2005;12: 355–369.
85. Perrozo R, Folkers G, Scapozza L. Thermodynamics of protein-ligand interactions: History, presence and future aspects. *J Recept Signal Transduct Res* 2004; 24:1–52.
86. Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A. ProTherm and ProNIT: Thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res* 2006;34:204–206.
87. Kitajima K, Ahmad S, Selvaraj S, Kubodera H, Sunada S, An J, Sarai A. Development of a protein-ligand interaction database, ProLINT, and its application to QSAR analysis. *Genome Inform* 2002;13:498–499.
88. Block P, Sotriffer CA, Dramburg I, Klebe G. AffinDB: A freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res* 2006;34:D522–D526.
89. Liu T, Lin Y, Wen X, Jorrisen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 2006;35:D198–D201.
90. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW. BIND-The Biomolecular Interaction Network Database. *Nucleic Acids Res* 2001;29:242–245.
91. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res* 2004;32:D115–D119.
92. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2002;30:52–55.

93. Cozzini P, Fornabaio M, Marabotti A, Abraham DJ, Kellogg GE, Mozzarelli A. Free energy of ligand binding to protein: evaluation of the contribution of water molecules by computational methods. *Curr Med Chem* 2004;11:3093–3118.
94. Neumann D, Kohlbacher O, Lenhof HP, Lehr CM. Lectin-sugar interaction. Calculated versus experimental binding energies. *Eur J Biochem* 2002;269: 1518–1524.
95. Wang J, Szewczuk Z, Yue SY, Tsuda Y, Konishi Y, Purisima EO. Calculation of relative binding free energies and configurational entropies: A structural and thermodynamic analysis of the nature of non-polar binding of thrombin inhibitors based on hirudin 55-65. *J Mol Biol* 1995;253:473–492.
96. Irwin JJ, Shoichet BK. ZINC—A free database of commercially available compounds for virtual screening. *J Chem Inf Model* 2005;45:177–182.
97. Bashton M, Nobeli I, Thornton JM. PROCOGNATE: A cognate and domain mapping for enzymes. *Nucleic Acids Res* 2008;36:D618–D622.
98. Pearl FM, Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, Sillitoe I, Thornton J, Orengo CA. The CATH database: An extended protein family resource for structural and functional genomics. *Nucleic Acids Res* 2003;31: 452–455.
99. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–540.
100. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. *Nucleic Acids Res* 2002;30:276–280.
101. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res* 2000;28:304–305.
102. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* 2004;32:431–433.
103. Wallace AC, Laskowski RA, Singh J, Thornton JM. Molecular recognition by proteins: Protein-ligand interactions from a structural perspective. *Biochem Soc Trans* 1996;24:280–284.
104. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol* 1997;272:133–143.
105. Aytuna AS, Keskin O, Gursoy A. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* 2005;21:2850–2855.
106. Leach AR, Shoichet BK, Peishoff CE. Prediction of protein-ligand interactions. Docking and scoring: Successes and gaps. *J Med Chem* 2006;49:5851–5855.
107. Lybrand TP. Ligand-protein docking and rational drug design. *Curr Opin Struct Biol* 1995;5:224–228.
108. Lengauer T, Rarey M. Computational methods for biomolecular docking. *Curr Opin Struct Biol* 1996;6:402–406.
109. Taylor RD, Jewsbury PJ, Essex JW. A review of protein-small molecule docking methods. *J Comput Aided Mol Des* 2002;16:151–166.

110. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat Rev Drug Discov* 2004;3:935–949.
111. Ortiz RA, Gomez-Puertas P, Leo-Macias A, Lopez-Romero P, Lopez-Vinas E, Morreale A, Murcia M, Wang K. Computational approaches to model ligand selectivity in drug design. *Curr Top Med Chem* 2006;6:41–55.
112. Rosenfeld R, Vajda S, DeLisi C. Flexible docking and design. *Annu Rev Biophys Biomol Struct* 1995;24:677–700.
113. Abagyan R, Totrov M. High-throughput docking for lead generation. *Curr Opin Chem Biol* 2001;5:375–382.
114. Alvarez JC. High-throughput docking as a source of novel drug leads. *Curr Opin Chem Biol* 2004;8:365–370.
115. Chandonia J-M, Brenner SE. The impact of structural genomics: Expectations and outcomes. *Science* 2006;311:347–351.
116. Seifert MH. ProPose: steered virtual screening by simultaneous protein ligand docking and ligand -ligand alignment. *J Chem Inf Model* 2005;45:449–460.
117. Ewing TJ, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* 2001;15:411–428.
118. Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R. Optimal docking area: A new method for predicting protein–protein interaction sites. *Proteins* 2005;58:134–143.
119. Alzate-Morales JH, Contreras R, Soriano A, Tuñon I, Silla E. A computational study of the protein-ligand interactions in CDK2 inhibitors: Using quantum mechanics/molecular mechanics interaction energy as a predictor of the biological activity. *Biophys J* 2007;92:430–439.
120. Alves CN, Martí S, Castillo R, Andrés J, Moliner V, Tuñón I, Silla E. A quantum mechanics/molecular mechanics study of the protein-ligand interaction for inhibitors of HIV-1 integrase. *Chemistry* 2007;3:7715–7724.
121. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein- ligand interactions. *J Mol Biol* 2000;295:337–356.
122. Muegge I, Martin YC. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J Med Chem* 1999;42:791–804.
123. Ozrin VD, Subbotin MV, Nikitin SM. PLASS: Protein-ligand affinity statistical score- A knowledge-based force-field model of interaction derived from the PDB. *J Comput Aided Mol Des* 2004;18:261–270.
124. Halperin I, Buyong M, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* 2002; 47:409–443.
125. Bissantz C, Folkers G, Rognan D. 2000Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem* 2000;43:4759–4767.
126. Tame JR. Scoring functions: A view from the bench. *J Comput Aided Mol Des* 1999;13:99–108.
127. Böhm HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des* 1994;8:243–256.

128. Raha K, Merz KM Jr. Large-scale validation of a quantum mechanics based scoring function: Predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes. *J Med Chem* 2005;48:4558–4575.

129. McInnes C. Improved lead-finding for kinase targets using high-throughput docking. *Curr Opin Drug Discov Devel* 2006;9:339–347.

130. Böhm HJ. LUDI: Rule-based automatic design of new substituents for enzyme inhibitor leads. *Comput Aided Mol Des* 1992;6:61–78.

131. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* 1997;267:727–748.

132. Trosset JY, Scheraga HA. PRODOCK: Software package for protein modeling and docking. *J Comput Chem* 1999;20:412–427.

133. Liu M, Wang S. MCDOCK: A Monte Carlo simulation approach to the molecular docking problem. *J Comput Aided Mol Des* 1999;13:435–451.

134. Fahmy A, Wagner G. TreeDock: A tool for protein docking based on minimizing van der Waals energies. *J Am Chem Soc* 2002;124:1241–1250.

135. Burkhard P, Taylor P, Walkinshaw MD. An example of a protein ligand found by database mining: Description of the docking method and its verification by a 2.3 Å X-ray structure of a thrombin-ligand complex. *J Mol Biol* 1998;277:449–466.

136. Cole JC, Murray CW, Nissink JW, Taylor RD, Taylor R. Comparing protein-ligand docking programs is difficult. *Proteins* 2005;60:325–332.

137. Stahl M, Bohm HJ. Development of filter functions for protein-ligand docking. *J Mol Graph Model* 1998;16:121–132.

10

ANALYSIS OF TOXICOGENOMIC DATABASES

LYLE D. BURGOON

Table of Contents

10.1 INTRODUCTION

Toxicogenomics may improve quantitative safety and risk assessments by providing a wealth of data that investigators can correlate with mechanism of action and toxic responses through phenotypic anchoring. Organizations using toxicogenomics require data management solutions to manage the overabundance of data generated by toxicogenomic studies, as well as the sample annotation and complementary toxicology and pathology data required for interpretation.

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

Sample annotation allows investigators to identify trends in their data that correlate with phenotypes or which may explain spurious results. Sample annotation includes animal husbandry, caging information, sex, age, body weight, tissue/organ information, gross pathology, treatment/exposure protocols, surgical details, *in vitro* culture conditions, species, and strain information.

Complementary toxicology data allow investigators to correlate toxicogenomic data with standard observable toxicological phenotypes. Investigators can begin to differentiate adaptive responses from mechanistic changes in gene expression through phenotypic anchoring. Complementary toxicology data include changes in body weight, changes in the rate of body weight gain, clinical chemistry, histopathology, changes in morphology, tumorigenesis, and cytotoxicity assays.

Toxicogenomic information management systems (TIMS), specialized relational databases, track these data for organizations. Under ideal conditions, organizations unite their databases into federations or through warehouses to provide a unified data sharing environment. Data federations use specialized software that accepts queries from users and maps the query across several databases. This allows a user to make complicated queries across databases with little knowledge of their structure or the data they contain. Data warehouses take snapshots of their member databases and integrate the data into a larger database. Data integration across an organization within federations and warehouses allows multiple users to have access to the same information for data mining and decision-making purposes.

Many journals require authors to deposit their transcriptomic data within a public repository as a condition of publication. Repositories hold the promise of serving as open source intelligence (OSINT) points that organizations may leverage to obtain new knowledge. For instance, in the future, an organization may obtain all of the transcriptomic data for the members of a particular drug class to identify potential off-target affects.

Several challenges exist with respect to using TIMS and repositories for pharmaceutical data mining. This chapter will discuss these challenges and will address data mining methods using toxicogenomic databases. The chapter will also give several examples of databases and repositories and will discuss the need for toxicogenomic data standards.

10.2 TOXICOGENOMIC DATABASES AND REPOSITORIES

The major toxicogenomic database developers, representing academic, government, and industry interests, have been active since the beginning of the era. Two of the first databases to emerge were dbZach (http://dbzach.fst.msu.edu/; initially, the database of testis expressed transcripts (dbTEST), later renamed as TIMS dbZach) [1,2] and the United States Food and Drug

Administration (U.S. FDA)'s ArrayTrack (http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/) [3,4].

Once toxicogenomic data began to emerge within the literature, the toxicogenomic-specific data repository, Chemical Effects in Biological Systems (CEBS) knowledge base [5], became available. This repository became necessary as the database developers and members of the community realized that the existing microarray repositories lacked support for the minimum information required to interpret a toxicogenomic experiment.

10.2.1 TIMS

Organizations that adopted toxicogenomic methods quickly realized the need to develop data management strategies. Many of these groups decided to maintain the *status quo*: individual data producers/investigators would be required to manage their own data. Other groups decided to utilize their information technology (IT) expertise to adopt database approaches. Two of these database systems (TIMS dbZach and ArrayTrack) have emerged as full-service solutions, now termed TIMS.

The TIMS solutions emerged to address problems associated with data being scattered throughout an organization, issues with data access and data sharing. For instance, organizations that lack data management strategies must back up data located at each terminal, instead of a single database, increasing the risk of irreparable data loss. In addition, in order for individuals within the organization to share data, they must first locate who is in charge of the data and must request access to it. If the user in charge of the data grants the request, both users must figure out a way to share the data with each other.

These issues with data access and sharing create internal data silos that may reinforce negative organizational politics. The problem with data silos is the creation of "gatekeeper" effects, where a single individual controls access to a segment of information/data. This creates a situation where the gatekeepers may feel that the data they control is theirs and are not the property of the organization. For example, if an investigator is trying to identify potential within-class toxic effects, but the data for each compound exists on several different systems controlled by different individuals, the investigator must communicate with each gatekeeper to access all of the information that they require. This breeds inefficiencies with respect to identifying the gatekeepers, requesting permission to obtain organizational data, and the associated wait times. Although less likely to occur within smaller organizational structures with a powerful centralized management, these types of counterproductive silos exist within all types of organizations, from small academic research laboratories to large multinational corporations.

Database systems, including TIMS, prevent the gatekeeper effect by centralizing the access decisions for similar types of organizational data. For instance, a single TIMS could manage the data for an entire division or for an

entire laboratory, depending upon the organizational structure. TIMS built on scalable architectures, including server hardware and appropriate database software, would grow with the organization's needs. TIMS also incorporate best practices for data security and risk management. Rather than requiring a complex data recovery plan for every terminal where data may reside, a more simplified plan would cover the single source of data—the TIMS.

There are, however, drawbacks to implementing TIMS. If an organization hosts/implements an in-house TIMS, they will require a database administrator to maintain the system, ensure it runs at peak performance, and secure the data. TIMS generally require specialized software and user training. Unless purchasing or acquiring an existing solution, the organization may have to create one, which can require significant time and expense. However, investigators have to weigh the return on investment (ROI) from implementing a TIMS against the cost of decreased or impaired user access to data, the existence of data silos, and impending inefficiencies.

10.2.1.1 TIMS dbZach and ArrayTrack TIMS dbZach and ArrayTrack support local analysis and interpretation of toxicogenomic data, including gene, protein, and metabolite expression data [1–4]. These TIMS systems consist of database back-end and front-end applications for users to query and upload to the database. Although an academic group created TIMS dbZach and the FDA created ArrayTrack, both systems aim to facilitate the data management and analysis needs of individual organizations.

TIMS dbZach grew out of the need to manage the entire microarray process, from construction to gene expression analysis through functional annotation and phenotypic anchoring. To accommodate growth, the system uses a modular design, with each module corresponding to a specific theme, such as subsystems for clone management, microarray data management, gene annotation, sample annotation, and toxicology. The modular design allows the developers to design modifications and upgrades with minimal impact on the rest of the system. This has facilitated the creation of new subsystems to manage orthologous gene relationships, metabolomics, and gene regulation [1,2]. Currently TIMS dbZach is available through arrangement with Michigan State University (http://dbzach.fst.msu.edu).

ArrayTrack consists of a microarray database and an analysis suite developed by the National Center for Toxicology Research at the U.S. FDA (http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/). ArrayTrack, similar to TIMS dbZach, stores the microarray data and allows users to perform functional analysis by linking the data to pathways and gene ontology data. Currently, the FDA uses ArrayTrack in the review of genomic data submitted by study sponsors. ArrayTrack also accepts data uploads in the SimpleTox format, based on the Standard for Exchange of Nonclinical Data (SEND) v2.3 and the Study Data Tabulation Model (SDTM) [3,4].

10.2.2 Toxicogenomic Data Repositories

Investigators who intend to publish toxicogenomic data may be required to submit their data to a repository. Generally, most journals require authors to submit microarray data to either the Gene Expression Omnibus (GEO, National Institutes of Health) or ArrayExpress (European Bioinformatics Institute). The sentiment of the toxicogenomic community, however, has been that neither GEO [6] nor ArrayExpress [7] captures all of the information required to interpret a toxicogenomic experiment [8].

For instance, GEO and ArrayExpress both handle the descriptions of the biological specimen and the assays well; however, what they lack are the subject characteristics and handling of the study design and execution—specifically the procedures and their timeline. All of these study characteristics are essential attributes of any toxicology experiment that may place an experiment within the proper context and may influence the data interpretation. These concerns, among others, motivated the creation of the CEBS knowledge base [5,9,10].

CEBS builds upon the data management provided by the other microarray repositories. Specifically, CEBS includes the ability to manage complex experimental timelines and exposure procedures, as well as histopathology and clinical pathology data. This allows users to view the gene expression data in light of the exposure parameters and any relevant toxicity data. Currently, CEBS contains 11 mouse and 22 rat studies across 136 chemicals and 32 microarray studies (http://cebs.niehs.nih.gov/, accessed January 30, 2008).

10.3 TOXICOGENOMIC DATA STANDARDS

When mentioning gene expression data standards, most scientists think of Minimum Information about a Microarray Experiment (MIAME). The Microarray Gene Expression Data Society (MGED) created MIAME to guide scientific investigators, journal editors, and reviewers in the minimum amount of information required for a scientist to reproduce a microarray experiment [11]. It is important to note that MIAME is not a standard, *per se*, but rather a guidance. Unfortunately, many scientists, journal editors, and reviewers regard it to be the *de facto* microarray data standard [12].

Regardless of its standing, several toxicologists associated with MGED expanded upon the MIAME document to create a toxicogenomic-specific version, called MIAME/Tox (http://www.mged.org/MIAME1.1-DenverDraft.DOC). MIAME/Tox expanded many of the existing MIAME definitions within the experiment description section to include toxicologically relevant examples. For instance, MIAME/Tox suggests the inclusion of necropsy, histopathology, and clinical pathology data.

Recently, a diverse group of toxicogenomic scientists from government, academia, and industry proposed a preliminary checklist for toxicology data.

Although not yet a standard, this proposal outlines the details that the authors feel every toxicology study should include [8]. Among the details are specifications of the subject and procedure characteristics, study design, execution, and timeline, as well as any clinical pathology and histopathology data. The authors also state that they would prefer to see complementary, corroborating data for each study (e.g., clinical pathology and histopathology, or gene expression and histopathology).

10.3.1 Regulatory Guidance from the U.S. FDA and U.S. Environmental Protection Agency (EPA)

The U.S. FDA and U.S. EPA have created guidance documents outlining their plans for using genomic data in their decision-making process. It is important to note that these documents are not regulatory standards and are only industry guidance that reflects each agency's current thinking on the topic of pharmaco- and toxicogenomic data submissions.

Both agencies have stated that they will accept toxicogenomic data. However, neither agency believes that toxicogenomic data alone is sufficient to make a regulatory decision [13–16]. Generally, FDA regulatory mandates only exist if using known valid biomarkers, or probable valid biomarkers when the sponsor uses the data in the decision-making process (http://www.fda.gov/cder/genomics/QuickReference.pdf and http://www.fda.gov/cder/guidance/6400fnlAttch.pdf). The FDA encourages voluntary submission of genomic data through their Voluntary Genomic Data Submission program. The EPA still needs to work out how it will handle toxicogenomic data. The EPA's current position, that there is a relationship between toxicogenomic data and adverse outcomes, remains unclear. Thus, changes in gene expression are not subject to reporting under the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) and the Toxic Substances Control Act (TSCA) [14]. In addition, EPA reserves the right to use toxicogenomic data on a case-by-case basis when determining whether to list/delist a chemical from the Toxic Release Inventory [14].

10.3.2 Standards versus Guidelines

MIAME/Tox inherits all of the strengths and weaknesses of the parent MIAME document—a well-built guidance that the toxicogenomics community has forced to become a standard. There is generally widespread confusion as to the difference between the terms "guidance" and "standard," with the community using both terms interchangeably, or with significant indifference. What the community does not realize is the impact that this indifference has created and the problems associated with misrepresenting and misinterpreting guidance as a standard [12,17].

Standards define a strict set of properties that implementers must meet for compliance. However, guidelines define the best practice. Standards denote

minimum requirements, allowing an implementer to be either within or out of compliance. Guidance documents differ from standards by not making any minimum requirements. Thus, when someone implements guidance, their data cannot be within or out of compliance.

Consider the GEO and ArrayExpress microarray data repositories. As of the time of writing, both repositories claimed to be "MIAME compliant." These organizations perceive MIAME to be a standard by their use of the term compliant, and use the term to help editors, reviewers, and authors to justify their use of these resources (e.g., when a journal requires adherence to MIAME). In spite of their compliance with MIAME, Table 10.1 shows differences in their implementations and begins to demonstrate the danger inherent with elevating guidance to a standard.

By applying the MIAME guidance as a standard, the National Center for Biotechnology Information (creator of GEO) and the European Bioinformatics Institute (creator of the ArrayExpress) have created a division within the microarray community. Specifically, users generally must choose whether they will submit to the GEO or ArrayExpress repository to meet their journal's publication requirements governing microarray data deposition [17]. Although it is possible for scientists to format their data for deposition in both systems, this would require far more time than to submit to only one system. If MGED had written MIAME as a proper standard, this would not be a problem as GEO and ArrayExpress could share data.

This demonstrates the point that toxicogenomics requires real standards. With real standards in place, with clear requirements, groups can begin to create databases that can automatically share data. Consider what is required for two computer systems to share data in a meaningful manner between organizations. Both systems must recognize and expect the same data in order to communicate—they have to agree on a communications protocol. For two preclinical toxicology database systems to intercommunicate, they must agree to use the same terminology and to expect drug dose information in a specific number and unit format, and alanine aminotransferase activity (ALT) data either numerically or as text ("high," "low," or "normal"). If one system reports ALT data textually based on an internal standard, and another system reports the data as a number and unit, the two systems are not interoperable.

Just as an organization would not purchase a piece of software written to run on Windows for a Linux computer, they need to make sure that their choice of database software will enable them to use it in their downstream analyses, or to be included in any regulatory filings. This means organizations must consider how they plan to use their toxicogenomic data. If the organization plans to deposit data into CEBS, it would be helpful to choose a system that has an automated upload path to CEBS (i.e., CEBS and that database system have pre-agreed on a set standard they will follow for communication). Until toxicogenomic standards emerge, investigators must ensure their analysis and data management software interoperate with their proposed regulatory and data deposition workflows.

TABLE 10.1 Comparison of ArrayExpress and Gene Expression Omnibus (GEO) MIAME Implementations

MIAME Standard (MIAME Checklist)	ArrayExpress (Based on MIAMExpress)	GEO	Comments
The goal of the experiment—one line maximum (e.g., the title from the related publication)	Experiment name	Series title	
A brief description of the experiment (e.g., the abstract from the related publication)	Experiment description	Series summary	
Keywords, for example, time course, cell type comparison, array CGH (The use of MGED ontology terms is recommended.)	Experimental design type	Series type	
Experimental factors—the parameters or conditions tested, such as time, dose, or genetic variation (The use of MGED ontology terms is recommended.)	Experimental factors	Specified through "variable subsets" of the series submission	
Experimental design—relationships between samples, treatments, extracts, labeling, and arrays (e.g., a diagram or table)	Inferred	Inferred	
Quality control steps taken (e.g., replicates or dye swaps)	Quality-related indicators (allow user specification of quality control steps outside of dye swaps and replication)	Specified through "repeats" of the series submission (only allows replicates and dye swaps)	Some authors may utilize other quality control methods, such as comparison to historically defined data sets of known quality. This is supported by ArrayExpress but not in GEO.
Links to the publication, any supplemental websites, or database accession numbers	Publication of experiment	Series PubMed ID	

The origin of each biological sample (e.g., name of the organism, the provider of the sample) and its characteristics (e.g., gender, age, developmental stage, strain, or disease state)	Sample description; ArrayExpress only requires a sample name, organism, and sample type. The other fields are not required.	Sample organism (required), sample name (required), and sample characteristics; no listing of required information for sample characteristics given	Authors to choose what information is submitted.
Manipulation of biological samples and protocols used (e.g., growth conditions, treatments, separation techniques)	Growth conditions and sample treatment protocols	Sample protocol fields	There is no standardized format for protocols in either system. Authors enter the level of detail they feel is sufficient. For example, some editors/reviewers may require information on the source of a gavage needle used for treatment, while others may feel the gauge is sufficient.
Experimental factor value for each experimental factor, for each sample (e.g., "time = 30 minutes" for a sample in a time course experiment).	Additional qualifiers	Specified through "variable subsets" of the series submission	
Technical protocols for preparing the hybridization extract (e.g., the RNA or DNA extraction and purification protocol) and labeling	Extract protocol and labeled extract protocol	Extract and labeling protocols	There is no standardized format for protocols in either system. Authors enter the level of detail they feel is sufficient.
External controls (spikes), if used	NA—would probably have to be specified in protocol. ArrayExpress does track feature locations for spike-in controls.	NA—would probably have to be specified in protocol	No specific reporting requirement exists. When used, investigators should specify what features on the array probe are spike-ins, at what level probe is spiked into the sample, and the purpose of the spike-in control. None of these requirements are specified by MIAME.

TABLE 10.1 *Continued*

MIAME Standard (MIAME Checklist)	ArrayExpress (Based on MIAMExpress)	GEO	Comments
The protocol and conditions used for hybridization, blocking, and washing, including any postprocessing steps such as staining	Hybridization protocol	Hybridization protocol	There is no standardized format for protocols in either system. Authors enter the level of detail they feel is sufficient.
The raw data, i.e., scanner or imager and feature extraction output (providing the images is optional). The data should be related to the respective array designs (typically each row of the imager output should be related to a feature on the array—see Array Designs).	Raw data file	Sample table	GEO prefers dual channel data from loop designs to be submitted as single-channel data; however, the assignment of samples to microarrays is lost, which precludes certain types of analyses. GEO only requires the normalized value (single- and dual-channel loop-type designs), or normalized ratios (single channel), and the feature identifier. ArrayExpress requires the actual raw data file (.EXP and .CEL for Affymetrix, and .txt or .gpr files from other analysis platforms). GEO allows other data fields to also be included, such as background intensity data.

The normalized and summarized data, i.e., set of quantifications from several arrays upon which the authors base their conclusions (for gene expression experiments also known as gene expression data matrix and may consist of averaged normalized log ratios). The data should be related to the respective array designs (typically each row of the summarized data will be related to one biological annotation, such as a gene name).	Normalized data file	Sample table	ArrayExpress supports submission of averaged data matrices (also known as final gene expression data matrices). This would be a file containing the normalized average expression for each gene under a specified condition. In contrast, GEO does not support averaged data matrices in their submission, preferring data of the type listed above. An averaged data matrix format is acceptable under the MIAME convention.
Image scanning hardware and software, and processing procedures and parameters	Scanning protocol	Scan protocol	There is no standardized format for protocols in either system. Authors enter the level of detail they feel is sufficient.
Normalization, transformation, and data selection procedures and parameters	Normalization protocol	Data processing	There is no standardized format for protocols in either system. Authors enter the level of detail they feel is sufficient.
General array design, including the platform type (whether the array is a spotted glass array, an *in situ* synthesized array, etc.); surface and coating specifications and spotting protocols used (for custom made arrays), or product identifiers (the name or make, catalog reference numbers) for commercially available arrays	Array definition file and array protocol	Platform descriptive fields	The data required is generally the same between the two systems.

TABLE 10.1 *Continued*

MIAME Standard (MIAME Checklist)	ArrayExpress (Based on MIAMExpress)	GEO	Comments
For each feature (spot) on the array, its location on the array (e.g., metacolumn, metarow, column, row) and the reporter present in the location (Note that the same reporter may be present on several features.)	Array definition file	Complete feature location annotation is not required; specification is supported through a nonstandard category specification.	MIAME requires feature location annotation to be provided; however, ArrayExpress and GEO differ in their interpretation. ArrayExpress requires full location information (major grid coordinates and minor grid coordinates) while GEO does not require this information to be reported
For each reporter, unambiguous characteristics of the reporter molecule, including reporter role—control or measurement; the sequence for oligonucleotide-based reporters; the source, preparation, and database accession number for long (e.g., cDNA or polymerase chain reaction (PCR) product based) reporters; primers for PCR product-based reporters	Array definition file; primers for PCR-based reporters not supported	None are required, but may be specified through standard and nonstandard category specifications	GEO has the flexibility to support all of these entries but does not require any of them. ArrayExpress does not require primers for PCR-based reporters to be reported.
Appropriate biological annotation for each reporter, for instance, a gene identifier or name (Note that different reporters can have the same biological annotation.)	Array definition file	None are required, but may be specified through standard and nonstandard category specifications	As biological annotation often changes with sequencing and annotation projects mature, these data should always be associated with the source database, the build of the source database, and the date it was accessed.

NA = not applicable.

10.4 DATA EXTRACTION AND DATA MINING

Useful databases share an important trait: the ability to extract data from it. Users accomplish this by querying the database using the structured query language (SQL). Database developers, software vendors, and others create graphical user interfaces (GUIs) and web interfaces to allow users to query the database without having to know SQL. However, users with SQL and analytical experience, as well as database access, may be able to use programs, such as R and SAS, to combine data extraction with data mining.

TIMS dbZach and ArrayTrack are distributed with GUIs to facilitate data extraction and analysis. The TIMS dbZach GUIs provide data extraction methods for sample and gene annotation analysis in other softwares. The GUIs extract the data in tab-delimited text files that are readable by other softwares, such as R, SAS, GeneSpring, and Microsoft Excel. ArrayTrack allows for similar data extraction; however, its intent is for the application to be a complete analytical suite. CEBS uses a web GUI for all user interactions, including data download.

Data mining does not occur within the database itself, but generally through specialized software. This software may interface directly with the database or it may require the user to extract the data from the database. For instance, the programming languages R and SAS provide direct communication to most database engines, including Oracle, Microsoft SQL Server, and MySQL. In some instances, database developers include specialized codes, referred to as stored procedures or functions, within the database to speed commonly used queries. For instance, the dbZach installation at Michigan State University includes stored procedures for reporting microarray data and for monitoring database use. This allows users to use more simplified codes to query data from dbZach using R and SAS.

10.5 IS A TIMS RIGHT FOR YOU?

It is essential to understand the client's needs and the current operating environment when designing a data management solution. As noted previously, designers need to consider several factors, including the organization's goals, the regulatory context, the data mining software used, the operating system and hardware used, whether the system integration into the organization's IT infrastructure is necessary, and whether or not deposition in a repository is necessary. The designer must consider the requirements and constraints with respect to personnel expertise and additional training that may be required. For instance, does the organization have database administrators with the proper expertise to manage these databases?

Although a toxicogenomic database may be an asset in the business context (i.e., an item that adds value to the organization, or may generate revenue in some way), organizations must consider the ROI. For low-throughput orga-

nizations, where toxicogenomics is not a key business activity, a database may be more of a liability, especially with respect to resources used to train personnel and acquisition of hardware. However, larger organizations that use toxicogenomics on a regular basis may see benefits from using a database. For instance, our laboratory has seen strong ROI from automating portions of the data analysis pipeline, clone management during the construction process, development of quality assurance and quality control protocols, and from cross-study analyses that the database facilitated.

TIMS have helped several organizations with respect to their data management and analysis needs. As toxicogenomic standards develop and as the regulatory community adopts toxicogenomic technologies, TIMS systems will align. Community data sharing facilitated by the use of repositories such as CEBS will increase, and automated data sharing pipelines will emerge. This will facilitate the integration and fusion of toxicogenomic data, including histopathology and clinical pathology, and other classic toxicology assays.

REFERENCES

1. Burgoon LD, Boutros PC, Dere E, Zacharewski TR. dbZach: A MIAME-compliant toxicogenomic supportive relational database. *Toxicol Sci* 2006;90: 558–568.
2. Burgoon LD, Zacharewski TR. dbZach toxicogenomic information management system. *Pharmacogenomics* 2007;8:287–291.
3. Tong W, Harris S, Cao X, Fang H, Shi L, Sun H, Fuscoe J, Harris A, Hong H, Xie Q, Perkins R, Casciano D. Development of public toxicogenomics software for microarray data management and analysis. *Mutat Res* 2004;549:241–253.
4. Tong W, Cao X, Harris S, Sun H, Fang H, Fuscoe J, Harris A, Hong H, Xie Q, Perkins R, Shi L, Casciano D. ArrayTrack—Supporting toxicogenomic research at the U.S. Food and Drug Administration National Center for Toxicological Research. *Environ Health Perspect* 2003;111:1819–1826.
5. Waters M, Boorman G, Bushel P, Cunningham M, Irwin R, Merrick A, Olden K, Paules R, Selkirk J, Stasiewicz S, Weis B, Van Houten B, Walker N, Tennant R. Systems toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base. *EHP Toxicogenomics* 2003;111:15–28.
6. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;30: 207–210.
7. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone S. ArrayExpress—A public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2003;31:68–71.
8. Fostel JM, Burgoon L, Zwickl C, Lord P, Christopher CJ, Bushel PR, Cunningham M, Fan L, Edwards SW, Hester S, Stevens J, Tong W, Waters M, Yang C, Tennant R. Toward a checklist for exchange and interpretation of data from a toxicology study. *Toxicol Sci* 2007;99:26–34.

9. Fostel J, Choi D, Zwickl C, Morrison N, Rashid A, Hasan A, Bao W, Richars A, Tong W, Bushel PR, Brown R, Bruno M, Cunningham ML, Dix D, Eastin W, Frade C, Garcia A, Heinloth A, Irwin R, Madenspacher J, Merrick BA, Papoian T, Paules R, Rocca-Serra P, Sansone AS, Stevens J, Tower K, Yang C, Waters M. Chemical Effects in Biological Systems—Data Dictionary (CEBS-DD): A compendium of terms for the capture and integration of biological study design description, conventional phenotypes, and 'omics data. *Toxicol Sci* 2005;88:585–601.
10. Xirasagar S, Gustafson SF, Huang C-C, Pan Q, Fostel J, Boyer P, Merrick BA, Tomer KB, Chan DD, Yost KJ III, Choi D, Xiao N, Stasiewicz S, Bushel P, Waters MD. Chemical Effects in Biological Systems (CEBS) object model for toxicology data, SysTox-OM: Design and application. *Bioinformatics* 2006:22:874–882.
11. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum Information about a Microarray Experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001;29:365–371.
12. Burgoon LD. Clearing the standards landscape: the semantics of terminology and their impact on toxicogenomics. *Toxicol Sci* 2007;99:403–412.
13. Dix DJ, Gallagher K, Benson WH, Groskinsky BL, McClintock JT, Dearfield KL, Farland WH. A framework for the use of genomics data at the EPA. *Nat Biotechnol* 2006;24:1108–1111.
14. Benson WH, Gallagher K, McClintock JT. U.S. Environmental Protection Agency's activities to prepare for regulatory and risk assessment applications of genomics information. *Environ Mol Mutagen* 2007;48:59–362.
15. Goodsaid F, Frueh FW. Implementing the U.S. FDA guidance on pharmacogenomic data submissions. *Environ Mol Mutagen* 2007;48:354–358.
16. Frueh FW. Impact of microarray data quality on genomic data submissions to the FDA. *Nat Biotechnol* 2006;24:1105–1107.
17. Burgoon LD. The need for standards, not guidelines, in biological data reporting and sharing. *Nat. Biotechnol* 2006;24:1369–1373.

11

BRIDGING THE PHARMACEUTICAL SHORTFALL: INFORMATICS APPROACHES TO THE DISCOVERY OF VACCINES, ANTIGENS, EPITOPES, AND ADJUVANTS

MATTHEW N. DAVIES AND DARREN R. FLOWER

Table of Contents

11.1 INTRODUCTION

It is a truth universally acknowledged that mass vaccination, with the possible exception of public sanitation, is the most effective prophylactic for infectious disease. Over 70 infectious diseases commonly affect the human species. Many of these have been targeted successfully with vaccines, and there are now over

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

50 licensed vaccines, half in common use. Most prevent childhood infections or are used by travelers to tropical or subtropical regions; a significant minority are used to combat disease in developing nations. In the First World, the annual mortality for diseases such as smallpox, polio, diphtheria, or measles is less than 0.1%. Activity in the vaccine arena, neglected for several decades, is now frantic. Dozens of vaccine candidates have passed through phase II clinical trials, and during the past decade, vaccines in late development have numbered over 150. Unlike antibiotics, resistance to vaccines is negligible.

Despite such palpable success, major problems remain. No licensed vaccines exist for HIV [1] and malaria [2], two of the World Health Organization (WHO)'s three big global killers, and there are no realistic hopes for such vaccines appearing in the short to medium term. The only vaccine licensed for the third major world disease, tuberculosis, has only limited efficacy [3]. Moreover, the mortality and morbidity for several diseases, which are targeted by vaccines, remain high, for example, influenza, with an annual global estimate of half a million deaths. Add to this the 35 new, previously unknown infectious diseases identified in the past 25 years: HIV, Marburg's disease, severe acute respiratory syndrome (SARS), dengue, West Nile, and over 190 human infections with the potentially pandemic H5N1 influenza. It is commonly believed that new contagious diseases will continue to emerge in the 21st century. The world of the 21st century is threatened by parasitic diseases such as malaria, visceral leishmaniasis, tuberculosis, and emerging zoonotic infections, such as H5N1; antibiotic-resistant bacteria; and bioterrorism—a threat compounded by a growing world population, overcrowded cities, increased travel, climate change, and intensive food production.

An inability to exploit new disease targets and increased regulatory pressure has reduced research and development pipelines within the pharmaceutical industry. At the same time, the industry has faced growing competition from generics [4] and so-called me-too drugs [5]. Together this has led the industry to question the business models it has exploited so well over the last 50 years: selling drugs of often marginal therapeutic advantage predominantly to the First World. Plugging the gap left by this dwindling pipeline has become a priority. Targeting infectious diseases is part of this. Infection can be controlled using an intelligent combination of anti-infective drugs, both antivirals and antibiotics, vaccines, and diagnostics. Over the last decade, there has been a huge increase in the number of drugs targeting viruses. This was partly driven by attempts to contain pandemic HIV. Increased understanding of viral life cycles has identified new viral target proteins, including neuraminidases needed for viral release from the cell, proteases that cleave viral polyproteins, RNA- and DNA-dependent helicases and polymerases, and enzymes responsible for viral genome replication. Most recently, understanding viral entry into cells has led to the development of so-called fusion inhibitors [6] such as the anti-AIDS drug enfuvirtide. The postwar golden age of antibiotics ended long ago. Now, widespread misuse of antibiotics has engendered equally widely spread antibiotic resistance. Despite success in discovering new vac-

cines and antivirals, there has been little or no major success in developing novel antibiotics. Diseases can also be addressed using immunotherapy and biopharmaceuticals, such as therapeutic antibodies [7].

Persistent infection, which includes HIV, hepatitis B, hepatitis C, and tuberculosis, occurs when a pathogen evades or subverts T-cell responses and is a key therapeutic target. At the other extreme are benign yet economically important infections, such as the common cold. Respiratory tract infections remain the major cause of community morbidity and hospitalization in the developed world, accounting for about 60% of general practitioner referrals and causing the loss of a huge number of working days. Sporadic or epidemic respiratory infections are causing by over 200 distinct viruses, including coronaviruses, rhinoviruses, respiratory syncytial virus (better known as RSV), parainfluenza virus, influenza A and B, and cytomegalovirus. Antiallergy vaccination also offers great potential for successful commercial exploitation. This often relies on allergen-specific short-term immunotherapy (STI) [8], where patients are administered increasing quantities of allergen to augment their natural tolerance. STI, though often effective, is very time consuming and is not universally applicable. Recombinant hypoallergenic allergens are also of interest as they can target specific immune cells. New agents for the prophylaxis and treatment of allergic disease are legion: recombinant proteins, peptides, immunomodulatory therapy, and DNA vaccines, which are particularly promising tools. Several antiallergy DNA vaccines are being developed, including optimized expression of allergen genes, CpG enrichment of delivery vectors, and the targeting of hypoallergenic DNA vaccines. Vaccines against the common cold or antiallergy vaccines lie close to so-called lifestyle vaccines. None of these vaccines necessarily saves lives but does reduce hugely important economic effects of disease morbidity. Lifestyle vaccines target dental caries and drug addiction, as well as genetic and civilization diseases, such as obesity. Genetic diseases arise from Mendelian or multifactorial inheritance. Multifactorial diseases arise from mutations in many different genes and have a major environmental component. Heart disease, diabetes, and asthma are in part all multifactorial disorders.

The other significant area where vaccination strategies are being investigated is cancer. An example is GARDASIL, the new human papillomavirus vaccine [9], licensed in 2006 with the aim of preventing 4000 deaths annually from cervical cancer. Cancer is the second greatest cause of death in the developed world after cardiovascular disease; yet most of the 250,000 deaths from cervical cancer occur in the third world. Cancer treatment typically involves a combination of chemotherapy, radiotherapy, and surgery. While treating primary tumors this way is largely successful, preventing the metastatic spread of disease is not. Cancer vaccines are attractive, both clinically and commercially, since they exploit immunity's ability to recognize and destroy tumors. Tumor cells express several mutated or differentially expressed antigens, enabling the immune system to discriminate between nonmalignant and cancerous cells. Tumor antigens form the basis of both subunit and

epitope-based vaccines. Host immune system responses to tumor antigen cancer vaccines are often weak, necessitating the use of adjuvants.

In 2000, the annual sales of vaccines stood at approximately $5 billion; 6 years on and the global vaccine marketplace had grown to $10.8 billion in 2006. Of course, such a figure needs to be set against the total size of the pharmaceutical market. In 2000, the total sales for all human therapies (small molecules, vaccines, therapeutic antibodies, etc.) were about $350 billion. By 2004, global sales had reached $550 billion. This represented a 7% increase on 2003 sales, which in turn was a 9% rise compared to 2002. At the same time, the farm livestock health market was worth around $18 billion and the companion animal health market was valued at about $3 billion. Currently, vaccines form only a very small part of the wider marketplace for medicines and pharmaceutical therapy. Compared to drugs designed to battle cholesterol, high blood pressure, and depression, vaccines have long been the poor relations of the pharmaceutical industry, and it is still true that vaccines remain a neglected corner of the global drug industry. Indeed, at $10.8 billion, vaccines make less than the $13 billion generated by Lipitor, which, as we saw, is currently the world's top seller. Likewise, we see a similar phenomenon if we compare vaccines to the protein therapeutics market, which was valued at $57 billion in 2005. The market for therapeutic antibodies was worth an estimated $13.6 billion, accounting for more than 24% of the total biotech market. However, sales of vaccines have been growing at or about 10–12%, compared to a more modest annual figure of 5–6% for small-molecule drugs. Annual growth in the vaccine sector is expected to approach 20% during the next 5 years.

Viewed commercially, vaccines have many attractive characteristics; compared to small-molecule drugs, vaccines are more likely to escape the development pipeline and to reach the market, with 70% of programs gaining regulatory approval. Vaccines also enjoy long product half-lives since vaccine generics are virtually nonexistent. About 90% of all vaccines are sold directly to governments and public health authorities, so they have much smaller marketing costs. However, until relatively recently, vaccines were not considered to be an attractive business, since pricing was unattractive, profit margins slight, and there was an unvarying threat of litigation.

Vaccination, until relatively recently, has been a highly empirical science, relying on tried-and-tested yet poorly understood approaches to vaccine development. As a consequence of this, relatively few effective vaccines were developed and deployed during the 150 years following Jenner. What low-hanging fruit there was could be picked with ease, but most targets remained tantalizingly out of reach. A vaccine is a molecular or supramolecular agent that can elicit specific protective immunity and can ultimately mitigate the effect of subsequent infection. Vaccination is the use of a vaccine, in whatever form, to produce prophylatic active immunity in a host organism. Vaccines have taken many forms. Until recently, they have been attenuated or inactivated whole pathogen vaccines such as antituberculosis Bacille

Calmette-Guérin (BCG) or Sabin's vaccine against polio. Safety difficulties have lead to the subsequent development of other strategies for vaccine development. The most successful alternative has focused on the antigen or subunit vaccine, such as the recombinant hepatitis B vaccine [10]. Vaccines based around sets of epitopes have also gained ground in recent years. They can be delivered into the host in many ways: as naked DNA vaccines, using live viral or bacterial vectors, and via antigen-loaded antigen-presenting cells. Adjuvants are substances, such as alum, that are used with weak vaccines to increase immune responses [11].

Immunomics is a post-genomic systems biology approach to immunology that explores mechanistic aspects of the immune system [12]. It subsumes immunoinformatics and computational vaccinology, combining several fields, including genomics, proteomics, immunology, and clinical medicine. To date, a key focus of immunomics has been the development of algorithms for the design and discovery of new vaccines. The success of a vaccine can be measured by its strength, its specificity, the duration of the immune response, and its capacity to create immunological memory. While it is possible to assess these properties in the laboratory, it is not feasible to do on the scale of a large pathogenic genome. With more and more pathogen genomes being fully or partially determined, developing *in silico* methods able to identify potential vaccine candidates has become an imperative. To that end, many computational techniques have been applied to vaccine design and delivery. Here we outline currently available techniques and software for vaccine discovery as well as examples of how such algorithms can be applied. We concentrate on four areas: antigen prediction, epitope prediction, vector design, and adjuvant identification.

11.2 PREDICTING ANTIGENS

The word antigen has a wide meaning in immunology. We use it here to mean a protein, specifically one from a pathogenic microorganism, that evokes a measurable immune response. Pathogenic proteins in bacterial are often acquired, through a process summarized by the epithet horizontal transfer, in groups. Such groups are known as pathogenicity islands. The unusual G + C content of genes and particularly large gene clusters is tantamount to a signature characteristic of genes acquired by horizontal transfer. Genome analysis at the nucleic acid level can thus allow the discovery of pathogenicity islands and the virulence genes they encode.

Perhaps the most obvious antigens are virulence factors (VFs): proteins that enable a pathogen to colonize a host or to induce disease. Analysis of pathogens, such as *Vibrio cholerae* or *Streptococcus pyogenes*, has identified coordinated "systems" of toxins and VFs which may comprise over 40 distinct proteins. Traditionally, VFs have been classified as adherence/colonization factors, invasions, exotoxins, transporters, iron-binding siderophores, and

miscellaneous cell surface factors. A broader definition groups VFs into three: "true" VF genes, VFs associated with the expression of true VF genes, and VF "lifestyle" genes required for colonization of the host [13].

Several databases that archive VFs exist. The Virulence Factor Database (VFDB) contains 16 characterized bacterial genomes with an emphasis on functional and structural biology and can be searched using text, Basic Local Alignment Search Tool (BLAST), or functional queries [14]. The ClinMalDB U.S. database was established following the discovery of multigene families encoding VFs within the subtelomeric regions of *Plasmodium falciparum* [15] and *Plasmodium vivax* [16]. TVFac (Los Alamos National Laboratory Toxin and VFDB) contains genetic information on over 250 organisms and separate records for thousands of virulence genes and associated factors. The Fish Pathogen Database, set up by the Bacteriology and Fish Diseases Laboratory, has identified over 500 virulence genes using fish as a model system. Pathogens studied include *Aeromonas hydrophila*, *Edwardsiella tarda*, and many *Vibrio* species. *Candida albicans* virulence factor (CandiVF) is a small species-specific database that contains VFs, which may be searched using BLAST or a HLA-DR hotspot prediction server [17]. PHI-base is a noteworthy development since it seeks to integrate a wide range of VFs from a variety of pathogens of plants and animals [18]. Obviously, antigens need not be VFs, and another nascent database is intending to capture a wider tranche of data. We are helping to develop the AntigenDB database (http://www.imtech.res.in/raghava/antigendb/), which will aid considerably this endeavor.

Historically, antigens have been supposed to be secreted or exposed membrane proteins accessible to surveillance of the immune system. Subcellular location prediction is thus a key approach to predicting antigens. There are two basic kinds of prediction method: manual construction of rules of what determines subcellular location and the application of data-driven machine learning methods, which determine factors that discriminate between proteins from different known locations. Accuracy differs markedly between different methods and different compartments, mostly due to a paucity of data. Data used to discriminate between compartments include the amino acid composition of the whole protein; sequence-derived features of the protein, such as hydrophobic regions; the presence of certain specific motifs; or a combination thereof.

Different organisms evince different locations. PSORT is a knowledge-based, multicategory prediction method, composed of several programs, for subcellular location [19]; it is often regarded as a gold standard. PSORT I predicts 17 different subcellular compartments and was trained on 295 different proteins, while PSORT II predicts 10 locations and was trained on 1080 yeast proteins. Using a test set of 940 plant proteins and 2738 nonplant proteins, the accuracy of PSORT I and II was 69.8% and 83.2%, respectively. There are several specialized versions of PSORT. iPSORT deals specifically with secreted, mitochondrial, and chloroplast locations; its accuracy is 83.4% for plants and 88.5% for nonplants. PSORT-B only predicts bacterial sub-

cellular locations. It reports precision values of 96.5% and recall values of 74.8%. PSORT-B is a multicategory method that combines six algorithms using a Bayesian network.

Among binary approaches, arguably the best method is SignalP, which employs neural networks and predicts N-terminal Spase-I-cleaved secretion signal sequences and their cleavage site [20]. The signal predicted is the type II signal peptide common to both eukaryotic and prokaryotic organisms, for which there is a wealth of data, in terms of both quality and quantity. A recent enhancement of SignalP is a hidden Markov model (HMM) version able to discriminate uncleaved signal anchors from cleaved signal peptides.

One of the limitations of SignalP is overprediction, as it is unable to discriminate between several very similar signal sequences, regularly predicting membrane proteins and lipopteins as type II signals. Many other kinds of signal sequence exist. A number of methods have been developed to predict lipoproteins, for example. The prediction of proteins that are translocated via the twin arginine translocation (TAT) dependent pathway is also important but is not addressed yet in any depth.

We have developed VaxiJen (http://www.jenner.ac.uk/VaxiJen/), which implements a statistical model able to discriminate between candidate vaccines and nonantigens, using an alignment-free representation of the protein sequence [21]. Rather than concentrate on epitope and nonepitope regions, the method used bacterial, viral, and tumor protein data sets to derive statistical models for predicting whole-protein antigenicity. The models showed prediction accuracy up to 89%, indicating a far higher degree of accuracy than has been obtained previously, for example, for B-cell epitope prediction. Such a method is an imperfect beginning; future research will yield significantly more insight as the number of known protective antigens increases.

11.3 REVERSE VACCINOLOGY

Reverse vaccinology is a principal means of identifying subunit vaccines and involves a considerable computational contribution. Conventional experimental approaches cultivate pathogens under laboratory conditions, dissecting them into their components, with proteins displaying protective immunity identified as antigens. However, it is not always possible to cultivate a particular pathogen in the lab nor are all proteins expressed during infection are easily expressed *in vitro*, meaning that candidate vaccines can be missed. Reverse vaccinology, by contrast, analyzes a pathogen genome to identify potential antigens and is typically more effective for prokaryotic than eukaryotic organisms.

Initially, an algorithm capable of identifying open reading frames (ORFs) scans the pathogenic genome. Programs that can do this include ORF Finder [22], Glimmer [23], and GS-Finder [24]. Once all ORFs have been identified,

proteins with the characteristics of secreted or surface molecules must be identified. Unlike the relatively straightforward task of identifying ORFs, selecting proteins liable to immune system surveillance is challenging. Programs such as ProDom [25], Pfam [26], and PROSITE [27] can identify sequence motifs characteristic of certain protein families and can thus help predict if a protein belongs to an extracellular family of proteins.

The NERVE program has been developed to further automate and refine the process of reverse vaccinology, in particular the process of identifying surface proteins [28]. In NERVE, the processing of potential ORFs is a six-step process. It begins with the prediction of subcellular localization, followed by the calculation of probability of the protein being adhesion, the identification of transmembrane (TM) domains, a comparison with the human proteome and then with that of the selected pathogen, after which the protein is assigned a putative function. The vaccine candidates are then filtered and ranked based upon these calculations. While it is generally accepted that determining ORFs is a relatively straightforward process, the algorithm used to define extracellular proteins from other proteins needs to be carefully selected. One of the most effective programs that can be used for this purpose is HensBC, a recursive algorithm for predicting the subcellular location of proteins [29]. The program constructs a hierarchical ensemble of classifiers by applying a series of if–then rules. HensBC is able to assign proteins to one of four different types (cytoplasmic, mitochondrial, nuclear, or extracellular) with approximately 80% accuracy for gram-negative bacterial proteins. The algorithm is nonspecialized and can be applied to any genome. Any protein identified as being extracellular could be a potential vaccine candidate.

The technique of reverse vaccinology was pioneered by a group investigating *Neisseria meningitidis*, the pathogen responsible for sepsis and meningococcal meningitis. Vaccines based upon the capsular proteins have been developed for all of the serotypes with the exception of subgroup B. The *N. meningitidis* genome was scanned for potential ORFs [30,31]. Out of the 570 proteins that were identified, 350 could be successfully expressed *in vitro*, and 85 of these were determined to be surface exposed. Seven identified proteins conferred immunity over a broad range of strains within the natural *N. meningitidis* population, demonstrating the viability of *in silico* analysis as an aid to finding candidates for the clinical development of a MenB vaccine. Other examples of the successful application of reverse vaccinology is *Streptococcus pneumoniae*, a major cause of sepsis, pneumonia, meningitis, and otitis media in young children [32,33]. Mining of the genome identified 130 potential ORFs with significant homology to other bacterial surface proteins and VFs. One hundred eight of 130 ORFs were successfully expressed and purified; six proteins were found to induce protective antibodies against pneumococcal challenge in a mouse sepsis model. All six of these candidates showed a high degree of cross reactivity against the majority of capsular antigens expressed *in vivo* and which are believed to be immunogenic in humans.

Another example is *Porphyromonas gingivalis* is a gram-negative anaerobic bacterium present in subgingival plaques present in chronic adult periodontitis, an inflammatory disease of the gums. Shotgun sequences of the genome identified approximately 370 ORFS [34]. Seventy-four of these had significant global homology to known surface proteins or an association with virulence. Forty-six had significant similarity with other bacterial outer membrane proteins. Forty-nine proteins were identified as surface proteins using PSORT and 22 through motif analysis. This generated 120 unique protein sequences, 40 of which were shown to be positive for at least one of the sera. These were used to vaccinate mice, with only two of the antigens demonstrating significant protection. *Chlamydia pneumoniae* is an obligate intracellular bacterium associated with respiratory infections and cardiovascular and atherosclerotic diseases. One hundred forty-one ORFS were selected through *in silico* analysis [35], and 53 putative surface-exposed proteins identified. If reverse vaccinology is applied appropriately in vaccine design, it can save enormous amounts of money, time, and wasted labor.

11.4 EPITOPE PREDICTION

Complex microbial pathogens, such as *Mycobacterium tuberculosis*, can interact within the immune system in a multitude of ways [36]. For a vaccine to be effective, it must invoke a strong response from both T cells and B cells; therefore, epitope mapping is a central issue in their design. *In silico* prediction methods can accelerate epitope discovery greatly. B-cell and T-cell epitope mapping has led to the predictive scanning of pathogen genomes for potential epitopes [37]. There are over 4000 proteins in the tuberculosis genome; this means that experimental analysis of host–pathogen interactions would be prohibitive in terms of time, labor, and expense.

T-cell epitopes are antigenic peptide fragments derived from a pathogen that, when bound to a major histocompatibility complex (MHC) molecule, interact with T-cell receptors after transport to the surface of an antigen-presenting cell. If sufficient quantities of the epitope are presented, the T cell may trigger an adaptive immune response specific for the pathogen. MHC class I and class II molecules form complexes with different types of peptide. The class I molecule binds a peptide of 8–15 amino acids in length within a single closed groove. The peptide is secured largely through interactions with anchoring residues at the N and C termini of the peptide, while the central region is more flexible [38]. Class II peptides vary in length from 12–25 amino acids and are bound by the protrusion of peptide side chains into cavities within the groove and through a series of hydrogen bonds formed between the main-chain peptide atoms and the side-chain atoms of the MHC molecule [39]. Unlike the class I molecule, where the binding site is closed at either end, the peptide can extend out of both open ends of the binding groove.

B cells generate antibodies when stimulated by helper T cells as part of the adaptive immune response. The antibodies act to bind and neutralize pathogenic material from a virus or bacterium. Individual antibodies are composed of two sets of heavy and light chains. Each B cell produces a unique antibody due to the effects of somatic hypermutation and gene segment rearrangement. Those cells, within the primary repertoire whose antibodies convey antigen recognition, are selected for clonal expansion, an iterative process of directed hypermutation and antigen-mediated selection. This facilitates the rapid maturation of antigen-specific antibodies with a high affinity for a specific epitope. A B cell appropriate to deal with a specific infection is selected and cloned to deal with the primary infection, and a population of the B cell is then maintained in the body to combat secondary infection. It is the capacity to produce a huge variety of different antibodies that allows the immune system to deal with a broad range of infections.

Experimentally determined IC_{50} and BL_{50} affinity data have been used to develop a variety of MHC-binding prediction algorithms, which can distinguish binders from nonbinders based on the peptide sequence. These include motif-based systems, support vector machines (SVMs) [40–42], HMMs [43], quantitative structure-activity relationship (QSAR) analysis [44], and structure-based approaches [45–47]. MHC-binding motifs are a straightforward and easily comprehended method of epitope detection, yet produce many false-positive and many false-negative results. SVMs are machine learning algorithms based on statistical theory that seeks to separate data into two distinct classes (in this case binders and nonbinders). HMMs are statistical models where the system being modeled is assumed to be a Markov process with unknown parameters. In an HMM, the internal state is not visible directly, but variables influenced by the state are. HMMs aim to determine the hidden parameters from observable ones. An HMM profile can be used to determine those sequences with "binderlike" qualities. QSAR analysis techniques have been used to refine the peptide interactions with the MHC class I groove by incrementally improving and optimizing the individual residue-to-residue interactions within the binding groove. This has led to the design of so-called superbinders that minimize the entropic disruption in the groove and are therefore able to stabilize even disfavored residues within so-called anchor positions. Finally, molecular dynamics has been used to quantify the energetic interactions between the MHC molecule and peptide for both classes I and II by analysis of the three-dimensional structure of the MHC–peptide complex.

Several programs are available that can help design and optimize vaccines. In this section, some of the most effective algorithms for each form of vaccine design are discussed. For T-cell epitope prediction, many programs are available. A sensible approach for a new user would be to use MHCBench [48], an interface developed specifically for evaluating the various MHC-binding peptide prediction algorithms. MHCBench allows users to compare the performance of various programs with both threshold-dependent and

-independent parameters. The server can also be extended to include new methods for different MHC alleles.

B-cell prediction is more problematic due to the difficulties in correctly defining both linear and discontinuous epitopes from the rest of the protein. The epitope of a B cell is defined by the discrete surface region of an antigenic protein bound by the variable domain of an antibody. The production of specific antibodies for an infection can boost host immunity in the case of both intracellular and extracellular pathogens. The antibody's binding region is composed of three hypervariable loops that can vary in both length and sequence so that the antibodies generated by an individual cell present a unique interface [49]. All antibodies contain two antigen-binding sites, composed of complementary determining (CDR) loops. The three CDR loops of the heavy and light chains form the "paratope," the protein surface that binds to the antigen. The molecular surface that makes specific contact with the residues of the paratope is termed an "epitope." A B-cell epitope can be an entire molecule or a region of a larger structure. The study of the paratope–epitope interaction is a crucial part of immunochemistry, a branch of chemistry that involves the study of the reactions and components on the immune system.

Despite the extreme variability of the region, the antibody-binding site is more hydrophobic than most protein surfaces with a significant predilection for tyrosine residues. B-cell epitopes can be divided into continuous (linear) and discontinuous (conformational), the latter being regions of the antigen separated within the sequence but brought together in the folded protein to form a three-dimensional interface. Another problem with B-cell epitopes relates to the fact that they are commonly divided into two groups: continuous epitopes and discontinuous epitopes. Continuous epitopes correspond to short peptide fragments of a few amino acid residues that can be shown to cross react with antibodies raised against the intact protein. Since the residues involved in antibody binding represent a continuous segment of the primary sequence of the protein, they are also referred to as "linear" or "sequential" epitopes. Studies have shown that this class of epitope often contains residues that are not implicated in antibody interaction, while some residues play a more important role than others in antibody binding. Discontinuous epitopes are composed of amino acid residues that are not sequential in the primary sequence of a protein antigen but are brought into spatial proximity by the three-dimensional folding of the peptide chain [50].

There is considerable interest in developing reliable methods for predicting B-cell epitopes. However, to date, the amino acid distribution of the complementary antigen surface has been difficult to characterize, presenting no unique sequential or structural features upon which to base a predictive system. It is partly for this reason that the B-cell epitope has lagged far behind T-cell prediction in terms of accuracy but also because much of the data upon which predictions are based remain open to question due to the poorly under-

stood recognition properties of cross-reactive antibodies. One of the central problems with B-cell epitope prediction is that the epitopes themselves are entirely context dependent. The surface of a protein is, by definition, a continuous landscape of potential epitopes that is without borders. Therefore, both epitope and paratope are fuzzy recognition sites, forming not a single arrangement of specific amino acids but a series of alternative conformations. In this instance, a binary classification of binder and nonbinder may simply not reflect the nature of the interaction. A factor also to be considered is that the average paratope consists of only a third of the residues within the CDR loops, suggesting the remaining two-thirds could potentially bind to an antigen with an entirely different protein surface.

Often, a short length of amino acids can be classified as a continuous epitope, though in fact it may be a component of a larger discontinuous epitope; this can be a result of the peptide representing a sufficient proportion of the discontinuous epitope to enable cross reaction with the antibody. Since the majority of antibodies raised against complete proteins do not cross react with peptide fragments derived from the same protein, it is thought that the majority of epitopes are discontinuous. It is estimated that approximately 10% of epitopes on a globular protein antigen are truly continuous in nature. In spite of this, the majority of research into B-cell epitope prediction has focused largely on linear peptides on the grounds that they are discrete sequences and are easier to analyze. This can only be resolved by examination of the three-dimensional structure of the protein where the distinction between the continuous and discontinuous forms is not relevant.

Initial research into B-cell epitope prediction looked for common patterns of binding or "motifs" that characterize epitopes from nonepitopes. Unfortunately, the wide variety of different epitope surfaces that can be bound made it impossible to determine any such motifs. More sophisticated machine learning approaches such as artificial neural networks have also been applied but never with an accuracy exceeding 60%. More recently, structural analysis of known antigens has been used to determine the surface accessibility of residues as a measure of the probability that they are part of an epitope site. Despite these fundamental limitations, several B-cell epitope prediction programs are available, including Discotope [51], 3DEX [52], and CEP [53]. Both conformational epitope prediction (CEP) and Discotope measure the surface accessibility of residues, although neither has been developed to the point where they can identify coherent epitope regions rather than individual residues. A recent review of B-cell epitope software [54] calculated the A_{ROC} curves for the evaluated methods were about 0.6 (indicating 60% accuracy) for DiscoTope, ConSurf (which identifies functional regions in proteins), and PPI-PRED (protein–protein interface analysis) methods, while protein–protein docking methods were in the region of 65% accuracy, never exceeding 70%. The remaining prediction methods assessed were all close to random. In spite of this, the increasing number of available antigen–antibody structures combined with sophisticated techniques for structural analysis suggests a more

methodical approach to the study interface will yield a better understanding of what surfaces can and cannot form stable epitopes. The proposed research will take several different approaches to this problem, which will lead to a more comprehensive understanding of antibody–antigen interactions.

11.5 DESIGNING DELIVERY VECTORS

Safe and effective methods of gene delivery have been sought for 30 years. Viral delivery of genes has effectively targeted inter alia hemophilia, coronary heart disease, muscular dystrophy, arthritis, and cancer. Despite their immanent capacity to transfer genes into cells, concerns over safety, manufacturing, restricted targeting ability, and plasmid size have limited deployment of effective and generic gene therapy approaches. This remains a key objective for vaccinology. Vectors for gene therapy and vaccines differ in their requirements, yet both must overcome issues of targeting, plasmid cargo, and adverse immunogenicity. For example, up to 10% of the vaccinia genome can be replaced by DNA coding for antigens from other pathogens. The resulting vector generates strong antibody and T-cell responses and is protective. Viruses commonly used as vectors include poxviruses, adenovirus, varicella, polio, and influenza. Bacterial vectors include both *Mycobacterium bovis* and salmonella. Adding extra DNA coding for large molecule adjuvants greatly can exacerbate antibody or T-cell responses.

Successful transfection is hampered by DNA degradation within and outside the cell, by inadequate cell penetration, by poor intracellular trafficking, and by inefficient nuclear localization. The material can enter cells in many ways. Following clathrin-dependent endocytosis or endocytosis via lipid raft and/or membrane microdomains, the material transfers from early endosomes to sorting endosomes, where it may be exocytosed or may transfer to late endosomes. From late endosomes, material transfers to lysosomes for acidic and enzymatic digestion. Gene delivery requires both vector escape from digestion in late endosomes and nuclear translocation. Caveolin-dependent endocytosis, phagocytosis, and macropinocytosis do not transfer the material to the endolysosomal pathway. Some internalized material is released into the cytosol through unknown mechanisms. However, creating vectors with such desirable properties is difficult, and their effectiveness may be compromised by their capacity to downregulate other immune responses. The efficient and rational design of effective vaccine vectors is an area where informatic techniques could play a large role.

Similar to, yet simpler than, viral vectors are so-called DNA vaccines; they are plasmids capable of expressing antigenic peptide within the host [55]. They are an attractive alternative to conventional vaccines, generating both a cellular and a humoral immune response, which are effective versus intracellular pathogens. The efficiency of a DNA vaccine has been successfully enhanced using codon optimization [56], CpG motif engineering [57,58], and the intro-

duction of promoter sequences [59,60]. Codon optimization has been the most effective in enhancing protein expression efficiency. Codons optimal for translation are those recognized by abundant tRNAs [61]. Within a phylogenetic group, codon frequency is highly correlated with gene expression levels. Immunogenicity depends upon effective translation and transcription of the antigen; it is possible to enhance this by selecting optimal codons for the vaccine.

The most comprehensive approach to vaccine optimization is taken by DyNAVacs, an integrative bioinformatics tool that optimizes codons for heterologous expression of genes in bacteria, yeasts, and plants [62]. The program is also capable of mapping restriction enzyme sites, primer design, and designing therapeutic genes. The program calculates the optimal code for each amino acid encoded by a stretch of DNA by using a codon usage table, which contains codon frequencies for a variety of different genomes.

A similar technique, CpG optimization, may be used to optimize the codons in respect to CG dinucleotides. Pattern recognition receptors that form part of the innate immune system can often distinguish prokaryotic DNA from eukaryotic DNAs by detecting unmethylated CpG dinucleotides, in particular base contexts, which are termed "CpG motifs." The presence of such motifs in the sequence can be highly advantageous so long as it does not interfere with the process of codon optimization.

11.6 ADJUVANT DISCOVERY

Another technique for optimizing the efficacy of vaccines is to develop an efficient adjuvant. An adjuvant is defined as any chemical that is able to enhance an immune response when applied simultaneously with a vaccine and thus improves the efficacy of vaccination [63,64]. It is possible that some adjuvants act as immune potentiators, triggering an early innate immune response that enhances the vaccine effectiveness by increasing the vaccine uptake. Adjuvants may also enhance vaccination by improving the depot effect, the colocalization of the antigen and immune potentiators, by delaying the spread of the antigen from the site of infection so that absorption occurs over a prolonged period [65]. Aluminum hydroxide or alum is the only adjuvant currently licensed in humans. Aluminum-based adjuvants prolong antigen persistence due to the depot effect, stimulating the production of IgG1 and IgE antibodies [66] and triggering the secretion of interleukin-4. There are also several small-molecule, druglike adjuvants, such as imiquimod, resiquimod, and other imidazoquinolines [67–69]. Other small molecules that have been investigated for adjuvant properties include monophosphoryl lipid A, muramyl dipeptide, QS21, polylactide co-glycolide (PLG) and Seppic ISA-51[70]. In many cases, the adjuvant molecules have displayed toxic properties or have shown poor adsorption, making them unsuitable for use. Thus, there is a great demand for new compounds that can be used as adjuvants.

Chemokine receptors are a family of G protein-coupled receptors (GPCRs) that transduce chemokines, leukocyte chemoattractant peptides that are secreted by several cell types in response to inflammatory stimuli [71–73]. GPCRs are a superfamily of transmembrane proteins responsible for the transduction of a variety of endogenous extracellular signals into an intracellular response [74–76]. Activation of the chemokine receptors triggers an inflammatory response by inducing migration of the leukocytes from circulation to the site of injury or infection. The receptors play a pivotal role in angiogenesis, hematopoiesis, and brain and heart development, and there is also evidence that CCR5 precipitates the entry of HIV-1 into CD4+ T cells by the binding of the viral envelope protein gp120 [77,78]. There are 18 chemokine receptors and over 45 known chemokine ligands. The chemokines can be divided into the CC and CXC family; the former contains two cysteine residues adjacent within the protein sequence, while in the latter, they are separated by a single amino acid. CCR4 is a chemokine receptor expressed on Th2-type CD4+ T cells and has been linked to allergic inflammation diseases such as asthma, atopic dermatitis, and allergic rhinitis. There are two chemokines that bind the CCR4 receptor exclusively: CCL22 and CCL17 [79]. Inhibition of the two ligands has been shown to reduce the migration of T cells to sites of inflammation, suggesting than any CCR4 antagonist could provide an effective treatment for allergic reactions, specifically in the treatment of asthma. Anti-CCL17 and anti-CCL22 antibodies have both been observed to have efficacy, the property that enables a molecule to impart a pharmacological response, in murine asthma models.

It is possible for the CCR4 receptor to act as an adjuvant due to its expression by regulatory T cells (Tregs) that normally downregulate an immune response [80]. The Tregs inhibit dendritic cell maturation and thus downregulate expression of the costimulatory molecule. A successful CCR4 antagonist would therefore be able to enhance human T-cell proliferation in an *in vitro* immune response model by blocking the Treg proliferation. This suggests that an effective CCR4 antagonist would have the properties of an adjuvant. A combination of virtual screening and experimental validation has been used to identify several potential adjuvants capable of inhibiting the proliferation of Tregs. Small-molecule adjuvant discovery is amenable to techniques used routinely by the pharmaceutical industry. Three-dimensional virtual screening is a fast and effective way of identifying molecules by docking a succession of ligands into a defined binding site [81]. A large database of small molecules can be screened quickly and efficiently in this way. Using "targeted" libraries containing a specific subset of molecules is often more effective. It is possible to use "privileged fragments" to construct combinatorial libraries, those which are expected to have an increased probability of success. A pharmacophore is a specific three-dimensional map of biological properties common to all active confirmations of a set of ligands exhibiting a particular activity that can be used to discover new molecules with similar properties. Several small mol-

ecules have been investigated for adjuvant properties in this way [82]. More recently, molecules that selectively interfere with chemokine-mediated T-cell migration have shown the potential to act as adjuvants by downregulating the expression of costimulatory molecules, limiting T-cell activation. Small-molecule chemokine receptor antagonists have been identified and have shown to be effective at blocking chemokine function *in vivo* [83,84], although to date, no compound has reached a phase II clinical trial.

11.7 DISCUSSION

Vaccine design and development is an inherently laborious process, but the programs and techniques outlined here have the potential to simplify the process greatly. The techniques described also have the potential to identify candidate proteins that would be overlooked by conventional experimentation. Reverse vaccinology has, in particular, proved effective in the discovery of antigenic subunit vaccines that would otherwise remain undiscovered.

It is sometimes difficult for outsiders to assess properly the relative merits of *in silico* vaccine design compared to mainstream experimental studies. The potential, albeit largely unrealised, is huge, but only if people are willing to take up the technology and use it appropriately. People's expectations of computational work are often largely unrealistic and highly tendentious. Some expect perfection and are soon disappointed, rapidly becoming vehement critics. Others are highly critical from the start and are nearly impossible to reconcile with informatic methods. Neither appraisal is correct, however. Informatic methods do not replace, or even seek to replace, experimental work, only to help rationalize experiments, saving time and effort. They are slaves to the data used to generate them. They require a degree of intellectual effort equivalent in scale yet different in kind to that of so-called experimental science. The two disciplines, experimental and informatics, are thus complementary albeit distinct.

Like the discovery process of small molecules, vaccines also suffer from the process of attrition. Few notional vaccines ever get tested in the laboratory. Few candidate vaccines successful in laboratory tests on small animals ever get tested in man. Few vaccines entering phase I trials ever reach phase III. Not all phase III candidates reach the marketplace. The development of a new drug or vaccine is a risky business, but is ultimately a beneficial one; despite all the cant and hypocrisy that surrounds and permeates these endeavors, in the end, lives are saved or lives are improved. The pharmaceutical industry is doubtless brazen and profit hungry. This, I am afraid, is a necessary and probably unavoidable evil. Without the industry, the public health tools that drugs and vaccines represent would not exist and all those lives would not be saved. Ignoring the nuances and counter arguments, when push comes to shove, as the saying goes, the bottom line is as simple as that.

REFERENCES

1. Girard MP, Osmanov SK, Kieny MP. A review of vaccine research and development: The human immunodeficiency virus (HIV). *Vaccine* 2006;24:4062–4081.
2. Vekemans J, Ballou WR. *Plasmodium falciparum* malaria vaccines in development. *Expert Rev. Vaccines* 2008;7:223–240.
3. de Lisle GW, Wards BJ, Buddle BM, Collins DM. The efficacy of live tuberculosis vaccines after presensitization with *Mycobacterium avium*. *Tuberculosis (Edinb)* 2005;85:73–79.
4. Kumet R, Gelenberg AJ. The effectiveness of generic agents in psychopharmacologic treatment. *Essent Psychopharmacol* 2005;6:104–111.
5. Austin PC, Mamdani MM, Juurlink DN. How many "me-too" drugs are enough? The case of physician preferences for specific statins. *Ann Pharmacother* 2006;40: 1047–1051.
6. Fung HB, Guo Y. Enfuvirtide: A fusion inhibitor for the treatment of HIV infection. *Clin Ther* 2004;26:352–378.
7. Yi H, Zhang J, Zhao Y. The effects of antibody treatment on regulatory CD4(+) CD25(+) T cells. *Transpl Immunol* 2008;19:37–44.
8. Palma-Carlos AG, Santos AS, Branco-Ferreira M, Pregal AL, Palma-Carlos ML, Bruno ME, Falagiani P, Riva G. Clinical efficacy and safety of preseasonal sublingual immunotherapy with grass pollen carbamylated allergoid in rhinitic patients. A double-blind, placebo-controlled study. *Allergol Immunopathol (Madr)* 2006;34: 194–198.
9. Hung CF, Ma B, Monie A, Tsen SW, Wu TC. Therapeutic human papillomavirus vaccines: Current clinical trials and future directions. *Expert Opin Biol Ther* 2008; 8:421–439.
10. Ebo DG, Bridts CH, Stevens WJ. IgE-mediated large local reaction from recombinant hepatitis B vaccine. *Allergy* 2008;63:483–484.
11. O'Hagan DT, MacKichan ML, Singh M. Recent developments in adjuvants for vaccines against infectious diseases. *Biomol Eng* 2001;18:69–85.
12. De Groot AS. Immunomics: Discovering new targets for vaccines and therapeutics. *Drug Discov Today* 2006;11:203–209.
13. Guzmán E, Romeu A, Garcia-Vallve S. Completely sequenced genomes of pathogenic bacteria: A review. *Enferm Infecc Microbiol Clin* 2008;26:88–98.
14. Yang J, Chen L, Sun L, Yu J, Jin Q. VFDB 2008 release: An enhanced web-based resource for comparative pathogenomics. *Nucleic Acids Res* 2008;36(Database issue):D539–D542.
15. Mok BW, Ribacke U, Winter G, Yip BH, Tan CS, Fernandez V, Chen Q, Nilsson P, Wahlgren M. Comparative transcriptomal analysis of isogenic *Plasmodium falciparum* clones of distinct antigenic and adhesive phenotypes. *Mol Biochem Parasitol* 2007;151:184–192.
16. Merino EF, Fernandez-Becerra C, Durham AM, Ferreira JE, Tumilasci VF, d'Arc-Neves J, da Silva-Nunes M, Ferreira MU, Wickramarachchi T, Udagama-Randeniya P, Handunnetti SM, Del Portillo HA. Multi-character population study of the vir subtelomeric multigene superfamily of *Plasmodium vivax*, a major human malaria parasite. *Mol Biochem Parasitol* 2006;149:10–16.

17. Tongchusak S, Brusic V, Chaiyaroj SC. Promiscuous T cell epitope prediction of *Candida albicans* secretory aspartyl protienase family of proteins. *Infect Genet Evol* 2008;8(4):467–473.
18. Winnenburg R, Urban M, Beacham A, Baldwin TK, Holland S, Lindeberg M, Hansen H, Rawlings C, Hammond-Kosack KE, Köhler J. PHI-base update: Additions to the pathogen host interaction database. *Nucleic Acids Res* 2008; 36(Database issue):D572–D576.
19. Rey S, Acab M, Gardy JL, Laird MR, deFays K, Lambert C, Brinkman FS. PSORTdb: A protein subcellular localization database for bacteria. *Nucleic Acids Res* 2005;33:(Database issue):D164–D168.
20. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2007;2:953–971.
21. Doytchinova IA, Flower DR. VaxiJen: A server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* 2007;8:4.
22. Rombel IT, Sykes KF, Rayner S, Johnston SA. ORF Finder: A vector for high-throughput gene identification. *Gene* 2003:282:33–41.
23. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 1999;27:4636–4641.
24. Ou HY, Guo FB, Zhang CT. GS-Finder: A program to find bacterial gene start sites with a self-training method. *Int J Biochem Cell Biol* 2004;36:535–534.
25. Servant F, Bru C, Carrère S, Courcelle E, Gouzy J, Peyruc D, Kahn D. ProDom: Automated clustering of homologous domains. *Brief Bioinform* 2002;3:246–251.
26. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam protein families database. *Nucleic Acids Res* 2000:28:263–266.
27. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A. The PROSITE database. *Nucleic Acids Res* 2002;30:235–238.
28. Vivona S, Bernante F, Filippini F. NERVE: New Enhanced Reverse Vaccinology Environment. *BMC Biotechnol* 2006;6:35.
29. Bulashevska A, Eils R. Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. *BMC Bioinformatics* 2006;7:298.
30. Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW, Peden JF, Dodson RJ, Nelson WC, Gwinn ML, DeBoy R, Peterson JD, Hickey EK, Haft DH, Salzberg SL, White O, Fleischmann RD, Dougherty BA, Mason T, Ciecko A, Parksey DS, Blair E, Cittone H, Clark EB, Cotton MD, Utterback TR, Khouri H, Qin H, Vamathevan J, Gill J, Scarlato V, Masignani V, Pizza M, Grandi G, Sun L, Smith HO, Fraser CM, Moxon ER, Rappuoli R, Venter JC. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* 2000;287:1809–1815.
31. Pizza M, Scarlato V, Masignani V, Giuliani MM, Aricò B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, Galeotti CL, Luzzi E, Manetti R, Marchetti E, Mora M, Nuti S, Ratti G, Santini L, Savino S, Scarselli M, Storni E, Zuo P, Broeker M, Hundt E, Knapp B, Blair E, Mason T, Tettelin H, Hood DW, Jeffries AC, Saunders NJ, Granoff DM, Venter JC, Moxon ER, Grandi G, Rappuoli R. Whole genome sequencing to identify vaccine candidates against serogroup *B. meningococcus*. *Science* 2000;287:1816–1820.

32. Wizemann TM, Heinrichs JH, Adamou JE, Erwin AL, Kunsch C, Choi GH, Barash SC, Rosen CA, Masure HR, Tuomanen E, Gayle A, Brewah YA, Walsh W, Barren P, Lathigra R, Hanson M, Langermann S, Johnson S, Koenig S. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. *Infect Immun* 2001;69:1593–1598.

33. Maione D, Margarit I, Rinaudo CD, Masignani V, Mora M, Scarselli M, Tettelin H, Brettoni C, Iacobini ET, Rosini R, D'Agostino N, Miorin L, Buccato S, Mariani M, Galli G, Nogarotto R, Nardi Dei V, Vegni F, Fraser C, Mancuso G, Teti G, Madoff LC, Paoletti LC, Rappuoli R, Kasper DL, Telford JL, Grandi G. Identification of a universal Group B streptococcus vaccine by multiple genome screen. *Science* 2005;309:148–150.

34. Ross BC, Czajkowski L, Hocking D, Margetts M, Webb E, Rothel L, Patterson M, Agius C, Camuglia S, Reynolds E, Littlejohn T, Gaeta B, Ng A, Kuczek ES, Mattick JS, Gearing D, Barr IG. Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis*. *Vaccine* 2001;19:4135–4412.

35. Montigiani S, Falugi F, Scarselli M, Finco O, Petracca R, Galli G, Mariani M, Manetti R, Agnusdei M, Cevenini R, Donati M, Nogarotto R, Norais N, Garaguso I, Nuti S, Saletti G, Rosa D, Ratti G, Grandi G. Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*. *Infect Immun* 2002;70:368–379.

36. McMurry J, Sbai H, Gennaro ML, Carter EJ, Martin W, De Groot AS. Analyzing *Mycobacterium tuberculosis* proteomes for candidate vaccine epitopes. *Tuberculosis (Edinb)* 2005;85:195–205.

37. Pizza M, Scarlato V, Masignani V, Giuliani MM, Aricò B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, Galeotti CL, Luzzi E, Manetti R, Marchetti E, Mora M, Nuti S, Ratti G, Santini L, Savino S, Scarselli M, Storni E, Zuo P, Broeker M, Hundt E, Knapp B, Blair E, Mason T, Tettelin H, Hood DW, Jeffries AC, Saunders NJ, Granoff DM, Venter JC, Moxon ER, Grandi G, Rappuoli R. Identification of vaccine candidates against serogroup *B. meningococcus* by whole-genome sequencing. *Science* 2000;287:1816–1820.

38. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S. SYFPEITHI: Database for MHC ligands and peptide motifs. *Immunogenetics* 1999;50:213–219.

39. Jardetzky TS, Brown JH, Gorga JC, Stern LJ, Urban RG, Strominger JL, Wiley DC. Crystallographic analysis of endogenous peptides associated with HLADR1 suggests a common, polyproline II-like conformation for bound peptides. *Proc Natl Acad Sci USA* 1996;93:734–738.

40. Donnes P, Elofsson A. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics* 2002;3:25.

41. Liu W, Meng X, Xu Q, Flower DR, Li T. Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics* 2006;7:182.

42. Wan J, Liu W, Xu Q, Ren Y, Flower DR, Li T. SVRMHC prediction server for MHC-binding peptides. *BMC Bioinformatics* 2006;7:463.

43. Noguchi H, Kato R, Hanai T, Matsubara Y, Honda H, Brusic V, Kobayashi T. Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. *J Biosci Bioeng* 2002;94:264–270.

44. Doytchinova IA, Walshe V, Borrow P, Flower DR. Towards the chemometric dissection of peptide-HLA-A*0201 binding affinity: Comparison of local and global QSAR models. *J Comput Aided Mol Des* 2005;19:203–212.

45. Davies MN, Hattotuwagama CK, Moss DS, Drew MG, Flower DR. Statistical deconvolution of enthalpic energetic contributions to MHC-peptide binding affinity. *BMC Struct Biol* 2006;6:5–17.

46. Davies, M.N., Sansom, C.E., Beazley, C., Moss, D.S. (2003). A novel predictive technique for the MHC class II peptide-binding interaction. *Mol. Med.*, 9, 220–225.

47. Wan S, Coveney P, Flower DR. Large-scale molecular dynamics simulations of HLA-A*0201 complexed with a tumor-specific antigenic peptide: Can the alpha3 and beta2m domains be neglected? *J Comput Chem* 2004;25:1803–1813.

48. Salomon J, Flower DR. Predicting class II MHC-peptide binding: A kernel based approach using similarity scores. *BMC Bioinformatics* 2006;7:501.

49. Blythe MJ, Flower DR. Benchmarking B Cell epitope prediction: Underperformance of existing methods. *Protein Sci* 2004;14:246–248.

50. Greenbaum JA, Andersen PH, Blythe M, Bui HH, Cachau RE, Crowe J, Davies M, Kolaskar AS, Lund O, Morrison S, Mumey B, Ofran Y, Pellequer JL, Pinilla C, Ponomarenko JV, Raghava GP, van Regenmortel MH, Roggen EL, Sette A, Schlessinger A, Sollner J, Zand M, Peters B. Towards a consensus on datasets and evaluation metrics for developing B Cell epitope prediction tools. *J Mol Recognit* 2007;20:75–82.

51. Andersen PH, Nielsen M, Lund O. Prediction of residues in discontinuous B Cell epitopes using protein 3D structures. *Protein Sci* 2006;15:2558–2567.

52. Schreiber A, Humbert M, Benz A, Dietrich U. 3D-Epitope-Explorer (3DEX): Localization of conformational epitopes within three-dimensional structures of proteins. *J Comput Chem* 2005;26:879–887.

53. Kulkarni-Kale U, Bhosle S, Kolaskar AS. CEP: A conformational epitope prediction server. *Nucleic Acids Res* 2005;33:W168–W171.

54. Ponomarenko JV, Bourne PE. Antibody-protein interactions: Benchmark datasets and prediction tools evaluation. *BMC Struct Biol* 2007;7:64.

55. Babiuk LA, Babiuk SL, Loehr BI, van Drunnen Littel-van den Hurk S. Nucleic acid vaccines: Research tool or commercial reality. *Vet Immunol Immunopathol* 2000;76:1–23.

56. Babiuk LA, Pontarollo R, Babiuk S, Loehr B, van Drunen Littel-van den Hurk, S. (2003). Induction of immune responses by DNA vaccines in large animals. *Vaccine* 2003;21:649–658.

57. Uchijima M, Yoshida A, Nagata T, Koide Y. Optimization of codon usage of plasmid DNA vaccine is required for the effective MHC class I-restricted T Cell responses against an intracellular bacterium. *J Immunol* 1998;161:5594–5599.

58. Klinman DM, Yamshchikov G, Ishigatsubo Y. Contribution of CpG motifs to the immunogenicity of DNA vaccines. *J Immunol* 1997;158:3635–3639.

59. Booth JS, Nichani AK, Benjamin P, Dar A, Krieg AM, Babiuk LA, Mutwiri GK. Innate immune responses induced by classes of CpG oligodeoxynucleotides in ovine lymph node and blood mononuclear cells. *Vet Immunol Immunopathol* 2006;115:24–34.

60. Lee AH, Suh YS, Sung JH, Yang SH, Sung YC. Comparison of various expression plasmids for the induction of immune response by DNA immunization. *Mol Cells* 1997;7:495–501.
61. Xu ZL, Mizuguchi H, Ishii-Watabe A, Uchida E, Mayumi T, Hayakawa T. Optimization of transcriptional regulatory elements for constructing plasmid vectors. *Gene* 2001;272:149–156.
62. Henry I, Sharp PM. Predicting gene expression level from codon usage bias. *Mol Biol Evol* 2007;24:10–12.
63. Harish N, Gupta R, Agarwal P, Scaria V, Pillai B. DyNAVacS: An integrative tool for optimized DNA vaccine design. *Nucleic Acids Res* 2006;34(Web Server issue):W264–W266.
64. Singh M, O'Hagan DT. Recent advances in vaccine adjuvants. *Pharm Res* 2002;19:715–728.
65. Stills HF Jr. Adjuvants and antibody production: Dispelling the myths associated with Freund's complete and other adjuvants. *ILAR J* 2005;46:280–293.
66. Gupta RK. Aluminum compounds as vaccine adjuvants. *Adv Drug Deliv Rev* 1998;32:155–172.
67. Singh M, Srivastava I. Advances in vaccine adjuvants for infectious diseases. *Curr HIV Res* 2003;1:309–320.
68. Schijns VEJC. Mechanisms of vaccine adjuvant activity: Initiation and regulation of immune responses by vaccine adjuvants. *Vaccine* 2003;21:829–831.
69. Iellem A, Colantonio L, Bhakta S, Sozzani S, Mantovani A, Sinigaglia F, D'Ambrosio D. Unique chemotactic response profile and specific expression of chemokine receptors CCR4 and CCR8 by CD4+CD25+ regulatory T cells. *J Exp Med* 2001;194:847–854.
70. Schijns VE. Mechanisms of vaccine adjuvant activity: initiation and regulation of immune responses by vaccine adjuvants. *Vaccine* 2003;21:829–831.
71. Charoenvit Y, Goel N, Whelan M, Rosenthal KS, Zimmerman DH. 2004CEL-1000—A peptide with adjuvant activity for Th1 immune responses. *Vaccine* 2004;22:2368–2373.
72. Hedrick JA, Zlotnik A. Chemokines and lymphocyte biology. *Curr Opin Immunol* 1996;8:343–347.
73. Luster AD. Chemokines—Chemotactic cytokines that mediate inflammation. *N Engl J Med* 1998;338:436–445.
74. Locati M, Murphy PM. Chemokines and chemokine receptors: Biology and clinical relevance in inflammation and AIDS. *Annu Rev Med* 1999;50:425–440.
75. Christopoulos A, Kenakin T. G protein-coupled receptor allosterism and complexing. *Pharmacol Rev* 2002;54:323–374.
76. Gether U, Asmar F, Meinild AK, Rasmussen SG. Structural basis for activation of G-protein-coupled receptors. *Pharmacol Toxicol* 2002;91:304–312.
77. Bissantz C. Conformational changes of G protein-coupled receptors during their activation by agonist binding. *J Recept Signal Transduct Res* 2003;23:123–153.
78. Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, Di Marzio P, Marmon S, Sutton RE, Hill CM, Davis CB, Peiper SC, Schall TJ, Littman DR, Landau NR. Identification of a major co-receptor for primary isolates of HIV-1. *Nature* 1996;381:661–666.

79. Feng Y, Broder CC, Kennedy PE, Berger EA. HIV-1 entry cofactor: Functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science* 1996;272:872–877.

80. Chvatchko Y, Hoogewerf AJ, Meyer A, Alouani S, Juillard P, Buser R, Conquet F, Proudfoot AE, Wells TN, Power CA. A key role for CC chemokine receptor 4 in lipopolysaccharide-induced endotoxic shock. *J Exp Med* 2000;191:1755–1764.

81. Lieberam I, Forster I. The murine beta-chemokine TARC is expressed by subsets of dendritic cells and attracts primed $CD4^+$ T cells. *Eur J Immunol* 1999;29: 2684–2694.

82. Schellhammer I, Rarey M. FlexX-Scan: Fast, structure-based virtual screening. *Proteins* 2004;57:504–517.

83. Charoenvit Y, Brice GT, Bacon D, Majam V, Williams J, Abot E, Ganeshan H, Sedegah M, Doolan DL, Carucci DJ, Zimmerman DH. A small peptide (CEL-1000) derived from the beta-chain of the human major histocompatibility complex class II molecule induces complete protection against malaria in an antigen-independent manner. *Antimicrob Agents Chemother* 2004;48:2455–2463.

84. Godessart N. Chemokine receptors: Attractive targets for drug discovery. *Ann N Y Acad Sci* 2005;1051:647–657.

PART IV

DATA MINING METHODS IN CLINICAL DEVELOPMENT

12

DATA MINING IN PHARMACOVIGILANCE

Manfred Hauben and Andrew Bate

Table of Contents

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

12.1 INTRODUCTION

The stages in the drug development continuum collectively comprise a prolonged time span marked by the accumulation of increasing amounts of complex scientific information generated in the quest to understand drug efficacy and safety. Thus, discovery of drug information continues long after drug discovery and regulatory approval. Some of these data are useful and some are redundant. The challenge is to distill out the useful information from the useless information at each stage of development so as to facilitate the movement of helpful drugs through the development continuum so that the right drugs get to the right patient. To date, data mining has a played a role in each stage. Data mining has been used to support high-throughput screening [1], lead optimization [2], predictive toxicology [3], pharmacokinetic calculations [4], predicting treatment options [5], and adverse event detection both pre- [6] and post marketing [7]. Clearly, the discovery of knowledge of a drug extends well beyond the discovery of the drug and is a long-term commitment.

The effective application of medical therapy requires a judicious assessment of the patient under treatment, the treatment indication, the therapeutic benefits of the administered drug(s), and their side-effect profiles. Such integrated risk–benefit assessments necessarily take place at both the level of the individual patient and also as part of a public health remit on a population level. Therapeutic effects are the focus of many early controlled studies in clinical development, and the therapeutic profile is quite well defined at the time of marketing authorization. The other side of the benefit–risk assessment is somewhat more tricky in that only a fraction of the side effects have been completely defined at the time of marketing. Therefore, systems must be in place for continuous monitoring for new side effects of drugs, even for approved indications after marketing authorization.

With an increasing number of molecular-level therapeutic targets being identified and with demographic changes associated with increased comorbid illnesses and polypharmacy, it is not surprising that it is becoming more challenging for some organizations to implement the aforementioned continuous surveillance. Herein we describe how statistics and technology can be leveraged to support the process of drug safety surveillance.

12.2 THE NEED FOR POST-MARKETING DRUG SAFETY SURVEILLANCE

The evolution of modern drug safety surveillance thinking has often been driven by various public health tragedies. Understanding this connection is not only of historical interest but yields insights into some of the traditional approaches to surveillance that might have to be modified to meet modern-day challenges.

The first prominent episode was the 1937 elixir sulfanilamide incident in which Massengill & Co. produced a liquid preparation of sulfanilamide in which the active moiety was intentionally dissolved in diethylene glycol in demand for a liquid formulation of the drug, and while it passed tests of appearance and fragrance, it was unfortunately never tested for toxicity. At the time, there was no legal or regulatory mandates for safety/toxicity testing of new drugs and so, when the formulation was manufactured and distributed, not surprisingly (with hindsight), the highly toxic diethylene glycol resulted in the deaths of more than 100 people [8]. Tragically, international safety incidents due to the use of diethylene glycol continue periodically to modern times [9–11].

The next major safety incident catalyzed the institution of post-marketing surveillance (PMS) requirements that are still used today. In 1961, the *Lancet* published a letter by McBride [12], an Australian physician, noting that congenital anomalies are present in 1.5% of births overall but almost 20% of pregnancies in women given thalidomide as a sedative or as an antiemetic. The thalidomide-treated mothers delivered babies with multiple severe congenital abnormalities involving mesenchymal-derived musculoskeletal structures. This is the paradigm of the "astute clinician model" [13] in which the observational acumen of the clinician results in the detection of an event(s) that is clinically and/or quantitatively distinctive.

In the wake of the thalidomide disaster, it was clear that there were inadequate systems for the ongoing surveillance of medicinal products after drug launches, and it was agreed that such a disaster should never be allowed to happen again. As a consequence, surveillance systems were set up in several countries. The first systematic collection of reports occurred in Australia, Italy, the Netherlands, New Zealand, Sweden, the United Kingdom, the United States, and West Germany, and in 1968, 10 countries from Australasia, Europe, and North America agreed to pool all their data in a World Health Organization (WHO)-sponsored project with the intention of identifying rare but serious reactions as early as possible. This project became the WHO Program for International Drug Monitoring, and this pooling of spontaneously reported data in a central database continued. The number of member countries and the rate of the increase of reports has continued, and currently, there are 70 countries that contribute data and nearly 4 million case reports in the WHO database of suspected adverse drug reactions (ADRs) [14]. In parallel to these organizations, pharmaceutical companies have also developed in-house databases of case reports involving their own drugs. Most pharmaceutical companies' databases are a fraction of the size of the above databases, but some larger organizations with large product portfolios have very large databases that are of the same order of magnitude of size. Thus, most monitoring of approved medicinal products reflects a parallel and interactive collaboration between government, industry, transnational, drug monitoring centers, with other stakeholders such as patient organizations playing an increasingly active role.

12.3 THE RELATIONSHIP BETWEEN DATA QUANTITY AND QUALITY

The fundamental lessons of history and contemporary pharmacovigilance (PhV) reinforce the important reality that the unraveling of the safety profile of a drug is a continuous process that begins with the early drug development phases (see Chapter 10) and lasts as long as the medicine is dispensed to patients. As the knowledge of a drug accumulates, so does the quantity of information on the product; however, the increasing data are of variable quality and completeness. The first human studies ("first-in-human") are a critical juncture in which there is great concern about the safety of the patients because of the absence of any human experience with medicine. As such, the numbers of patients are quite small, and the patients are typically healthy volunteers without significant medical or comedication history and are monitored very closely in in-patient study units. If the drug passes this and subsequent tests, the increased understanding of the safety profile leads to a greater comfort level with administering the drug to human beings and therefore progresses to studies that employ an increasing number of subjects who are not always healthy volunteers, may be taking comedications, and suffer comorbid illnesses. So, with increasing understanding comes increased numbers and trial subjects increasingly similar to the potential patients anticipated to be the main beneficiaries of treatment. Nonetheless, even the largest randomized studies are very structured and impose significant constraints on the number, size, and complexity of patients in order to be logistically feasible and to allow for the application of inferential statistics. Finally, when a critical evidentiary mass is reached, the drug may be approved. The natural progression from low to high in terms of patient numbers and complexity and from greater to less in the intensity of individual patient monitoring takes a quantum leap after marketing authorization. After that milestone, the number of patients treated and their medical complexity may explode, with patient monitoring becoming unavoidably more relaxed and variable. Consequently, it is well understood in the specialist community that inevitably some knowledge about the safety profile of a product will only be established after a product has received marketing authorization. This is not the case with the general public and perhaps explains the disproportionate media impact that concerns of possible rare side effects can have, often with very limited or erroneous evidential basis such as the inappropriate measles, mumps, rubella (MMR) autism scare [15].

Given the above relationships and evolving information streams, postmarketing drug surveillance is a keystone surveillance activity that aims

1. to protect patients from inappropriate drug use,
2. to reassure patients that their health is protected,
3. to protect a product from inappropriate and unfounded safety concerns,

4. to provide an avenue for patients to express their concerns about medicinal treatments,
5. to help develop new products without the harmful profiles of current best therapeutic interventions, and
6. to help develop new indications of products based on unpredicted side effects when in routine use.

As an indication of the ongoing importance of this activity in the post-approval phase of drug development, Table 12.1 presents data on 12 post-approval drug withdrawals in the United States, which occurred between 1997 and 2001 as listed in the January–February 2002 Food and Drug Administration (FDA) consumer magazine [16].

Drug safety surveillance objectives (1)–(4) listed above are probably the most widely but it is worth mentioning that point number 5 is not entirely theoretical. New indications have emerged from observation of adverse effects. The first suggestion that the central α-adrenergic antihypertensive agent clonidine, originally synthesized as a nasal decongestant, might have useful blood pressure-lowering properties, appeared when a member of the original nasal decongestant trial group allowed his secretary to self-administer clonidine intranasally for a cold. She subsequently developed low blood pressure, bradycardia, and slept for 24 hours (the self-administered dose amounted to an overdose with 20 tablets). This observation was replicated and reinforced during the initial trials [17]. The hair-restorative properties of topical minoxidil were pursued based on the observed side effect of hypertrichosis with oral

TABLE 12.1 Post-Approval Drug Withdrawals in the United States, 1997–2001

Drug Name	Use	Adverse Risk	Year Approved	Year Withdrawn
Cerivastatin	LDL reduction	Rhabdomyolysis	1997	2001
Rapaccuronium Br	Anesthesia	Bronchospasm	1999	2001
Alosetron	Irritable bowel	Ischemic colitis	2000	2000
Cisapride	Heartburn	Arrhythmia	1993	1993
Phenylpropanolamine	Decongestant	Stroke	Pre-1962	2000
Troglitazone	Type 2 diabetes	Liver toxicity	1997	2000
Astemizole	Antihistamine	Arrhythmia	1988	1999
Grepafloxcin	Antibiotic	Arrhythmia	1997	1999
Mibefradil	High BP and angina	Arrhythmia	1997	1998
Bromfenac	Analgesia	Liver toxicity	1997	1998
Terfenadine	Antihistamine	Arrhythmia	1985	1998
Fenfluramine	Appetite suppressant	Valve disease	1973	1997
Dexfenfluramine	Appetite suppressant	Valve disease	1996	1997

LDL = low density lipoprotein; BP = blood pressure.

minoxidil [18], and the development of phosphodiesterase (PDE)-5 inhibitors for erectile dysfunction sprang from the observation of penile erections as a common side effect in multiple-dose phase I trials [19]. Identifying unanticipated therapeutic effects of drugs and new indications by systematically screening safety databases is still in the embryonic stage [20]. Drug safety data on products could and should be increasingly used to help in future drug development.

Drug development is often considered finished when the efficacy of a product has been demonstrated beyond reasonable doubt, with limited focus on safety. For example, consider the following definition of drug discovery from the School of Pharmacy, University of California [21]: "research process following drug discovery that takes a molecule with desired biological effects in animal models and prepares it as a drug that can be used in humans." Even when a broader definition of drug development is considered and includes randomized clinical trials in humans, this is still only considered to the point of approval, with some increased focus on safety, as well as effectiveness rather than efficacy.

Traditional segmentation of the drug life cycle into discrete phases may foster parochial views that limit the full potential of drug discovery. In fact, these segments are not discrete but overlap, and there is much feedback and communication between them. It is well established that much important useful information about the side effect of a medicinal product is only established after drug launch. In addition to side-effect information as a source of previously unexpected new indications as described above, a better understanding of the side-effect profiles of products on the market can be used to determine and investigate possible side effects (or even likely spurious associations) of new medical therapies, issues of central concern in PhV, and risk management planning. It may also help in the prioritization of candidates based on their likely approval post marketing, and may possibly reduce the number of good candidates dropped because of unwarranted concern about apparent markers for side effects if these findings were seen for similar products but did not in the end lead to the anticipated side-effect profile. Similarly, there is a huge amount of potentially relevant data collected while a company's earlier product or competitor products are marketed, which could be relevant during early drug development.

12.4 SIGNAL DETECTION—THE FRONT LINE OF PhV

12.4.1 PhV

PhV has been defined as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem" [22]. It has often been used synonymously with PMS or with drug safety monitoring. The historic equation of "PhV" with "PMS" relates to the fact that clinical trials in support of drug applications, with their

necessary constraints on size, duration, and patient heterogeneity cannot reliably capture the full range of ADRs. Therefore, ADRs that are rare or occur only after prolonged latency are often unknown at the time of initial approval. However, just as the drug discovery process is continuous with no rigid boundaries despite the classic segmentation used to depict drug development, PhV is becoming more holistic and integrative and is commencing earlier in the drug development process.

PhV entails activities founded on a complex knowledge base involving clinical, informatics, and statistical domains. A quote attributed to Edward Shortliffe describing medical decision making would probably strike a chord with those working in the complex and often uncertain world of PhV as an apt description: "making acceptable decisions in an imperfectly understood problem space often using incomplete or erroneous information."

12.4.2 Signal Detection in PhV

The "front line" of PhV consists of signal detection—the expeditious identification of early clues of potential ADRs that may be novel by virtue of their nature, severity, and/or frequency.

There is an extensive suite of activities, strategies, techniques, and data streams linked with this surveillance activity, the "front-end" goal of which is to expeditiously detect potential "signals" of possible novel safety phenomena. When a credible signal of a new adverse event is detected, it triggers an evaluation, which usually begins with a detailed review of individual case reports of the association, which are submitted to spontaneous reporting system (SRS) databases as described below. The initial investigation of a signal may determine that a causal relationship is sufficiently likely to warrant some action (e.g., labeling amendment), that the relationship is most likely noncausal, or that it is unclear, but continued monitoring and/or further studies are indicated.

While there is semantic ambiguity, imprecision, and variability in the use of the term "signal," one commonly used definition is that of the WHO: "reported information on a possible causal relationship between an adverse event and drug, the relationship being unknown or incompletely documented previously. Usually more than a single report is required to generate a signal, depending on seriousness of the event and quality of information" [23]. Another comprehensive definition [24] that emphasizes the need for rational thought prior to concluding a signal exists is "a set of data constituting a hypothesis that is relevant to the rational and safe use of a medicine. Such data are usually clinical, pharmacological, pathological, or epidemiological in nature. A signal consists of data and arguments." Credible signals that are returned to analysts as a result of the PhV activities are then subjected to various analytical investigations, which hopefully provide convergent lines of evidence that illuminate the possible relationship, either strengthening/weakening it, and ultimately confirming/refuting it under ideal circumstances,

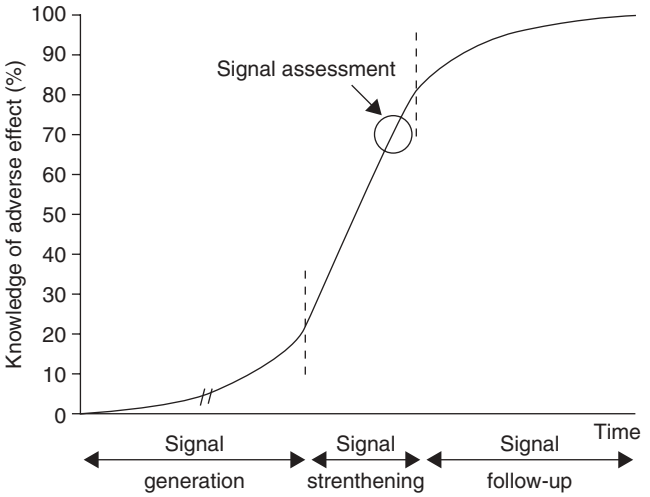


Figure 12.1 The lengthy process of the discovery of a drug-induced disorder, from the earliest suspicion via a credible signal to a fully explained and understood phenomenon.

though PhV often involves decision making in the setting of residual uncertainty. The process from initial detection of a possible signal to confirmation/refutation with sufficient certitude for decision making is rarely based on a single technique or tool and is not typically a rapid discrete process or precisely delineated steps. It often involves iterative assessment of clinical and numerical data streams over time. Meyboom et al. [25] has depicted the process with a logistic-type curve (see Fig. 12.1 [with kind permission from Adis International]).

Collectively, these activities form a major component of "phase IV research" or "post-marketing research," which has been defined as a "generic term used to describe all activities after drug approval by the regulatory agency …" [26]. While SRS data have been a cornerstone of PhV for decades, data sets from claims databases, epidemiological databases, and clinical trial databases all play key roles in PhV primarily in the steps after signal detection as subsequent assessments that strengthen or weaken a signal. The role of these databases in initial signal detection is increasing and is likely to increase further [27]. Nonetheless, SRS data will continue to play a central role [27].

12.5 TARGETS, TOOLS, AND DATA SETS

To fully appreciate the landscape or "application domain" [28] of post-approval safety surveillance and the role that data mining can play in it, a review of its elements is in order. This domain consists of the adverse events under surveillance, the mechanism by which adverse drug effects (ADEs,

defined as any injury resulting from drug therapy) are reported, the available data sets that permanently record the observed and reported occurrences of these events, and the methods and tools used to interrogate these data sets.

12.6 THE SAMPLE SPACE OF ADVERSE EVENTS

PhV is rather unique among surveillance systems in terms of the range and variety of the disease under surveillance. The variety is in terms of pathophysiological mechanisms, clinical phenotypes, and quantitative representation, by which we mean their frequency/risk of occurrence in exposed versus unexposed populations.

The diversity is perhaps most striking with respect to clinical phenotypes. With increasing numbers of drugs targeting increasing numbers of identified molecular targets within complex signal transduction cascades, adverse drug reactions (ADRs) may rival syphilis and miliary tuberculosis, which are often called "great imitators" in medicine due to their extremely protean clinical phenotypes. The full range of specific clinical presentations is beyond the scope of this article, but in addition to the more widely appreciated ADRs such as allergic reactions, hepatitis, rashes, and gastrointestinal disturbances, medicines may also induce kidney stones, biliary stones, many forms of vasculitis, pneumothorax, tendon rupture, myopia, pyloric stenosis, hiccups, hypothermia, noncardiogenic pulmonary edema, and cardiomyopathy, to name just a few. In some instances, the ADR may go exactly counter to what one would expect from the pharmacological properties or intended purpose of the drug, for example, anaphylactic reactions to corticosteroids or hypertensive reactions from drugs given to lower blood pressure. The latter have been referred to as paradoxical reactions [29]. This underscores the importance of the prepared mind expecting the unexpected [30]. The clinical and mechanistic variety of ADEs has inspired the development of various ADR classification schemes and conceptual frameworks [31].

The quantitative dimension of ADR variety refers to the fact that some ADEs are relatively common and some are rare. For example, headache, rash, abdominal pain, and diarrhea are rather common events in the general population, in populations under treatment with drugs, and in SRS databases. Other events are rare in general populations, drug-treated populations, and SRS databases such as pure-red cell aplasia, aplastic anemia, agranulocytosis, and Creutzfeld–Jacob disease, to name a few. The relative frequency of events in treated populations versus the general/untreated populations and in SRS databases influences the optimum method and likelihood of detection. An event that is extremely rare in all populations is so striking that it is likely amenable to detection through clinical observation ("acute clinician paradigm). An illustration is iatrogenic Creutzfeld–Jakob disease, a very rare and fatal spongiform encephalopathy now known to be caused by a prion. Three cases in association with the human growth hormone treatment in the 1980s were considered so

striking and unlikely to be due to a chance association that the National Hormone and Pituitary Program, the main source of human growth hormone (hGH) in the United States, was terminated [32]. Common events that are induced by drugs only rarely are difficult to detect by any means because the risk attributable to the drug is so small relative to the baseline risk. Other combinations of frequency in treated and untreated populations are displayed in Table 12.2 [31]. While this table is constructed from the perspective of substantiating adverse events, the implications are similar for signal detection. For example, it may be quite difficult to detect a small increase in risk of an event that is very common in the background population.

It is possible to blend clinical and quantitative elements to arrive at additional classifications that may be then used to positively influence monitoring strategy. For example, so-called designated medical events (DMEs) are rare (quantitative), serious (clinical), and have a high drug-attributable risk (i.e., a significant proportion of occurrences of these events are drug induced) (quantitative). Although there is no single universally accepted DME list, Table 12.3 shows some ADRs generally considered DMEs.

Similarly, the WHO critical term listing is a list of ADR terms defined as "adverse reaction terms referring to, or possibly being indicative of, serious disease states, which have been regarded as particularly important to follow up. A serious disease is one that may be fatal, life-threatening, or causing prolonged inpatient hospitalization, or resulting in persistent or significant

TABLE 12.2 Quantitative Range of Events of Interest in PhV, as Listed in Reference 31

Incidence in Patients Taking the Drug	Background Incidence of the Event	Example	Ease of Proving the Association
Common	Rare	Phocomelia due to thalidomide	Easy clinical observation
Rare	Rare	ASA and Reye's syndrome	Less easy clinical observation
Common	Common	ACEI and cough	Difficult, large observational study
Uncommon	Moderately common	HRT and breast carcinoma	Very difficult, large trial
Rare	Common	None established	Impossible

TABLE 12.3 Some Examples of DMEs

Aplastic anemia	Steven–Johnson syndrome
Agranulocytosis	Torsade des pointes
Hepatic failure	Toxic epidermal necrolysis

disability or incapacity." The critical term listing has been an integral part of the WHO's signal-detection strategy since its introduction in the 1970s. Given these characteristics and the comparative consequences of false alarms versus missed signals, they are often considered sentinel events irrespective of drug, meaning that one to three cases may be considered an alert [33]. A similar concept is the targeted medical event (TME) similar to a DME, but it is based on clinical/pharmacological characteristics specific to a drug, its treatment indication(s), and/or the patient-specific characteristics.

12.7 REPORTING MECHANISM

While the rules, regulations, and procedures governing the spontaneous reporting of ADRs vary, there are basic commonalities:

1. With the exception of pharmaceutical companies that are legally bound to report ADRs to health authorities, it is a voluntary activity by the source reporter (e.g., healthcare practitioner, patient). This is the basis for the term "spontaneous reporting" or "spontaneous report."
2. The reporter does not have to be certain that the drug caused the event–any suspicion, however tentative, is sufficient for spontaneous reporting.
3. There must be an identifiable drug, patient, and event. However, source documentation to verify the reports does not have to be submitted by the reporter, though pharmaceutical companies and, to a lesser extent, health authorities typically request this information.
4. The total number of people exposed to the drug and the total number that experiences/did not experience an event are unknown. In other words, the complete numerator and denominator figures that are a prerequisite for quantifying risk are unavailable, and it is not appropriate to use SRS data to estimate absolute or relative risk for the occurrence of ADRs.

There is no clear probability structure underlying the overall sampling scheme since these reports are anecdotal and are voluntarily submitted. Thus, there are differential influences, including confounding factors (discussed in detail below) and various reporting artifacts, that may result in some drugs, events, and/or drug–event combinations being preferentially reported or not reported. Finally, the data elements within individual reports are subject to considerable qualitative and quantitative deficits in the form of missing or incorrect information. Some of these may be combined to result in a phenomenon such as duplicate reporting, which is particularly problematic with SRSs.

The many quantitative and qualitative defects associated with spontaneously reported data, including duplicate reporting, along with a sometimes

Box 12.1 Inman's Seven Deadly Sins of Reporting ADRs

- Fear of litigation
- Lethargy/indifference about contributing to the general advancement of knowledge
- Ambition to collect and publish a personal case series
- Guilt at having caused an ADR
- Complacency—the mistaken belief that only safe drugs are licensed
- Ignorance of the need for reporting
- Diffidence about reporting a mere suspicion

redundant drug nomenclature, result in the need for considerable preprocessing of the data prior to data mining analysis. Current preprocessing procedures are demonstrably imperfect. For example, while the scale of the problem is unclear, duplicate reports, which escape contemporary duplicate detection procedures, can complicate clinical and quantitative analysis of the data [34]. New approaches to preprocessing, such as the application of a hit-miss model adapted from record linkage [35], can be used to weight reports by similarity and to cluster those that are particularly similar. This better identification of likely duplicates saves resources for signal detection by reducing the number of false-positive leads.

Many factors can influence the entire process from observation of an adverse event, attribution/misattribution to a drug(s), to completing and submitting a report. (It is important to emphasize that individual reports may reflect attribution/misattribution since studies have documented high rates of misattribution to drugs [36].) Some of these factors are cultural/behavioral/attitudinal and result in substantial underreporting, which can range enormously. Inman has delineated the "seven deadly sins" of ADR reporting that exert an inhibitory influence on reporting behavior (Box 12.1).

12.8 THE ANATOMY OF SRS DATABASES

One of the most challenging aspects of PhV are the large repositories of spontaneous reports that are routinely employed to monitor the safety of marketed drug products by health authorities and large pharmaceutical companies. These are maintained by health authorities, transnational drug monitoring centers, and pharmaceutical companies. Understanding the anatomy of the individual records and overall architecture of such databases, especially their size and sparsity, is key to understanding the challenges faced in monitoring

TABLE 12.4 Representation of Two Records in an SRS Database

Age	Sex	Drug 1	Drug 2	Drug 3	Drug 15,000	AE1	AE2	AE3	AE 16,000
42	M	Yes	No	Yes	No	Yes	No	Yes	Yes
36	F	No	Yes	Yes	Yes	Yes	Yes	No	No

AE = adverse effect.

drug safety and important considerations for the development of technologies to assist human reviewers.

To understand the basic data representation of the individual reports that comprise these databases, Table 12.4 shows two fictitious but entirely typical entries corresponding to two reports in the SRS database.

Of note is the fact that every record may be considered to consist of very high dimensional information with each demographic variable, drug, and adverse event corresponding to a dimension. There are also additional variables related to medical history, and in addition, spontaneous reports sometimes have a narrative that may range from extremely scant to extremely detailed, possibly including source document information such as hospital discharge summaries and diagnostic laboratory records.

Two features of SRS data loom large in understanding the challenges presented by such data sets and the unique challenges encountered when trying to apply statistical methods to the analysis of such data.

Perhaps most apparent is the size of the larger SRS databases. Large health authorities, pharmaceutical companies with large product portfolios, and large transnational drug monitoring centers maintain huge database containing millions of records that are augmented with hundreds of thousands of reports per year. Smaller organizations with smaller databases may face similar problems when scaled to a lower number of reviewers. Perhaps slightly less well known is the number of drugs and adverse events that are encoded in the database. With 15,000 unique drug names and up to 16,000+ adverse event codes in the coding dictionaries and thesaurus used to memorialize the data, the number of potential combinations is huge, at 240 million potential combinations. A further complication is that the ADE dictionaries are hypergranular, meaning that many literally distinct event codes may be used for a given medical concept. Third, large SRS databases are very sparse, by which we mean most potential drug–event combinations are never reported, and of those that are, the majority may have only one or two reports. Finally, related to bullet point 4 under "reporting mechanism," each reported drug–ADE is only a subset of all occurrences of that drug–ADE, and there is no information in the data set on the number of times the drug was prescribed and ADE did not occur, nor on the rate of ADE in a nonexposed population.

12.9 METHODS IN DRUG SAFETY SURVEILLANCE

Basically, drug safety surveillance methods can be divided into two categories. One is essentially heuristic, using rules of thumb based on astute clinical observations and sound public health principles. The other are structured quantitative methods, the focus of this chapter. Given their clinical and quantitative variety, it is not surprising that some ADRs are detected via clinical observations, while other ADRS may be first recognized purely because of what seems to be a quantitatively higher-than-expected reporting frequency, i.e., after accumulation of a critical mass of cases, although later clinical review remains essential [33] and some by a combination of both strategies. Determining what constitutes a critical mass given the enormous limitations of the data and the data-generating mechanism, and the desire to maintain a rational balance of sensitivity and specificity, is the key conundrum of quantitative approaches to signal detection, which we will discuss in detail below.

12.10 TRADITIONAL APPROACHES TO DRUG SAFETY SURVEILLANCE

Since the institution of SRS databases, PV has relied heavily on the "astute clinician model" and on heuristics based on domain expertise and commonsense public health principles. Essentially, certain case reports or case series will appear "striking" to the data analyst and be considered for further investigation. A case or case series may appear striking to an observer for various reasons including the clinical nature of the event itself (e.g., the passage of a solid renal calculus composed of a drug), striking chronological features (e.g., the well-documented stereotypical recurrence of certain objective ADEs after multiple drug administrations, the first occurrence of a very unusual event, or ADEs with cogent arguments for biological plausibility. Other features that may flag a case/case series as striking or likely to be informative have been discussed in detail [24,33,37–40]). Although the striking case *approach* is commonly used, how common/uncommon "striking *cases*" actually are is unknown.

However, determining that a case/case series is striking should always be determined with a refined understanding of the relevant pathophysiology, and first instincts can be misleading. For example, it is not uncommon to see reports of noncytotoxic drug-induced hair loss within a week or two of commencing a drug. At first blush, such a rapid onset may seem compelling evidence of drug causation, but in fact, such a time frame is incompatible with the known physiology of the human hair follicle, which involves time cycles of follicle/hair growth, growth arrest, quiescence, and regrowth. This overlaps the field of causality assessment of ADEs, which is beyond the scope of this chapter but for which there is an abundant literature.

An important example of a commonly used heuristic is the maintenance of lists of DMEs that serve as sentinel ("worst first") events. As stated above,

because these are rare, serious, and have a high drug-attributable risk (not necessarily limited to specific drugs/drug classes), there is more of a premium placed on sensitivity versus specificity, and as few as one to three cases get extra attention. When similar considerations are linked to the pharmacology, treatment indications, or treatment populations of specific drugs, a TME may be used in the same way as a DME list. In a sense, these are related to what Amery has called the "striking case method" [40].Often, review of spontaneous reports will be triggered by concern emerging from other sources, such as an isolated case report in the literature or an unexpected occurrence of events in a clinical trial.

12.11 QUANTITATIVE APPROACHES

This conceptual foundation for quantitative approaches was formulated by David Finney in a seminal paper titled "The Design and Monitor of Drug Use" [41]. It was first routinely operationalized by Dr. Ed Napke as a "pigeonhole" cabinet in 1968 for the Canadian adverse event reporting system [42]. Each pigeonhole was a slot representing the intersection of a drug/event row and column. Reports involving that drug and event were filed in the respective slot. Colored tabs were attached to reports involving events deemed severe or unusual. Accumulations of colored tabs in certain pigeonholes provided a visual clue that the reporting frequency of the association might be quantitatively distinctive, which in turn might trigger further investigation. This is consistent with a fundamental process of safety surveillance and assessment—determining if the occurrence of an ADE exceeds what one would expect by chance. In contemporary drug safety surveillance, the extent to which the number of reports observed exceeds this expectation is expressed as a ratio measure of disproportionality, generically known as an observed-to-expected ratio ("O/E") or relative reporting ("RR").

Contemporary quantitative methods, also known as data mining algorithms (DMAs), construct and present to the user virtual pigeonhole cabinets but employ more structured statistical approaches instead of subjective visual cues for distinguishing which adverse events are quantitatively interesting.

This is illustrated in Table 12.5 as a cross tabulation of all possible drugs and events in which the number of reports of the first through the Mth AE is tabulated for the first through the Nth drugs. Each cell is the number of reports of the m_{th} AE reported for the n_{th} drug. This is, in effect, a huge, modern-day pigeonhole cabinet.

There is a way to condense Table 12.5 in accordance with the gray color-coding scheme, into a 2×2 contingency table that perhaps facilitates an understanding of association patterns between a drug and an event and the calculation of O/Es (Table 12.6).

The gray color-coding scheme that matches the expanded representation is intended to make quite obvious a fundamental characteristic of SRS data. The

TABLE 12.5 The SRS Database as a Virtual "Pigeonhole Cabinet"

	AE_1	AE_2	AE_3	AE_4	AE_M	Total Reports with Drug
Drug 1	N_{11}	N_{12}	N_{13}	N_{14}	N_{1M}	N_1
Drug 2	N_{21}	N_{22}	N_{23}	N_{24}	N_{2M}	N_2
Drug 3	N_{31}					N_3
Drug 4	N_{41}					N_4
Drug *n*	N_{n1}					N_N
Total reports of event	$N_{.1}$					$N_{..}$

TABLE 12.6 2×2 Contingency Table

	Event of Interest +	Event of Interest –
Drug of interest +		
Drug of interest –		

combination of interest (red) actually represents a small fraction of the data and even a small fraction of the adverse event data for that drug (red plus yellow). By far, the largest subset of the data (brown) represents other drugs and other events from the combination of interest. This has practical implications when we consider the measures of association that we calculate via such cross-classification tables. These will be discussed in detail below. Also, for future reference, note that the 2×2 table in sense loses or masks important information. For example, note that the information in the brown zone of the fully expanded contingency table involves numerous different drugs and events, yet this is all collapsed into a single cell—in other words, all "other" events and "other" drugs are each collapsed into a separate single category.

The importance of the 2×2 table is that it provides convenient bookkeeping device by which we can tabulate the number of reports of a given drug-event combination (DEC) of interest and create a rational and structured model of what that number would be if it purely reflected the play of chance if the drug and event were truly independent of each other in the database.

TABLE 12.7 Common Measures of Association for 2 × 2 Tables Used in Disproportionality Analysis

Measure of Association	Formula	Probabilistic Interpretation	Chance Expectation
Relative reporting (RR)[a]	$\frac{A(A+B+C+D)}{(A+C)(A+B)}$	$\frac{\Pr(ae \mid \text{drug})}{\Pr(ae)}$	1
Proportional reporting rate ratio (PRR)	$\frac{A(C+D)}{C(A+B)}$	$\frac{\Pr(ae \mid \text{drug})}{\Pr(ae \mid \tilde{}\text{drug})}$	1
Reporting odds ratio (ROR)	$\frac{AD}{CB}$	$\frac{\Pr(ae \mid \text{drug})\Pr(\tilde{}ae \mid \tilde{}\text{drug})}{\Pr(\tilde{}ae \mid \text{drug})\Pr(ae \mid \tilde{}\text{drug})}$	1
		$\frac{\text{Log}_2 \Pr(ae \mid \text{drug})}{\Pr(ae)}$	0

[a] *C* and RR formulated in a Bayesian framework in BCPNN and MGPS, respectively.

Expressed a little differently, we can use these tables to determine the probability that a randomly drawn report will list both the drug and the event if they are unrelated in the database. The greater the actual number of observed reports exceeds this expected number, the more interesting it potentially becomes (Table 12.7).

Confidence intervals should also be calculated around the measures or a chi-squared test performed.

It is crucial to appreciate that a number of reports, exceeding that expected by chance, can never prove causality, and, considered alone, do not qualify as a credible signal. We illustrate this with examples below. There are several causes of a statistically disproportionate reporting frequency (a so-called signal of disproportionate reporting [SDR]) [43]. First, there will be variations in reporting that are essentially stochastic in nature and are especially problematic with rarely reported ADEs. For example, one can imagine that a misclassified report can have a much bigger impact if it is the only report or one of two reports than if it is one of a hundred reports. So all other factors being equal (which of course they rarely are), one may have more confidence in an O/E of 10 if that represents 100 observed compared to 10 expected versus 1 observed compared to 0.01 expected. The important sources of systematic bias inherent to the spontaneously reported data (i.e., the aforementioned confounders, biases, and reporting artifacts) may be entirely or partially responsible for many SDRs. Contemporary data mining methods cannot currently effectively address these systematic biases, hence the need for clinical review of DMA outputs.

For much of the database, the background noise associated with variability of sparse data can present a challenge to discerning true signals. There are two basic approaches to controlling the variability. One is based on classical

or frequentist notions of statistical unexpectedness, and the other is based on Bayesian statistics.

12.12 CLASSICAL OR FREQUENTIST APPROACHES

In this case, classical statistical notions of unexpectedness are used to help improve the signal-to-noise ratio. The common feature of these approaches is that they rely only on information contained in the specific 2×2 table corresponding to the DEC of interest. For example, when calculating a proportional reporting ratio (PRR) for a given 2×2 table, the analyst may also specify additional threshold criteria of at least three reports and an associated χ^2 value of >3.85 (corresponding to a p value of ≤0.05). A limitation in such a binary approach (i.e., a separating threshold dividing ADRs into two classes: SDR+ versus SDR–) is that even with very small observed counts, if the expected count is small enough, the χ^2 value will be greater than 3.85, and the statistic will fail to screen out such associations, which may be false positives. A similar approach may be used with the p value of each statistic. Alternatively, the standard error may be used to determine a credibility interval/lower limit (5% threshold) of the 90% confidence interval of the statistic. This reduces the number of associations presented to the analyst and mitigates stochastic fluctuations.

Of course, there is no restriction against using higher thresholds of statistical unexpectedness or using a ranking versus a binary classification approach. By ranking classification, we mean there is no discrete threshold of interestingness, but rather, the associations are somehow ranked according to how quantitatively interesting they are relative to one another. One form of ranking implementation is a bivariate plot of the disproportionality metric (e.g., the PRR and the reporting odds ratio [ROR]) versus the measure of statistical unexpectedness. Analysts would then view the DECs in the upper right-hand corner as most quantitatively interesting, since they are both very disproportionate and are much less likely to represent stochastic fluctuations, with the least quantitatively interesting DECs in the lower left corner. Figures 12.2 and 12.3 provide examples from the European Medicines Agency (EMEA) and the Swedish Medical Products Agency (MPA), which currently use PRRs for routine signal detection. The limitations of spontaneous reports means that caution is needed to not place inappropriate focus on the ranking order, but instead to see it, as with thresholds, as a pragmatic approach to focus on clinical review on issues most likely to represent emerging drug safety issues.

12.12.1 Overview: The Bayesian Approach

The challenge of sparsity in spontaneous report data sets was one of the impetuses for the development of Bayesian methodologies since in other arenas, Bayesian approaches have demonstrated superiority to frequentist approaches

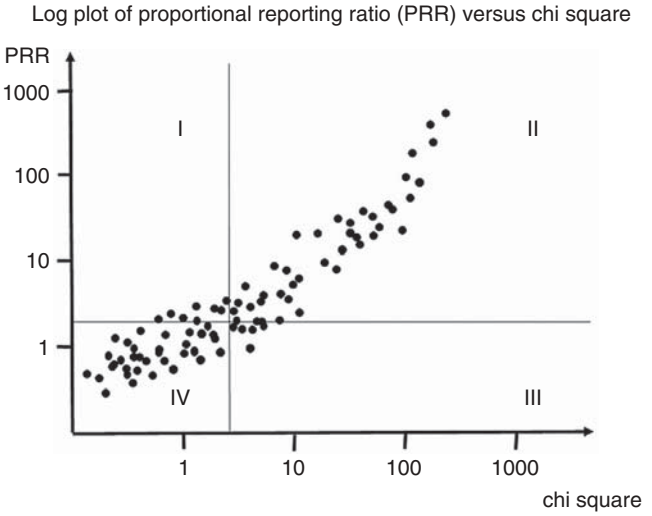


Figure 12.2 PRR versus chi-square (χ^2) bivariate plot (courtesy of Gunilla Sjolin-Forsberg, Medical Products Agency [MPA], Sweden).

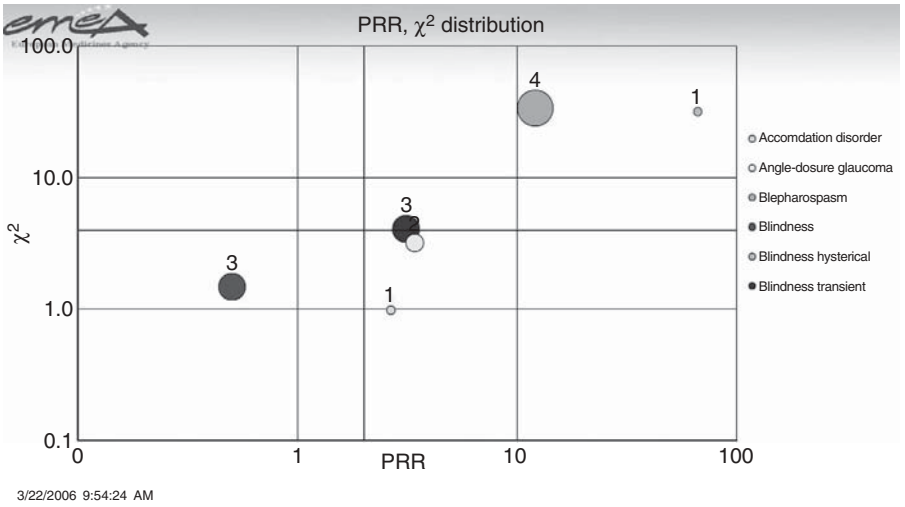


Figure 12.3 PRR versus χ^2 bivariate plot and case count (courtesy of Francois Maignen, European Medicines Agency [EMEA]).

when the available information is extremely limited. There are currently two major Bayesian techniques used for data mining in PhV, the Bayesian confidence propagation neural network (BCPNN) [44] and the multi-item Gamma–Poisson shrinker (MGPS) [45].

Bayesian methods, first adapted to drug safety signal detection by the WHO Unified Monte Carlo (UMC) [44], may be viewed as composite of two approaches to calculating an O/E ratio for each drug–event combination. One approach views each DEC as representing a realization of a unique process and that the huge numbers of spontaneously reported DECs have unrelated sources of variability. An alternative is to view all of the reported drug–event combinations as realizations of the same random process and just take an overall or grand mean of these O/E ratios based on marginal reporting frequencies/probabilities—basically a null 2×2 table. Neither view is absolutely correct, hence their combination in a Bayesian approach. This approach appeals to our prior knowledge and plausible belief that given the sparsity of the data, the numerous reporting artifacts, and confounders, most ADEs are not being reported unexpectedly frequently when stochastic fluctuations are taken into account and do not have implications for public safety.

The grand or null mean reflects our "prior belief" or first guess about the O/E for any ADE, and in effect "shrinks" or pulls high local O/Es supported by minimal data toward this prior belief. This is the basis for the term "Bayesian shrinkage." This grand mean O/E is also referred to as the "moderating prior," which in fact is not a single value but reflects a range of plausible values, each with an associated probability manifested as a probability distribution of possible O/Es. This amount of shrinkage is inversely related to the amount of data on the ADR of interest. In other words, for rarely reported ADRs, the null O/E is very influential on the weighted average, but as reports accumulate, this influence diminishes until a critical mass of cases is achieved and the effect of the moderating prior is then swamped by the local O/E [46].

With rare exceptions [47], shrinkage methods are presented as having only positive effects in their intended domains of application. However, caution is required. Bradley Effron, a renowned statistician and proponent of empirical Bayesian methodologies, puts it this way: "If you use an empirical Bayes estimate, everything gets pulled toward the central bulge. You have to grit your teeth and believe the fact that even though any one estimate may be off, overall you're getting a lot of improvement. That's what we have to get people, including ourselves to believe" [48]. In other words, Bayesian methods increase the signal-to-noise ratio but not perfectly, and may result in loss of credible signals with the noise, either absolutely or relatively in terms of timing. However, these comments refer to the use of the methods in isolation and as we discuss below, both Bayesian and frequentist strategies should be used in combination with other filters as part of an overall signal-detection process.

12.12.2 The Principal Bayesian Methods: BCPNN and MGPS

There are currently two major Bayesian methodologies based on 2×2 tables: the BCPNN and the multi-item Gamma–Poisson shrinker. Fundamentally, the difference between the two approaches is the manner in which the moderating

prior is derived. The BCPNN uses a Bayesian approach, while MGPS uses an empirical Bayesian approach.

The Bayesian BCPNN effectively constructs a null 2×2 table for each possible ADR that has a count in cell a = 0.5 and for which the cell counts in all the cells of the null 2×2 table conform to the prior belief that the drug and event are independent (O/E = 1, information component [IC] = $\log_2$O/E = 0). Thus, it amounts to an extra batch of data consisting of 0.5 reports for which the drug and event are independent. The constraint on cell count "a" of 0.5 is titrated to achieve a desired level of shrinkage in the WHO database, and other databases might justify other values.

The empirical Bayesian MGPS allows the existing data to determine the null 2×2 table and consequently the amount of shrinkage. This amounts to pooling, or borrowing information, from all possible 2×2 tables to determine the moderating prior and then forming a weighted composite of the null O/E and the "local" O/E of the individual 2×2 table. As the data is used to determine the null 2×2 table, rather than a prior belief, the null 2×2 table may have a mean O/E that is different from one, which in turn determines the extent of the shrinkage (Fig. 12.4).

As a specific illustration of what shrinkage actually "looks like" in a real example, Table 12.8 displays frequentist (PRR, χ^2) and empirical Bayesian (lower 90% confidence limit of the logarithmized EBGM—denoted EB05) disproportionality metrics for the association of amiodarone and basal cell carcinoma in the U.S. FDA database. Note that based on three reports in 1991, the frequentist PRR protocol returns very high disproportionalities, while the

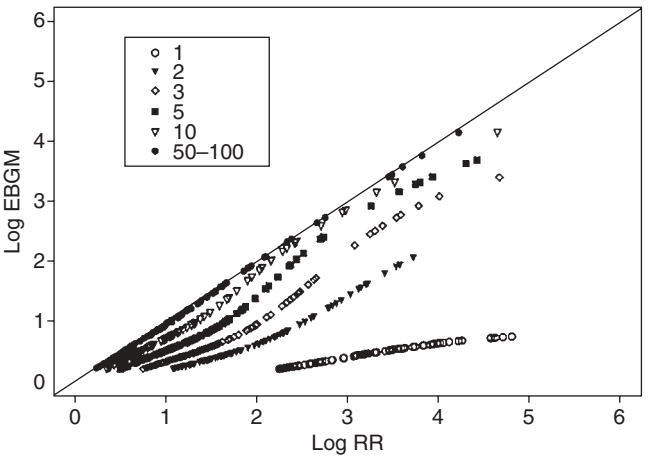


Figure 12.4 Scatterplot of frequentist (log RR) versus Bayesian Measure (log EBGM) measure of disporportionality as a function of the number of reports (courtesy of David Madigan, PhD, Department of Statistics, Columbia University). EBGM = empirical Bayes geometric mean.

TABLE 12.8 Temporal Evolution of PRR and EB05

	Skin Carcinoma			Basal Cell Carcinoma			Combined Carcinoma		
Year	N	PRR	EB05	N	PRR	EB05	N	PRR	EB05
1990	0	0	0	0	0	0	0	0	0.09
1991	3	17.61	1.91	0	0	0	3	20.88	1.93
1992	3	12.17	1.22	0	0	0	3	13.63	1.49
1993	4	10.64	1.58	0	0	0	4	11.77	2.06
1994	4	17.07	1.05	0	0	0	4	7.98	1.45
1995	5	5.26	1.06	0	0	0	5	6.05	1.42
1996	8	5.5	1.6	0	0	0.06	8	6.33	2.34
1997	8	3.89	1.23	0	0	0.05	8	4.47	1.56
1998	8	2.77	0.99	0	0	0.04	8	3.1	1.04
1999	8	2.17	0.84	0	0	0.18	8	2.36	0.83
2000	8	1.78	0.73	1	1.23	0.3	9	2.08	0.8
2001	8	1.52	0.65	2	1.53	0.3	10	1.87	0.78
2002	8	1.14	0.53	6	2.65	0.8	14	1.88	0.91
2003	9	1.15	0.55	6	1.72	0.62	15	1.64	0.84
2004	10	1.13	0.57	9	1.66	0.73	19	1.66	0.92

empirical Bayesian approach pulls or shrinks the results very nearly to one. Also note that as additional reports are submitted to the database, the PRR values decrease dramatically.

The data mining exercise, as it does not take account of the clinical details of each case, should not be considered a form of causality assessment. However, neither should the fact that an initially high O/E decreases over time with increasing numbers of reports be considered to disprove causality. Clinical review is required before deciding whether an SDR can be considered a signal. However, in the absence of clinical factors suggestive of causality, such patterns often represent stochastic fluctuations or "noise." Bayesian approaches can improve the signal-to-noise ratio, but in some instances, such methods, and indeed any that reduces the noise, will filter out/diminish true signals along with the noise. When examining this combination, there are features that argue for and against a possible association. As with any combination, the numbers above cannot be considered in a biological vacuum; the clinical features in the individual cases provide the required context.

Figures 12.5 and 12.6 illustrate additional data mining outputs in a graphical format to promote familiarity with the capabilities of these tools and to reinforce an intuitive understanding of these calculations. As a point of orientation, the mining is performed with "cumulative subsetting," in which the database is effectively rolled back in time so that we track its growth on the abscissa. So, 1984 represents the accumulated data to that point, and so on with subsequent years. The ordinate represents the value of the corresponding disproportionality metric. In some graphs, confidence or posterior intervals

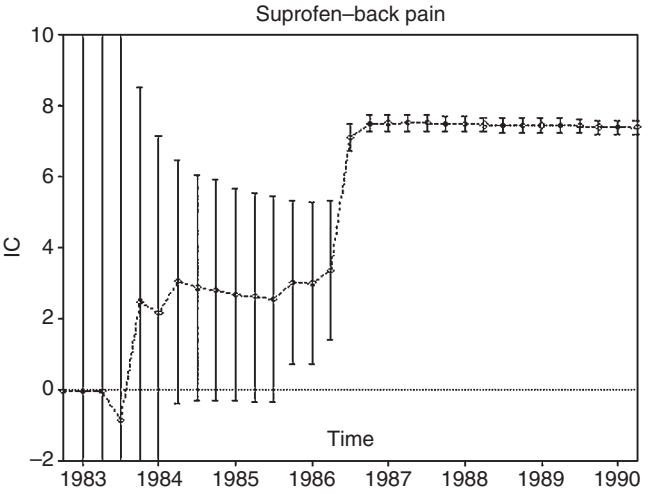


Figure 12.5 Cumulative change in time for the association of Suprofen–back pain based on spontaneous reports in the WHO database.

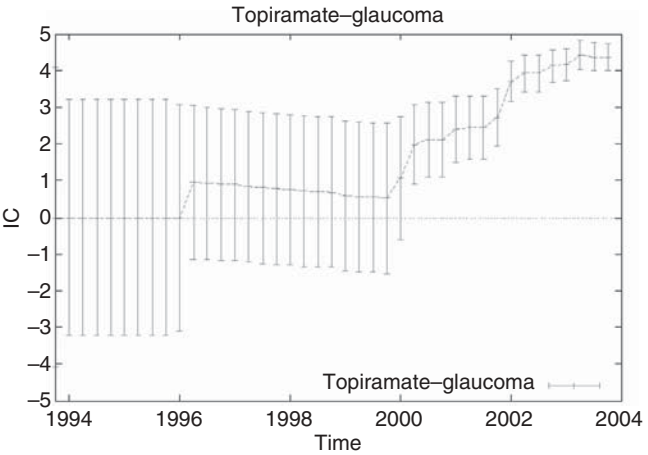


Figure 12.6 IC time scan of topiramate–glaucoma in WHO UMC database.

are provided. Note that in the graphs of IC, the threshold value is 0 since it is a logarithmic metric.

Initially, an IC of zero with wide posterior intervals is calculated as the prior assumption of independence determines the weighted composite in the absence of any reports with suprofen in the database, though there are reports of back pain reported with other drugs. In the middle of 1983, reports of suprofen and other adverse events were submitted. These additional reports

increase the expected count of suprofen–back pain without incrementing the observed count of zero, so that the O/E becomes less than one, making the logarithm of this metric (the IC) drop below zero with somewhat narrowed posterior intervals. In the last quarter of 1983, the first report of suprofen–back pain is submitted and the IC becomes positive (with wide credibility intervals) because the expected count for this DEC is so low (a total of 46 reports with suprofen). In the fourth quarter of 1985, the third suspected report of this DEC is submitted, pushing the lower 95% of the posterior interval over zero, which is a threshold criteria at WHO UMC. The IC continues to increase to over seven as more reports of the DEC are submitted. The posterior intervals become narrow because of large observed and expected ratio.

The above examples are retrospective data mining exercises for which it is difficult to conclusively affirm that data mining would have actually resulted in earlier detection and confirmation of these events. In contrast, the figure illustrates an example where data mining prospectively identified a drug–event combination that was subsequently adjudicated as being sufficiently probable for action.

In this instance, the association met quantitative threshold criteria in the second quarter of 2000, which was published in the WHO signal report published in April 2001. The first literature report of this association appeared in July, and the FDA issued a "Dear Healthcare Professional" letter in October.

While the two Bayesian methods in routine use in PhV are the BCPNN and MGPS, variants of these approaches have also been suggested; see, for example, the Bayesian-based false discovery rate approach suggested by Gould [49]. It should be noted that a Bayesian approach need not necessarily be complex, and some suggestions for possible simple alternatives are proposed [50,51].

12.13 EVALUATING AND VALIDATING DATA MINING PERFORMANCE IN PhV

Assessing the performance of these methods is extremely challenging. The discourse around it can get quite contentious and we are somewhat skeptical that any validation study will satisfy every interested party.

There are many published data mining exercises that yield some insight into data mining performance. Validation may be based on authentic SRS data or simulated SRS databases. Those using authentic data may be either retrospective or prospective in nature. The majority of published validation exercises involve retrospective evaluations using a screening paradigm. By this we mean that a reference set of true-positive and true-negative adverse events is compiled, and the data mining outputs are adjudicated against this reference set with performance metrics consisting of sensitivity, specificity, predictive values, and receiver operating characteristic (ROC) curves. A smaller number of published validation exercises involve the use of simulated data sets.

These have involved specific PhV scenarios, such as adverse events in black box warnings and associated with drug withdrawals. However, it is important at the outset to summarize the considerable challenges to validating and assessing the comparative and incremental utility of these tools. We briefly delineate these here and will subsequently discuss some in more detail below:

1. There is a large space of available choices of varying degrees of arbitrariness that are available to the analyst (see Table 12.9). This is a double-edged sword. It maximizes exploratory capacity but complicates the design and comparison of data mining procedures and makes data mining exercises susceptible to confirmation bias, in which an analysis is fitted or selected based on the fit of results to preexisting expectations.

   Some of these choices actually influence the data mining output, while others influence the significance and action taken for a given set of results, while some could affect both. Each set of choices is a defining configuration for a given data mining run, and each configuration may result in different data mining outputs and/or responses.

   Analytical choices affect the actual numerical calculations. For example, data mining using suspect versus suspect plus concomitant drugs, stratified versus nonstratified analysis, analyzing each reporting year individually versus cumulative analyses, and restricting/excluding parts of the data to change expected counts can each result in different numerical outputs.

   There are deployment choices that may not affect the numerical calculations but determine the impact and significance of the data mining findings. For example, one organization may use a specific metric/threshold as a signaling criterion, while another organization uses the same metric/threshold combination but in combination with additional "triage logic" [53] based on common-sense public health notions. In each instance, the data mining is the same and the SDRs produced are the same, but if the association does not meet triage criteria, the finding will result in further action only by the first organization. Other deployment choices include whether to use the DMA as a binary classifier (i.e., SDR present/absent) versus a triage/ranking classifier, to use the DMA in parallel versus in series with conventional procedures, and to use as a supplement or as a substitute for conventional procedures.

   Perhaps the most fundamental choice once a decision is made to data mine with a specific DMA for purposes of binary classification is which metric/threshold to use. Currently, there is a small set of commonly used/endorsed thresholds of disproportionality, statistical unexpectedness and/or reporting frequency. These are often selected somewhat arbitrarily, based on individual organizational experience, validation studies of varying levels of rigor and generalizability, and/or the specific task at hand. Performance of these thresholds may be quite situation dependent

TABLE 12.9 Implementation Choices in the Use of a DMA on Spontaneous Reports[a]

Deployment Choices	Analytical Choices
Pharmacovigilance activity	Data/data source
Initial signal detection	Public databases
Modifying an index of suspicion	Proprietary database
Position of DMA within the organization	Full database
In series with conventional procedures	Database restriction to lower background reporting of adverse events
In parallel with conventional procedures	Dictionary architecture/case definitions
As replacement for conventional procedures	WHOART versus MedDRA
Classification activity	Level of specificity of terminology (e.g., preferred versus higher- or lower-level term)
Binary (SDR versus no SDR)	User-defined combinations of preferred terms
Triage/ranking (no cutoff defining SDR prioritization criteria)	Standardized MedDRA queries (SMQs)
Time of initiation in product life cycle	Drugs analyzed
New drugs (high premium on sensitivity?)	Suspect
Old drugs	Suspect plus concomitant
Single drugs versus between-drug comparisons	Drug specificity (e.g., substance or salt, or therapy group)
Metrics/thresholds[b]	Algorithm
Discrete metrics	PRR
Thresholds	ROR
Credibility intervals	BCPNN
Case count thresholds	MGPS
	Stratified versus unstratified analysis
	Age
	Gender
	Year of report
	Country of origin
	Other
	Dimensionality
	2-D (i.e., drug–event pairs)
	3-D (e.g., drug–drug–event or interaction)
	Temporality
	Cross-sectional analysis
	Time-trend analysis
	Metrics/thresholds[b]
	Discrete metrics (e.g., IC, EB05, EBGM)
	Threshold
	Interval metrics
	Case count thresholds

[a]Adapted from Hauben and Bate [52].

[b]Depending on the mode of data presentation, these may be considered as elements of both deployment choices and analytical choices.

WHOART = World Health Organization Adverse Reactions Terminology; MedDRA = Medical Dictionary for Regulatory Activities.

and involves a trade-off between generating too many false negatives and identifying the truly positive.

2. Another important source of debate relates to the lack of consensus on gold standards for adjudicating causality in PhV. The point at which causality is definitely established is also rarely clear, certainly retrospectively. This pertains in part to which associations may serve as true/false/positive/negative reference standards. A wide range of views have been expressed about adjudicating causality and what may or may not serve as reference drug–event associations, but it should be noted that the target environment of PhV is innately probabilistic in nature with residual uncertainty being the rule rather than the exception. Therefore, some have suggested a more flexible view that may make more sense for real-world PhV scenarios. For example, events that are probably causal in nature, even if not guaranteed to be causal, or events for which further investigation is warranted at a given time point, even if the association is ultimately discounted, may be considered events that are true positives in the sense of being events worth detecting.
3. The lack of a decision-theoretic calculus of opportunity costs and utilities is associated with the trade-off in sensitivity and specificity. Thus, some exercises use overall accuracy as the ultimate benchmark. Such an approach contains the implicit assumption that the opportunity costs and consequences of false-positive and false-negative findings are equal when in fact, in certain scenarios, a certain level of some types of misclassification errors may be desirable. Similarly, there are no clear decision rules for selecting between a less efficient algorithm that produces a higher overall number of credible associations versus a more efficient algorithm that identifies less credible associations overall. This is quite important because it is understandable that analysts may have a preference for approaches that are less labor intensive. It is tempting to focus exclusively on reducing false-positive and false-negative findings when in fact, finding an optimum balance between sensitivity and specificity is the target. The optimum balance may be highly situation dependent. For example, the relative premium of sensitivity versus specificity may be a function of the nature of the event with a relatively higher premium on sensitivity versus specificity during the early stages of a product's life cycle. For older products with well-established safety profiles and a large corpus of data, there may be a relatively higher premium on specificity.

There are additional performance factors not normally accounted for in published data mining exercises that may have practical implications in real-world PhV scenarios, which often involve a dynamically evolving hypothesis. One example is computational cost. Computational cost or expediency is quite variable with the frequentist methods being most computationally expedient, while other DMAs, such as logistic regression and MGPS, require intensive

computations and often require multiple scans of the entire database with resulting prolonged run times [54]. Computational cost may have practical implications and should be considered with many other factors in algorithm selection.

There are many additional sources of variability in data mining outputs related to multiple factors, including inherent mathematical properties of the algorithms themselves [55]. With the numerous aforementioned sources of residual uncertainty and challenges in accurately quantifying DMA performance, it is impossible to declare the universal superiority of a single method or a single metric/threshold combination as superior in all PhV scenarios.

The difficulties with validation of new techniques are of course not unique to the assessment of performance of DMAs in PMS. A similar debate occurs in the development of new biomarkers. With biomarkers, there is both technical/analytic validation and clinical validation. Lester [56] makes the point that in terms of clinical validation, there is no substitute for ongoing experience, and the more a test is used, the more it will be seen as "validated." Lester [56] quotes a recent FDA Science Forum, where it was suggested that rather than validation it should be more of reaching a "comfort zone," referring to clinical validation. While this also applies to data mining, this does not afford us the luxury of being complacent about assessing the incremental benefit of using DMAs routinely.

There is now almost a decade of experience with development, testing, and implementation of data mining in PhV. Such an approach has improved signal-detection practices to a marked degree at some organizations such as the WHO. For other organizations, the experience to date suggests that data mining as currently implemented may have a benefit in some situations, but that this benefit may be modest and that conventional PhV often identifies credible associations in advance of data mining. However, for many organizations, data mining identified associations that are already known, are under evaluation, or are noncausal after evaluation. This should not be interpreted to mean that DMAs are not useful. Rather, it suggests that the most useful view would fall between the extremes of "unbridled optimism" to "considerable pessimism" noted by Bate and Edwards [57] and that we must carefully consider both the strengths and weakness of these methods. Any cautionary emphasis in tone or content should not be construed as a condemnation of the quantitative methods but rather as a concern that tools with an impressive mathematical foundation may desensitize users to the rate-limiting nature of SRS data and may consequently amplify the potential for its misuse.

12.14 PRACTICAL IMPLEMENTATION

Most organizations and researchers maintain a position between the above two extremes, recognizing that disproportionality analysis represents a credible addition to the PhV tool kit that has enhanced the signal-detection performance of major PhV organizations when used responsibly and in light

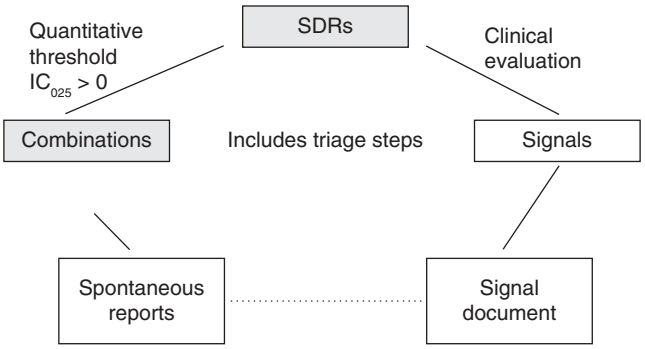


Figure 12.7 Outline of UMC signaling procedure.

of their inherent limitations and the profound rate-limiting defects in the data. Accordingly, these organizations use a comprehensive of strategies, tools, and data streams that include both clinical and computational approaches. Expressed a little differently, most organizations use these tools as supplements to, and not substitutes for, traditional signal-detection practices. Therefore, quantitative calculations in combination with various forms of discerning parameters/triage logic are used [53,58]. A schematic outline of the signaling procedure at one major PhV organization, the WHO UMC, is shown in Figure 12.7, taken from Reference 59.

12.15 THE NEED FOR COMPLEX METHODS

Some of the next generation of research in statistical methods in drug safety will focus on the use of more complex methods to make use of the information resolution that is lost with current methods. First, note that 2×2 tables result in a loss of information. If you could "unpack" cells B, C, and D in the 2×2 table, you would be reminded that these single categories lump together huge numbers of drugs as "other drugs" and numerous events as "other events." Each of these drugs and events have their own relationships, which may be important to understanding safety phenomena, such as drug–drug interactions and bystander effects, in which a drug may be associated in the 2×2 table because it is frequently coprescribed with another drug known to have that side effect. Furthermore, with only two variables displayed (drug and event), it is difficult to assess both complex interdependencies and the independent contribution of other covariates that could be confounding factors or effect modifiers.

There is an additional impetus for the development of more complex methods. The exposition and examples to this point have focused on associations involving a single drug and single event, sometimes referred to as a two-dimensional (2-D) phenomenon. While 2-D associations account for the

bulk of phenomena encountered in day-to-day PhV, there are more complex higher-dimensional phenomenon of great public health importance. For example, instead of 2-D drug–event associations, we may have associations involving multiple interacting drugs (e.g., $drug_1$–$drug_2$–event) or drug-induced syndromes in which a constellation of signs/symptoms is associated with the drug (e.g drug–$event_1$–$event_2$–$event_3$). Not only are these phenomena important to detect, but once identified, it is important to define the full range of their clinical phenotypes and to distinguish distinct but clinically overlapping syndromes. For example, neuroleptic malignant syndrome and serotonin syndrome are distinct entities, but both overlap and involve neuromuscular and autonomic features. Another textbook example of complex ADRs are drug-induced embryopathies.

Intuitively, such higher-dimensional phenomena should be more challenging to detect from ADR listing because the "prepared mind" has to make multiple cognitive links by inspecting the data. However, it is important to emphasize that, among ADRs, drug–drug interactions may be particularly amenable to detection based on a sound understanding of clinical pharmacological principles and the extensive pharmacological data available at approval, and thus may be more amenable to knowledge-based inspection. Quantitative methods adapted for drug–drug interaction detection have been applied to spontaneous reports with limited practical success. Methods have been based on logistic regression [60] and extension of measure of disproportionality to focus on an unexpected three-way dependency compared to that expected from two-way dependencies [45,61]. A three-way reporting dependency exists if the probability of a randomly selected report listing all three elements (e.g., drug–drug–event) is greater than the probability of a randomly selected report listing the most strongly dependent pairs among the former triplet (e.g., drug–drug or drug–event) The limited success has, at least in part, been due to the methods' focus on a multiplicative model; recent research has shown that an additive model can be more effective for spontaneous report screening [51,62].

Among the other information that is invisible in a 2×2 table are data on variables that may be confounding factors (also known as "lurking variables") or effect modifiers. Some of these can be relatively easy to observe in certain circumstances such as confounding by age, gender, and year of report. However, the number of potential confounding factors and effect modifiers, both recorded and unrecorded, presents difficulties in that they can result in spurious or masked associations [63]. Furthermore, the interplay of multiple variables can potentially reveal complex drug–drug interactions and drug-induced syndromes.

Since the simplest phenomenon of this nature is confounding, we illustrate with an elementary hypothetical example. Consider a fictitious drug, Rosinex, which causes nausea [64]. Suppose that 90% of the individuals taking Rosinex experience nausea, whereas 10% of the individuals not taking Rosinex experience nausea. Further, suppose that Rosinex makes one susceptible to eye infections. Consequently, due to standard practice guidelines, 90% of the

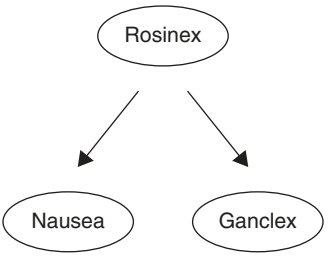


Figure 12.8 Graphical causal model. Rosinex causes nausea and also causes individuals to take Ganclex. Taking Ganclex has no effect on the probability of experiencing nausea.

TABLE 12.10 $2 \times 2 \times 2$ Contingency Table from an SRS Database That is Consistent with These Probabilities and with the Causal Model

		Nausea	No Nausea	Total
Rosinex	Ganclex	81	9	90
Rosinex	No Ganclex	9	1	10
No Rosinex	Ganclex	1	9	10
No Rosinex	No Ganclex	90	810	900

Rosinex users also take a prophylactic antibiotic called Ganclex, whereas about 1% of the non-Rosinex users take Ganclex. Ganclex does not cause nausea. Figure 12.8 shows a causal model that describes the situation. Table 12.10 shows data that a+9re consistent with this description.

Considering only Ganclex and nausea, the observed count is 82 as compared to an expected value of about 18, leading to an RR of over 4! The EBGM score would be similar. So, even though Ganclex has no causal relationship with nausea, the data mining approach based on 2×2 tables would generate a Ganclex–nausea SDR.

This is a simple example of a more general phenomenon. In general, particular patterns of association between observed and unobserved variables can lead to essentially arbitrary measures of association involving the observed variables. These measures can contradict the true unknown underlying causal model that generated the data. For example, in addition to drug–drug interaction detection, other coreporting of pairs of drugs needs to be highlighted to prevent an "innocent bystander" being inappropriately associated with an apparent ADR, in fact caused by a coprescribed and reported drug [65]. Screening out for confounders can be done, but adjustment by too many variables can lead to the missing of signals in the application of data mining [63].

There is a clear need to find patterns involving many more variables on the spontaneous reports. One example is clustering of the different adverse events

Cij	i	A1202	A0116	A0725	A0154	A0092	A0791	A0163	A0091	A0093	A0224	A0151	A0210	A0043	A0280	A0576	A0156	A0507
j	Ci/Cj	723	585	517	357	348	270	217	174	174	145	143	125	108	108	92	66	60
A1202	723	–	109	171	23	29	126	17	20	11	39	24	27	22	8	43	6	3
A0116	585	109	–	121	67	43	88	26	20	13	16	40	24	26	19	17	8	6
A0725	517	171	121	–	33	38	109	18	25	14	47	30	43	35	9	38	3	4
A0154	357	23	67	33	–	24	9	8	6	11	11	12	8	20	8	3	5	1
A0092	348	29	43	38	24	–	25	39	7	9	14	10	8	11	5	10	16	2
A0791	270	126	88	109	9	25	–	7	13	5	25	14	19	11	5	47	5	5
A0163	217	17	26	18	8	39	7	–	6	6	5	10	5	5	4	2	3	2
A0091	174	20	20	25	6	7	13	6	–	19	9	2	5	6	2	1	5	6
A0093	174	11	13	14	11	9	5	6	19	–	3	8	8	1	3	1	1	7
A0224	145	39	16	47	11	14	25	5	9	3	–	6	29	18	2	7	1	1
A0151	143	24	40	30	12	10	14	10	2	8	6	–	1	5	4	3	2	5
A0210	125	27	24	43	8	8	19	5	5	8	29	1	–	4	2	6	3	1
A0043	108	22	26	35	20	11	11	5	6	1	18	5	4	–	5	3	2	1
A0280	108	8	19	9	8	5	5	4	2	3	2	4	2	5	–	1	1	2
A0576	92	43	17	38	3	10	47	2	1	1	7	3	6	3	1	–	2	1
A0156	66	6	8	3	5	16	5	3	5	1	1	2	3	2	1	2	–	2
A0507	60	3	6	4	1	2	5	2	6	7	1	5	1	1	2	1	2	–

Figure 12.9 ADR codes and specific ADR terms as a result of application of a recurrent Bayesian confidence propagation neural network (BCPNN) to the WHO database.

listed on similar reports This can represent several patterns of interest including symptoms that constitute a syndrome. In an ADR database, the sparse nature of the data means that rarely, if ever, will all constituent symptoms of a syndrome be listed on any single case report. The individual ADR terms that make up a syndrome will not even necessarily show strong associations (positive scores of measure of disproportionality) with the drug causing the syndrome. The symptoms will occur sometimes with the drug in small groups of terms and have strong associations to other more common drug-related symptoms in the syndrome. Searching for coreporting of all symptoms has limited use, and more sophisticated methods are needed to find such relationships. A recurrent BCPNN has been applied to the WHO database of suspected ADRs [66]. This method is able to highlight clusters of ADR terms reported for specific drugs such as the following cluster of ADR terms highlighted within reporting of haloperidol suspected ADRs (Fig. 12.9).

The columns and rows list the same ADR codes that refer to specific ADR terms. The numbers in the body of the table are the number of suspected haloperidol where the pairs of ADR terms in the row and column are colisted. White squares represent pairs of ADR terms between which there is a positive IC value, the blue squares a negative IC value. The highlighted ADRs in the first pattern were neuroleptic malignant syndrome (NMS), hypertonia, fever, tremor, confusion, increased creatine phosphokinase, agitation, coma, convulsions, tachycardia, stupor, hypertension, increased sweating, dysphagia, leukocytosis, urinary incontinence, and apnea. Only one ADR term (A0116—hypertonia) had a positive IC with all other terms in the pattern; also, this list does not simply correspond to the most reported ADRs (not the highest IC value terms) for haloperidol. All ADRs are symptoms associated with NMS in standard literature sources, with the exception of dysphagia, for which published case reports exist of a possible link to NMS.

12.16 DISCUSSION

DMAs are routinely used in PhV by some organizations. In comparison to methods routinely used in other parts of the pharmaceutical industry, data mining approaches in drug safety seem relatively unsophisticated. In practice, this perception may partly reflect a failure to consider important elements in the relevant application domain. We suggest that the subtleties of the methods and the difficulties examining performance make the use of these methods challenging. The complexity of the data also argues for simple methods to maximize transparency and to make the volatility of quantitative outputs to data quality issues visible. We hope that despite these limitations and residual uncertainties, it is clear that computer-based quantitative methods have expanded the range of credible options available to major PhV organizations facing the challenge of processing vast and rapidly increasing quantities of complex and diverse data in the setting of constrained resources to the benefit of patients.

Not every organization will benefit from the application of DMAs for the task of PhV, and of those that do, some may benefit more than others. However, having a comprehensive menu of signal-detection tools and strategies that includes both clinical and quantitative approaches will allow organizations to customize a suite of signal-detection procedures that is best suited for their situation. While proprietary software will inevitably be aggressively promoted as a one-size-fits-all solution, it is likely that all the available quantitative methods represent viable options when intelligently deployed, and that the more important question for organizations using these tools is how to optimize the deployment of whichever tool they select, as part of a holistic approach to signal detection using multiple methods and data streams.

Stephen Hawking has said that a publisher warned him that for every equation he included in his books, sales would drop by half. We have observed an opposite effect in our field, namely, a tendency to be overawed by more complex methods that may desensitize users to the limitations and complexities of the data that are not necessarily overcome by more elaborate mathematical frameworks. We have no doubt that with increasing interest in accelerating statistical research in drug safety [67], there will be increasing experience with more sophisticated methods, and we will be better able to answer if and when more complex methods are more effective.

REFERENCES

1. Diller DJ, Hobbs DW. Deriving knowledge through data mining high-throughput screening data. *J Med Chem* 2004;47:6373–6383.
2. Weaver, DC. Applying data mining techniques to library design, lead generation and lead optimization. *Curr Opin Chem Biol* 2004;8:264–270.

3. Li H, Ung CY, Yap CW, Xue Y, Li ZR, Cao ZW, Chen YZ. Prediction of genotoxicity of chemical compounds by statistical learning methods. *Chem Res Toxicol* 2005;18:1071–1080.
4. Ette EI, Williams P, Fadiran E, Ajayi FO, Onyiah LC. The process of knowledge discovery from large pharmacokinetic data sets. *J Clin Pharmacol* 2001;41:25–34.
5. Shah SC, Kusiak A, O'Donnell MA. Patient-recognition data-mining model for BCG-plus interferon immunotherapy bladder cancer treatment. *Comput Biol Med* 2006;36:634–655.
6. Cerrito P. Application of data mining for examining polypharmacy and adverse effects in cardiology patients. *Cardiovasc Toxicol* 2001;1:177–179.
7. Bate A, Lindquist M, Edwards IR, Orre R. A data mining approach for signal detection and analysis. *Drug Saf* 2002;25:393–397.
8. US Food and Drug Administration. *Taste of Raspberries, Taste of Death. The 1937 Elixir Sulfanilamide Incident*. 1981. Available at http://www.fda.gov/oc/history/elixir.html (accessed February 19, 2008).
9. Woolf AD. The Haitian diethylene glycol poisoning tragedy: A dark wood revisited. *JAMA* 1998;279:1215–1216.
10. Singh J, Dutta AK, Khare S, Dubey NK, Harit AK, Jain NK, Wadhwa TC, Gupta SR, Dhariwal AC, Jain DC, Bhatia R, Sokhey J. Diethylene glycol poisoning in Gurgaon, India. *Bull World Health Organ* 2001;79:88–95.
11. Wax PM. It's happening again–Another diethylene glycol mass poisoning. *J Toxicol Clin Toxicol* 1996;34:517–520.
12. McBride WG. Thalidomide and congenital abnormalities. *Lancet* 1961;278:1358.
13. Carey JC. Editor's note. *Am J Med Genet* 2002;111:54.
14. Lindquist M. VigiBase, the WHO Global ICSR Database System: Basic facts. *Drug Inf J* 2008;l42:409–419.
15. Wakefield AJ. MMR vaccination and autism. *Lancet* 1999;354:949–950.
16. US Food and Drug Administration. Untitled. *FDA Consumer Magazine*, January–February, 2002.
17. Stähle H. A historical perspective: Development of clonidine. *Best Prac Res Clin Anaesthesiol* 2000;14:237–246.
18. Vanderveen EE, Ellis CN, Kang S, Case P, Headington JT, Voorhees JJ, Swanson NA. Topical minoxidil for hair regrowth. *J Am Acad Dermatol* 1984;11:416–421.
19. Ghofrani HA, Osterloh IH, Grimminger F. Sildenafil: From angina to erectile dysfunction to pulmonary hypertension and beyond. *Nat Rev Drug Discov* 2006;5:689–702.
20. Brinker A, Beitz J. Use of a spontaneous adverse drug events database for identification of unanticipated drug benefits. *Clin Pharmacol Ther* 2002;71:99–102.
21. School of Pharmacy, U.o.C. School of Pharmacy, University of California. 2007. Available at http://pharmacy.ucsf.edu/glossary/d/.
22. World Health Organization. The Importance of Pharmacovigilance: Safety Monitoring of Medicinal Products. WHO, 2002.
23. Edwards IR, Biriell C. Harmonisation in pharmacovigilance. *Drug Saf* 1994;10:93–102.

24. Meyboom RH, Egberts AC, Edwards IR, Hekster YA, de Koning FH, Gribnau FW. Principles of signal detection in pharmacovigilance. *Drug Saf* 1997;16:355–365.
25. Meyboom RH, Hekster YA, Egberts AC, Gribnau FW, Edwards IR. Causal or casual? The role of causality assessment in pharmacovigilance. *Drug Saf* 1997;17:374–389.
26. Glasser SP, Salas M, Delzell E. Importance and challenges of studying marketed drugs: What is a phase IV study? Common clinical research designs, registries, and self-reporting systems. *J Clin Pharmacol* 2007;47:1074–1086.
27. Reidenberg MM. Improving how we evaluate the toxicity of approved drugs. *Clin Pharmacol Ther* 2006;80:1–6.
28. Bate A, Lindquist M, Edwards IR. The application of knowledge discovery in databases to post-marketing drug safety: Example of the WHO database. *Fund Clin Pharmacol* 2008;22:127–140.
29. Hauben M, Aronson JK. Paradoxical reactions: Under-recognized adverse effects of drugs. *Drug Saf* 2006;29:970.
30. Trontell A. Expecting the unexpected–Drug safety, pharmacovigilance, and the prepared mind. *N Engl J Med* 2004;351:1385–1387.
31. Aronson JK, Ferner R. Clarification of terminology in drug safety. *Drug Saf* 2005;28:851–870.
32. Lewis A, Yu M, DeArmond SJ, Dillon WP, Miller BL, Geschwind MD. Human growth hormone-related iatrogenic Creutzfeldt–Jakob disease with abnormal imaging. *Arch Neurol* 2006;63:288–190.
33. Edwards IR, Lindquist M, Wiholm BE, Napke E. Quality criteria for early signals of possible adverse drug reactions. *Lancet* 1990;336:156–158.
34. Hauben M, Reich L, DeMicco J, Kim K. Extreme duplication in the US FDA adverse events reporting system database. *Drug Saf* 2007;30:551–554.
35. Noren GN, Orre R, Bate A, Edwards IR. Duplicate detection in adverse drug reaction surveillance. *Data Mining and Knowledge Discovery* 2007;14(3): 305–328.
36. Guruprasad PA, Rowlins MD, Day CP. Accuracy of hepatic adverse drug reaction reporting in one English health region. *BMJ* 1999;11:1041.
37. Hauben M, Aronson JK. Gold standards in pharmacovigilance: The use of definitive anecdotal reports of adverse drug reactions as pure gold and high-grade ore. *Drug Saf* 2007;30:645–655.
38. Meyboom RH, Lindquist M, Egberts AC, Edwards IR. Signal selection and follow-up in pharmacovigilance. *Drug Saf* 2002;25:459–465.
39. Venulet J. Possible strategies for early recognition of potential drug safety problems. *Adverse Drug React Acute Poisoning Rev* 1988;7:39–47.
40. Amery WK. Signal generation for spontaneous adverse event reports. *Pharmacoepidemiol Drug Saf* 1999;8:147–150.
41. Finney DJ. The design and logic of a monitor of drug use. *J Chron Dis* 1965;18: 77–98.
42. Napke E. Present ADR monitoring methods. In: *Drug Monitoring*, edited by Mann RD, Andrews EB, pp. 1–10. London: Academic Press, 1977.

43. Hauben M, Reich L. Communication of findings in pharmacovigilance: Use of the term "signal" and the need for precision in its use. *Eur J Clin Pharmacol* 2005;61:479–480.
44. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, De Freitas RM. A Bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol* 1998;54:315–321.
45. DuMouchel W, Pregibon D. Empirical Bayes screening for multi-item associations. In: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, edited by Ganter B, Stumme G, and Wille R, pp. 67–76. New York : ACM, 2001.
46. Norén GN, Bate A, Orre R, Edwards IR. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. *Stat Med* 2006;25:3740–3757.
47. Tate RL. A cautionary note on shrinkage estimates of school and teacher effects. *Fla J Educ Res* 2004;42:1–21.
48. Holmes S, Morris C, Tibshrani B. Effron: A conversation with good friends. *Stat Sci* 2003;18:268–281.
49. Gould AL. Accounting for multiplicity in the evaluation of "signals" obtained by data mining from spontaneous report adverse event databases. *Biomed J* 2007;49: 151–165.
50. Madigan D. Discussion—Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am Stat* 1999;53:198–200.
51. Norén GN, Sundberg R, Bate A, Edwards R. A statistical methodology for drug–drug surveillance. *Stat Med* 2006;25(21):3740–3757.
52. Hauben M, Bate A. Data mining in drug safety. In: *Side Effects of Drugs*, edited by Aronson JK, pp. xxxiii–xlvi. Amsterdam: Elsevior Science, 2007.
53. Stahl M, Lindquist M, Edwards IR, Brown EG. Introducing triage logic as a new strategy for the detection of signals in the WHO Drug Monitoring Database. *Pharmacoepidemiol Drug Saf* 2004;13:355–363.
54. Hauben M, Madigan D, Hochberg AM, Reisinger SJ, O'Hara DJ. Data mining in pharmacovigilance: Computational cost as a neglected performance parameter. *Int J Pharm Med* 2007;21:319–323.
55. Hauben M, Reich L, Gerrits CM, Younus M. Illusions of objectivity and a recommendation for reporting data mining results. *Eur J Clin Pharmacol* 2007;63:517–521.
56. Lester DS. The many lives of the biomarker. *Drug Inf J* 2007;41:551–553.
57. Bate A, Edwards IR. Data mining in spontaneous reports. *Basic Clin Pharmacol Toxicol* 2006;98:330–335.
58. Hauben M. Signal detection in the pharmaceutical industry: Integrating clinical and computational approaches. *Drug Saf* 2007;30:627–630.
59. Lindquist M, Edwards IR, Bate A, Fucik H, Nunes AM, Stahl M. From association to alert—A revised approach to international signal analysis. *Pharmacoepidemiol Drug Saf* 1999;8:S15–S25.
60. Van Puijenbroek EP, Egberts AC, Meyboom RH, Leufkens HG. Signalling possible drug-drug interactions in a spontaneous reporting system: Delay of withdrawal bleeding during concomitant use of oral contraceptives and itraconazole. *Br J Clin Pharmacol* 1999;47:689–693.

61. Orre R, Lansner A, Bate A, Lindquist M. Bayesian neural networks with confidence estimations applied to data mining. *Comput Stat Data Anal* 2000;34:473–493.
62. Thakrar BT, Grundschober SB, Doessegger L. Detecting signals of drug-drug interactions in a spontaneous reports database. *Br J Clin Pharmacol* 2007;64: 489–495.
63. Hopstadius J, Norén GN, Bate A, Edwards IR. Impact of stratification in adverse drug reaction surveillance. *Drug Saf* 2008;31:1035–1048.
64. Hauben M, Madigan D, Gerrits CM, Walsh L, Van Puijenbroek EP. The role of data mining in pharmacovigilance. *Expert Opin Drug Saf* 2005;4:929–948.
65. Purcell P, Barty S. Statistical techniques for signal generation: The Australian experience. *Drug Saf* 2002;25:415–421.
66. Orre R, Bate A, Norén GN, Swahn E, Arnborg S, Edwards IR. A bayesian recurrent neural network for unsupervised pattern recognition in large incomplete data sets. *Int J Neural Syst* 2005;15:207–222.
67. Bilker W, Gogolak V, Goldsmith D, Hauben M, Herrera G, Hochberg A, Jolley S, Kulldorff M, Madigan D, Nelson R, Shapiro A, Shmueli G. Accelerating statistical research in drug safety. *Pharmacoepidemiol Drug Saf* 2006;15:687–688.

13

DATA MINING METHODS AS TOOLS FOR PREDICTING INDIVIDUAL DRUG RESPONSE

AUDREY SABBAGH AND PIERRE DARLU

Table of Contents

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

13.1 THE PROMISE OF PHARMACOGENOMICS

There are often large differences among individuals in the way they respond to medications in terms of both host toxicity and treatment efficacy. Individual variability in drug response varies from potentially life-threatening adverse drug reactions (ADRs) to equally serious lack of therapeutic efficacy. Serious adverse drug reactions (SADRs) are estimated to be the fourth leading cause of death in the United States and each year, about 2 million patients in the United States experience an SADR when using marketed drugs, resulting in 100,000 deaths [1]. Similar numbers have been estimated for other Western countries [2]. The resulting cost burden is enormous, representing tens of billions of dollars, and has an impact on both the healthcare and pharmaceutical industries internationally [3,4]. Moreover, SADR can lead to drug withdrawals, depriving some patients of otherwise beneficial drugs: between 1976 and 2007, 28 drugs were withdrawn from the U.S. market for safety reasons [5,6]. Regulators, drug companies, physicians, and their patients would all like tools to better predict the apparently unpredictable.

A variety of factors, including age, sex, diet, state of health, and concomitant therapy, can influence a person's response to drug therapy. However, it has become clear during the past 50 years that genetics can account for a large part of this interindividual variability (Fig. 13.1). Clinical observations of inherited differences in drug effects were first documented in the 1950s [7–9], leading to the birth of pharmacogenetics.

The field of pharmacogenetics seeks to identify genetic determinants of drug response, including both those that are inherited and those that arise within tumors. Once a drug is administered to a patient, it is absorbed and distributed to its site of action, where it interacts with targets (such as receptors and enzymes), undergoes metabolism, and is then excreted. Each of these processes could potentially involve clinically significant genetic variation. Absorption, distribution, metabolism, and excretion can all influence pharmacokinetics, that is, the final concentration of the drug reaching its target. Genetic variation can also occur in the drug target itself or in signaling cascades downstream from the target, in the latter case involving pharmacodynamic factors (Fig. 13.1).

Initially, pharmacogenetic studies focused their attention on variations in single candidate genes chosen on the basis of our knowledge about the medication's pharmacokinetics and mechanisms of action. They especially focused on drug metabolizing enzymes (DME) since genetic variation of drug metabolism has long been considered as one of the major causes of interindividual variation in drug effects. However, contemporary studies increasingly involve entire "pathways" encoding proteins that influence both pharmacokinetics and pharmacodynamics, as well as genome-wide approaches. For this reason, some authors have suggested the term pharmacogenomics as a replacement for pharmacogenetics. The precise distinction between pharmacogenetics and pharmacogenomics remains unclear, but the term "pharmacogenomics"

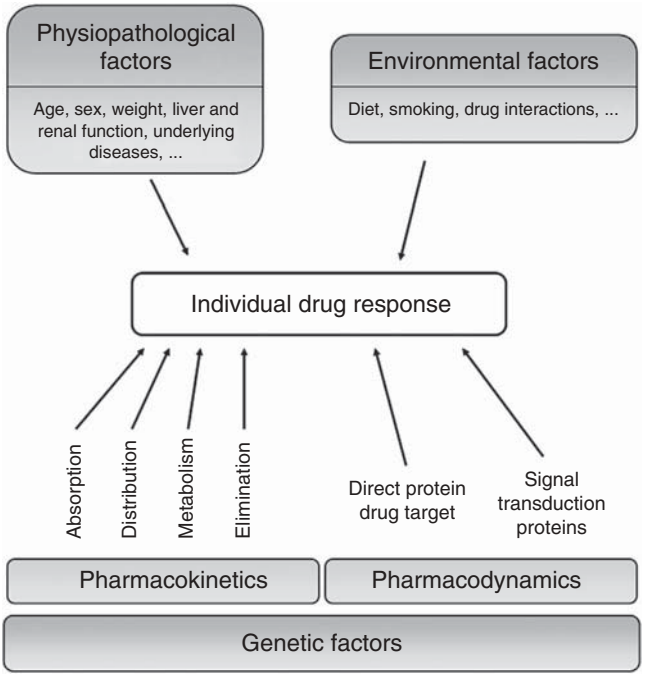


Figure 13.1 The multifactorial nature of drug response. Most drug effects are determined by the interplay of several gene products that influence the pharmacokinetics and pharmacodynamics of medications as well as by nongenetic factors such as age, sex, state of health, and lifestyle.

is usually employed to refer to the new genomic methodologies used to identify the network of genes that govern an individual's response to drug therapy, such as genome-wide scans, haplotype tagging, gene expression profiling, and proteomics [10].

Until recently, much of the research efforts in human genomics have focused on the genetic determinants of complex diseases rather than on the genetics of drug response. Yet, pharmacogenomics is likely to provide more immediate clinical returns than the study of common disease predisposition [10]. When an association between a pharmacogenetic variant and a drug response phenotype is identified, it can be of direct diagnostic use: such genetic predictors can be used to avoid rare ADRs, to adjust dose or to select which of several alternative drugs has the highest efficacy. By contrast, when a new variant predisposing to a complex disease is identified, it may indicate a new therapeutic target, but it takes a long time to develop new medicines to hit this target. Finding such variants is therefore less immediately usable than identifying pharmacogenomic variants that, through diagnostic testing, can rapidly increase the efficacy and safety of existing therapies. Another aim of pharmacogenomics is also to discover new therapeutic targets. Therefore, the

enormous potential of pharmacogenomics and its high clinical relevance make it extremely attractive, and there is an increasing effort to discover new pharmacogenomic variants.

By understanding the genetic factors that govern variable drug response, pharmacogenomics seeks to reduce the variation in how people respond to medicines by tailoring therapy to individual genetic makeup (Fig. 13.2). The ultimate goal is to yield a powerful set of molecular diagnostic methods that will become routine tools with which clinicians will select medications and drug doses for individual patients, with the goal of enhancing efficacy and safety. Some tests that incorporate pharmacogenetic data into clinical practice are now available [12], with many more to follow. In addition to optimizing the use of currently prescribed medications, pharmacogenomics may also offer new strategies and efficiencies in the drug development process. If nonresponders (NRs) or toxic responders can be prospectively identified by genotyping, it may be possible to reduce the number of subjects needed in phase II and phase III clinical trials by eliminating those who will not (cannot) respond due to inherited differences in DMEs or drug targets [11].

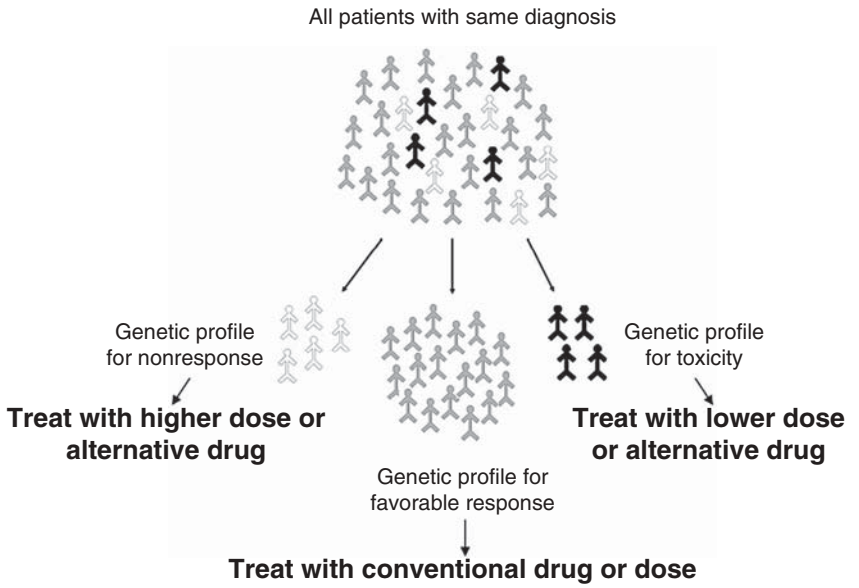


Figure 13.2 The promise of pharmacogenomic testing. By applying the results of pharmacogenomic research to clinical practice, physicians will be able to use information from patients' DNA to determine how patients are likely to respond to a particular medicine. The end result will be the optimal selection of medications and their dosages based on the individual patient and not treatment based on the average experience from the entire universe of patients with a similar diagnosis (modified from Reference 11).

13.2 COMBINATORIAL PHARMACOGENOMICS

Many of the first pharmacogenetic traits that were identified were monogenic—that is, they involved only a single gene. There are several examples of common pharmacogenetic variants that have essentially Mendelian effects on drug response, such as NAT2, cytochrome P450 2D6 (CYP2D6), and thiopurine S-methyltransferase (TPMT) (Fig. 13.3). These monogenic traits, all involving drug metabolism, have a marked effect on pharmacokinetics, such that individuals who inherit an enzyme deficiency must be treated with markedly different doses of the affected medications. However, in most instances, the overall pharmacological effects of medications are more often polygenic traits determined by multiple polymorphisms in many genes that influence both the pharmacokinetics and pharmacodynamics of medications. Such more complex traits are more difficult to elucidate in clinical studies.

A recent review by Wilke et al. [4] highlights the potential genetic complexity of drug response and susceptibility to ADRs. Most drugs are indeed metabolized by several different enzymes, can be transported by different types of proteins, and ultimately interact with one or more targets. If several steps in this type of pathway were to display genetic variation, multiple overlapping distributions would quickly replace multimodal frequency distributions like those shown in Fig. 13.3.

13.2.1 DME–DME Interactions

Many drugs are eliminated from the body by more than one metabolic pathway. The complex nature of each gene–gene (DME–DME) interaction is

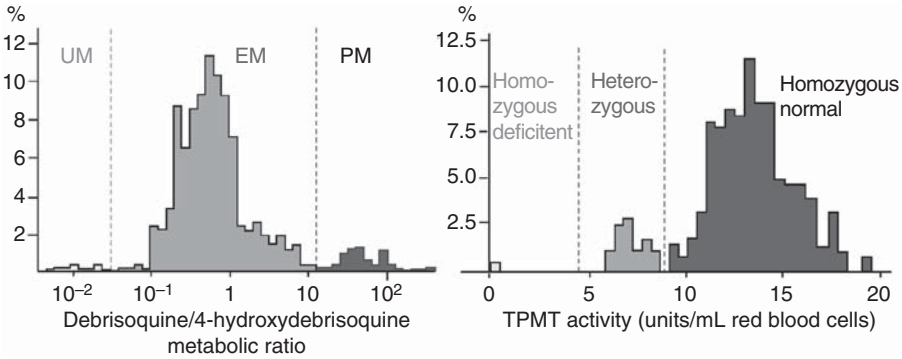


Figure 13.3 (A) Frequency distribution of the debrisoquine/4-hydroxydebrisoquine ratio in 1011 Swedish subjects. Poor metabolizers (PMs), extensive metabolizers (EMs), and ultrarapid metabolizer (UMs) (modified from References 13–15). (B) Frequency distribution of the erythrocyte activity of the thiopurine S-methyltransferase (TPMT) in 298 unrelated subjects (164 males and 134 females), classified depending on their genetic polymorphism at the *TPMT* gene (modified from Reference 16).

then partly defined by the respective contribution of each gene product to the overall biological activity of the drug. Variables to be considered include the balance between metabolic activation and inactivation, the balance between phase I (oxidative) and phase II (conjugative) activity, and the relative potency of each metabolite with respect to the specific clinical phenotype being studied [4].

Consider the example of phenytoin, an important first-line antiepileptic drug known to be metabolized by several phase I DMEs [17]. At least two of these enzymes, cytochrome P450 2C9 (CYP2C9) and CYP2C19, are highly polymorphic in most human populations, and pharmacokinetic studies have demonstrated that abnormal *CYP2C9* and *CYP2C19* alleles are associated with altered circulating drug levels [18,19]. Therefore, although a nonfunctional variant of either gene might not clearly predispose to increased risk of ADR *in vivo*, a combination of two nonfunctional variants in both genes might. Moreover, if the relative balance between different routes of phase I metabolism is likely to affect the overall clinical efficacy of drugs, the situation is further compounded when one considers the impact of phase II metabolism. It is conceivable that a polymorphism causing subtle alterations in phase I enzyme activity that may not predispose patients to an ADR when considered alone could precipitate a phenotypic change in the presence of an otherwise subclinical phase II polymorphism. Such a relationship can only be elucidated through combinatorial analyses that account for variability in both enzyme systems.

13.2.2 Interactions between Pharmacokinetic Factors

Functional genetic polymorphisms are not limited exclusively to drug metabolism and can affect the full spectrum of drug disposition, including a growing list of transporters that influence drug absorption, distribution, and excretion. The *ABCB1* (multidrug resistance 1 [*MDR1*]) gene product P-glycoprotein is the most widely studied drug transporter: it has a recognized role in the bioavailability and biliary, intestinal, and renal excretion of numerous drugs and has been a particular focus of attention as a putative mechanism of drug resistance [20]. Statins, which undergo oxidative phase I metabolism by polymorphic members of the cytochrome P450 family, additional modifications through phase II conjugation, and differential tissue distribution via membrane transporters including ABCB1, provide a striking example of a situation in which genetic variation may affect a multitude of kinetic processes [4].

13.2.3 Interactions between Pharmacokinetic and Pharmacodynamic Factors

Genetic variation in pharmacodynamic processes can also lead to clinically recognizable differences in treatment outcome. This additional layer of complexity takes us far beyond monogenic traits into a situation that most inves-

tigators believe will represent a substantial, if not the major, component of the discipline's future—polygenic variation in both pharmacokinetics and pharmacodynamics. A recent example that involves genes that influence both the pharmacokinetics and pharmacodynamics of the anticoagulant drug warfarin illustrates that point.

Warfarin is a widely used coumarin anticoagulant that is difficult to use because of the wide variation in dose required to achieve a therapeutic effect, a narrow therapeutic range, and the risk of bleeding. At least 30 genes may be involved in the mode of action of warfarin, but the most important ones affecting the pharmacokinetic and pharmacodynamic parameters of warfarin are *CYP2C9* and vitamin K epoxide reductase complex subunit 1 (*VKORC1*). The *CYP2C9* gene product is the main enzyme involved in warfarin metabolism and *VKORC1* encodes the direct protein target of the drug. These two genes, together with environmental factors, have been shown to account for around 50–60% of the variance in warfarin dose requirement [21]. This example represents, probably in simplified form, the type of multifactorial model that many investigators expect to observe with increasing frequency in the future.

Another example illustrates the extent to which a combinatorial approach that considers multiple interacting genes could be beneficial. Arranz et al. [22] performed an association study using a multiple candidate gene approach to gain insight into the genetic contribution to response variability to clozapine in schizophrenic patients. They investigated 19 genetic polymorphisms in 10 neurotransmitter receptor-related genes and looked for the combination of polymorphisms that gives the best predictive value of response to clozapine. A combination of the six polymorphisms showing the strongest association with response provided a positive predictive value of 76%, a negative predictive value of 82%, with a sensitivity of 96% for identifying schizophrenic patients showing improvement with clozapine, and a specificity of 38% for the identification of patients who did not show a substantial improvement in response to clozapine treatment.

In conclusion, the behavior of most drugs may depend on a wide range of gene products (DMEs, transporters, targets, and others), and in many cases, the importance of polymorphisms in one of the relevant genes might depend on polymorphisms in other genes. It is therefore important that researchers investigate potential gene–gene interactions, or epistasis, and gene–environment interactions, which are increasingly recognized phenomena in the field of human genetics and in pharmacogenomics.

13.3 IDENTIFYING USEFUL MARKER COMBINATIONS FOR THE PREDICTION OF INDIVIDUAL DRUG RESPONSE

The objective of pharmacogenomic research is to identify a genetic marker, or a set of genetic markers, that can predict how a given person will respond

to a given medicine. A significant challenge for pharmacogenetic researchers is therefore to identify and apply useful statistical methods for finding such predictive marker combinations.

13.3.1 Logistic Regression

In clinical studies comparing the genotype frequencies between responder and nonresponder individuals for a given treatment, it is tempting to use logistic regression to model the relationship between a set of multilocus genotypes and the treatment outcome. The logistic regression model consists of a weighted sum of predictors linked to the outcome variable by the logit function. Weights are determined in such a way that the resulting sum discriminates in the best possible way between responder and nonresponder individuals by showing large values for the former and low values for the latter. If genetic markers are to be considered as potential predictors, the logistic regression model would be a weighted sum of genotype codes, where, for example, the three genotypes at a biallelic marker such as a single-nucleotide polymorphism (SNP) are assigned codes of 0, 1, or 2. This parametric statistical method is often applied in genetic epidemiology to analyze the effect of genetic and environmental predictors on a dichotomic outcome, such as drug response.

13.3.2 The Need for Higher-Order Computational Methods

The logistic regression approach, however, suffers from several shortcomings. First, logistic regression is not appropriate to detect complex gene–gene interactions (e.g., situations where some gene variants act with others additively, in a multiplicative way, or with a compensatory effect) since, like other traditional regression methods, it relies on the basic assumption of linear combinations only [23,24]. Second, the rapid increase in the availability of large numbers of genetic markers makes the number of potential predictors very large and, when combined with the generally much smaller number of observations, creates a statistical problem that has been referred to as the "curse of dimensionality" [25]. Because the number of possible interaction terms grows exponentially as each additional main effect is included in the model, logistic regression, like most parametric statistical methods, is limited in its ability to deal with interaction data involving many simultaneous factors [4]. It has been shown through simulation studies that having fewer than 10 outcome events per independent variable can lead to biased estimates of the regression coefficients and a consequent increase in type I and type II errors [26].

Therefore, higher-order computational methods are needed to select from the large amount of genetic and environmental predictors, a small group of predictors, and/or interactions between predictors that have a significant effect on the treatment outcome.

13.4 DATA MINING TOOLS AVAILABLE TO PREDICT INDIVIDUAL DRUG RESPONSE FROM GENETIC DATA

Several pattern recognition tools, described 15–20 years ago, have recently been applied in epidemiological genetic studies and have proven to be highly successful for modeling the relationship between combinations of polymorphisms and clinical end points [27]. Compared to traditional techniques of analysis such as logistic regression, these nonparametric statistical methods offer the possibility to model complex nonlinear relationship between phenotype and genotype, without the explicit construction of a complicated statistical model. Hence, recent applications in the field of genetic epidemiology shifted toward data mining approaches, and a dramatic burst of methods occurred during the last decade (see, for review, References 28 and 29).

Because a comprehensive discussion of all available methods is beyond the scope of this chapter, we detail many of the most popular methods. In particular, we focus on methods based on partitioning the data along a tree with various optimization criteria, methods based on combinatorial procedures searching for the best combination of input genetic variables as predictive of the phenotype of interest, or neural network methods, which attempt to classify phenotype by training successive layers through an activation function, the genetic data being introduced as input.

13.4.1 Tree-Based Methods

Classification tree methods, also called recursive partitioning (RP) methods, are tree-shaped structures representing sets of decisions that generate rules of classification of a data set, the final purpose being that the terminal leaves of the tree contain observations that are the most homogeneous in terms of drug response and that are linked to the genetic markers selected along the tree branching process. The first step of the tree reconstruction is to find the best screened genetic marker (SNP, haplotype, or genotype) that allows splitting the sample into two homogeneous subgroups contrasted for their drug response phenotype. This process of splitting is recursively continued until it meets a certain criterion or stops before the last leaves contain too few individuals. Then, the nodes of the tree are explored backward by a pruning procedure to test their significance, removing those that are not significant at a prespecified p level, by a χ^2 test for instance [30–32]. These classification tree methods have the advantage of allowing a large number of input predictors, such as genetic polymorphisms or SNPs, but are not suitable for identifying the possible effects of interactions between input variables when the marginal effects on the drug response are not significant.

Cross-validation methods are used to estimate the prediction error of the constructed decision tree. The data set is randomly divided into n groups (typically 10). Only $n - 1$ groups are selected to construct the classification

tree, and the remaining group is used to evaluate the prediction accuracy. This procedure is repeated a large number of times to finally obtain an averaged accuracy and the predicted drug response for each individual.

Random forest methods [33,34] are also based on classification trees that are comparable to the previous one. However, the tree is not built by following a deterministic way. Indeed, multiple bootstrapped samples are first built up. The individuals that are not randomly drawn during this resampling process will be used further to test the prediction error of the tree. Second, at each split, a random selection of predictors (as SNPs) is performed, and the selected ones are used to carry on dividing the tree. This procedure is repeatedly done a large amount of time. Finally, one counts how often, among the number of random trees, the phenotypes of the left-out individuals, which are different from one bootstrap to another, are allocated to the different class of predictors. Finally, the largest class is viewed as the best predictor of a given phenotype. The random forest approach has been improved to take into account imbalanced data [35].

Random forests generally exhibit a substantial performance improvement over the single tree classifier. Moreover, the predictive importance for each predictor variable can be scored by measuring the increase in misclassification occurring when the values of the predictor are randomly permuted. These importance scores account for the contribution of a variable to the prediction of the response in the presence of all other variables in the model. They consequently take into account interactions among variables and make interaction variables more likely to be given high importance relative to other variables.

13.4.2 Combinatorial Methods

The goal of this family of methods is to search over all possible combinations of polymorphisms to find the combination(s) that best predict the outcome of interest. These methods are particularly suited for the identification of possible gene–gene interactions since no marginal main effects are needed during the training/model-building stage. Moreover, they have the advantage of being nonparametric and free of a specified genetic model.

The multifactor dimensionality reduction (MDR [36]) method is specifically designed for the identification of polymorphism combinations associated with a binary outcome (as responders versus NRs or as "+"/ versus "–" phenotype). It uses a data reduction strategy for collapsing high-dimensional genetic data into a single multilocus attribute by classifying combinations of multilocus genotypes into high-risk and low-risk groups based on a comparison of the ratios of the numbers of "+" and "–" individuals. Part of the task of MDR is to select the appropriate combination of genotypes to be used in the collapsed multilocus attribute. The new, one-dimensional multilocus variable is then evaluated for its ability to classify and predict the clinical end points through cross-validation and permutation testing. The first step consists

in randomly dividing the whole sample of individuals into n groups for further cross validation. The second step, using only $n - 1$ groups, consists in selecting in turn k polymorphic markers (e.g., SNPs or categorical predictors), among N, each having several possible classes. Then, within each cell obtained by crossing all classes of the k markers, the ratio of "+" and "–" individuals is evaluated. The next step is to pool all the cells having a ratio higher than a specified value (e.g., 1), reducing the dimensionality from k to 1 (with two classes, i.e., high and low ratios). Then, the ability to predict the status is estimated by testing the subsample initially left apart for this purpose. The proportion of wrongly classified individuals in this subsample is used as a prediction error. The procedure is repetitively performed, by randomization of the initial groups, to get an average prediction error. Moreover, all the combinations of k markers can be explored, k starting from 2 to the largest value compatible with computational facilities. The combination of markers giving the smallest average prediction error is considered as the one giving the strongest association with the drug response. The statistical significance can be obtained by usual permutation tests. Through simulated data, the MDR approach was shown to be quite powerful in the presence of genotyping errors or missing data, far less when proportions of phenocopy or genetic heterogeneity are large.

The MDR approach has been recently extended to take into account imbalanced data sets [37], to handle quantitative traits instead of binary ones, and to adjust for covariates such as sex, age, or ethnicity (generalized MDR [38]). Moreover, to facilitate the processing of large data sets, a more efficient MDR algorithm has been developed to allow an unlimited number of study subjects, total variables, and variable states, thereby enabling a 150-fold decrease in runtime for equivalent analyses [39].

Other approaches were proposed to predict a quantitative phenotype from multilocus data sets. For instance, combinatorial partitioning method (CPM) is a method looking for the optimal partitions of a set of genotypes minimizing the within-partition variance and maximizing the among-partitions variance of the trait [40]. It considers all possible loci combinations and evaluates the amount of phenotype variability explained by partitions of multilocus genotypes into sets of genotypic partitions. Those sets of genotypic partitions that explain a significant amount of phenotypic variability are retained and validated using cross-validation procedures. However, CPM has to examine a number of partitions that enormously increase with the number of selected loci. To overcome the computationally intensive search technique used by CPM, Culverhouse et al. [41] developed the restricted partitioning method (RPM). RPM uses a search procedure that does not require to exhaustively compare all partitions, by first performing a multiple comparison test and then by merging, within a new partition, the partitions showing no statistical differences in their mean quantitative trait.

Detection of informative combined effects (DICE) is another approach combining the advantages of the regressive approaches, in terms of modeling

and interpretation of effects, with those of data exploration tools [42]. It aims at exploring all the major and combined effects of a set of polymorphisms (and other nongenetic covariates) on a phenotype of any kind (binary, quantitative, or censored). The DICE algorithm explores, using a forward procedure, a set of competing regressive models for which an information criterion, Akaike's information criterion [43], is derived. The model space is explored sequentially and in a systematic way, considering the inclusion of main and combined (additive and interactive) effects of the covariates and leading to the identification of the most informative and parsimonious model(s) that minimize(s) the information criteria. This model-building approach is more similar to a traditional logistic regression framework than the other combinatorial methods.

13.4.3 Artificial Neural Networks

Artificial neural networks (ANNs) is another tool that can be used to predict individual drug response using a set of predictor variables such as SNPs, genotypes, or haplotypes. The network is structured in an input layer composed of units, each of them corresponding to a specific predictor. Each input unit is connected to one or several units that belong to a first hidden layer. These previous units can themselves be connected to units of a second hidden layer. The number of layers can be increased to optimize the prediction. Finally, the last hidden layer has its units connected to the output layer that contains the predicted value. The process consists in repeatedly introducing in the network several sets of input data with their associated output data, progressively adjusting the weight coefficient allocated to links that connect units of successive layers. The process stops when there is no more improvement in prediction accuracy. An ANN model can be constructed for each combination of predictor variables (SNPs). Lastly, the most parsimonious combination of predictor showing the best performance is retained and its statistical significance can be estimated using a permutation test strategy [34,44]. The ANN has several features that are well adapted for pharmacogenetics since it can handle large quantities of data, does not require any particular genetic model, and can test interaction between variables.

The data mining tools discussed above can be readily applied in pharmacogenetic candidate gene studies as well as in genome-wide scans. However, for large-scale or genome-wide studies, it may be beneficial to combine several approaches into a multistep methodological framework. For example, the first step could involve a data mining tool to select from the large amount of genetic polymorphisms a small group of predictors and/or interactions between predictors that have a significant effect on the treatment outcome. Subsequently, parameters for the selected predictors can be estimated by a traditional statistical method such as logistic regression analysis to put the model in a more interpretable or familiar framework.

13.5 APPLICATIONS OF DATA MINING TOOLS IN PHARMACOGENOMICS

The empirical approach for the development of a genomic signature in pharmacogenomics consists in measuring a large number of "features" (genetic or nongenetic variables) on each treated patient belonging to a body of "training data" and then selecting the combination of features that are most significantly correlated with patient response. The features could, for example, be SNPs as measured by genotyping the patients' lymphocytes for a panel of candidate genes, or gene expression levels as determined by a whole genome microarray expression profile of the patient's tumor. As it is important that the genomic signature classifier be reproducibly measurable and accurate in predicting treatment outcome, it must be validated on an independent test set to confirm that it would generalize to new data. Ideally, for a genetics-based predictive assay to be useful, both its sensitivity and specificity need to be as close to 100% as possible. Of course, what might be considered adequate levels of sensitivity and specificity will depend on the particular medicine and treatment outcome being evaluated. We review here some of the most successful applications of data mining approaches in the field of pharmacogenomics for the development of genomic signatures using both genetic polymorphism data and gene expression data.

13.5.1 Development of Pharmacogenomic Classifiers from Single-Nucleotide Germline Polymorphisms

Serretti and Smeraldi [45] reported the first attempt to use ANN in pharmacogenetic analyses. They applied this technique to short-term antidepressant response in mood disorders. One hundred twenty-one depressed inpatients treated with fluvoxamine were included in this study. All patients were evaluated at baseline and weekly thereafter until the sixth week using the 21-item Hamilton Rating Scale for Depression (HAM-D-21) [46]. A decrease in HAM-D scores to 8 or less was considered the response criterion. According to this criterion, 81 patients were classified as responders and 40 patients as NRs. Two gene polymorphisms located in the transcriptional control region upstream of the 5-HTT coding sequence (*SERTPR*) and in the tryptophan hydroxylase (*TPH*) gene have previously been shown to be significantly associated with drug response, and the authors wanted to reanalyze the data by using a neural network strategy to evaluate the possible nonlinear interactions between these two gene polymorphisms. A multilayer perceptron network composed by one hidden layer with seven nodes was chosen. The inputs to the first layer of the neural network consisted of *SERTPR* and *TPH* genotypes, while the target outputs consisted of the response status. The network was then trained to attempt to predict response from genotypes. They performed both training and testing on the entire data set, and the statistical significance of any observed association between outputs and affection status

was estimated using a permutation test. A total of 77.5% of responders and 51.2% of NRs were correctly classified (empirical p value = 0.0082). They performed a comparison with traditional techniques: a discriminant function analysis correctly classified 34.1% of responders and 68.1% of NRs ($F = 8.16$, $p = 0.0005$). The authors concluded that ANN may be a valid technique for the analysis of gene polymorphisms in pharmacogenetic studies.

An ANN-based approach was also used by Lin et al. [47] to address gene–gene and gene–environment interactions in interferon therapy for chronic hepatitis C (CHC) patients by using genetic and clinical factors. The combination of pegylated interferon and ribavirin is the gold standard for treating CHC patients. However, treatment failure occurs in about half of the patients, and therapy often brings significant adverse reactions to some patients. Therefore, considering side effects and treatment cost, the prediction of treatment response as early as possible, ideally before treatment, is of major interest. In their study, Lin et al. [47] collected blood samples from 523 CHC patients who had received interferon and ribavirin combination therapy, including 350 sustained responders and 173 NRs. Based on the treatment strategy for CHC patients, they focused on candidate genes involved in pathways related to interferon signaling and immunomodulation. A total of 20 SNPs were selected from six candidate genes (*ADAR*, *CASP5*, *ICSBP1*, *IFI44*, *PIK3CG*, and *TAP2*). They implemented a feedforward neural network to model the responsiveness of interferon and the backpropagation algorithm was used for the learning scheme. Inputs were the genetic and clinical factors including SNP markers, viral genotype, viral load, age, and gender. Outputs were the interferon-responding status. The prediction accuracy of each model was estimated using a fivefold cross-validation procedure, and a permutation test was applied to measure the statistical significance of an association between predictors and drug response. All possible combinations of N factors were evaluated sequentially, and the N-factor model displaying the highest prediction accuracy was retained. *IFI44* was found in the significant two-, three-, and four-locus gene–gene effect models as well as in the significant two- and three-factor gene–environment effect models. Furthermore, viral genotype remained in the best two-, three-, and four-factor gene–environment models. Thus, these results strongly support the hypothesis that *IFI44*, a member of the family of interferon-inducible genes, and viral genotype may play a role in the pharmacogenomics of interferon treatment. In addition, their approach identified a panel of 10 factors that may be more significant than the others for further investigation. Hence, their results suggest that an ANN-based approach may provide a useful tool to deal with the complex nonlinear relationship between genetic and clinical factors and the responsiveness of interferon.

Culverhouse et al. developed the RPM method to improve computation time as compared to CPM. To assess the properties of RPM on real data, Culverhouse et al. [41] applied the RPM algorithm on data involving the metabolism of irinotecan, a drug in common use in chemotherapy for a variety

of cancers. The data analyzed consisted of 10 quantitative measures of irinotecan metabolism and genotypes for 10 SNPs in seven candidate genes (*ABCB1*, *ABCC1*, *ABBC1*, *XRCC1*, *CYP3A4*, *CYP3A5*, and *UGT1A1*) thought to be related to irinotecan metabolism for 65 unrelated individuals. Initial analysis of these data using standard statistical methods found no significant correlation between any single locus genotypes and any of the quantitative traits. RPM was then applied with the hope of revealing nonlinear gene–gene interactions associated with one of the quantitative phenotypes. They restricted their analysis to two-way interactions since performing tests for higher-order interactions would be expected to have very low power because of the sparseness of the multilocus observations and the need to correct for many more tests. The empirical p values were obtained from permutation tests based on a null distribution of 5000 points computed separately for each pair of loci. Nine combinations of two SNPs were found to be significantly associated with one of the quantitative traits studied ($p < 0.05$).

In a recent study, Warren et al. [48] applied a tree-based data mining method to a pharmacogenetic study of the hypersensitivity reaction (HSR) associated with treatment with abacavir, an effective antiretroviral drug that is used to treat HIV-infected patients. Thirty-two genetic markers resulting from replication of genome scan discoveries, plus six markers found during candidate gene studies, were chosen as potential contributors to polygenic or epistatic effects leading to susceptibility to abacavir HSR. Among these markers, *HLA-B*5701* possessed the highest performance characteristics, with a sensitivity of 56.4% and a specificity of 99.1%. Although specificity was quite high, sensitivity was only moderate. Therefore, RP was applied to evaluate combinations of three or more markers with respect to their usefulness in estimating HSR risk. The goal was to identify a marker set with sufficient sensitivity and specificity to be clinically useful. One thousand random trees were generated using data from 349 white subjects, including 118 patients who developed presumed HSR (cases) and 231 patients who did not (controls). None of the RP trees produced displayed performance characteristics with both high sensitivity and high specificity. However, the four most predictive RP trees resulted in performance characteristics slightly better than *HLA-B*5701* alone. Furthermore, RP results enabled the genetic delineation of multiple risk categories. For instance, if one of the most predictive RP tree was applied to a population of white abacavir-treated patients, it was estimated that 17.4% of patients would be assigned to a group with a 0.2% risk of HSR. An additional 75.0% would be assigned to a group with a 2.7% risk of HSR. The remaining 3.6% of the patients would have an HSR risk of 21.3%, or higher, including 2.5% of all patients whose estimated risk would be 100%. Hence, in contrast to traditional diagnostic tests that typically classify patients into one of two groups, the RP algorithm is able to identify subsets of a patient population for which the estimated risk may be extremely low—in the case of prediction of adverse events, a protective effect—or very high.

Data mining approaches can also be applied at the level of single genes. For example, in situations where the drug response phenotype is mainly determined by a single gene, one may want to define the most parsimonious combination of polymorphisms within this gene that could predict individual drug response with a high accuracy. As an illustration of the application of data mining tools to data sets of polymorphisms typed in a single gene, we present here the results of a previous study where we investigated the ability of several pattern recognition methods to identify the most informative markers in the *CYP2D6* gene for the prediction of CYP2D6 metabolizer status [32]. CYP2D6 plays a crucial role in the metabolism of xenobiotics and processes about 20% of all commonly prescribed drugs. Genetic polymorphism at the *CYP2D6* gene locus is responsible for pronounced interindividual and interethnic differences in the catalytic activity of the enzyme (Fig. 13.3). Since therapeutic efficacy and adverse events in treatment with many drugs depend on CYP2D6 activity, it is anticipated that genotyping of *CYP2D6* may become an important tool in individually optimized drug therapy. However, in many respects, the *CYP2D6* gene represents a challenge for genotyping because (1) it is extremely polymorphic, with over 90 known allelic variants and subvariants reported to date, and (2) the polymorphisms reported are not only single nucleotide in nature but are also gene deletion, duplication, and pseudogene derivatives. Therefore, efficient genetics-based assays, in which only the most informative markers for phenotype prediction would be screened, are needed to simplify the analyses while keeping a high predictive capacity. The goal of our study was to define which set of *CYP2D6* polymorphisms should be routinely identified to allow a sufficiently reliable but still practicable estimation of an individual's metabolic capacity. To address this issue, four data mining tools (classification trees, random forests, ANN, and MDR) were applied to *CYP2D6* genetic data from eight population samples of various ethnic origin. Marker selection was performed separately in each population sample in order to design ethnic-specific pharmacogenetic tests that take into account population-specific genetic features for *CYP2D6*. All possible combinations of polymorphisms in *CYP2D6* were evaluated for their ability to correctly classify and predict individual metabolizer status from the provided multilocus genotypes. The prediction accuracy of each combination was estimated through cross-validation procedures, and the most parsimonious combination of polymorphisms showing the highest prediction accuracy was selected in each sample. Our results showed that the number of polymorphisms required to predict CYP2D6 metabolic phenotype with a high predictive accuracy can be dramatically reduced owing to the strong haplotype block structure observed at *CYP2D6*. ANN and MDR provided nearly identical results and performed better than the tree-based methods. For almost all samples, the ANN and MDR methods enabled a two-third reduction in the number of markers, for a prediction accuracy still above 99% (Fig. 13.4). The most informative polymorphisms for phenotype prediction appeared to differ across samples of different ethnic origin. Nonetheless, a certain agreement among

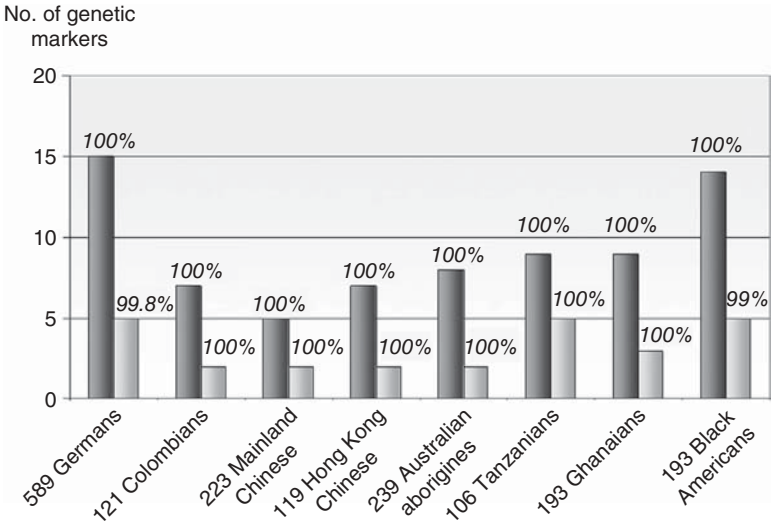


Figure 13.4 Results provided by the MDR method when applied to the eight *CYP2D6* data sets [32]. MDR was used to evaluate the ability of each combination of markers to discriminate poor and intermediate metabolizers versus rapid and ultrarapid ones. The best model retained was given by the most parsimonious combination of markers showing the highest prediction accuracy. Bars in dark grey indicate the initial number of polymorphic markers considered in each sample, and bars in light grey indicate the number of markers included in the selected model. Percentages in italics indicate the prediction accuracy achieved by the combination of all polymorphic markers considered in a sample (bars in dark grey) or the prediction accuracy achieved by the combination of markers included in the selected model (bars in light grey). Prediction accuracy was defined as the ratio between the number of individuals correctly classified and the total number of classified individuals in a sample, and was estimated by using a fivefold cross-validation procedure. All selected models were significant at the 0.001 level.

populations of common ancestry was noted. Therefore, data mining methods appear as promising tools to improve the efficiency of genotyping tests in pharmacogenomics with the ultimate goal of prescreening patients for individual therapy selection with minimum genotyping effort.

13.5.2 Development of Pharmacogenomic Classifiers from Gene Expression Data

Another area where predictive data mining has been applied is the analysis of gene expression data. Such data can be measured with DNA microarrays, which offer a powerful and effective technology for studying the expression patterns of thousands of distinct genes simultaneously, or with techniques that

rely on a real-time polymerase chain reaction (RT-PCR) when the expression of only a few genes needs to be measured with greater precision. Gene expression predictive classifiers of response to treatment are generated by correlating gene expression data, derived from biopsies taken before preoperative systemic therapy, with clinical and/or pathological response to the given treatment. This strategy has already been applied successfully in several studies.

In a recent study, Asselah et al. [49] examined the liver gene expression profiles of CHC patients receiving pegylated interferon plus ribavirin with the aim of identifying a liver gene signature that would be able to predict sustained virological response prior to drug therapy. They indeed hypothesized that NRs and sustained virological responders (SVRs) might have different liver gene expression patterns prior to treatment. A total of 58 genes associated with liver gene expression dysregulation during CHC were selected from the literature. Quantitative RT-PCR assays were used to analyze the mRNA expression of these 58 selected genes in liver biopsy specimens taken from the patients before treatment. Prediction models were then built using a supervised learning classifier, the *k*-nearest neighbor algorithm, for gene signature discovery. A gene signature was first built on a training set of 40 patients with CHC including 14 NRs and 26 SVRs, and it was then validated on an independent validation set of 29 patients including 9 NRs and 20 SVRs. The *k*-nearest neighbor algorithm identified a two-gene classifier (including *IFI27* and *CXCL9*), which accurately predicted treatment response in 79.3% (23/29) of patients from the validation set, with a predictive accuracy of 100% (9/9) and of 70% (14/20) in NRs and SVRs, respectively. Hence, the results of this study demonstrated that NRs and SVRs have different liver gene expression profiles before treatment and that treatment response can be predicted with a two-gene signature. Moreover, since the two genes included in the signature encode molecules secreted in the serum, it may provide a logical functional approach for the development of serum markers to predict treatment response.

It is worth noting that the majority of studies that have attempted to define a gene expression signature predictive of a treatment outcome using data mining tools are related to the field of oncology, where there is a strong need for defining individualized therapeutic strategies. Most anticancer drugs are indeed characterized by a very narrow therapeutic index and severe consequences of over- or underdosing in the form of, respectively, life-threatening ADRs or increased risk of treatment failure. Such studies look for somatic predictors of drug response by examining gene expression profiles within tumors. For example, Heuser et al. [50] investigated whether resistance to chemotherapy in acute myeloid leukemia (AML) could be represented by gene expression profiles, and which genes are associated with resistance. In AML, resistance to induction chemotherapy indeed occurs in 20–50% of patients. Accurate prediction of a patient's individual risk is thus required to determine the appropriate treatment. In order to identify genes predictive of *in vivo* drug resistance, Heuser et al. [50] used cDNA microarrays containing ~41,000 features to compare the gene expression profile of AML blasts

between 22 patients with good response and 11 patients with poor response to induction chemotherapy. These 33 patients were used as a training set and 104 patients with newly diagnosed AML, completely independent of the above mentioned 33 patients, were used as a test set for validation. Supervised prediction analysis was performed with the method of nearest shrunken centroids, a clustering-based method [51] for the top differentially expressed genes between good and poor responders. Prediction analysis using 10-fold cross validation revealed that response to induction chemotherapy could be predicted with an accuracy of 80%. Moreover, when applied to the independent test set, the treatment response signature divided samples into two subgroups with significantly inferior response rate (43.5% versus 66.7%, $p = 0.04$), significantly shorter event-free and overall survival ($p = 0.01$ and $p = 0.03$, respectively) in the poor-response compared to in the good-response signature group. These data indicate that resistance to chemotherapy is at least, in part, an intrinsic feature of AML blasts and can be evaluated by gene expression profiling prior to treatment.

13.6 CONCLUSION

Predictability testing is one of the main aims of pharmacogenomics. The issue of selecting the most informative genetic markers for the prediction of a treatment outcome is therefore of high clinical relevance and requires appropriate search methods due to the increased dimensionality associated with looking at multiple genotypes. Data mining approaches appear as promising tools for finding such predictive marker combinations and should facilitate the design of cost-effective and accurate genetics-based predictive assays. With gene–gene interactions playing an important role in individual drug response and with the increasing availability of genome-wide SNP data, the logical next step is a genome-wide, gene–gene interaction analysis. Yet, the data mining tools that could consider hundreds of thousands of SNPs and gene expression profiles of thousands of patients do not exist yet. A major challenge to computer scientists is therefore to make these tools available and to design efficient heuristics to surpass the prohibitively complex exhaustive search for gene interactions.

REFERENCES

1. Giacomini KM, Krauss RM, Roden DM, Eichelbaum M, Hayden MR, Nakamura Y. When good drugs go bad. *Nature* 2007;446:975–977.
2. Severino G, Del Zompo M. Adverse drug reactions: Role of pharmacogenomics. *Pharmacol Res* 2004;49:363–373.
3. Wilke RA, Lin DW, Roden DM, Watkins PB, Flockhart D, Zineh I, Giacomini KM, Krauss RM. Identifying genetic risk factors for serious adverse drug reactions: Current progress and challenges. *Nat Rev Drug Discov* 2007;6:904–916.

4. Wilke RA, Reif DM, Moore JH. Combinatorial pharmacogenetics. *Nat Rev Drug Discov* 2005;4:911–918.
5. Tufts Center for the Study of Drug Development. Drug safety withdrawals in the US not linked to speed of FDA approval. *Tufts CSDD Impact Report* 2005;7(5): 1–4.
6. CDER 2000 report to the nation: Improving public health through human drugs. Center for Drug Evaluation and Research, Food and Drug Administration, 2000.
7. Hughes HB, Biehl JP, Jones AP, Schmidt LH. Metabolism of isoniazid in man as related to the occurrence of peripheral neuritis. *Am Rev Tuberc* 1954;70:266–273.
8. Alving AS, Carson PE, Flanagan CL, Ickes CE. Enzymatic deficiency in primaquine-sensitive erythrocytes. *Science* 1956;124:484–485.
9. Kalow W. Familial incidence of low pseudocholinesterase level. *Lancet* 1956;II: 576–577.
10. Goldstein DB, Tate SK, Sisodiya SM. Pharmacogenetics goes genomic. *Nat Rev Genet* 2003;4:937–947.
11. Evans WE, Johnson JA. Pharmacogenomics: The inherited basis for interindividual differences in drug response. *Annu Rev Genomics Hum Genet* 2001;2: 9–39.
12. Andersson T, Flockhart DA, Goldstein DB, Huang SM, Kroetz DL, Milos PM, Ratain MJ, Thummel K. Drug-metabolizing enzymes: Evidence for clinical utility of pharmacogenomic tests. *Clin Pharmacol Ther* 2005;78:559–581.
13. Weinshilboum RM, Wang L. Pharmacogenetics and pharmacogenomics: Development, science, and translation. *Annu Rev Genomics Hum Genet* 2006;7: 223–245.
14. Bertilsson L, Lou YQ, Du YL, Liu Y. Pronounced differences between native Chinese and Swedish populations in the polymorphic hydroxylations of debrisoquin and S-mephenytoin. *Clin Pharmacol Ther* 1992;51:388–397
15. Bertilsson L, Lou YQ, Du YL, Liu Y. Pronounced differences between native Chinese and Swedish populations in the polymorphic hydroxylations of debrisoquin and S-mephenytoin. *Clin Pharmacol Ther* 1994;55:648.
16. Wang L, Weinshilboum RM. Thiopurine S-methyltransferase pharmacogenetics: Insight, challenges and future directions. *Oncogene* 2006;25:1629–3168.
17. Anderson GD. Pharmacogenetics and enzyme induction/inhibition properties of antiepileptic drugs. *Neurology* 2004;63:S3–S8.
18. Kerb R, Aynacioglu AS, Brockmöller J, Schlagenhaufer R, Bauer S, Szekeres T, Hamwi A, Fritzer-Szekeres M, Baumgartner C, Ongen HZ, Güzelbey P, Roots I, Brinkmann U. The predictive value of MDR1, CYP2C9, and CYP2C19 polymorphisms for phenytoin plasma levels. *Pharmacogenomics J* 2001;1:204–210.
19. Hung CC, Lin CJ, Chen CC, Chang CJ, Liou HH. Dosage recommendation of phenytoin for patients with epilepsy with different CYP2C9/CYP2C19 polymorphisms. *Ther Drug Monit* 2004;26:534–540.
20. Leschziner GD, Andrew T, Pirmohamed M, Johnson MR. ABCB1 genotype and PGP expression, function and therapeutic drug response: A critical review and recommendations for future research. *Pharmacogenomics J* 2007;7:154–179.

21. Wadelius M, Pirmohamed M. Pharmacogenetics of warfarin: Current status and future challenges. *Pharmacogenomics J* 2007;7:99–111.
22. Arranz M, Munro J, Birkett J, Bollonna A, Mancama D, Sodhi M, Lesch KP, Meyer JF, Sham P, Collier DA, Murray RM, Kerwin RW. Pharmacogenetic prediction of clozapine response. *Lancet* 2000;355:1615–1616.
23. Moore JH, Lamb JM, Brown NJ, Vaughan DE. A comparison of combinatorial partitioning and linear regression for the detection of epistatic effects of the ACE I/D and PAI-1 4G/5G polymorphisms on plasma PAI-1 levels. *Clin Genet* 2002;62:74–79.
24. Schaid DJ, Olson JM, Gauderman WJ, Elston RC. Regression models for linkage: Issues of traits, covariates, heterogeneity, and interaction. *Hum Hered* 2003;55: 86–96.
25. Bellman R. *Adaptive Control Processes: A Guided Tour*. Princeton: Princeton University Press, 1961.
26. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49:1373–1379.
27. Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet* 2003;4:701–709.
28. Heidema AG, Boer JMA, Nagelkerke N, Mariman ECM, Van der A DL, Feskens EJM. The challenge for genetic epidemiologists: How to analyse large numbers of SNPs in relation to complex diseases. *BMC Genet* 2006;7:23.
29. Motsinger AA, Ritchie MD, Reif DM. Novel methods for detecting epistasis in pharmacogenomics studies. *Pharmacogenomics* 2007;8:1229–1241.
30. Zhang H, Bonney G. Use of classification trees for association studies. *Genet Epidemiol* 2000;19:323–332.
31. Sabbagh A, Darlu P. SNP selection at the NAT2 locus for an accurate prediction of the acetylation phenotype. *Genet Med* 2006;8:76–85.
32. Sabbagh A, Darlu P. Data-mining methods as useful tools for predicting individual drug response: Application to CYP2D6 data. *Hum Hered* 2006;62:119–134.
33. Breiman L. Random forests. *Machine Learning* 2001;45:5–32.
34. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: Exploiting interactions using random forests. *BMC Genet* 2004;5:32.
35. Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. *Mach Learn* 2001;45:5–32.
36. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. *American Journal of Human Genetics* 2001;69:138–147.
37. Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet Epidemiol* 2007;31:306–315.
38. Lou XY, Chen GB, Yan L, Ma JZ, Elston RC, Li MD. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am J Hum Genet* 2007;80:1125–1137.

39. Bush WS, Dudek SM, Ritchie MD. Parallel multifactor dimensionality reduction: A tool for the large-scale analysis of gene-gene interactions. *Bioinformatics* 2006;22:2173–2174.
40. Nelson MR, Kardia SLR, Ferrell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 2001;11:458–470.
41. Culverhouse R, Klein T, Shannon W. Detecting epistatic interactions contributing to quantitative traits. *Genet Epidemiol* 2004;27:141–152.
42. Tahri-Daizadeh N, Trgouet DA, Nicaud V, Manuel N, Cambien F, Tiret L. Automated detection of informative combined effects in genetic association studies of complex traits. *Genome Res* 2003;13:1852–1860.
43. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr* 1974;19:716–723.
44. North BV, Curtis D, Cassell PG, Hitman GA, Sham PC. Assessing optimal neural network architecture for identifying disease-associated multi-marker genotypes using permutation test, and application to calpain 10 polymorphisms associated with diabetes. *Ann Hum Genet* 2003;67:348–356.
45. Serretti A, Smeraldi E. Neural network analysis in pharmacogenetics of mood disorders. *BMC Med Genet* 2004;5:27.
46. Hamilton M. Development of a rating scale for primary depressive illness. *Brit J Soc Clin Psychol* 1967;6:278–296.
47. Lin E, Hwang Y, Chen EY. Gene-gene and gene-environment interactions in interferon therapy for chronic hepatitis C. *Pharmacogenomics* 2007;8:1327–1335.
48. Warren LL, Hughes AR, Lai EH, Zaykin DV, Haneline SA, Bansal AT, Wooster AW, Spreen WR, Hernandez JE, Scott TR, Roses AD, Mosteller M. CNA30027 and CNA30032 study teams. Use of pairwise marker combination and recursive partitioning in a pharmacogenetic genome-wide scan. *Pharmacogenomics J* 2007;7:180–189.
49. Asselah T, Bieche I, Narguet S, Sabbagh A, Laurendeau I, Ripault MP, Boyer N, Martinot-Peignoux M, Valla D, Vidaud M, Marcellin P. Liver gene expression signature to predict response to pegylated interferon plus ribavirin combination therapy in patients with chronic hepatitis C. *Gut* 2008;57:516–524.
50. Heuser M, Wingen L, Steinemann D, Cario G, von Neuhoff N, Tauscher M, Bullinger L, Krauter J, Heil G, Döhner H, Schlegelberger B, Ganser A. Gene-expression profiles and their association with drug resistance in adult acute myeloid leukemia. *Haematologica* 2005;90:1484–1492.
51. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA* 2002;99: 6567–6572.

14

DATA MINING METHODS IN PHARMACEUTICAL FORMULATION

RAYMOND C. ROWE AND ELIZABETH A. COLBOURN

Table of Contents

14.1 INTRODUCTION

Before a new drug can be released in the market, it needs to be formulated to produce a quality product that is acceptable to both regulatory bodies and patients and that can be manufactured on a large scale. There are many

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

formulation types depending on the route of administration of the active drug:

- *Capsules.* These are primarily intended for oral administration and are solid preparations with hard or soft shells composed of gelatine or hydroxypropyl methyl cellulose and small amounts of other ingredients such as plasticizers, fillers, and coloring agents. Their contents may be powders, granules, pellets, liquids, or pastes.
- *Oral liquids.* These consist of solutions, suspensions, or emulsions of one or more active ingredients mixed with preservatives, antioxidants, dispersing agents, suspending agents, thickeners, emulsifiers, solubilizers, wetting agents, colors, and flavors in a suitable vehicle, generally water. They may be supplied ready for use or may be prepared before use from a concentrate or from granules or powders by the addition of water.
- *Tablets.* These are solid preparations each containing a single dose of one or more active drugs mixed with a filler/diluent, a disintegrant, a binder, a lubricant, and other ingredients such as colors., flavors, surfactants, and glidants. Tablets are prepared by compacting powders or granules in a punch and die and can exist in a variety of shapes and sizes. Tablets can also be formulated using a variety of polymers to provide a range of drug release profiles from rapid release over minutes to prolonged release over many hours. Tablets may also be coated either with sugar or with polymer films. The latter may be applied to enhance identification, in which case colored pigments may be added to increase stability, in which case opacifying agents may be added, or to provide varying release profiles throughout the gastrointestinal tract.
- *Parenterals.* These are sterile preparations intended for administration by injection, infusion, or implantation. Injections are sterile solutions, emulsions, or suspensions comprising the active drug together with suitable pH adjusters, tonicity adjusters, solubilizers, antioxidants, chelating agents, and preservatives in an appropriate vehicle, water or oil based. If there are stability issues, the formulation may be prepared as a freeze-dried sterile powder to which the appropriate sterile vehicle is added prior to administration. Infusions are sterile aqueous solutions or emulsions intended for administration in large volumes. Implants are sterile solid preparations designed to release their active drug over an extended period of time.
- *Topicals.* These are semisolid preparations such as creams, ointments, or gels intended to be applied to the skin or to certain mucous membranes for local action. They may be single or multiphase, comprising one or more active drugs mixed with emulsifiers, oils, soaps, gelling agents, or waxes with a continuous phase of either water or oil.

- *Eye preparations.* These are specifically intended for administration to the eye in the form of solutions, lotions, or ointments. All preparations must be sterile.
- *Suppositories and pessaries.* These are preparations intended for either rectal or vaginal administration of drugs. They are formulated using a suitable base that melts at body temperature.
- *Inhalation preparations.* These can be solutions, suspensions, or powders intended to be inhaled as aerosols for administration to the lung.

The development of a commercial product is a time-consuming and complicated process, as the design space is multidimensional and virtually impossible to conceptualize. It requires the optimization of both the formulation and the manufacturing process to produce a product with the required properties since these are determined not only by the ratios in which the ingredients are combined but also by the processing conditions used. Although relationships between ingredient levels, processing conditions, and product performance may be known anecdotally, rarely can they be quantified, and hence formulation is often undertaken as an iterative process. Generally, one or more drugs are mixed with various ingredients (excipients) and, as development progresses, the choice of these excipients and their levels as well as the manufacturing process are changed and optimized as a result of intensive, time-consuming experimentation. This in turn results in the generation of large amounts of data, the processing of which is challenging.

Traditionally, formulators have tended to use statistical techniques such as a response surface methodology to investigate the design space, but optimization by this method can be misleading especially if the formulation is complex. Recent advances in mathematics and computer science have resulted in the development of other data mining techniques that can be used to remedy the situation—neural networks (for modeling the design space), genetic algorithms (for optimizing the formulation and manufacturing process), and neurofuzzy logic and decision trees (for exploring the relationships within the design space and for generating understandable rules that can be used in future work). This chapter reviews the current situation and provides some worked examples to illustrate the concept.

14.2 METHODOLOGY

Modeling the design space in formulation is an ideal application for neural networks.

Neural networks, like humans, learn directly from input data. The learning algorithms take two main forms. Unsupervised learning, where the network is presented with input data and learns to recognize patterns in the data, is useful for organizing amounts of data into a smaller number of clusters. In supervised learning, which is analogous to "teaching" the network, the network

is presented with a series of matching input and output examples, and it learns the relationships connecting the inputs to the outputs. Supervised learning has proved most useful for formulation, where the goal is to determine cause-and-effect links between inputs (ingredients and processing conditions) and outputs (measured properties).

The basic component of a neural network is the neuron, a simple mathematical processing unit that takes one or more inputs and produces an output. For each neuron, every input has an associated weight that defines its relative importance in the network, and the neuron simply computes the weighted sum of all the outputs and calculates an output. This is then modified by means of a transformation function (sometimes called a transfer or activation function) before being forwarded to another neuron. This simple processing unit is known as a perceptron, a feedforward system in which the transfer of data is in the forward direction, from inputs to outputs only.

A neural network consists of many neurons organized into a structure called the network architecture. Although there are many possible network architectures, one of the most popular and successful is the multilayer perceptron (MLP) network. This consists of identical neurons all interconnected and organized in layers, with those in one layer connected to those in the next layer so that the outputs in one layer become the inputs in the subsequent layer. Data flow into the network via the input layer, pass through one or more hidden layers, and finally exit via the output layer, as shown in Figure 14.1. In theory, any number of hidden layers may be added, but in practice, multiple layers are necessary only for those applications with extensive nonlinear behavior, and they result in extended computation time. It is generally accepted that the performance of a well-designed MLP model is comparable with that achieved by classical statistical techniques.

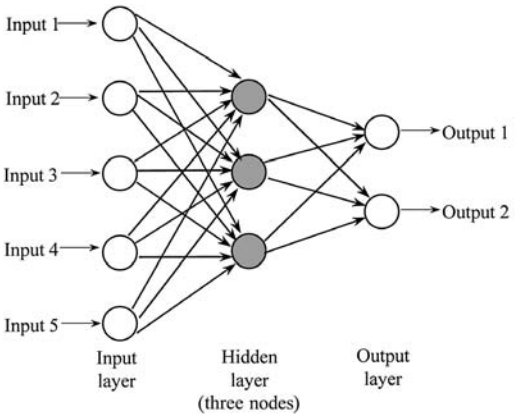


Figure 14.1 Diagram of a multilayer perceptron neural network with one hidden layer.

Unlike conventional computer programs, which are explicitly programmed, supervised neural networks are "trained" with previous examples. The network is presented with example data, and the weights of inputs feeding into each neuron are adjusted iteratively until the output for a specific network is close to the desired output. The method used to adjust the weights is generally called backpropagation, because the size of the error is fed back into the calculation for the weight changes. There are a number of possible backpropagation algorithms, most with adjustable parameters designed to increase the rate and degree of convergence between the calculated and desired (actual) outputs. Although training can be a relatively slow process especially if there are large amounts of data, once trained, neural networks are inherently fast in execution.

Genetic algorithms are an ideal optimization technique based on the concepts of biological evolution. Like the biological equivalent, genetic algorithms require a definition of "fitness," which is assessed according to how well the solution meets user-specified goals. Genetic algorithms work with a population of individuals, each of which is a candidate solution to the problem. Each individual's fitness is assessed and unless an optimum solution is found, a further generation of possible solutions is produced by combining large chunks of the fittest initial solutions by a crossover operation (mimicking mating and reproduction). As in biological evolution, the population will evolve slowly, and only the fittest (i.e., best) solutions survive and are carried forward. Ultimately, after many generations, an optimum solution will be found.

Genetic algorithms are especially useful for complex multidimensional problems with local minima as well as the global minimum. Unlike conventional, more directed searches (like steepest descent and conjugate gradient methods), they are capable of finding the global minimum reliably. Effective use of genetic algorithms requires rapid feedback of the success or failure of the possible solutions, as judged by the fitness criteria. Hence, the combination of a genetic algorithm with a neural network is ideal. Such a combination (illustrated in Fig. 14.2) is used in INForm, a software package available from Intelligensys Ltd., UK, in which formulations can be modeled using a neural network and then optimized using genetic algorithms.

In defining the concept of fitness, it is possible to give each property a different degree of importance in the optimization, using a weighting factor. This allows conflicting objectives to be examined and, combined with fuzzy logic, provides a useful framework for describing complex formulation goals. Fuzzy logic, as the name implies, blurs the clear-cut "true" and "false" of conventional "crisp" logic by assigning a noninteger number that describes the "membership" in a particular set as somewhere between 0 (false) and 1 (true). Therefore, in additional to the "black and white" of conventional logic, fuzzy logic allows "shades of gray" to be described intuitively and accurately. So, if the formulator is seeking a tablet with a disintegration time of less than 300 seconds, one with a disintegration time of, say, 310 seconds will not be rejected

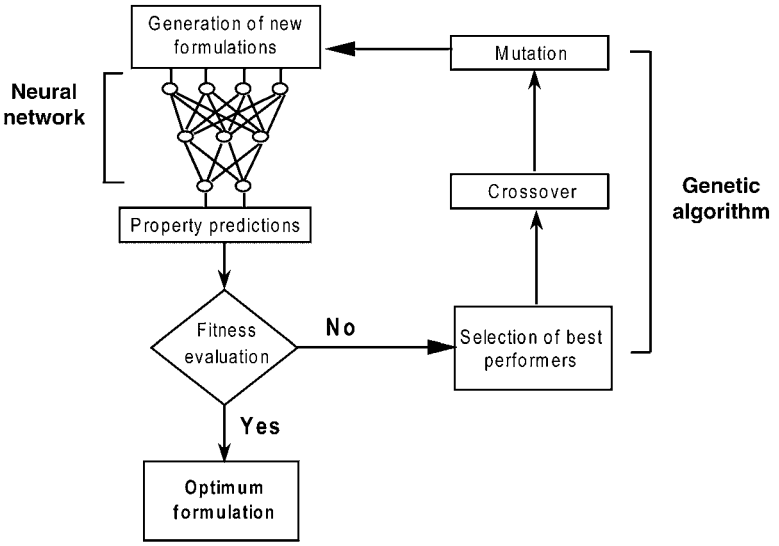


Figure 14.2 Diagram of a genetic algorithm linked to a neural network for modeling and optimization.

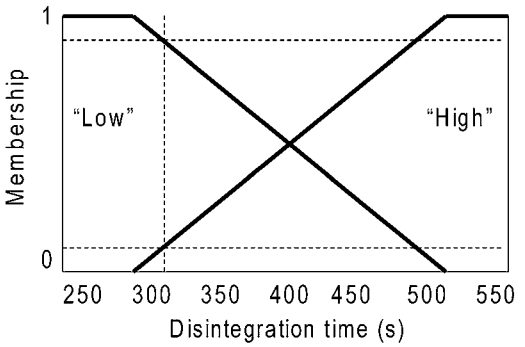


Figure 14.3 Fuzzy logic representation of the disintegration time of a tablet as low or high.

out of hand but will be assigned a desirability of somewhat less than 100% (as shown in Fig. 14.3) according to its membership in the low set.

More recently, coupling fuzzy logic with neural networks has led to the development of neurofuzzy computing, a novel technology that combines the ability of neural networks to learn directly from data, with fuzzy logic's capacity to express the results clearly in linguistic form. Essentially, the neurofuzzy architecture is a neural network with two additional layers for fuzzification and defuzzification. This has led to a powerful new modeling capability that not only develops models that express the key cause-and-effect relationships within a formulation data set but also allows these to be expressed as simple

actionable rules in the form IF (ingredient) … THEN (property), with an associated "confidence level." Neurofuzzy computing underpins FormRules, a software package from Intelligensys Ltd., UK that allows rules to be extracted directly from formulation data.

Whereas neurofuzzy logic is ideally suited to the case where the measured properties take nonintegral numerical values, they do not work well when the properties are "classifications," e.g., a simple pass/fail criterion. In the case where the properties lie within specific discrete classes, decision trees are more effective in encapsulating the information buried in the data. A number of decision tree algorithms have been proposed over the past few years; some deal effectively with numerical inputs, while others are designed more specifically to treat the case where the inputs, as well as the properties, lie in defined classes. Especially powerful algorithms that have been used successfully are ID3 and its successors C4.5 and C5, developed by Quinlan [1,2]. The C4.5 and C5 algorithms are capable of dealing with numerical as well as "classified" inputs.

The ID3 and C4.5/C5 algorithms are based on the concept of information entropy, and both use a training data set of previously classified examples from which to "learn." The aim is to split the data set in a way that gives the maximum information gain (difference in entropy). Subsequent splits are made on subgroups arising from the initial split, producing a treelike structure. Simpler decision trees are preferred over more complex ones, with tree "pruning" used in C4.5 and C5 to remove branches that do not provide useful information.

14.3 APPLICATIONS

The past decade has seen a dramatic increase in the number of reported applications of data mining tools being used in pharmaceutical formulation [3–7]. Applications now cover a variety of formulations—for example, immediate and controlled release tablets, skin creams, hydrogel ointments, liposomes and emulsions, and film coatings. The following examples are by no means exhaustive, but show where data mining tools have been used successfully.

14.3.1 Tablet Formulations (Immediate Release)

Two papers in the mid-1990s reported the earliest data mining studies on immediate release tablets. In the first, tablet formulations of hydrochorothiazide [8] were modeled using neural networks in an attempt to maximize tablet strength and to select the best lubricant. In the other, a tablet formulation of caffeine was modeled [9] in order to relate both formulation and processing variables with granule and tablet properties.

Both these studies were successful in demonstrating that neural networks performed better than conventional statistical methods. In a later paper [10],

the data from the caffeine tablet formulation were subsequently reanalyzed using a combination of neural networks and genetic algorithms. This study showed that the optimum formulation depended on the relative importance placed on the output properties and on constraints applied both to the levels of the ingredients used in the formulation and to the processing variables. Many "optimum formulations" could be produced, depending on the trade-offs that could be accepted for different aspects of product performance. In a more recent paper [11], the same data have been studied using neurofuzzy computing. Useful rules relating the disintegration time to both formulation and processing variables were automatically generated.

In a series of papers, personnel from Novartis and the University of Basel in Switzerland have highlighted the pros and cons of neural networks for modeling immediate release tablets [12–15]. In other studies, neural networks have been found useful in modeling tablet formulations of antacids [16], plant extracts [17], theophylline [18], and diltiazem [19]. In a recent paper, Lindberg and Colbourn [20] have used neural networks, genetic algorithms, and neurofuzzy to successfully analyze historical data from three different immediate release tablet formulations. Using a data set published in the literature [13], Shao et al. [21,22] have critically compared the three data mining technologies of neural networks, neurofuzzy logic, and decision trees in both data modeling and rule generation. As expected, each has its own strengths and weaknesses.

Application of the technology is not limited to the tableting process alone. Pigmented film coating formulations have recently been modeled and optimized to enhance opacity and to reduce film cracking using neural networks combined with genetic algorithms [23,24] as well as being studied using neurofuzzy techniques [25]. In the latter investigation, the rules discovered were consistent with known theory.

14.3.2 Tablet Formulations (Controlled Release)

In this domain, the first studies were carried out in the early 1990s by Hussain and coworkers at the University of Cincinnati [26]. They modeled the *in vitro* release characteristics of a number of drugs from matrices consisting of a variety of hydrophilic polymers and found that in the majority of cases, neural networks with a single hidden layer had a reasonable performance in predicting drug release profiles. Later studies using similar formulations [27] have confirmed these findings as have recent studies in Japan [28].

Neural networks have also been used in Slovenia to model the release characteristics of diclofenac [29], in China to study the release of nifedipine and nomodipine [30], and in Yugoslavia to model the release of aspirin [31]. More recently, work in this area has been extended to model osmotic pumps in China [32] and enteric coated tablets in Ireland [33].

14.3.3 Topical Formulations

Topical formulations by their very nature are usually multicomponent, and it is not surprising that neural networks have been applied to deal with this complexity. The first work was performed on hydrogel formulations containing anti-inflammatory drugs in Japan in 1997 [34], followed up by further studies in 1999 [35] and in 2001 [36]. Lipophilic semisolid emulsion systems have been studied in Slovenia [37,38], and transdermal delivery formulations of melatonin in Florida [39]. In all cases, the superiority of neural networks over conventional statistics has been reported.

14.3.4 Other Formulations

Neural networks have been applied to the modeling of pellet formulations to control the release of theophylline [40] and to control the rate of degradation of omeprazole [41]. They have also been applied to the preparation of acrylic microspheres [42] and to model the release of insulin from an implant [43]. In a recent study from Brazil, the release of hydrocortisone from a biodegradable matrix has been successfully modeled [44]. Recent work has focused on the modeling of estradiol release from membranes [45] and the formulation of solid dispersions [46].

14.4 WORKED EXAMPLES

To illustrate what sorts of information can be extracted from various formulation data sets, four different studies have been undertaken using commercially available software. The packages employed were INForm and FormRules. INForm uses MLP neural networks to model the data and incorporates a range of backpropagation algorithms. Additionally, it integrates a genetic algorithm approach to optimization. FormRules is based on neurofuzzy logic, using the adaptive spline modeling of data (ASMOD) [47] algorithm. In this approach, models of varying complexity are developed, and a model assessment criterion is used to select the simplest model that best represents the data. Several different model assessment criteria are used in FormRules; the most common are structural risk minimization (SRM) and minimum descriptor length (MDL). The output from FormRules is a linguistic rule of the form IF … AND … THEN, with a "confidence level" analogous to the membership function, which is defined relative to the maximum and minimum values that the property can take.

14.4.1 Controlled Release Tablets

In the first of these studies, a controlled release tablet formulation is investigated, using data discussed in a paper by Chen et al. [18]. Their tablet formula-

tion uses two polymers, together with dextrose and a lubricant, in varying amounts. The amount of drug was held constant. In addition to the four ingredients, three other variables (tablet hardness, percent moisture, and particle size) more related to processing aspects were also varied. Twenty-two unique formulations were made, and the *in vitro* amount of drug released at various time intervals, from 1 to 24 hours, was measured for each. At first sight, it may seem surprising that good information can be extracted when there are seven input variables and only 22 formulations; however, using neurofuzzy logic, useful knowledge can be gained especially at the shorter release times where the release measurements are more accurate.

Using neurofuzzy techniques, separate models were developed for each specific release time. Analysis of variance (ANOVA) statistics were used to assess the quality of the models, and these showed that very good models could be found for short and intermediate release times. Only relatively poor models (ANOVA R^2 value less than 0.7) could be found for the longest release times, and a closer examination of the data revealed considerable scatter, reflecting the difficulties of making accurate measurements at these times. Nevertheless, the models were sufficiently reliable for information to be extracted from them for release times of up to 16 hours.

The neurofuzzy data mining exercise shows clearly that just one of the polymers, described as Polymer A by Chen et al., dominates the short-term (1–2 hours) release, and that when the amount of this polymer is high, then the amount of drug released is low. At long times (above 10 hours), the amount of Polymer B controls the amount of drug released, with release being lowest when the amount of Polymer B is high. At intermediate times, both Polymers A and B control the amount of drug released.

Detailed examination of the rules shows that the amount of dextrose has no significant effect on the amount of drug released at any given time. Particle size and tablet hardness also have a negligible effect, while the moisture percentage and the amount of lubricant have a minor effect on the release at intermediate times. Indeed, the data mining study highlighted an interaction between the amount of Polymer A and the lubricant on the 8-hour release; this is shown in the full rule set for the 8-hour release given in Table 14.1,

TABLE 14.1 Rules Extracted from Data for Amount of Drug Released after 8 Hours

1. IF Polymer B is LOW, then 8-hour release is HIGH (0.91).
2. IF Polymer B is HIGH, then 8-hour release is LOW (1.00).
3. IF Polymer A is LOW AND lubricant is LOW, then 8-hour release is HIGH (1.00).
4. IF Polymer A is LOW AND lubricant is HIGH, then 8-hour release is LOW (1.00).
5. IF Polymer A is HIGH AND lubricant is LOW, then 8-hour release is LOW (1.00).
6. IF Polymer A is HIGH AND lubricant is HIGH, then 8-hour release is LOW (0.77).
7. IF % moisture is LOW, then 8-hour release is HIGH (0.70).
8. IF % moisture is HIGH, then 8-hour release is LOW (1.00).

where some of the rules, those describing the interactions, are of the form IF … AND … THEN. Rules 3 and 4 in Table 14.1 show that the lubricant affects the release significantly when the amount of Polymer A is low; its influence when the amount of Polymer A is high is less marked, affecting only the confidence levels for the rules. In effect, rules 5 and 6 say that when the amounts of Polymer A and lubricant are both high, then the amount of drug released at 8 hours is not as low as when the amount of Polymer A is high, but the amount of lubricant is low.

By using the information that only Polymers A and B, the lubricant, and the percentage of water are important in controlling the release, more conventional neural network models can be generated and used in conjunction with optimization methods (in the present case, genetic algorithms) to generate formulations that give a specific desired release profile.

14.4.2 Immediate Release Tablets

In the second worked example, data published by Kesavan and Peck [9] have been analyzed. Their tablet formulation consisted of

- anhydrous caffeine (40% w/w) as a model active drug,
- dicalcium phosphate dihydrate (Ditab) or lactose (44.5–47.5% w/w) as a diluent,
- polyvinylpyrrolidone (PVP) (2.0–5.0% w/w) as a binder,
- corn starch (10% w/w) as a disintegrant, and
- magnesium stearate (0.5% w/w) as a lubricant.

Two types of granulation equipment—fluidized bed and high shear mixing—were used, and the binder was added either wet or dry. Thirty-two different experiments were available.

This data set is interesting for data mining because three of the input variables (the diluent type, the type of granulation equipment, and the binder addition) are "classified" rather than numerical values. However, all properties took numerical values, so the data set is not ideally suited to treatment using decision trees. Therefore, a neurofuzzy treatment was helpful in identifying the key relationships in the data. Models of varying complexity can be developed by changing the model selection criterion, and the work reported here used the strictest criterion (SRM), which gives the simplest models.

The four tablet properties that were measured were hardness, friability, thickness, and disintegration time, and separate models were evolved for each. Tablet hardness depended most strongly on the two process conditions, i.e., the method of granulation and on whether the binder was added dry or in solution. There was a lesser dependence on the selection of diluent, with lactose leading to softer tablets, in line with the expectations of experienced formulators. The rules for tablet hardness are shown in Table 14.2.

TABLE 14.2 Rules Governing Hardness Discovered with Data Mining

Rule Set 1			
Granulation Equipment	Binder Addition	Hardness	Confidence (%)
Fluidized bed	Dry	Low	100
Fluidized bed	Wet	Low	92
High shear mixer	Dry	High	91
High shear mixer	Wet	High	52
Rule Set 2			
Diluent	Hardness	Confidence (%)	
Lactose	Low	63	
Ditab	High	91	

TABLE 14.3 Rules Governing Disintegration Time

Rule Set 1
IF the diluent is lactose, THEN the disintegration time is LOW (0.87). IF the diluent is Ditab, THEN the disintegration time is HIGH (1.00).
Rule Set 2
IF PVP % is LOW, THEN the disintegration time is LOW (1.00). IF PVP % is MID, THEN the disintegration time is HIGH (0.70). IF PVP % is HIGH, THEN the disintegration time is HIGH (0.82).
Rule Set 3
IF the granulation equipment is a fluidized bed AND the binder addition is dry, THEN the disintegration time is LOW (0.58). IF the granulation equipment is a fluidized bed AND the binder addition is wet, THEN the disintegration time is HIGH (0.71). IF the granulation equipment is high shear mixing AND the binder addition is dry, THEN the disintegration time is LOW (0.51). IF the granulation equipment is high shear mixing AND the binder addition is wet, THEN the disintegration time is LOW (1.00).

Disintegration time, the other property of major importance, depends on all the variables except the diluent percentage. The most important variable is the diluent type, with lactose (which has been found to yield softer tablets) giving shorter disintegration times. The full rule set is given in Table 14.3.

The second rule set in Table 14.3 shows that the main effect of PVP is to lower disintegration times, when it is added in relatively small quantities. With

PVP in medium or larger quantities, the disintegration time is high. There is a complex interaction between the choice of granulation equipment and whether the binder is added dry or in solution, as shown in the third rule set. It is also worth noting the third rule of this set, which shows that when high shear mixing is used and binder addition is dry, then the disintegration time is low with a confidence of only 51%. This rule actually illustrates that the disintegration time is neither low nor high (hence confidence of about 50% that it is low) but lies in the middle of the range.

Generally, all of these rules are in line with those expected by expert formulators, with lactose giving softer tablets that disintegrate more quickly.

In their experimental work, Kesavan and Peck measured granule properties as well as properties of the finished tablet. One issue of interest in data mining is whether the granule properties are important in predicting the properties of the finished tablet or whether models can be developed to link the formulation variables directly to the tablet properties. Various options have been investigated, and it has been found that good cause-and-effect models can be discovered without involving the granule properties as inputs for the tablet property models. This is a valuable insight since it means that changes in the formulation can be linked directly to tablet properties without requiring intermediate measurements on the granules. Indeed, it allows an optimization of the tablet properties directly from the formulation. For example, if the formulator is seeking a hard tablet that disintegrates quickly (which, as the neurofuzzy data mining study shows, is difficult to achieve), then the various trade-offs can be examined.

In this application, the neural network was trained using a network architecture with a four-node hidden layer, and the goal was to look at the trade-offs between hardness and other properties (disintegration time, friability, and thickness). Various weightings of the different properties were investigated in an attempt to find hard tablets that disintegrated quickly.

These studies show clearly that tablet hardness can be achieved only by sacrificing disintegration time. Furthermore, the optimized formulation shows that the percentage of diluent lies at the top of the experimental range, while the PVP concentration lies at the bottom of the range. This suggests that the experimental range should be expanded in order to look for a better formulation.

14.4.3 Drug-Loaded Nanoparticles

The above examples have both been concerned with tablets. However, data mining as the name suggests is "data driven" and as long as there are data available, then the techniques can be applied. One new issue that formulators recently have had to face is the delivery of novel peptides and proteins now being developed by biotechnology, for example, using drug-loaded nanoparticles. Data have been published by Attivi et al. [48] on insulin-loaded

TABLE 14.4 Rules for Size of Nanoparticles

PH	PCL/RS Ratio	Size	Confidence (%)
Low	Low	Low	100
High	Low	Low	97
Medium	High	Low	96
Medium	Low	Low	94
Low	High	Low	81
High	High	High	100

nanoparticles produced by a water-in-oil-in-water emulsification in an aqueous solution of polyvinyl alcohol (PVA), followed by a drying process. The study varied only three inputs; these were the ratio of the polymers used in the nanosphere, referred to as the PCL/RS ratio (ratio of poly(epsiloncaprolactone) to Eudragit RS), the volume of the PVA aqueous solution, and the pH of the aqueous PVA solution. This restriction on the number of inputs was imposed largely because the authors were performing a statistical study, which (unlike the case for neurofuzzy data mining) can become very complex with larger numbers of variables. Eighteen unique formulations were produced, and five properties (size, polydispersity index, zeta potential, amount of entrapped insulin, and amount of insulin released after 7 hours) were measured for each.

In the neurofuzzy study, the model selection criterion was selected for each model so that good ANOVA statistics were obtained. In all cases, the model statistics from the neurofuzzy study were as good as, or slightly better than, the original statistical study of Attivi et al. [48].

Particle size was found to depend on the PCL/RS ratio and on pH; these are the same variables as were found by the statistical study. There is an interaction between these two variables, as illustrated by the rules given in Table 14.4.

Table 14.4 shows that, like the particle size, polydispersity depends on the PCL/RS ratio and on pH. Again, the same variables were found to be important here as in the statistical study.

The model for zeta potential is a more interesting case, since a good model cold not be obtained using statistics. However, the MDL model assessment criterion allowed a model to be developed, although this was quite complex as Figure 14.4 shows. Again the PCL/RS ratio and the pH are the only contributing factors.

The model for entrapped insulin depended primarily on the volume of PVA, although pH also had a role to play. The "combined rules" show that IF the volume of PVA is LOW AND IF pH is MID, THEN the entrapped insulin is HIGH. There is a maximum in the amount of entrapped insulin when pH is in the midrange, and it decreases when pH is either LOW or HIGH.

There were repeated data points in the published formulations, and these show a fair degree of scatter. Therefore, not surprisingly, using SRM as the

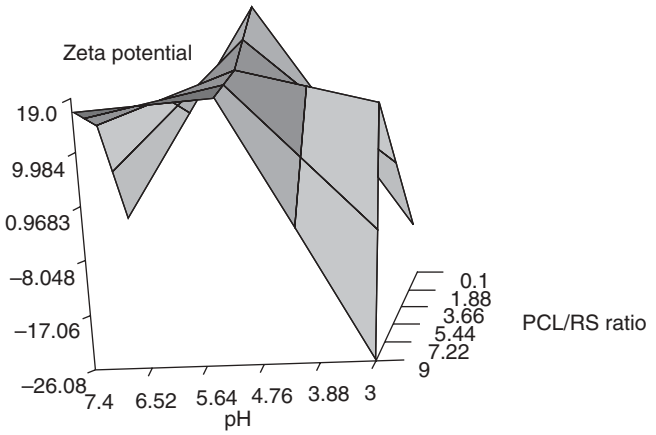


Figure 14.4 Graphical representation of model for zeta potential.

model selection criterion gave poor results for the released insulin. A reasonable model was developed using the MDL model assessment criterion, although the R^2 value was lower than that reported from the statistical study. The PCL/RS ratio was the most important variable, with the volume of PVA playing a minor role.

14.4.4 Suspensions

The final worked example concerns a redispersible suspension of rifampicin and was performed using data published by Elkheshen et al. [49]. In this study, the ingredients and the range over which they were allowed to vary (as percentages of the constituted suspension) were

- sucrose (30%, 45%, or 60%),
- Avicel (1%, 1.5%, or 2%),
- Aerosil (0%, 0.5%, or 1%), and
- aerosol (0%, 0.05%, or 0.01%).

In addition, there were other ingredients (rifampicin, sodium citrate, citric acid, sodium benzoate, and flavor) that were not varied in the experiments.

Twenty-one experiments were performed using a 2^4 factorial design with five replicates of the center point. The properties measured were bulk density, flowability of the powder, viscosity of the suspension after 24 hours, sedimentation volume as percentage of the initial volume, and percentage ease of redispersibility.

Good models (ANOVA R^2 values in excess of 0.85) could be developed for all of the properties. Bulk density was shown by FormRules to depend on

TABLE 14.5 Rules for Bulk Density from Neurofuzzy Data Mining

Sucrose Concentration	Aerosil Concentration	Bulk Density	Confidence (%)
Low	Low	High	81
Low	High	Low	72
High	Low	High	78
High	High	High	89

the percentages of sucrose and Aerosil. This is consistent with the results of Elkheshen et al. [49] from their statistical study. In addition, FormRules produced "rules" of the form given in Table 14.5.

As the rules in Table 14.5 show, there is an interaction between the amount of sucrose and the amount of Aerosil. At low levels of sucrose, Aerosil has a marked effect. This is not the case when the percentage of sucrose is at the high end of the range. In that case, Aerosil has only a small effect. Just the confidence level is affected; the conclusion is that bulk density will be high if sucrose % is high.

For flowability, FormRules showed that only the percentage of Aerosil was important. This is consistent with the statistical study—although the statistical work also suggested that sucrose might play a role. Above about 0.5%, adding more Aerosil does not significantly increase flowability.

Viscosity was more complex, depending on the amounts of sucrose, Avicel, and Aerosil. Adding sucrose increased the viscosity in a linear fashion, as did adding Avicel. Aerosil had a complex effect, decreasing viscosity at low concentrations, but increasing it at higher ones.

Sedimentation volume was affected primarily by the amount of Aerosil. If the percentage of Aerosil is low, then the percentage of sedimentation is low. Sedimentation is higher when the amount of Aerosil is increased.

Redispersibility percentage was also affected mainly by Aerosil, with an approximately linear relationship between the amount of Aerosil and the redispersibility. This is consistent with the statistical results of Elkheshen et al. More complex models could be developed by changing the model selection criterion, showing that all four inputs have some role to play; however, this leads to rules that are very complicated to interpret, thus losing some of the key benefits of data mining.

14.5 BENEFITS AND ISSUES

Although there is a great deal of interest in data mining, quantified information on the benefits has been harder to find. From the applications and worked examples discussed earlier in this chapter, benefits that can be seen include

- effective use of incomplete data sets;
- rapid analysis of data;

- ability to accommodate more data and to retrain the network (refine the model);
- effective exploration of the total design space, irrespective of complexity;
- ability to accommodate constraints and preferences; and
- ability to generate understandable rules.

Business benefits specifically for the domain of product formulation have been given as [50]

- enhancement of product quality and performance at low cost,
- shorter time to market,
- development of new products,
- improved customer response,
- improved confidence, and
- improved competitive edge.

As this new technology moves from the realm of academe into practical application, there are also issues regarding the implementation of neural computing. Early adopters found problems related to software and lack of development skills; this has been reduced as commercial packages have come into wider use, and there is less need for bespoke in-house systems with their high programming and maintenance burden. However, even when commercial packages are used, there are a number of features that should be present before data mining can be used to advantage. Reasonable quantities of reliable data should be available in order to train an adequate model, and these must encapsulate the cause-and-effect relationships within the problem. The selection of data mining technique will depend on whether the properties are numerical or "categorical," with decision trees most appropriate for the latter. The greatest benefits are achieved for multidimensional problems, where it is difficult to express any analytic model and difficult to abstract the rules by any other mechanism. It helps if the problem is of practical importance, part of the organization's essential activity, and meets a real business need. Pharmaceutical formulation meets these criteria well, and data mining can be expected to provide significant benefits in the industry in the future.

It is interesting to note that the field is not stagnant. New applications of the technologies discussed above are being developed routinely. New approaches, such as mining "fractured" data, are being evaluated [51]. New technologies are being investigated, with papers using model trees [52] and genetic programming [53] being published very recently. In addition, the knowledge generated from various data mining exercises has been integrated into decision support systems [50,54].

REFERENCES

1. Quinlan JR. *Expert Systems in the Micro Electronic Age*. Edinburgh: Edinburgh University Press, 1979.
2. Quinlan JR. *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc., 1993.
3. Achanta AS, Kowalsk JG, Rhodes CT. Artificial neural networks: Implications for pharmaceutical sciences. *Drug Dev Ind Pharm* 1995;21:119–155.
4. Bourquin J, Schmidli H, van Hoogevest P, Leuenberger H. Basic concepts of artificial neural networks (ANN) modelling in the application to pharmaceutical development. *Pharm Dev Technol* 1997;2:95–109.
5. Rowe RC, Colbourn EA. *Applications of Neural Computing in Formulation*, pp. 4–7. London: Pharmaceutical Visions Spring Edition, 2002.
6. Sun Y, Peng Y, Chen Y, Shukla AJ. Application of artificial neural networks in the design of controlled release drug delivery systems. *Adv Drug Deliv Rev* 2003;55: 1201–1215.
7. Takayama K, Fujikawa M, Obata Y, Morishita M. Neural network based optimization of drug formulations. *Adv Drug Deliv Rev* 2003;55:1217–1231.
8. Turkoglu J, Ozarslan R, Sakr A. Artificial neural network analysis of a direct compression tabletting study. *Eur J Pharm Biopharm* 1995;41:315–322.
9. Kesavan JG, Peck GE. Pharmaceutical granulation and tablet formulation using neural networks. *Pharm Dev Technol* 1996;1:391–404.
10. Colbourn EA, Rowe RC. Modelling and optimization of a tablet formulation using neural networks and genetic algorithms. *Pharm Tech Eur* 1996;8:46–55.
11. Rowe RC, Colbourn EA. Generating rules for tablet formulations. *Pharm Tech Eur* 2000;12:24–27.
12. Bourquin J, Schmidli H, van Hoogeves P, Leuenberger H. Application of artificial neural networks (ANN) in the development of solid dosage forms. *Pharm Dev Technol* 1997;2:111–121.
13. Bourquin J, Schmidli H, van Hoogevest P, Leuenberger H. Comparison of artificial neural networks (ANN) with classical modelling technologies using different experimental designs and data from a galenical study on a solid dosage form. *Eur J Pharm Sci* 1998;6:287–300.
14. Bourquin J, Schmidli H, van Hoogevest P, Leuenberger H. Advantages of artificial neural networks (ANNs) as alternative modelling technique for data sets showing non-linear relationships using data from a galenical study on a solid dosage form. *Eur J Pharm Sci* 1998;7:5–16.
15. Bourquin J, Schmidli H, van Hoogevest P, Leuenberger H. Pitfalls of artificial neural networks (ANN) modelling technique for data sets containing outlier measurements using a study of mixture properties of a direct compressed tablet dosage form. *Eur J Pharm Sci* 1998;7:17–28.
16. Do QM, Dang GV, Le NQ. Drawing up and optimizing the formulation of Malumix tablets by an artificial intelligence system (CAD/Chem). *Tap Chi Duoc Hoc* 2000;6:16–19.
17. Rocksloh K, Rapp F-R, Abu Abed S, Mueller W, Reher M, Gauglitz G, Schmidt PC. Optimization of crushing strength and disintegration time of a high dose plant extract tablet by neural networks. *Drug Dev Ind Pharm* 1999;25:1015–1025.

18. Chen U, Thosor SS, Forbess RA, Kemper MS, Rubinovitz RL, Shukla AJ. Prediction of drug content and hardness of intact tablets using artificial neural networks and near-infrared spectroscopy. *Drug Dev Ind Pharm* 2001;27:623–631.
19. Sathe PM, Venit J. Comparison of neural networks and multiple linear regression as dissolution predictors. *Drug Dev Ind Pharm* 2003;29:349–355.
20. Lindberg NO, Colbourn EA. Use of artificial neural networks and genetic algorithms—Experiences from a tablet formulation. *Pharm Tech Eur* 2004;16:35–39.
21. Shao Q, Rowe RC, York P. Comparison of neurofuzzy logic and neural networks in modeling experimental data of an immediate release tablet formulation. *Eur J Pharm Sci* 2006;28:394–404.
22. Shao Q, Rowe RC, York P. Comparison of neurofuzzy logic and decision trees in discovering knowledge from experimental data of an immediate release tablet formulation. *Eur J Pharm Sci* 2007;31:129–136.
23. Plumb AP, Rowe RC, York P, Doherty C. The effect of experimental design in the modelling of a tablet coating formulation using artificial neural networks. *Eur J Pharm Sci* 2002;16:281–288.
24. Plumb AP, Rowe RC, York P, Doherty C. Effect of varying optimization parameters in optimization by guided evolutionary simulated annealing (GESA) using a tablet film coat on an example formulation. *Eur J Pharm Sci* 2003;18:259–266.
25. Rowe RC, Woolgar CG. Neurofuzzy logic in tablet film coating formulation. *Pharm Sci Technol Today* 1999;2:495–497.
26. Hussain AS, Yu X, Johnson RD. Application of neural computing in pharmaceutical product development. *Pharm Res* 1991;8:1248–1252.
27. Hussain AS, Shivanand Y, Johnson RD. Application of neural computing in pharmaceutical product development: Computer aided formulation design. *Drug Dev Ind Pharm* 1996;20:1739–1752.
28. Takahara J, Takayama K, Nagai T. Multi-objective simultaneous optimization technique based on an artificial neural network in sustained release formulations. *J Control Release* 1997;49:11–20.
29. Zupancic Bozic D, Vrecar F, Kozjek F. Optimization of diclofenac sodium dissolution from sustained release formulations using an artificial neural network. *Eur J Pharm Sci* 1997;5:163–169.
30. Sheng H, Wang P, Tu J-S, Yuan L, Pin Q-N. Applications of artificial neural networks to the design of sustained release matrix tablets. *Chin J Pharm* 1998;29:352–354.
31. Ibric S, Jovanovic M, Djuric A, Parojcic J, Petrovic SD, Solomun L, Stupor B. Artificial neural networks in the modelling and optimization of aspirin extended release tablets with Eudragit L100 as matrix substance. *Pharm Sci Technol* 2003;4:62–70.
32. Wu T, Pao W, Chen J, Shang R. Formulation optimization technique based on artificial neural network in salbutamol sulfate osmotic pump tablets. *Drug Dev Ind Pharm* 2000;26:211–215.
33. Leane MM, Cumming I, Corrigan O. The use of artificial neural networks for the selection of the most appropriate formulation and processing variables in order to predict the in vitro dissolution of sustained release minitablets. *Pharm Sci Tech* 2003;4:218–229.

34. Takahara J, Takayama K, Isowa K, Nagai T. Multi-objective simultaneous optimization based on artificial neural network in a ketoprofen hydrogel formula containing σ-ethylmenthol as a percutaneous absorption enhancer. *Int J Pharm* 1997;158:203–210.
35. Takayama K, Takahara J, Fujikawa M, Ichikawa H, Nagai T. Formula optimization based on artificial neural networks in transdermal drug delivery. *J Control Release* 1999;62:161–170.
36. Wu P-C, Obata Y, Fijukawa M, Li CJ, Higashiyama K, Takayama K. Simultaneous optimization based on artificial neural networks in ketoprofen hydrogel formula containing s-ethyl-3-burylcyclohexanol as a percutaneous absorption enhancer. *J Pharm Sci* 2001;90:1004–1014.
37. Agatonovic-Kustrin S, Alany RG. Role of genetic algorithms and artificial neural networks in predicting phase behaviour of colloidal delivery systems. *Pharm Res* 2001;18:1049–1055.
38. Agatonovic-Kustrin S, Glass BD, Wisch MH, Alany RG. Prediction of a stable microemulsion formulation for the oral delivery of a combination of antitubercular drugs using ANN technology. *Pharm Res* 2003;20:1760–1765.
39. Kandimalla KK, Kanikkannon N, Singh M. Optimization of a vehicle mixture for the transdermal delivery of melatonin using artificial neural networks and response surface method. *J Control Release* 1999;61:71–82.
40. Peh KK, Lim CP, Qwek SS, Khoti KH. Use of artificial networks to predict drug dissolution profiles and evaluation of network performance using similarity profile. *Pharm Res* 2000;17:1386–1398.
41. Turkoglu M, Varol H, Celikok M. Tableting and stability evaluation of enteric-coated omeprazole pellets. *Eur J Pharm Biopharm* 2004;57:277–286.
42. Yuksel N, Turkoglu M, Baykara T. Modelling of the solvent evaporation method for the preparation of controlled release acrylic microspheres using neural networks. *J Microencapsul* 2000;17:541–551.
43. Surini S, Akiyama H, Morishita M, Nagai T, Takayama K. Release phenomena of insulin from an implantable device composed of a polyion complex of chitosan and sodium hyaluronate. *J Control Release* 2003;90:291–301.
44. Reis MAA, Sinisterra RO, Belchior JC. An alternative approach based on artificial neural networks to study controlled drug release. *J Pharm Sci* 2004;93:418–428.
45. Simon L, Fernandes M. Neural network-based prediction and optimization of estradiol release from ethylene-vinyl acetate membranes. *Comput Chem Eng* 2004;28:2407–2419.
46. Mendyk A, Jachowicz R. Neural network as a decision support system in the development of pharmaceutical formulation-focus on solid dispersions. *Expert Syst Appl* 2005;28:285–294.
47. Bossley KM, Mills DJ, Brown M, Harris CJ. Construction and design of parsimonious neurofuzzy systems. In: *Advances in Neural Networks for Control Systems. Advances in Industrial Control*, edited by Irwin G, Hunt K, Warwick K, pp. 153–177. Berlin: Springer-Verlag, 1995.
48. Attivi D, Wehrie P, Ubrich N, Damge C, Hoffman M, Maincent P. Formulation of insulin-loaded polymeric nanoparticles using response surface methodology. *Drug Dev Ind Pharm* 2005;31:179–189.

49. Elkheshen SA, Badawi SS, Badawi AA. Optimization of a reconstitutable suspension of rifampicin using 24 factorial design. *Drug Dev Ind Pharm* 1996;22:623–630.
50. Rowe RC, Roberts RJ. *Intelligent Software for Product Formulation*. London: Taylor and Francis, 1998.
51. Shao Q, Rowe RC, York P. Investigation of an artificial intelligence technology—Model trees. Novel applications for an immediate release tablet formulation database. *Eur J Pharm Sci* 2007;31:137–144.
52. Shao Q, Rowe RC, York P. Data mining of fractured experimental data using neurofuzzy logic—Discovering and integrating knowledge hidden in multiple formulation databases for a fluid bed granulation process. *J Pharm Sci* 2008; 97:2091–2101.
53. Do DQ, Rowe RC, York P. Modelling drug dissolution from controlled release products using genetic programming. *Int J Pharm* 2008;351:194–200.
54. Mendyk A, Jachowicz R. Unified methodology of neural analysis in decision support systems built for pharmaceutical technology. *Expert Syst Appl* 2007;32: 1124–1131.

PART V

DATA MINING ALGORITHMS AND TECHNOLOGIES

15

DIMENSIONALITY REDUCTION TECHNIQUES FOR PHARMACEUTICAL DATA MINING

IGOR V. PLETNEV, YAN A. IVANENKOV, AND ALEXEY V. TARASOV

Table of Contents

15.1 INTRODUCTION

It has become increasingly evident that modern science and technology presents an ever-growing challenge of dealing with huge data sets. Some representative examples are molecular properties of compounds in combinatorial libraries, expression profiles of thousands of genes, multimedia documents and files, marketing databases, and so on.

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

Automatic extraction of the knowledge embedded within the large volumes of data—that is, *data mining*—is a prevalent research theme. Unfortunately, the inherent structures and relations that are sought for may not be easily recognizable due to a complexity of high-dimensional data. This is particularly true when chemical and biologic problems are considered. Therefore, the need for advanced tools transforming high-dimensional data into a lower dimensionality space (*dimensionality reduction*) cannot be overestimated.

This chapter describes a number of advanced dimensionality reduction techniques, both linear and nonlinear (Fig. 15.1). The following linear methods are described: principal component analysis (PCA) [1], linear discriminant analysis (LDA) [2], and factor analysis (FA) [3]. Among nonlinear methods are kernel PCA (KPCA) [4], diffusion maps (DMs) [5], multilayer autoencoders (MAs) [6], Laplacian eigenmaps [7], Hessian local linear embedding (HLLE) [8], local tangent space analysis (LTSA) [9], locally linear coordination (LLC) [10], multidimensional scaling (MDS) [11], local linear embedding (LLE) [12], and support vector machines (SVMs) [13]. Though this list includes the most popular in computational chemistry techniques, it is not exhaustive; some less important approaches (e.g., independent component analysis [ICA] [14]) are deliberately omitted. Also, because of the space limitations some nonlinear techniques (which may often be considered "flavors" of more general approaches) are not described. To only mention, they are principal curves [15], curvilinear component analysis (CCA) [16], generalized discriminant analysis (GDA) [17], kernel maps [18], maximum variance unfolding (MVU) [19], conformal eigenmaps (CEMs) [20], locality preserving projections (LPPs) [21], linear local tangent

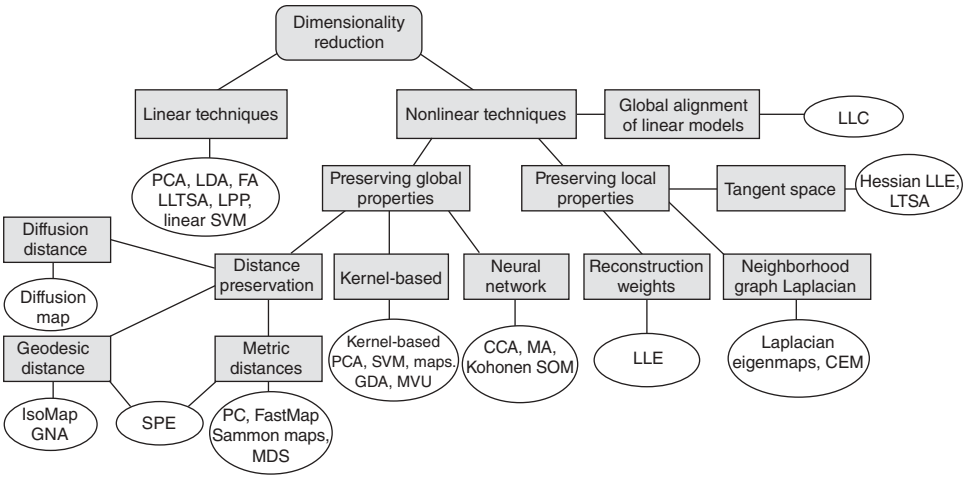


Figure 15.1 A functional taxonomy of advanced dimensionality reduction techniques. GNA = geodesic nullspace analysis; SOM = self-organizing map.

space alignment (LLTSA) [22], FastMap [23], geodesic nullspace analysis (GNA) [24], and various methods based on the global alignment of linear models [25–27].

Finally, it should be noted that several advanced mapping techniques (self-organizing Kohonen maps [28], nonlinear Sammon mapping [29], IsoMap [30,31], and stochastic proximity embedding [SPE] [32]) are described in more detail; see Chapter 16 of this book.

15.2 DIMENSIONALITY REDUCTION BASICS

In silico pharmacology is an explosively growing area that uses various computational techniques for capturing, analyzing, and integrating the biologic and medical data from many diverse sources. The term *in silico* is an indicative of procedures performed by a computer (silicon-based chip) and is reminiscent of common biologic terms *in vivo* and *in vitro*. Naturally, *in silico* approach presumes massive data mining, that is, extraction of the knowledge embedded within chemical, pharmaceutical, and biologic databases.

A particularly important aspect of data mining is finding an optimal data representation. Ultimately, we wish to be able to correctly recognize an inherent structure and intrinsic topology of data, which are dispersed irregularly within the high-dimension feature space, as well as to perceive relationships and associations among the studied objects.

Data structures and relationships are often described with the use of some similarity measure calculated either directly or through the characteristic features (descriptors) of objects. Unfortunately, the very essence of similarity measure concept is intimately connected with a number of problems, when applied to high-dimensional data.

First, the higher is the number of variables, the more probable is a possibility of intervariable correlations. While some computational algorithms are relatively insensitive to correlations, in general, redundant variables tend to bias the results of modeling. Moreover, if molecular descriptors are used directly for property prediction or object classification, overfitting can become a serious problem at the next stages of computational drug design.

Second, a common difficulty presented by huge data sets is that the principal variables, which determine the behavior of a system, are either not directly observable or are obscured by redundancies. As a result, visualization and concise analysis may become nearly impossible. Moreover, there is always the possibility that some critical information buried deeply under a pile of redundancies remains unnoticed.

Evidently, transforming raw high-dimensional data to the low-dimensional space of critical variables—*dimensionality reduction*—is a right tool to overcome the problems. For visualization applications, an ideal would be a mapping onto two-dimensional or three-dimensional surface.

The principal aim of dimensionality reduction is to preserve all or critical neighborhood properties (*data structure*). This presumes that the data points located close to each other in high-dimensional input space should also appear neighbors in the constructed low-dimensional feature space. Technically, there exist many approaches to dimensionality reduction. The simplest ones are *linear*; they are based on the linear transformation of the original high-dimensional space to target a low-dimensional one. Advanced *nonlinear* techniques use more complicated transforms. Evidently, nonlinear methods are more general and, in principle, applicable to a broader spectrum of tasks. This is why this chapter considers nonlinear techniques in more detail. However, it is worth mentioning that linear methods typically have mathematically strict formulation and, which is even more important, form a basis for sophisticated nonlinear techniques.

The classical linear methods widely used in chemoinformatics are PCA [33] and MDS [34]. PCA attempts to transform a set of correlated data into a smaller basis of orthogonal variables, with minimal loss in overall data variance. MDS produces a low-dimensional embedding that preserves original distances (=dissimilarity) between objects. Although these methods work sufficiently well in case of linear or *quasi*-linear subspaces, they completely fail to detect and reproduce nonlinear structures, curved manifolds, and arbitrarily shaped clusters. In addition, these methods, as many of stochastic partitioning techniques, can be most effectively used for the detailed analysis of a relatively small set of structurally related molecules. They are not well suited for the analysis of disproportionately large, structurally heterogeneous data sets, which are common in modern combinatorial techniques and high-throughput screening (HTS) systems. One additional problem of MDS is that it unfavorably scales quadratically with the number of input data points, which may require enormous computational resources. Therefore, there exists a continuing interest to novel approaches. Some examples of such advanced methods are agglomerative hierarchical clustering based on two-dimensional structural similarity measurement, recursive partitioning, and self-organizing mapping, as well as generative topographic mapping and truncated Newton optimization strategy could be effectively used [35–37]. Thus, a variety of different computational approaches were recently intended to apply neural net paradigm toward a nonlinear mapping (NLM). The immense advantage of neural nets lies in their extraordinary ability to allocate the positions of new data points in the low-dimensional space producing significantly higher predictive accuracy. A number of scientific studies have successfully applied the basic self-organizing principles, especially self-organizing Kohonen methodology for visualization and analysis of the diversity of various chemical databases [38–40]. Sammon mapping [29] is another advanced technique targeted for dimensionality reduction, which is currently widely used in different scientific areas, including modern *in silico* pharmacology. Although the practical uses of this method is also weakened by the mentioned restriction relating to large

data sets, it has several distinct advantages as against to MDS. The basic principles of Sammon algorithm are discussed in more detail below.

15.2.1 Clustering

Clustering is a common though simple computational technique widely used to partition a set of data points into groups (*clusters*) in such a manner that the objects in each group share the common characteristics, as expressed by some distance or similarity measure. In other words, the objects falling into the same cluster are similar to each other (*internally homogeneous*) and dissimilar to those in other clusters (*externally heterogeneous*).

Based on the way in which the clusters are formed, all clustering techniques may be commonly divided into two broad categories: *hierarchical*, which partitions the data by successively applying the same process to clusters formed in previous iterations, or *nonhierarchical* (*partitional*), which determines the clusters in a single step [41]. Hierarchical algorithms are more popular as they require very little or no *a priori* knowledge.

Hierarchical methods are divided in two minor classes: agglomerative (*bottom-up*) or divisive (*top-down*). In agglomerative analysis, clusters are formed by grouping samples into bigger and bigger clusters until all samples become members of a single cluster. Before the analysis, each sample forms its own, separate cluster. At the first stage, two samples are combined in the single cluster, at the second, the third sample is added to the growing cluster, and so on. Graphically, this process is illustrated by agglomerative dendrogram. Inversely, a divisive scheme starts with the whole set and successively splits it into smaller and smaller groups.

There are two more important details: the way of measuring the distance between samples (metrics) and the way of measuring the distance between samples and cluster (linkage rule). The popular options are Euclidean, squared Euclidean, and Manhattan city-block metrics in combination with complete linkage, Ward's linkage, and weighted/unweighted pair-group average linkage rules.

Cluster analysis already found numerous applications in various fields including chemoinformatics. Because of its close ties with molecular similarity, clustering is often a tool of choice in the diversity analysis allowing one to reduce the complexity of a large data set to a manageable size [42,43]. Technically, clustering compounds comprises four principal steps. Initially, a rational set of molecular descriptors is selected (and, typically, scaled). Then pairwise distances between molecules are calculated and collected into similarity matrix. After that, cluster analysis technique is used to iteratively assign objects to different clusters. Finally, the clustering is validated, visually and/or statistically.

Many efforts to visualize the results of hierarchical and nonhierarchical clustering have been made based on graph drawing and tree layout algorithms.

One popular option is the Jarvis–Patrick algorithm. This method begins by determining the *k*-nearest neighbors of each object within the collection. Members are placed in the same cluster if they have some predefined number of common nearest neighbors. The major advantage of this method lies in the speed of processing performance, but the main disadvantage is a regrettable tendency to generate either too many singletons or too few very large clusters depending on the stringency of the clustering criteria. Several key algorithms for graph construction, tree visualization, and navigation, including a section focused on hierarchical clusters, were comprehensively reviewed by Herman [44]; Strehl and Ghosh [45] inspected many computational algorithms for visualizing nonhierarchical clusters including similarity matrix plot. Still, a classical dendrogram remains the most popular cluster visualization method. This layout visually emphasizes both the neighbor relationship between data items in clusters (*horizontal*) and the number of levels in the cluster hierarchy (*vertical*). The basic limitation of radial space-filling and linear tree visualizations is the decreasing size and resolution of clusters with many nodes or deep down in the hierarchy.

To overcome these challenges, Agrafiotis et al. [46] recently developed a new radial space-filling method targeted for visualizing cluster hierarchies. It is based on radial space-filling system and nonlinear distortion function, which transforms the distance of a vertex from the focal point of the lens. This technique has been applied (Fig. 15.2A,B) [47,48] for the visualization and analysis of chemical diversity of combinatorial libraries and of conformational space of small organic molecules [49]. In the first case, the radial clustergram represented a virtual combinatorial library of 2500 structures produced by combining 50 amines and 50 aldehydes *via* reductive amination. Each product was accurately described by 117 topological descriptors, which were normalized and decorrelated using PCA—to an orthogonal set of 10 latent variables accounting for 95% of the total variance in the input data. The obtained radial clustergram (Fig. 15.2) is coded in grayscale by the average molecular weight (A) and log *P* (B) (dark gray corresponds to higher value); the two clusters designated as A and B are easily recognized. Note that color significantly changes at cluster boundaries, so one may reveal structurally related chemical families with distinct properties.

While all of the tested structures share a common topology, which explains their proximity in diversity space, compounds located in cluster "A" contain several halogens as well as at least one bromine atom, which increases both their molecular weight and log *P*. There are no molecules in the first cluster with bromine atom, and none of them carry more than one light halogen (F or Cl).

The second example (Fig. 15.2C,D) illustrates the application of radial clustergram for visualization of conformational space. The data set consisted of 100 random conformations of known HIV protease inhibitor, Amprenavir. Each pair of conformers was superimposed using a least-squares fitting procedure, and the resulting root mean square deviation (RMSD) was used as

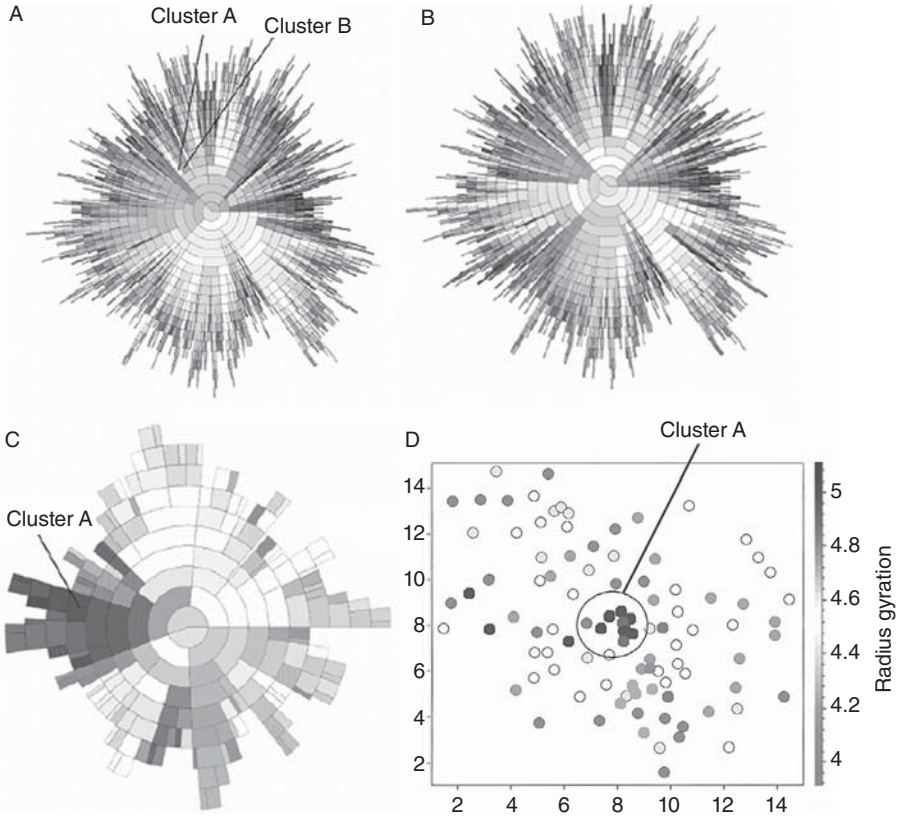


Figure 15.2 Radial clustergrams of a combinatorial library containing 2500 compounds; grayscale coding is by average molecular weight (A) and log *P* (B), radial clustergram (C), and SPE map (D) of conformational space around Amprenavir.

a measure of the similarity between the two conformations. The radial clustergram color coded by the radius of gyration (a measure of the extendedness or compactness of the conformation) is shown in Figure 15.2C, while Figure 15.2D shows a nonlinear SPE map of the resulting conformations (see Chapter 16, Section 16.2.5 for SPE description; the method embeds original data into a two-dimensional space in such a way that the distances of points on the map are proportional to the RMSD distances of respective conformations).

Among other modern algorithms related to clustering that should be mentioned are the maximin algorithm [50,51], stepwise elimination and cluster sampling [52], HookSpace method [53], minimum spanning trees [54], graph machines [55], singular value decomposition (SVD), and generalized SVD [56].

15.3 LINEAR TECHNIQUES FOR DIMENSIONALITY REDUCTION

15.3.1 PCA

PCA, also known as Karhunen–Loeve transformation in signal processing, is a quite simple though powerful and popular linear statistical technique with a long success history [57–60]. It allows one to guess an actual number of independent variables and, simultaneously, to transform the data to reduced space. PCA is widely used to eliminate strong linear correlations within the input data space as well as to data normalization and decorrelation.

By its essence, PCA constructs a low-dimensional representation of the input data, which describes as much of the variance of source data as possible [61]. Technically, it attempts to project a set of possibly correlated data into a space defined by a smaller set of orthogonal variables (principal components [PCs] or eigenvectors), which correspond to the maximum data variance. From a practical viewpoint, this method combines descriptors into a new, smaller set of noncorrelated (orthogonal) variables.

In mathematical terms, PCs are directly computed by diagonalizing the variance covariance matrix, m_{ij}, a square symmetric matrix containing the variances of the variables in the diagonal elements and the covariances in the off-diagonal elements: $m_{ij} = m_{ji} = \frac{1}{N}\sum (x_{ki} - \xi_i)(x_{kj} - \xi_j); \xi_i = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$, where ξ_i is the mean value of variable x_i, and N is the number of observations in the input data set. Using this strategy, PCA attempts to find a linear mapping basis M that maximizes $M^T \operatorname{cov}_{X-\bar{X}} M$ where $\operatorname{cov}_{X-\bar{X}}$ is the covariance matrix of zero mean data X (D-dimensional data matrix X). Such linear mapping can easily be formed by the d PCs derived from the covariance matrix m_{ij}. Principally, PCA attempts to maximize $M^T \operatorname{cov}_{X-\bar{X}} M$ with respect to M, under the constraint $|M| = 1$. This constraint, in turn, can be consistently enforced by introducing a Lagrange multiplier λ. Hence, an unconstrained maximization of $M^T \operatorname{cov}_{X-\bar{X}} M + \lambda(1 - M^T M)$ can be efficiently performed, then a stationary point of this expression can be regularly found assuming that $\operatorname{cov}_{X-\bar{X}} M = \lambda M$. Following this logic, PCA investigates the eigenproblem lying in $\operatorname{cov}_{X-\bar{X}} M = \lambda M$, which can be solved successfully for the d principal eigenvalue λ. The low-dimensional data representations encoded entirely by y_i (the ith row of the D-dimensional data matrix Y) of the sample point x_i (high-dimensional data points or the ith row of the D-dimensional data matrix X) can then be computed by mapping them onto the linear basis M, i.e., $Y = (X - \bar{X})M$. Considering that the eigenvectors of covariance matrix are the PCs while the eigenvalues are their respective variances, the number of PCs directly corresponds to the number of the original variables. In other words, PCA reduces the dimensionality of input data points by throwing off the insignificant PCs that contribute the least to the variance of the data set (i.e., PCs with the smallest eigenvalues) until the maximum variance approximates manually or machine predefined threshold, typically defined in the range of

90–95% of the original input value. Finally, the input vectors are transformed linearly using the transposed matrix. At the output of PCA processing, the obtained low-dimensional coordinates of each sample in the transformed frame represent linear combinations of the original, cross-correlated variables. The immense advantage of PCA algorithm lies in the following statement: there are no assumptions toward the underlying probability distribution of the original variables, while the distinct disadvantage can be related meaningfully to a heightened sensibility to possible outliers, missing data, and poor correlations among input variables due to irregular distribution within the input space.

Having due regard to all the advantages mentioned above, PCA is evidently not suitable for the study of complex chemoinformatics data characterized by a nonlinear structural topology. In this case, the modern advanced algorithms of dimensionality reduction should be used. For example, Das et al. [62] recently effectively applied a nonlinear dimensionality reduction technique based on the IsoMap algorithm [31] (see Chapter 16, Section 16.2.4).

15.3.2 LDA

LDA [2] is a linear statistical technique generally targeted for the separation of examined vector objects belonging to different categories. As PCA mainly operates on principle related to eigenvectors formation, LDA is generally based on a combination of the independent variables called discriminant function. The main idea of LDA lies in finding the maximal separation plan between external data points [2]. In contrast to the majority of dimensionality reduction algorithms described in this chapter, LDA can be regarded as a supervised technique. In essence, LDA finds a linear mapping image M that provides the maximum linear separation among the estimated classes in the low-dimensional representation of the input data. The major criteria that are primarily used to formulate a linear class separability in LDA are the *within*-class scatter Z_w and the *between*-class scatter Z_b, which are correspondingly defined as $Z_w = \sum_f p_f \operatorname{cov}_{X^f - \bar{X}^f}$, $Z_b = \operatorname{cov}_{X-\bar{X}} - Z_w$, where p_f is the class prior of class label f, $\operatorname{cov}_{X^f - \bar{\bar{X}}^f}$ is the covariance matrix of the zero-mean data point x_i directly assigned to class $f \in F$ (where F is the set of possible classes), while $\operatorname{cov}_{X-\bar{X}}$ is the covariance matrix of the zero-mean data assigned to X. In these terms, LDA attempts to optimize the critical ratio between Z_w and Z_b in the low-dimensional representation of the input data set, by finding a linear mapping image M that maximizes the so-called Fisher criterion: $\phi(M) = \dfrac{M^T Z_b M}{M^T Z_w M}$. The post maximization problem can be efficiently solved by computing the d principal eigenvectors of $Z_w^{-1} Z_b$ under the following requirement: $d < |F|$. The low-dimensional data representation Y of the input data points dispersed within the high-dimensional space X can be easily computed by embedding the input vector samples onto the linear basis M, i.e.,

$Y(X-\bar{X})M$. Similar to regression analysis, the mathematical probability of achieving adequate separation simply by chance increases proportionally with the number of variables used. Therefore, the ratio between input data points and descriptor variables should be fixed preferably to at least 3. In addition, a bias distribution within the input data samples may become a significant problem; therefore, the estimated categories should be uniformly distributed within the input space. Otherwise, a trivial separation may be simply achieved, despite the fact that the above-specified ratio is greater than 3.

Typically, the regular output of LDA is commonly expressed as percentage of compounds correctly and incorrectly classified. It should be noted that cross-validation strategy can also be fruitfully applied, as in the case of regression analysis, to cross-test the predictive ability of model. For the classification of new objects from an independent external test set, the critical threshold value is commonly used. Thus, if the calculated value of the discriminant function for the tested object is lower than the critical/threshold value, the object is directly assigned to the "inactive" category, if it is higher—to the active one. In addition, LDA can also be beneficially applied to handle more than two categories in accordance to a number of selected discriminant functions.

15.3.3 FA

FA is a statistical linear technique closely related to PCA that attempts to extract coherent subsets of variables that are relatively independent from one another [57]. Both methods rely principally on an eigenvalue analysis of the covariance matrix, and both use linear combinations of variables to explain a set of observations. However, PCA is mainly focused on the observation of variables themselves, and the combination of these variables is used entirely for simplifying their analysis and interpretation. Conversely, in FA, the observed variables are of little intrinsic value; what is of interest is the underlying factors ("hidden variables"). It is of paramount importance in many cases where an actually meaningful variable is not directly observable. The key goal of FA is to properly explain the possible correlations among the estimated variables referring to underlying factors, which are not directly observable. The factors are usually represented by linear combinations of original variables; they are thought to be a representative of the underlying process that has created the correlations. Factors may be associated with two or more of these variables (*common factors*) or with a single variable (*unique factor*). The relationship between the original variables and the derived factors is explicitly expressed in the form of factor loadings. Inherently, these loadings are statically indeterminate but at the same time they can be readily derived from the eigenvalues of covariance matrix. By using the rotation (of coordinate axes) procedure, each variable becomes highly loaded with one factor, and all the factor loadings are either large or nearly negligible. A number of different rotation methods are currently available, including varimax, quartimax, and

equimax. Varimax is the most widely used method maximizing the variance of the loadings.

FA has been widely used in different fields. For example, Cummins et al. [63] recently applied this method to reduce a set of 61 molecular descriptors to four factors, which were further used to compare the diversity of five chemical databases. It was also employed by Gibson et al. [60] in their comparative study of 100 different heterocyclic aromatic systems.

15.4 NONLINEAR TECHNIQUES FOR DIMENSIONALITY REDUCTION

As was already mentioned, conventional linear methods of data mining and visualization are inadequate to represent the extremely large, high-dimensional data sets that are frequently encountered in molecular diversity/similarity analyses. So the various nonlinear techniques became a tool of choice in modern chemoinformatics. Despite utilizing a variety of mathematical formulations, all of these methods are intended to preserve the global or local topology of the original data. In other words, the data points (or clusters) located near each other in a high-dimensional space should also be neighbors in a low-dimensional representation.

15.4.1 Global Techniques

15.4.1.1 KPCA KPCA is an advanced version of conventional (linear) PCA that is specifically adapted to the use of a kernel function [64]. Notably, recent years have seen a dramatic reformulation of several linear techniques with the use of the "kernel trick," which resulted in the development and expansion of, e.g., kernel ridge regression and SVMs (see Section 15.4.1.7). In contrast to the traditional linear PCA, KPCA computes the principal eigenvectors of the kernel matrix rather than those of the covariance matrix. The transformation of traditional PCA in the kernel basis through the kernel-based matrix by using different kernel functions is fairly straightforward. Since KPCA uses a kernel function, it evidently produces an NLM. In mathematical terms, KPCA computes the kernel matrix M of the data point x_i. The components of the kernel matrix are defined as $m_{ij} = f(x_i, x_j)$, where f is a kernel function [65]. The constructed kernel matrix M is then centered using the following modification of eigencomponents: $m_{ij} = m_{ij} - \frac{1}{n}\sum_l m_{il} - \frac{1}{n}\sum_l m_{jl} - \frac{1}{n^2}\sum_{lk} m_{lk}$. One should note that this operation directly corresponds to subtracting the mean value in traditional PCA; it makes sure that the features defined by the kernel function in the high-dimensional input space have zero mean. Then, the principal eigenvector v_i (defined unambiguously by the modified kernel matrix) is computed. Note that in a high-dimensional space, the eigenvector α_i of the covariance matrix constructed by m components represents scaled versions of the corresponding

eigenvector v_i of the kernel-based matrix: $\alpha_i = \frac{1}{\sqrt{\lambda_i}} v_i$. In order to obtain the low-dimensional data representation, the input data is projected onto the eigenvectors of the covariance matrix α_i. As a result, the low-dimensional data representation Y is constructed according to the following equation: $Y = \left\{ \sum_j \alpha_1 f(x_j, x), \sum_j \alpha_2 f(x_j, x), ..., \sum_j \alpha_k f(x_j, x) \right\}$, where f is the predefined kernel function (also used in the computation of the corresponding kernel matrix).

Since KPCA is a kernel-based method, the mapping performed by KPCA relies greatly on the choice of the kernel function f, which includes the linear kernel (making KPCA equal to traditional PCA), the polynomial kernel (homogeneous) and polynomial kernel (inhomogeneous), and the Gaussian radial basis function (RBF) and sigmoid kernel [65]. KPCA has been successfully applied in various fields, including speech recognition [66] and novelty detection [67], as well as *in silico* drug design [68]. The practical application of KPCA is inherently limited by the size of the kernel matrix, i.e., due to the squared relationship between the number of input samples and the number of kernel matrix components. To effectively overcome this drawback, some approaches have recently been proposed, for example in Reference 69.

15.4.1.2 DMs Originated from the dynamical system theory, DM framework is a spectral clustering algorithm, which is principally based on determination of the Markov random walk on the graph [5,70]. Following the algorithm, a specific measure of proximity between the input data points, also called diffusion distance, is implicitly defined through a number of time steps using a random walk strategy. In the low-dimensional representation of the data, the pairwise diffusion distances have assiduously abided by the initial ones as well as possible. In the DM structure, a completely regular graph of the data is primarily constructed, then the weight coefficient w_{ij} of the edges in the graph is computed accurately using the Gaussian kernel function, resulting in the construction of a matrix M with eigencomponents: $w_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$, where σ indicates the variance of the Gaussian distribution. Subsequently, the normalization of the obtained matrix M can be routinely performed in such a way that each row adds up to one. In result, a matrix denoted by $F^{(1)}$ is completely formed by the following entries: $f_{ij}^{(1)} = \frac{w_{ij}}{\sum_k w_{ik}}$. Since the DMs originate from dynamical systems theory, the obtained matrix $F^{(1)}$ can be reasonably considered as the Markov matrix that, in dynamical process, defines the forward transition probability matrix. Hence, the matrix $F^{(1)}$ represents the probability of a transition from the initial data point into the corresponding feature space image during a single time step, while the forward

probability matrix for t time steps, $F^{(t)}$ can be denoted by $(F^{(1)})^t$. Using the random walk forward probability $f_{ij}^{(t)}$, the diffusion distance can be uniquely defined by $Z^{(t)}(x_i, x_j) = \sum_k \frac{(f_{ik}^{(t)} - f_{jk}^{(t)})^2}{\varphi^{(0)}(x_k)}$. Here, the term $\varphi^{(0)}(x_k)$ directly attributes more weights to parts of the graph with a high density and can be easily that the by the following equation: $\varphi^{(0)}(x_i) = \frac{m_i}{\sum_j m_j}$, where m_i is the degree of node x_i defined by $m_i = \sum_j f_{ij}$. Based on the above equation, it is clearly seen that the pairs of data points with a high forward transition probability are characterized by short diffusion distances. The key underlying principle of such diffusion distance lies in the large number of paths passing through the graph that makes this algorithm more robust to excessive noise level as compared with, e.g., the geodesic distance. In the low-dimensional representation of the data D, diffusion tactic attempts to completely retain the estimated diffusion distances. In accordance to spectral theory related to the random walk, it can be clearly shown that the low-dimensional representation D that retains conceptually the diffusion distances can be completely formed by the d unique principal eigenvectors in the context of the corresponding eigenproblem: $F^{(t)}D = \lambda D$. As the graph is fully connected, the largest eigenvalue is trivial (viz, $\lambda_1 = 1$), and its eigenvector v_1 is then irretrievably discarded. Finally, the low-dimensional representation Y of the initial sample space X can be successfully performed by a number of principal eigenvalues, which are commonly used to normalize the corresponding eigenvectors. Thus, in this representation, the normalized eigenvectors, in turn, determine accurately the low-dimensional data representation following the definition of D by $\mathrm{D} = \{\lambda_2 v_2, \lambda_3 v_3, \ldots, \lambda_{d+1} v_{d+1}\}$.

15.4.1.3 MAs In contrast to a variety of the above-described methods of dimensionality reduction, MAs belong to a class of feedforward neural networks with an odd number of hidden layers [6]. While the middle hidden layer consists of f nodes, both the input and the output layers are commonly represented by F nodes. During the learning process, the key goal of this network is to minimize the mean squared error E_r observed between the input and the output neurons. In other terms, training the neural network on the data point x_i leads ultimately to a network in which the middle hidden layer provides the d-dimensional representation of the input vector samples preserving as much information in the high-dimensional space X as possible. When the data point x_i is used as the input neural signal, the low-dimensional representation y_i can be readily obtained by extracting node values, which constitute the middle hidden layer of the network. It should be highlighted that in the case of using the linear activation functions, the algorithm becomes very similar to PCA [71]. In order to allow the MAs to produce an NLM between the

high-dimensional and low-dimensional data points, the more powerful nonlinear activation functions can be effectively used, for example, sigmoid of hypertangent sigmoid. Due to the presence of a large number of synaptic weights, backpropagation strategy, widely applied for the neural net learning, converges slowly and is likely to get stuck in local minima. Fortunately, there are several approaches that can elegantly overcome this complication by performing a pretraining procedure using restricted Boltzmann machines (RBMs) [72]. In more detail, RBMs are neural networks composed entirely of the binary and stochastic units/neurons, where the internal connections among the hidden units are completely closed. Thus, RBMs can be successfully applied for training neural networks with many hidden layers using a learning approach based on simulated annealing algorithm. Thus, if the RBM-based pretraining procedure is performed, a fine-tuning of the total network weights can be immediately achieved using backpropagation learning strategy. As an alternative approach, genetic algorithms can also be effectively applied to train MAs [73].

15.4.1.4 MDS Among various approaches extensively applied in modern computational chemistry, molecular similarity is one of the most ubiquitous concepts [74]. This technique is widely used to analyze and categorize the chemical data of different types, rationalize the behavior and functions of organic molecules, and design novel chemical compounds with improved physical, chemical, and biologic properties. Usually, for the analysis of large collections of organic compounds, structural similarities can be uniquely defined by the symmetric matrix that contains all the pairwise relationships among the molecules presented in the external data set. However, it should be especially noted, that such pairwise similarity metric is not generally acceptable for numerical processing and visual inspection. A reasonable, workable solution to this methodological problem lies in embedding the input objects into a low-dimensional Euclidean space in a way that preserves the original pairwise proximities as faithfully as possible. There are at least two basic approaches, MDS and NLM, that effectively convert the input data points into a set of feature vectors that can subsequently be used for a variety of pattern recognition and classification tasks.

MDS is a positional-refinement linear technique of data mining that emerged from the practical need to visualize a set of objects described by means of the similarity or dissimilarity matrix. Initially, this method has originated in the field of psychology and can implicitly be traced back to the pioneering works of Torgerson [75] and Kruskal [76]. As described above, one of the toughest problems of dimensionality reduction methodology is to construct the acceptable and adequate representation of input data points in the low-dimensional feature space based generally on information related to the distances among these data samples buried deeply in the input space. Generally, this method consist mainly of the collection of statistical techniques that jointly attempt to map a set of input data points scattered randomly across the high-dimensional space and described by means of the

dissimilarity matrix into the low-dimensional display plane in a way that preserves their original pairwise interrelationships as closely as possible according to different similarity metrics, for example, classical Euclidean distances, d_{ij}, between the input points [75,76]. The quality of the mapping is usually expressed in the stress function, a measure of the error that occurred between the pairwise distances in the low-dimensional and high-dimensional representation of the data.

In mathematical terms, the main principle of MDS can be readily disclosed in the following way. Let us determine the set of κ input objects: the symmetric matrix d_{ij} is composed of dissimilarities between the examined objects, and a set of feature images is projected onto the m-dimensional display plane: $\{y_i, i = 1, 2, \ldots, k; y_i \in \mathcal{R}^m\}$. MDS painstakingly attempts to map vectors y_i onto the feature plane in such a way that their metric distances, usually Euclidean distances $d_{ij} = \|y_i - y_j\|$, approximate the corresponding values d_{ij} as closely as possible. Each learning iteration consists chiefly of calculating the similarity distances δ_{ij} observed between each pair of input points in a lower-dimensional trial configuration and, using a steepest descent algorithm, shifting the positions of those points so as to create a new configuration characterized by a smaller sum-of-squares difference between δ_{ij} and d_{ij}. For this purpose, at least four major functions are currently used; these include Kruskal's stress: $\zeta = \sqrt{\dfrac{\sum_{i<j} (\delta_{ij} - d_{ij})^2}{\sum_{i<j} \delta_{ij}^2}}$, Lingoes' alienation coefficient: $\zeta = \sqrt{1 - \dfrac{\sum_{i<j} (\delta_{ij} \cdot d_{ij})^2}{\sum_{i<j} \delta_{ij}^2}}$, raw stress function: $\phi(Y) = \sum_{ij} (\|x_i - x_j\| - \|y_i - y_j\|)^2$, and Sammon's cost function: $\phi(Y) = \dfrac{1}{\sum_{ij} \|x_i - x_j\|} \sum_{ij} \dfrac{(\|x_i - x_j\| - \|y_i - y_j\|)^2}{\|x_i - x_j\|}$, in which $\|x_i - x_j\|$ is the Euclidean distance between the high-dimensional data points x_i and x_j, while the term $\|y_i - y_j\|$ is the Euclidean distance between the low-dimensional feature points y_i and y_j. Sammon's cost function radically differs from the raw stress expression at the expense of putting greater emphasis on retaining all the distances observed among the initial space including "minor" distances that were originally small and often hastily assigned to insignificant category by other computational algorithms.

The actual embedding is commonly carried out in an iterative fashion by generating the initial set of coordinate y_i, computing the distance δ_{ij}, and finding a new set of the feature coordinate y_i using the steepest descent algorithm, such as Kruskal's linear regression or Guttman's rank-image permutation, as well as eigendecomposition of the pairwise distance matrix and the conjugate gradient or pseudo-Newton methods. Following the strategy, MDS learning cycle keeps repeating until the change in the applied stress function falls below some manually predefined or hardware-generated threshold value then learning procedure is completely terminated. There is a wide variety of MDS-based algorithms, involving different cost functions and optimization schemes [34].

Particularly, MDS is commonly used for visualization of multidimensional complex data sets arising in various scientific disciplines, e.g., in fMRI analysis [77] and in molecular modeling [78]. The enduring popularity of MDS has already led to the development of some advanced powerful computational techniques, such as SPE (see Section 15.4.1.7), stochastic neighbor embedding (SNE) [79], and FastMap [23].

Despite huge successes, the substantial computational cost of traditional MDS makes this technique particularly crude or completely inapplicable to large data sets. Thus, because of the quadratic dependence on the number of objects scaled, current MDS algorithms are notoriously slow and their application in modern combinatorial library designs is strictly limited to relatively small data sets. In more detail, the chronic failure of MDS lies mainly in the fact that it attempts to maximally preserve all the pairwise distances observed among the input data space, both local and significantly remote. Indeed, given a large set of input samples, the immense symmetric matrix composed of a number of interrelationships (*proximities*) between the input objects, and a set of images projected onto a D-dimensional display map, MDS diligently attempts to arrange each point of a future space across the whole future plane in such a way that their metric distances approximate the corresponding components of the initial matrix as closely as possible. It is typically accomplished by minimizing an error function that measures the discrepancy between the input and output distances, such as Kruskal's stress (see below). However, it has been widely known that many conventional similarity measures such as the Euclidean distance tend strongly to underestimate the proximity of data points within a nonlinear manifold, thereby leading to erroneous or anomalous embeddings [80,81].

To partly overcome these incurable problems, several advanced variants of the basic MDS algorithm were recently developed based generally on hybrid architecture. For example, a family of algorithms that cunningly combine NLM techniques with several types of neural networks, including feedforward nets, which in turn make it possible for the MDS to approximate very large data sets that are too complex and analytically intractable for conventional methodologies, were vividly described [49,82]. The developed approach directly employs an NLM algorithm to accurately project a randomly selected sample, and then "learns" the underlying transform recruiting one or more multilayer perceptrons (MLPs) trained following the basic backpropagation learning principle. The resulting nonlinear maps can be effectively used to train a series of neural networks, using the specific similarity channel connected to a small number of reference structures from the training set as the input signal. The distinct advantage of this approach is that it captures the NLM relationships in an explicit function, thereby allowing the scaling of additional patterns as they become available, without the need to reconstruct the entire map.

15.4.1.5 SVM Recently, a relatively novel method has become popular in machine learning community [13,83,84], which seems to be both powerful and

versatile. This is the so-called SVMs (originally proposed in the 1960s by Vladimir Vapnik), which exists in classification and regression flavors. Due to a solid theoretical basis, it is actively pursued now for applicability in various areas and already found numerous applications in chemistry, biochemistry, and *in silico* drug discovery.

There exists a number of excellent introductions into SVM methodology, so we will only summarize the main ideas and terms of the approach (following our description [85]).

A particularly important feature of SVM is that it explicitly relies on *statistical learning theory* and directly addresses an issue of avoiding overfitting. The key concept here is *structural risk minimization* (SRM) principle proposed by Vapnik and Chervonenkis in the early 1970s.

Suppose we have a set of m training data points $\{(\mathbf{x}_1, y_1) \ldots (\mathbf{x}_m, y_m)\}$ where $\mathbf{x}$ is a feature (descriptors; $\mathbf{X}$ is called input space) and y_m is a class label, typically, –1 and 1 in binary classification tasks. Suppose also that there exists an unknown probability distribution $P(\mathbf{x}, y)$, which describes a relation of features to classes. Classification attempts to associate the descriptors with classes by introducing prediction, or decision, function $f(\mathbf{x}, a)$, which value changes from –1 to +1, dependent on class. The decision function parameter a is to be found *via* minimization of the functional of expected error:

$$I(a) = \int Q(x, a, y) P(x, y) dxdy,$$

where $Q(x, a, y)$ is the so-called loss function. For example, choice of $Q = (y - f(x, a))^2$ corresponds to common least squares estimate.

Evident problem is that the integral depends on the unknown true distribution P defined for the whole input space while all that we actually have is some sampling from that distribution, the training set. So for practical purposes, the integral should be replaced with sum over training points only, the so-called *empirical risk*.

Notably, there could be a number of different functions that all give a good classification for training patterns but may differ in predictions. Evidently, one should select such a decision function that would perform best not only at training set examples but also on previously unseen data, that is, has the best *generalization* ability. According to SRM, this may be achieved by minimizing both empirical risk and *confidence interval*, the last term is proportional to the ratio of model complexity (measured with the so-called *Vapnik–Chervonenkis dimension*) to the number of training data points. Omitting mathematical details, SRM states that optimal classifier is given by a trade-off between reduction of the training error and limiting model complexity, whence limiting chances for overfitting.

Consider an example of classification in two-dimensional input space (Fig. 15.3). Given the depicted training set, both solid and dashed separation lines are acceptable; which one is better? Intuitively, it is clear that the better generalizing is given by the line that is less sensitive to small perturbations in

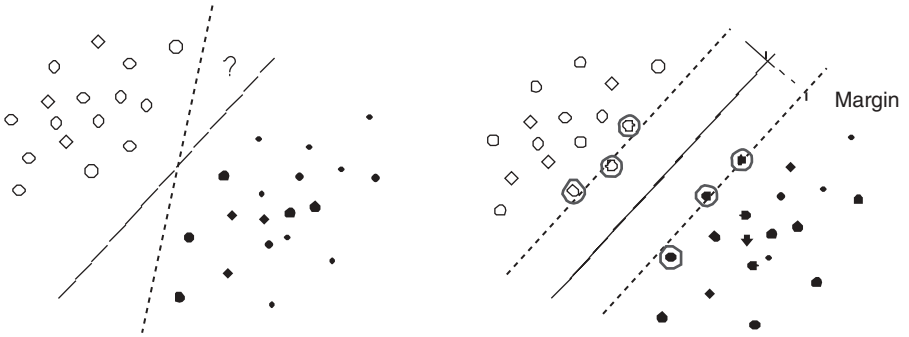


Figure 15.3 SVM classification principle. From the two possible linear discriminant planes (left), the best selection is one maximizing the margin (right).

placement of data points; this is a solid line on Figure 15.3. More broadly, the decision line or hyperplane must lie maximally far apart from the training points of different classes. This is exactly what follows from SRM and what constitutes the essence of SVM: the optimal classifier is the one providing the largest *margin* separating the classes (margin is defined as the sum of shortest distances from decision line to the closest points of both classes). Geometrically, the optimal line bisects the shortest line between the convex hulls of the two classes.

Notably, it appears that a relatively small number of data points, which are closest to the line (i.e., which lie on the margin; they are called *support vectors* [SVs]), are completely enough to determine the position of optimal separation line (*optimal separation hyperplane* [OSH] for high-dimension case).

Both SVs and OSH can be found by solving quadratic programming problem. If separating hyperplane is $Wx + b = 0$, which implies $y_i(Wx_i + \mathrm{b}) \geq 1$, $I = 1 \ldots m$, the decision is found by minimization Euclidian norm $\frac{1}{2}\|W\|^2$:

$$W = \sum_{i=1}^{m} y_i \alpha_i \cdot x_i .$$

Only if the corresponding Lagrange multiplier $\alpha_i > 0$, this x_i is an SV x. After minimization, decision function is written as

$$f(x) = sgn\left(\sum_{i=1}^{m} y_i \alpha_i \cdot x \cdot x_i + b \right).$$

Note that only a limited subset of training points, namely SVs, do contribute to the expression.

In linearly inseparable case, where no error-free classification can be achieved by hyperplane, there still exist two ways to proceed with SVM.

The first one is to modify linear SVM formulation to allow misclassification. Mathematically, this is achieved by introducing classification-error (*slack*) variables $\xi_i > 0$ and minimizing the combined quantity

$$\frac{1}{2}\|W\|^2 + C\sum_{i=1}^{m}\xi_i,$$

under the constraint defined as $y_i(Wx_i + b) \geq 1 - \xi_i$, $I = 1 \ \ldots \ m$. Here, the parameter C regulates a trade-off between minimization of training error and maximization of margin. Such approach is known as *soft margin technique*.

Another way is *nonlinear SVM*, which achieved a great deal of attention in the last decade. The most popular current approach is "transferring" data points from the initial descriptor space to space of higher dimension, which is derived by adding new degrees of freedom through nonlinear transformations of initial dimensions. The hope is that nonlinear in original space problem may become linear in higher dimensions, so that linear solution technique becomes applicable.

Importantly, direct transfer of the points from original to higher-dimensionality space is even not necessary, as all the SVM mathematics deals with dot products of variables ($\mathbf{x}_i$, $\mathbf{x}_j$) rather than with variable values $\mathbf{x}_i$, $\mathbf{x}_j$ itself. All which is necessary is to replace dot products ($\mathbf{x}_i$, $\mathbf{x}_j$) with their higher-dimensionality analogues, functions $K(\mathbf{x}_i, \mathbf{x}_j)$ expressed over *original* variable $\mathbf{x}$. The suitable function K is called *kernel*, and the whole approach is known as the *kernel trick*. Decision function in this case is written as

$$f(x) = sgn\left(\sum_{i=1}^{m} y_i\alpha_i \cdot K(x, x_i) + b\right).$$

The most common kinds of kernels are

$$K(x_i, x_j) = (x_i, x_j + 1)^d - \text{polynomial},$$
$$K(x_i, x_j) = \exp\left(-r\|x_i - x_j\|^2\right) - \text{RBF},$$
$$K(x_i, x_j) = \text{sigmoid}(\eta(x_i x_j) + a) - \text{two-layer perceptron}.$$

Finally, let us list the main SVM advantages:

1. We can build any complex classifier and the solution is guaranteed to be the global optimum (no danger of getting stuck at local minima). It is a consequence of quadratic programming approach and of the restriction of space of possible decisions.
2. There are few parameters to elucidate. Besides the main parameter C, only one additional parameter is needed to determine polynomial or RBF kernels, which typically (as can be judged from literature) demonstrate high classification power.

3. The final results are stable, reproducible, and largely independent of optimization algorithm. Absence of random constituent in SVM scheme guarantees that two users, which applied the same SVM model with the same parameters to the same data, will receive identical results (which is often not true with artificial neural networks).

In addition to topical application of SVM methodology toward various fields related to dimensionality reduction and pattern recognition, such as character recognition and text classification, SVM is currently widely used for the analysis of chemical data sets. For example, L'Heureux et al. [86] tested the ability of SVM, in comparison with well-known neural network techniques, to predict drug-likeness and agrochemical-likeness for large compound collections. For both kinds of data, SVM outperforms various neural networks using the same set of descriptors. Also, SVM was used for estimating the activity of carbonic anhydrase II (CA II) enzyme inhibitors; it was found that the prediction quality of the SVM model is better than that reported earlier for conventional quantitative structure activity relationships (QSAR).

Balakin et al. [35] effectively used a radial basis SVM classifier for the separation of compounds with different absorption, distribution, metabolism, and excretion (ADME) profiles within the Sammon maps (Fig. 15.4). To sum up, SVMs represent a powerful machine learning technique for object classification and regression analysis, and they offer state-of-the-art performance. However, the training of SVM is computationally expensive and relies on optimization.

15.4.2 Local Techniques

Advanced mapping techniques described above are intended to embed global structure of input data into the space of low dimensionality. In contrast to this formulation, local nonlinear methods of dimensionality reduction are based

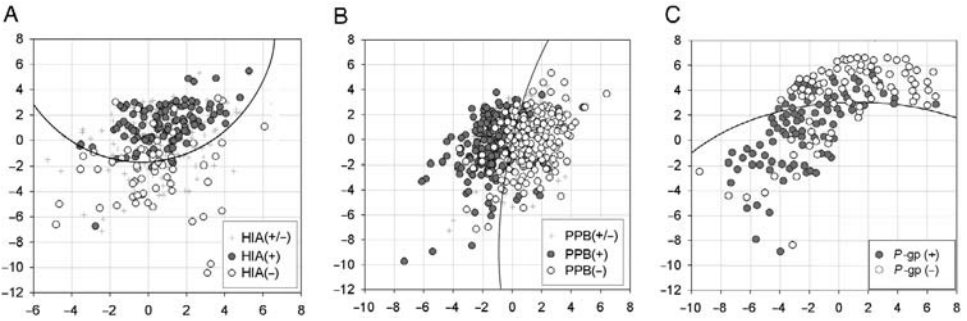


Figure 15.4 Nonlinear Sammon models developed for compounds with different ADME properties supported by SVM classification: (A) human intestinal absorption, (B) plasma protein binding affinity, and (C) *P*-gp substrate efficacy. HIA = human intestinal absorption; PPB = plasma protein binding.

solely on preserving structure of small neighborhood of each data point. The key local techniques include LLE, Laplacian eigenmaps, HLLE, and LTSA.

Of note, local techniques for dimensionality reduction can be freely viewed in the context of specific local kernel functions for KPCA. Therefore, these techniques can be cleverly redefined using the KPCA framework [87,88].

15.4.2.1 LLE LLE is a simple local technique for dimensionality reduction, which is generally similar to IsoMap algorithm in relying on the nearest neighborhood graph representation of input data points [12]. In contrast to IsoMap, LLE attempts to preserve solely the local structure of multivariate data. Therefore, the algorithm is much less sensitive to short-circuiting than IsoMap. Furthermore, the complete preservation of local properties often leads to successful embedding of nonconvex manifolds.

In formal terms, LLE tries to modify the local properties of the manifold around the processed data sample x_i by representing the data point as a linear combination w_i (the so-called reconstruction weight coefficients) of its k-nearest neighbor x_{ij}. Hence, using the data point x_i and a set of its nearest neighbors, LLE fits a hyperplane, making the bold assumption that the manifold is locally linear (the reconstruction weight w_i of the data point x_i is completely invariant to space rotation and translation, as well as rescaling). Due to the invariance, any linear mapping of the hyperplane into a low-dimensional space faithfully preserves the reconstruction weights within the space of lower dimensionality. In other words, if the local topology and geometry of the manifold are largely preserved in low-dimensional data representation, the reconstruction weight coefficient w_i, which explicitly manipulates the data point x_i and its adjacent array in high-dimensional data representation, also reconstructs the data point y_i from its neighbors in low-dimensional space.

The main idea of LLE is to find the optimal D-dimensional data representation F by adaptive minimization of the cost function: $\phi(F) = \sum_i \left(y_i - \sum_{j=1}^{k} w_{ij} y_{ij} \right)^2$. It can be clearly shown that $\phi(F) = (F - WF)^2 = F^T(I - W)^T(I - W)F$ is the common function that should to be further minimized. In this conventional formulation, I is the $n \times n$ identity matrix. Hence, the target coordinates of the low-dimensional representations y_i that minimize the cost function $\phi(F)$ can be easily found by computing the eigenvectors of $(I - W)^T(I - W)$ corresponding to the smallest d nonzero eigenvalues of the inproduct of $(I - W)$ from the solution set.

LLE has been successfully applied in various fields of data mining, including a super-resolution task [89] and sound source localization [90], as well as chemical data analysis [86]. Reportedly, LLE demonstrated poor performance in chemoinformatics. For example, this method has recently been reported to persistently fail in the visualization of even simple synthetic biomedical data sets [91]. It was also experimentally shown that in some cases, LLE performs

worse than IsoMap [92]. Probably, this may be attributed to the extreme sensitivity of LLE learning algorithm to "holes" in the manifolds [12].

15.4.2.2 HLLE HLLE is an advanced variant of the basic LLE technique that minimizes the curvilinear structure of high-dimensional manifold by embedding it into a low-dimensional space, making a hypothetical assumption that the low-dimensional data representation is locally isometric [8]. The basic principle of HLLE lies in the eigenanalysis of a matrix Ω that describes the curviness of the manifold determined around the processed data points, which is directly measured by means of the local Hessian. The key aspect of the local Hessian sack constructed in a local tangent space at the data point is invariance to differences in positions of the data points processed. It can be straightforwardly shown that the target low-dimensional coordinates can be easily found by performing an eigenanalysis of the core matrix Ω. The algorithm starts with identifying the k-nearest neighbors for each data point x_i based on Euclidean distance. Then, the local linearity of the manifold through the x_i nearest neighborhood is conservatively assumed. Hence, a principal basis describing a local tangent space at the data point x_i can be readily constructed using PCA performed across the k-nearest neighbor x_{ij}. In mathematical terms, a basis for the local tangent space for every data point x_i can be routinely determined by computing the d principal eigenvectors $M = \{m_1, m_2, \ldots, m_d\}$ of the covariance matrix $\operatorname{cov}|x_{ij} - \overline{x}_{ij}|$. It should be particularly noted that the above formulation strongly requires the following rigid restriction: $k \geq d$. Subsequently, an unbiased machine estimator for the Hessian sack of the manifold at point x_i in local tangent space coordinates is explicitly computed. For the practical realization of this computational task, the matrix Z_i is then meticulously formed. Containing (*in the columns*) all the cross products of M up to the dth order (*including a column with ones*), this matrix becomes orthonormalized after applying the Gram–Schmidt procedure. The expression of the tangent Hessian H_i can be further assayed by the transpose of the last $\frac{1}{2}d(d+1)$ columns of the orthonormalized matrix Z_i. Using Hessian estimators in local tangent coordinates, the core matrix Ω can then be easily constructed basedon Hessian entries $H_{im} = \sum_i \sum_j \left((H_i)_{jl} \times (H_i)_{jm}\right)$. Consequently, the target matrix contains information related to the curviness of high-dimensional data manifold. Thus, the eigenanalysis of the matrix Ω is performed mainly in order to find the low-dimensional data representation that appropriately minimizes the curviness of the manifold, while the eigenvectors corresponding to the d smallest nonzero eigenvalues of matrix Ω are selected and, in turn, construct the feature matrix Y, which contains a low-dimensional representation of the input data space.

15.4.2.3 Laplacian Eigenmaps Laplacian eigenmap algorithm [7] preserves local data structure by computing a low-dimensional representation of

the data in which the distances between a data point and its k-nearest neighbors are minimized as far as possible. To describe a local structure, the method uses a simple rule: The distance in the low-dimensional data representation between the data point and the first nearest neighbor contributes more to the cost function than the distance to the second nearest neighbor. Thus, the minimization of the cost function that can be formally defined as the key eigenproblem is effortlessly achieved in the context of spectral graph theory. Initially, the algorithm constructs the neighborhood graph G in which every data point x_i is directly connected to its k-nearest neighbors. Then, using the Gaussian kernel function, the weight of the edge can be easily computed for all the data points x_i and x_j constructing the graph G, thereby leading to a sparse adjacency matrix W. During the computation of the low-dimensional representation y_i, the core function can be strictly defined as $\phi(Y) = \sum_{ij} (y_i - y_j)^2 w_{ij}$, where the large weight w_{ij} corresponds to small distances between the processed data points x_i and x_j. Therefore, the potential difference between their low-dimensional representations y_i and y_j highly contributes to this cost function. As a consequence, nearby points in the high-dimensional space are also brought closer together in the low-dimensional representation.

In the context of the eigenproblem, the computation of the degree matrix M and the graph Laplacian L of the graph W jointly formulate the minimization task postulated before, so that the degree matrix M of W is a diagonal matrix, whose entries are the row sums of W (i.e., $m_{ij} = \sum_j w_{ij}$), whereas the graph Laplacian L can be easily computed using the following definition: $L = M - W$. Summarizing these basic postulates, the cost function can be further redefined in $\phi(Y) = \sum_{ij} (y_i - y_j)^2 w_{ij} = 2Y^T LY$. Therefore, the minimization of the $\phi(Y)$ can then be performed equivalently by minimizing the $Y^T LY$. Finally, for the d smallest nonzero eigenvalues, the low-dimensional data representation Y can subsequently be found by solving the generalized eigenvector problem defined as $Lv = \lambda Mv$. Summing up the aspects and advantages listed above, we can reasonably conclude that Laplacian eigenmaps represents at least not less powerful computational technique for low-dimensional data representation compared with LLE. It can be successfully applied in various fields of data mining, including chemoinformatics.

15.4.2.4 LTSA LTSA, a technique that is quite similar to HLLE, attempts to screen local properties within the high-dimensional data manifold using the local tangent space of each data point [9]. The fundamental principle of LTSA lies in the following statement: [being artificially restricted by the key assumption of local linearity of the manifold] there exists a linear mapping from a high-dimensional data point to its local tangent space; also, there exists a linear

mapping from the corresponding low-dimensional data point to the same local tangent space [9]. Thus, LTSA attempts to align these linear mappings in such a way that they construct a local tangent space of the manifold from a low-dimensional representation. In other words, the algorithm simultaneously searches for the feature coordinates of low-dimensional data representations as well as for the linear mappings of low-dimensional data points to the local tangent space of high-dimensional data. Similar to HLLE, the algorithm starts with computing specific bases (partly resembling Hessian sacks) for the local tangent spaces at data point x_i. This can be successfully achieved by applying PCA toward the k data point x_{ij} that are neighbors of the data point x_i results in a mapping (M_i) from the neighboring set of x_i to the local tangent space Ω_i. The most unique trait of this space lies in the existence of the linear mapping L_i from the local tangent space coordinates x_{ij} to the low-dimensional representations y_{ij}. Using this property, LTSA performs the following minimization: $\underbrace{\min}_{Y_i, L_i} \sum_i \|Y_i J_k - L_i \Omega_i\|^2$, where J_k is the centering matrix of size k [67]. It can be mathematically shown that the target solution of the posed minimization problem can be found readily using the eigenvectors of an alignment matrix B that correspond to the d smallest nonzero eigenvalues of B. For one turn, the components of the alignment matrix B can then be obtained as a result of iterative summation across all the matrices V_i starting from the initial values of $b_{ij} = 0$, for $\forall_{ij}$. It can also be shown that $BN_iN_i = BN_iN_i + J_k(I - V_iV_i^T)J_k$, where N_i is the selection matrix that contains the indices of the nearest neighbors around the data point x_i. Finally, the low-dimensional representation Y can be readily obtained by computation of the eigenvectors of the symmetric matrix $\frac{1}{2}(B + B^T)$ that correspond to the d smallest nonzero eigenvectors. LTSA have been successfully applied to solving various data mining tasks occurring widely in the chemoinformatic field, such as the analysis of protein microarray data [93].

15.4.2.5 Global Alignment of Linear Models In contrast to the sections presented previously, where we have willingly discussed two major approaches to construction of a low-dimensional data representation by preserving the global or local properties of input data, the current section briefly describes key mapping techniques widely used for performing the global alignment of linear models, computing the corresponding number of linear models, and constructing a low-dimensional data representation by aligning the linear models obtained.

Among a small number of methods targeted for the global alignment of linear models, LLC is a hugely promising technique broadly used for dimensionality reduction [10]. The bright idea of this method lies in computing the set of factor analyzers (see Section 16.3) by which the global alignment of the mixture of linear models can be subsequently achieved. The algorithm mainly proceeds in two principal steps: (1) computing the mixture of factor analyzers for the input data set by means of an expectation maximization

(EM) algorithm and (2) subsequent aligning of the constructed linear models in order to obtain a low-dimensional data representation using a variant of LLE. It should be especially noted that besides LLC, a similar technique called manifold charting has also been developed recently on the bias of this common principle [24].

Initially, LLC recruits a group of m factor analyzers using the EM algorithm [94]. Then, the obtained mixture outputs the local data representation y_{ij} and corresponding responsibility r_{ij} (where $j \in \{1, \ldots, m\}$) for every input data point x_i. In meticulous detail, the responsibility r_{ij} describes the connection between extent data point x_i and the linear model j, so it trivially satisfies $\sum_i r_{ij} = 1$. Using the set of estimated linear models and the corresponding responsibilities, responsibility-weighted data representations $w_{ij} = r_{ij}y_{ij}$ can be readily computed and stored in an $n \times mD$ block matrix W. The global alignment of the linear models is then performed based on matrix W and matrix M defined by $M = (I - F)^T(I - F)$, where F is the matrix containing the reconstruction weight coefficients produced by LLE (see Section 15.4.2.1), and I denotes the $n \times n$ identity matrix. LLC analyzes a set of linear models by solving the generalized eigenproblem $Av = \lambda Bv$ for the d smallest nonzero eigenvalues. In this equation, A denotes the inproduct of M^TW, whereas B denotes the inproduct of W. It can easily be shown that d eigenvector v_i computed from the matrix L uniquely defines a linear mapping projection from the responsibility-weighted data representation W to the underlying low-dimensional data representation Y. Finally, the low-dimensional data representation can be obtained immediately by computing the following equation: $Y = WL$.

REFERENCES

1. Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 1901;2:559–572.
2. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugen* 1936;7:179–188.
3. Gorsuch RL. Common factor analysis versus component analysis: Some well and little known facts. *Multivariate Behav Res* 1990;25:33–39.
4. Mika S, Scholkopf D, Smola AJ, Muller K-R, Scholz M, Ratsch G. Kernel PCA and de-noising in feature spaces. In: *Advances in Neural Information Processing Systems*, edited by Kearns MS, Solla S, Cohn D, pp. 536–542. Cambridge, MA: The MIT Press, 1999.
5. Nadler B, Lafon S, Coifman RR, Kevrekidis IG. Diffusion maps, spectral clustering and the reaction coordinates of dynamical systems. *Appl Comput Harm Anal* 2006;21:113–127.
6. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006;313:504–507.

7. Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: *Advances in Neural Information Processing Systems*, edited by Dietterich TB, Becker S, Ghahramani Z, Vol. 14, pp. 585–591. Cambridge, MA: The MIT Press, 2002.
8. Donoho DL, Grimes C. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. *Proc Natl Acad Sci USA* 2005;102:7426–7431.
9. Zhang Z, Zha H. Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment. *SIAM J Sci Comput* 2004;26:313–338.
10. Teh YW, Roweis ST. Automatic alignment of hidden representations. In: *Advances in Neural Information Processing Systems*, edited by Becker S, Thrun S, Obermayer K, Vol. 15, pp. 841–848. Cambridge, MA: The MIT Press, 2002.
11. Cox T, Cox M. *Multidimensional Scaling*. London: Chapman & Hall, 1994.
12. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science* 2000;290:2323–2326.
13. Vapnik V, Golowich S, Smola A. Support vector method for function approximation, regression estimation, and signal processing. *Adv Neural Inf Process Syst* 1996;9:281–287.
14. Bell AJ, Sejnowski TJ. An information maximization approach to blind separation and blind deconvolution. *Neural Comput* 1995;7:1129–1159.
15. Chang K-Y, Ghosh J. Principal curves for non-linear feature extraction and classification. In: *Applications of Artificial Neural Networks in Image Processing III*, edited by Nasrabadi NM, Katsaggelos AK, Vol. 3307, pp. 120–129. Bellingham, WA: SPIE, 1998.
16. Demartines P, Herault J. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans Neural Netw* 1997;8: 148–154.
17. Baudat G, Anouar F. Generalized discriminant analysis using a kernel approach. *Neural Comput* 2000;12:2385–2404.
18. Suykens JAK. Data visualization and dimensionality reduction using kernel maps with a reference point. Internal Report 07–22; ESAT-SISTA, Leuven, Belgium, 2007.
19. Weinberger KQ, Packer BD, Saul LK. Nonlinear dimensionality reduction by semi-definite programming and kernel matrix factorization. Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS-05), Barbados, West Indies, January 10, 2005.
20. Sha F, Saul LK. Analysis and extension of spectral methods for nonlinear dimensionality reduction. In: *Proceedings of the 22nd International Conference on Machine Learning*, Vol. 119, pp. 784–791. New York: ACM, 2005.
21. He X, Niyogi P. Locality preserving projections. In: *Advances in Neural Information Processing Systems*, Vol. 16, edited by Thrun S, Saul LK, Schulkopf B, p. 37. Cambridge, MA: The MIT Press, 2004.
22. Zhang T, Yang J, Zhao D, Ge X. Linear local tangent space alignment and application to face recognition. *Neurocomputing* 2007;70:1547–1533.
23. Faloutsos C, Lin K-I. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In: *Proceedings of 1995 ACM SIGMOD*, May 1995, San Jose, CA, pp. 163–174. New York: ACM, 1995.

24. Brand M. From subspaces to submanifolds. Proceedings of the 15th British Machine Vision Conference, London, 2004.
25. Brand M. Charting a manifold. In: *Advances in Neural Information Processing Systems*, edited by Becker S, Thrun S, Obermayer K, Vol. 15, pp. 985–992. Cambridge, MA: The MIT Press, 2002.
26. Roweis ST, Saul L, Hinton G. Global coordination of local linear models. In: *Advances in Neural Information Processing Systems*, edited by Dietterich TG, Becker S, Ghahramani Z, Vol. 14, pp. 889–896. Cambridge, MA: The MIT Press, 2001.
27. Verbeek J. Learning nonlinear image manifolds by global alignment of local linear models. *IEEE Trans Pattern Anal Mach Intell* 2006;28:1236–1250.
28. Kohonen T. The self-organizing map. *Proc IEEE* 1990;78:1464–1480.
29. Sammon JW. A non-linear mapping for data structure analysis. *IEEE Trans Comput* 1969;C-18:401–409.
30. Tenenbaum JB. Mapping a manifold of perceptual observations. In: *Advances in Neural Information Processing Systems*, edited by Jordan MI, Kearns MJ, Solla SA, Vol. 10, pp. 682–688. Cambridge, MA: The MIT Press, 1998.
31. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. *Science* 2000;290:2319–2323.
32. Agrafiotis DK. Stochastic algorithms for maximizing molecular diversity. *J Chem Inf Comput Sci* 1997;37:841–851.
33. Hotelling H, Educ J. Analysis of a complex of statistical variables into principal components. Part I. *J Educ Psychol* 1933;24:417–441.
34. Borg I, Groenen PJF. *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer, 1997.
35. Balakin KV, Ivanenkov YA, Savchuk NP, Ivaschenko AA, Ekins S. Comprehensive computational assessment of ADME properties using mapping techniques. *Curr Drug Discov Technol* 2005;2:99–113.
36. Yamashita F, Itoh T, Hara H, Hashida M. Visualization of large-scale aqueous solubility data using a novel hierarchical data visualization technique. *J Chem Inf Model* 2006;46:1054–1059.
37. Engels MFM, Gibbs AC, Jaeger EP, Verbinnen D, Lobanov VS, Agrafiotis DK. A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition. *J Chem Inf Model* 2006;46: 2651–2660.
38. Bernard P, Golbraikh A, Kireev D, Chrétien JR, Rozhkova N. Comparison of chemical databases: Analysis of molecular diversity with self organizing maps (SOM). *Analusis* 1998;26:333–341.
39. Kirew DB, Chrétien JR, Bernard P, Ros F. Application of Kohonen neural networks in classification of biologically active compounds. *SAR QSAR Environ Res* 1998;8:93–107.
40. Ros F, Audouze K, Pintore M, Chrétien JR. Hybrid systems for virtual screening: Interest of fuzzy clustering applied to olfaction. *SAR QSAR Environ Res* 2000; 11:281–300.
41. Jain AK, Murty MN, Flynn PJ. Data clustering: A review. *ACM Computing Surveys* 1999;31:264–323.

42. Brown RD, Martin YC. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J Chem Inf Comput Sci* 1996;36:572–584.
43. Bocker A, Derksen S, Schmidt E, Teckentrup A, Schneider G. A hierarchical clustering approach for large compound libraries. *J Chem Inf Model* 2005;45: 807–815.
44. Herman I. Graph visualization and navigation in information visualization: A survey. *IEEE Trans Vis Comput Graph* 2000;6:24–43.
45. Strehl A, Ghosh J. Relationship-based clustering and visualization for high-dimensional data mining. *INFORMS J Comput* 2003;15:208–230.
46. Agrafiotis DK, Bandyopadhyay D, Farnum M. Radial clustergrams: Visualizing the aggregate properties of hierarchical clusters. *J Chem Inf Model* 2007;47:69–75.
47. Agrafiotis DK. Diversity of chemical libraries. In: *The Encyclopedia of Computational Chemistry*, edited by Schleyer PR, Vol. 1, pp. 742–761. Chichester, U.K.: John Wiley & Sons, 1998.
48. Agrafiotis DK, Lobanov VS, Salemme FR. Combinatorial informatics in the post-genomics era. *Nat Rev Drug Discov* 2002;1:337–346.
49. Agrafiotis DK, Rassokhin DN, Lobanov VS. Multidimensional scaling and visualization of large molecular similarity tables. *J Comput Chem* 2001;22:488–500.
50. Lajiness MS. An evaluation of the performance of dissimilarity selection. In: *QSAR: Rational Approaches to the Design of Bioactive Compounds*, edited by Silipo C, Vittoria A, pp. 201–204. Amsterdam: Elsevier, 1991.
51. Hassan M, Bielawski JP, Hempel JC, Waldman M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol Divers* 1996;2:64–74.
52. Taylor R. Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. *J Chem Inf Comput Sci* 1995;35:59–67.
53. Boyd SM, Beverly M, Norskov L, Hubbard RE. Characterizing the geometrical diversity of functional groups in chemical databases. *J Comput Aided Mol Des* 1995;9:417–424.
54. Mount J, Ruppert J, Welch W, Jain A. IcePick: a flexible surface-based system for molecular diversity. *J Med Chem* 1999;42(1):60–66.
55. Goulon A, Duprat A, Dreyfus G. Graph machines and their applications to computer-aided drug design: A new approach to learning from structured data. In: *Lecture Notes in Computer Science*, edited by Calude CS, Vol. 4135, pp. 1–19. Berlin: Springer, 2006.
56. Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, O'Connell JX, Zhu S, Fero M, Sherlock G, Pollack JR, Brown PO, Botstein D, van de Rijn M. Molecular characterisation of soft tissue tumours: A gene expression study. *Lancet* 2002;359: 1301–1307.
57. Cooley W, Lohnes P. *Multivariate Data Analysis*. New York: Wiley, 1971.
58. Jolliffe IT. *Principal Component Analysis*, 2nd edn. New York: Springer-Verlag, 2002.
59. Martin EJ, Blaney JM, Siani MA, Spellmeyer DC, Wong AK, Moos WH. Measuring diversity: Experimental design of combinatorial libraries for drug discovery. *J Med Chem* 1995;38:1431–1436.

60. Gibson S, McGuire R, Rees DC. Principal components describing biological activities and molecular diversity of heterocyclic aromatic ring fragments. *J Med Chem* 1996;39:4065–4072.

61. Hotelling H. Analysis of a complex of statistical variables into principal components. Part II. *J Educ Psychol* 1933;24:498–520.

62. Das PD, Moll M, Stamati H, Kavraki LE, Clementi C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc Natl Acad Sci USA* 2006;103:9885–9890.

63. Cummins DJ, Andrews CW, Bentley JA, Cory M. Molecular diversity in chemical databases: Comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *J Chem Inf Comput Sci* 1996;36:750–763.

64. Scholkopf B, Smola AJ, Muller K-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput* 1998;10:1299–1319.

65. Shawe-Taylor J, Christianini N. *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge University Press, 2004.

66. Lima A, Zen H, Nankaku Y, Miyajima C, Tokuda K, Kitamura T. On the use of kernel PCA for feature extraction in speech recognition. *IEICE Trans Inf Syst* 2004;E87-D:2802–2811.

67. Hoffmann H. Kernel PCA for novelty detection. *Pattern Recognit* 2007;40:863–874.

68. Tomé AM, Teixeira AR, Lang EW, Martins da Silva A. Greedy kernel PCA applied to single-channel EEG recordings. *IEEE Eng Med Biol Soc* 2007;2007:5441–5444.

69. Tipping ME. Sparse kernel principal component analysis. In: *Advances in Neural Information Processing Systems*, edited by Leen TK, Dietterich TG, Tresp V, Vol. 13, pp. 633–639. Cambridge, MA: The MIT Press, 2000.

70. Lafon S, Lee AB. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Trans Pattern Anal Mach Intell* 2006;28:1393–1403.

71. Kung SY, Diamantaras KI, Taur JS. Adaptive principal component EXtraction (APEX) and applications. *IEEE Trans Signal Process* 1994;42:1202–1217.

72. Hinton GE, Osindero S, The Y. A fast learning algorithm for deep belief nets. *Neural Comput* 2006;18:1527–1554.

73. Raymer ML, Punch WF, Goodman ED, Kuhn LA, Jain AK. Dimensionality reduction using genetic algorithms. *IEEE Trans Evol Comput* 2000;4:164–171.

74. Johnson MA, Maggiora GM. *Concepts and Applications of Molecular Similarity*. New York: Wiley, 1990.

75. Torgeson WS. Multidimensional scaling: I. Theory and method. *Psychometrika* 1952;17:401–419.

76. Kruskal JB. Non-metric multidimensional scaling: A numerical method. *Phychometrika* 1964;29:115–129.

77. Tagaris GA, Richter W, Kim SG, Pellizzer G, Andersen P, Ugurbil K, Georgopoulos AP. Functional magnetic resonance imaging of mental rotation and memory scanning: A multidimensional scaling analysis of brain activation patterns. *Brain Res* 1998;26:106–12.

78. Venkatarajan MS, Braun W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physicalchemical properties. *J Mol Model* 2004;7:445–453.

79. Hinton GE, Roweis ST. Stochastic neighbor embedding. In: *Advances in Neural Information Processing Systems*, edited by Becker S, Thrun S, Obermayer K, Vol. 15, pp. 833–840. Cambridge, MA: The MIT Press, 2002.

80. Shepard RN, Carroll JD. Parametric representation of nonlinear data structures. In: *International Symposium on Multivariate Analysis*, edited by Krishnaiah PR, pp. 561–592. New York: Academic Press, 1965.

81. Martinetz T, Schulten K. Topology representing networks. *Neural Netw* 1994;7: 507–522.

82. Agrafiotis DK, Lobanov VS. Nonlinear mapping networks. *J Chem Info Comput Sci* 2000;40:1356–1362.

83. Vapnik V. *Estimation of Dependences Based on Empirical Data.* New York: Springer-Verlag, 1982.

84. Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

85. Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J Chem Inf Comput Sci* 2003;43:2048–2056.

86. L'Heureux PJ, Carreau J, Bengio Y, Delalleau O, Yue SY. Locally linear embedding for dimensionality reduction in QSAR. *J Comput Aided Mol Des* 2004;18: 475–482.

87. Bengio Y, Delalleau O, Le Roux N, Paiement J-F, Vincent P, Ouimet M. Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Comput* 2004;16:2197–2219.

88. Ham J, Lee D, Mika S, Scholkopf B. A kernel view of the dimensionality reduction of manifolds. Technical Report TR-110; Max Planck Institute for Biological Cybernetics, Tubingen, Germany, 2003.

89. Chang H, Yeung D-Y, Xiong Y. Super-resolution through neighbor embedding. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 275–282. Los Alamitos, CA: IEEE Computer Society, 2004.

90. Duraiswami R, Raykar VC. The manifolds of spatial hearing. In: *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Vol. 3, pp. 285–288. Philadelphia, PA, 2005.

91. Lim IS, Ciechomski PH, Sarni S, Thalmann D. Planar arrangement of high-dimensional biomedical data sets by Isomap coordinates. In: *Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems*, edited by Fyfe C, pp. 50–55. New York: Mount Sinai Medical School, 2003.

92. Jenkins OC, Mataric MJ. Deriving action and behavior primitives from human motion data. In: *International Conference on Intelligent Robots and Systems*, Vol. 3, edited by Deng Z, Neumann U, pp. 2551–2556, Lausanne, Switzerland, 2002.

93. Teng L, Li H, Fu X, Chen W, Shen I-F. Dimension reduction of microarray data based on local tangent space alignment. In: *Proceedings of the 4th IEEE*

International Conference on Cognitive Informatics, pp. 154–159. Washington, DC: IEEE Computer Society, 2005.

94. Ghahramani Z, Hinton GE. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1; Department of Computer Science, University of Toronto, 1996.

16

ADVANCED ARTIFICIAL INTELLIGENCE METHODS USED IN THE DESIGN OF PHARMACEUTICAL AGENTS

YAN A. IVANENKOV AND LUDMILA M. KHANDAROVA

Table of Contents

16.1 INTRODUCTION

In silico pharmacology is a growing scientific area that broadly covers the development of various computational techniques for capture, analysis, and integration of the biologic and medical data from various diverse sources [1,2]. The key advantage of these computational methods is the possibility to sig-

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

nificantly increase the number of potentially active molecules selected from databases as compared with a simple random selection [3]. Such models are currently routinely used for discovery and further optimization of novel compounds with specific activity toward different biologic targets; for addressing absorption, distribution, metabolism, and excretion (ADME) issues; cellular or organ-specific toxicity; physicochemical characterization; and so on. A variety of advanced computational algorithms and methods have been effectively applied recently in medicinal chemistry for dimensionality reduction and visualization of the chemical data of different types and structure [4], for example, in diversity analysis [5,6] and quantitative structure activity relationship (QSAR) modeling [7,8]. The majority of these computational models are commonly based on the basic principles of dimensionality reduction and mapping. In turn, dimensionality reduction is an essential computational technique for the analysis of a large-scale, streaming and tangled data.

Humans can visualize very complex data to differing degrees depending upon individual memory. However, probably since prehistoric times, humans have also relied on maps to visualize very complex coordinates and topologies, as well as their relationship to the world. It is only since relatively recent times that we have turned our attention to mapping the universe at the molecular scale, and specifically determining the different molecules that inhabit this "space" [9–11]. Starting with the molecules themselves is just the beginning as one would also need to consider the physicochemical properties and the interactions with different biologic systems, an incredibly complex and overwhelming amount of information. The classical methods of dimensionality reduction, for example, principal component analysis (PCA) and multidimensional scaling (MDS), are not specifically adapted for large data sets and "straight" mapping. Therefore, there is a growing interest in novel soft-computing approaches that might be applicable to the analysis of such data sets providing a comprehensive visualization. As we shall see, some advanced mapping methods derive from our understanding of the neural networks involved in image perception by the primary visual cortex of the human brain.

16.2 ADVANCED COMPUTATIONAL TECHNIQUES FOR CHEMICAL DATA MINING

16.2.1 Preamble

Among the various dimensionality reduction techniques that have been recently described in the scientific literature, nonlinear and self-organizing mappings (SOMs) are the unique techniques due to their conceptual simplicity and ability to effectively reproduce the topology of the input data space in a faithful and unbiased manner. The first method was initially designed to reproduce high-dimensional coordinates to the space of relatively low dimension based on distance measurement or similarity matrix, whereas the self-

organizing methodology implicitly uses the basic neural network principles, which can be successfully applied to construct a visual abstraction by means of rapid prototyping. In the first technique, the dimensionality reduction is generally achieved by reconstructing a low-dimensional coordinate set directly computed from a higher-dimensional representation and stored in the distance matrix; in the latter one, the original property vectors are mapped onto a two- or three-dimensional cell array arranged in a way that preserves the internal topology, whole structure, and density distribution of the original data set. Such representations can be successfully used for a variety of pattern recognition and classification tasks, including *in silico* drug design.

16.2.2 Nonlinear Sammon Mapping

Among various approaches extensively applied in modern computational chemistry, molecular similarity is one of the most ubiquitous concepts [12]. This technique is widely used to analyze and categorize the chemical data of different types, rationalize the behavior and functions of organic molecules, and design novel chemical compounds with improved physical, chemical, and biologic properties. Usually, for the analysis of large collections of organic compounds, structural similarities can be uniquely defined by the symmetric matrix that contains all the pairwise relationships among the molecules presented in the external data set. However, it should be noted that such a pairwise similarity metric is not generally acceptable for numerical processing and visual inspection. A reasonable, workable solution to this methodological problem lies in embedding the input objects into a low-dimensional Euclidean space in a way that preserves the original pairwise proximities as faithfully as possible. There are at least two basic approaches, MDS and nonlinear mapping (NLM), that effectively convert the input data points into a set of feature vectors that can subsequently be used for a variety of pattern recognition and classification tasks.

NLM is an advanced machine learning technique for improved data mining and visualization. This method, originally introduced by Sammon [13], represents a multivariate statistical technique closely related to MDS. Just like MDS, the main objective of the Sammon approach is to approximate local geometric and topological relationships on a visually intelligible two- or three-dimensional plot, whereas the fundamental idea of this method is to substantially reduce the high dimensionality of the initial data set into the low-dimension feature space, regardless of the number of dimensions from which it is constructed.

NLM can be metric or nonmetric, and is therefore equally applicable to a wide variety of input data. The basic difference between MDS and NLM is in the minimization procedure. The classical Sammon algorithm attempts to closely approximate global geometric relationships observed across the whole space of input vector samples basically in a two- or three-dimensional representation. Sammon mapping strongly resembles the MDS algorithm;

the process starts from a given finite set of n N-dimensional vector samples: $\{x_i,\ i = 1,2,\ \ldots\ ,\ k;\ x_i \in \mathcal{R}^N\}$. A distance function between input data points x_i and x_j is randomly selected in the initial space then is simply calculated by $d_{ij}^* = d(x_i, x_j)$; a target set of n images of x_i projected onto the M-dimensional feature space F: $\{y_i, i = 1,2, \ldots, k; y_i \in \mathcal{R}^M\}$ and a distance function between feature vectors y_i and y_j are also calculated by $d_{ij} = d(y_i, y_j)$. For conceptual distance measurement, several space metrics can be effectively used, such as Euclidean or Manhattan distances. The main idea of Sammon mapping is to optimally arrange the feature images y_i within the whole display feature plane in such a way that their Euclidean distances $d_{ij} = d(y_i, y_j)$ approximate as closely as possible to the corresponding original values $d_{ij}^* = d(x_i, x_j)$. This projection, which can only be made approximately, can be successfully carried out in an iterative fashion by minimizing the error function, $E(k)$, which in turn thoroughly estimates the deviation between similarity distances calculated for the original and projected/feature vector sets: $E(k) = \frac{1}{c}\sum_{i<j}^{n} \frac{\left[d_{ij}^* - d_{ij}\right]^2}{d_{ij}^*}$, where k is the iteration step, $c = \sum_{i<j}^{n} d_{ij}^*$, and $d_{ij} = \left(\sqrt{\sum_{k=1}^{M}[y_{ik} - y_{jk}]^2}\right)$. By direct analogy with the classical Newton minimization scheme, $E(k)$ is further minimized using both the instantaneous gradient and steepest descent algorithms. The minimization of this criterion commonly leads to a two- or three-dimensional plot, where the interobject distances are as similar as possible to the original distances. Thus, the initially (randomly) placed coordinate y_i is then iteratively updated following the gradient function $y_{pq}(k + 1) = y_{pq} - \eta\Delta_{pq}(k)$, where η is the learning rate parameter, and $\Delta_{pq}(k) = \frac{\frac{\partial E}{\partial y_{pq}}}{\left|\frac{\partial^2 E}{\partial y_{pq}^2}\right|}$ is the quotient of the division of the corresponding gradient component on the diagonal Hessian element determined on the kth iteration. With due respect to applying the stress function $E(k)$, corresponding gradient and Hessian components can be cleverly redefined in $\frac{\partial E}{\partial y_{pq}} = -\frac{2}{c}\sum_{j=1, j\neq p}^{n}\left[\frac{d_{pj}^* - d_{pj}}{d_{pj}d_{pj}^*}\right][y_{pq} - y_{jq}]$, whereas

$$\frac{\partial^2 E}{\partial y_{pq}^2} = -\frac{2}{c}\sum_{j=1, j\neq p}^{n}\frac{1}{d_{pj}d_{pj}^*} \times \left[\left(d_{pj}^* - d_{pj}\right) - \frac{(y_{pq} - y_{jq})^2}{d_{pj}}\left(1 + \frac{d_{pj}^* - d_{pj}}{d_{pj}}\right)\right].$$

The Sammon algorithm is arguably the most commonly used approach for accurate dimensionality reduction, but the main problem arising from the aforementioned techniques is that it too does not scale well with the size of the input data set. Although nonlinear scaling becomes more problematic as

the original dimensionality of the input space increases, the internal structure and topology of the data are very frequently reflected successfully on the resulting map. Several attempts have recently been undertaken to reduce the complexity and difficulty of the task. For example, Chang and Lee [14] proposed a heuristic relaxation approach in which a learning subject of the original objects (the frame) is progressively scaled using a Sammon-like methodology, and the remaining objects are then placed to the map by adjusting their distances to the frame objects already embedded into a low-dimensional feature plane. Several other modifications were also introduced and validated [15–17].

To resolve this unwanted problem, a new variant of the original Sammon algorithm was recently developed by Agrafiotis [5] based on the combined self-organized and nonlinear principles. The method belongs to the family of nonmetric algorithms, and therefore, it can be equally applicable to a wide variety of input data. Thus, it is especially useful when the (dis)similarity measure is not a true metric, i.e., it does not obey the fundamental distance postulates and, in particular, the triangle inequality, such as the Tanimoto coefficient. Although an "exact" projection is only possible when the distance matrix is positive, meaningful projections can be readily obtained even when this criterion is not completely satisfied. In this case, the quality of the space approximation is generally determined by a sum-of-squares error function such as Kruskal's stress: $\zeta = \sqrt{\dfrac{\sum_{i<j}^{k}\left(d_{ij}^{*} - d_{ij}\right)^2}{\sum_{i<j}^{k} d_{ij}^{2}}}$. The principal advantage of NLM over the Kohonen network is that they often provide much greater detail about the individual compounds and the corresponding interrelationships as demonstrated by the following example. The target projection was carried out entirely using a set of 12-dimensional autocorrelation descriptors and the Euclidean metric as a pairwise measure of dissimilarity among the examined structures including xanthene, cubane, and adamantane libraries [18]. The resulting two-dimensional plane is shown in Figure 16.1.

It is also quite possible to use a multilayer backpropagation neural network with n input and m output neurons ($m = 2, 3$). In this case, a nonlinear output can be directly used as neural net input resembling hybrid neural nets [19]. Results of numerical artificial simulation and real data show that the proposed technique is a promising approach to visualize multidimensional clusters by mapping the multidimensional data into a perceivable low-dimensional space. More recently, Agrafiotis and Lobanov slightly modified this method [20]. Instead of using the full data set, they have suggested to train a feedforward neural network to learn the projection obtained from conventional NLM of a subset of the total data. The trained network can subsequently be used to approximate the whole compound set. Based on modification of the key similarity function, several variants of the basic Sammon algorithm were recently

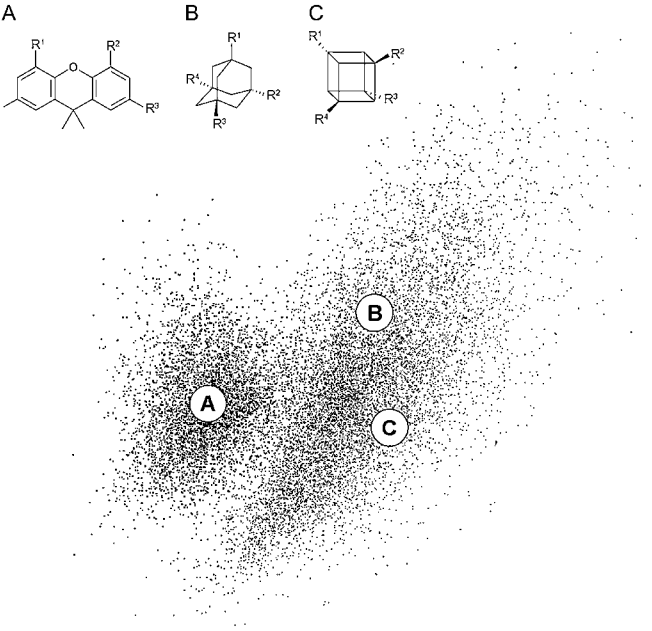


Figure 16.1 A nonlinear projection of the xanthene (A), admantane (B), and cubane (C) libraries.

developed by Agrafiotis [5]. As a result, a new family of projection algorithms that cleverly combine a stochastic search engine with a user-defined objective function that encodes any desirable selection criterion was developed and experimentally evaluated.

A wide number of different statistical problems can be effectively solved using the Sammon methodology. For example, a new method for analyzing protein sequences was also introduced by Agrafiotis [21] based on the Sammon NLM algorithm When applied to a family of homologous sequences, this method is quite able to capture the essential features of the similarity matrix and provides a faithful representation of chemical or evolutionary distance in a simple and intuitive way. The key merits of the new algorithm were clearly demonstrated using examples from the protein kinase family. This algorithm was also investigated as a means of visualizing and comparing large compound collections, represented generally by a set of various molecular descriptors [22].

Thus, it can be objectively concluded that the NLM strategy is very useful to represent a multidimensional data distribution in an intuitively intelligible manner. Unlike PCA, NLM preserves the spatial relationships among all the objects studied. However, NLM cannot be directly used to predict the position of new external objects because each axis of the constructed plot is not stationary and represents *per se* a nonlinear combination of the original variables.

Furthermore, the projection onto the two- or three-dimensional plane only makes sense if a majority of the variance is contained in few dimensions.

16.2.3 Self-Organizing Kohonen Maps

At least two different methods of self-organizing neural-based mapping are currently applied for dimensionality reduction as well as feature selection and topographic structure representation. The fundamental conception of self-organizing methodology initially originates from the experiments related to the investigation of the mechanism of image construction into realistic primary visual cortex of the human brain. Willshaw and von der Malsburg were pioneers in this field who developed one of the first computational models in which artificial neurons were tightly packed into the two interrelated lattices (Fig. 16.2A) [23].

As shown in Figure 16.2A, the "input" lattice is projected ingeniously onto the second two-dimensional plane by the corresponding synaptic route of weight coefficients. The first lattice is simply constructed by the presynaptic neurons, while the second lattice consists of postsynaptic neurons, which are not formally assigned in accordance with the common principle—"*winner takes all* (WTA)." Following both the *short-* and *long-range inhibitory mechanisms*, neuron weights attached to the postsynaptic surface are adjusted iteratively by the Hebb learning rule until the optimal values are reached. As a result, the increase of one synaptic weight directly leads to the decrease of others. Finally, it should be especially noted, that the described model is applicable solely for pattern recognition when the dimension of the input signal correlates closely to the dimension of the output feature image.

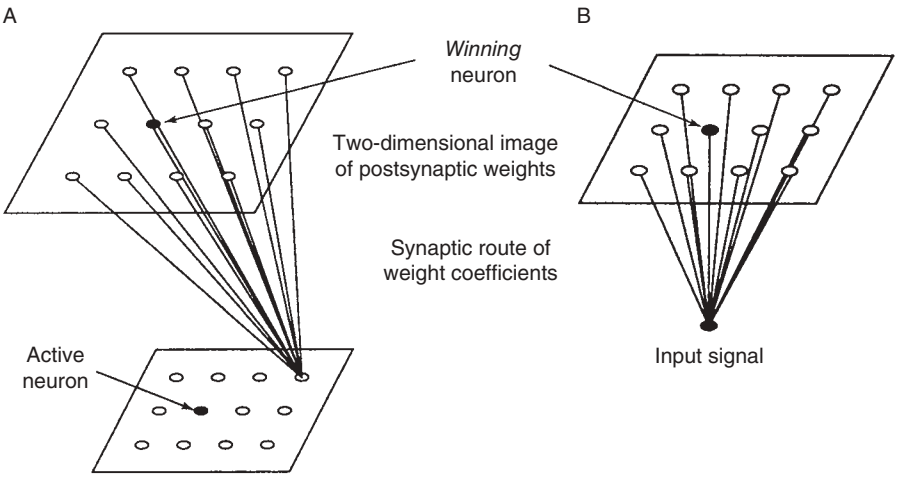


Figure 16.2 (A) A Willshaw–Malsburg's model of the self-organizing mapping; (B) a Kohonen-based approach for dimensionality reduction.

The second fundamental methodology of neural mapping based on the self-organizing strategy is schematically outlined in Figure 16.2B. This model, originally introduced by Teuvo Kohonen in 1988 [24], allows one to construct a low-dimensional topological representation of a high-dimensional data set by the optimal fixed amount of codebook feature vectors. Based primarily on the *vector quantization* (VQ) strategy, these weight vectors are adjusted iteratively to the components of the input vector objects producing an intuitively comprehensible two- or three-dimensional topological map. The SOM is one of the most popular and widely used neural network architectures. It is a powerful tool for visualization and data analysis that can be successfully applied in various scientific domains. Due to the limited space in the current chapter, we cannot dwell on this algorithm in too much detail. Briefly, the fundamental idea of SOMs lies in embedding a set of vector samples onto a two- or three-dimensional lattice in a way that preserves the relative topology and cluster structure of the original high-dimensional space. In the output, samples that are located close to each other in the input space should be closely embedded in the topologically isomorphic resulting space. Initially, all the Kohonen neurons receive identical input, and by means of lateral interactions, they compete among themselves.

The SOM algorithm has attracted a great deal of interest among researches in a wide variety of scientific fields. To the present day, a number of variants and different modifications of SOM have been developed and, perhaps most importantly, it has been extensively applied in various scientific disciplines ranging from engineering sciences to chemistry, medicine, biology, economics, and finance. From a biologic point of view, the Kohonen network is also biologically plausible just as in Willshaw–Malsburg's model [25,26]. In nature, the original prototype of this model is neatly presented in various brain structures to provide an ordered low-dimensional internal representation of the external complex information flow. Thus, self-organizing Kohonen maps were originally designed as an attempt to model intelligent information processing, i.e., the ability of the brain to form reduced representations of the most relevant facts and observations without significant loss of information about their interrelationships and a common topology. From a functional point of view, SOM resembles closely the VQ algorithm previously described by Linde et al. [27], which accurately approximates, in an unsupervised way, the probability density functions of a vector of input variables by a finite set of reference vectors with the sole purpose of providing a low-dimensional data representation by using a nearest-neighbor rule. In the Kohonen network, a learning process is firmly based on unsupervised logic then the target residential property is not considered within the training procedure. In contrast to supervised neural networks, SOM neurons are homogeneously arranged within the space spanned by a regular grid composed of many processing units (Kohonen neurons) in which the adaptation/learning process is generally performed by some predefined neighborhood rules. It produces both an iterative quantization of codebook vectors and an ordered representation of the original input data distribution.

In addition, since each neuron has well-defined, low-dimensional coordinates over the whole Kohonen lattice, SOM can also be properly considered as a backfitting two- or three-dimensional projection algorithm.

In an unsupervised environment, self-organizing maps can reasonably be assigned to a class of neural networks that are commonly based on a competitive learning principle also widely known as a self-organizing methodology or network [28]. In the basic variant of SOM, high-dimensional data are cleverly mapped onto a two-dimensional rectangular or hexagonal lattice of neurons in such a way as to preserve the internal topology and cluster structure of the original input space. The mapping implemented by the SOM algorithm can be mathematically formalized in the following manner: assume that the initial set of the input variables $\{x\}$ is formally defined as a real input vector $x = [x_1, x_2, \ldots, x_N]^T \in \mathcal{R}^N$ and that each element located in the SOM array is directly associated with the parametric reference vector (*synaptic weight vector*): $w = [w_1, w_2, \ldots, w_N]^T \in \mathcal{R}^N$. As a role, some predefined arbitrary values assigned to the initial synaptic weight coefficients are initially randomly generated (they should, however, not be too different from the data values to facilitate the convergence of the training process). The network is then continuously trained in an iterative fashion until a predefined threshold value is achieved or the final learning epoch is completed. Usually, a randomly chosen training sample, x_i is directly presented to the Kohonen network in a random order, then the metric distance from each neuron is readily computed. After the competition is over, the neuron that appears to be the closest to the input data sample is uniquely assigned to the "winning neuron" following the fundamental self-organizing principle—WTA. Subsequently, the weight vectors of this neuron are optimally adapted to the input sample. After the learning cycle is complete, each data point is presented again to the network, and the matching neuron is also determined in the same manner. Thus, during the competitive learning, the majority of Kohonen neurons (ideally, each neuron) being completely tuned to the different domains dispersed irregularly or systematically within the input space, and acts as a specific decoder of such domains. This process is repeated regularly until each training sample has been presented to the network, a phase referred to as a training epoch. After the training process is over, all the weight vectors are relaxed immediately and the constructed map becomes topologically ordered in accordance to the intrinsic structure of the input sample space.

As mentioned above, the unsupervised Kohonen methodology is closely similar to the fundamental principle of the VQ algorithm (Fig. 16.3). Thus, during the training process, the network weight vectors (*filled arrowheads*) move/quantized iteratively toward the topological centers of the input data distribution (*data points are drawn as vectors with open arrowheads*). As a result of the VQ process, it actually appears that the weight vectors before (Fig. 16.3A) and after training (Fig. 16.3B) principally differ in their location, resulting in two different clusters that are clearly formed by two neurons; $\times 1$ and $\times 2$ are two dimensions of the data space.

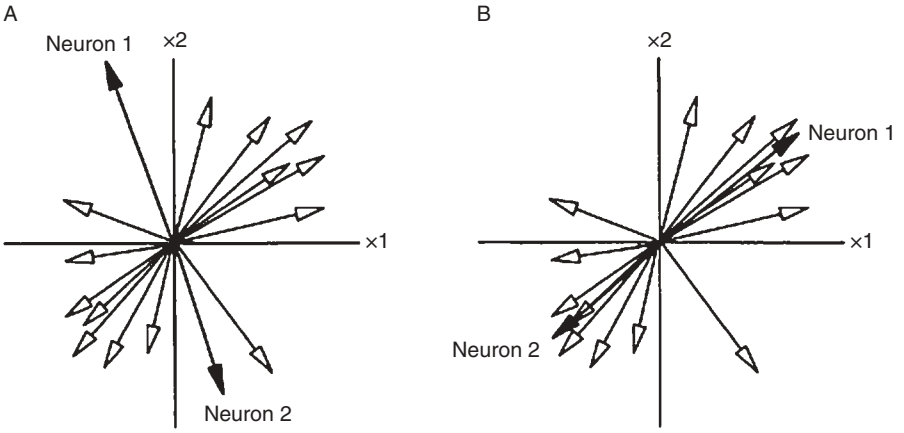


Figure 16.3 The successful implementation of the fundamental VQ principle underpinned by the unsupervised Kohonen logic.

The SOM algorithm, based on the basic Kohonen learning principle, can be formally presented in a step-by-step manner in the following way:

Step 1: At the beginning, the algorithm randomly selects a data point encoded by the vector x taken from the input data set.

Step 2: The corresponding low-dimensional image of the input vector x within the SOM array is then immediately defined using a function, which is dependent solely on the measure of distance observed between the input vector x and the related synaptic weight w. This criterion can be broadly defined as $c = \arg\min_i \{d(x, w_i)\}$, where $\{d(x, w)\}$ denotes a general distance measure, for example, Euclidean metrics, while c is the index of the unit (neuron) in the SOM lattice. In other words, the algorithm accurately determines the corresponding output element for which the weight vector w is conceptually closest to the presented input vector sample ("winner neuron", w_i^{win}), $\|w_i^{\text{win}} - x\| \leq \|w_i - x\|$, for all ith elements.

Step 3: Subsequently, a finite set of codebook vectors $\{w_i\}$ is collectively driven into the space of the x input patterns to approximate them by minimizing some reconstruction error measure/function. Let $p(x)$ be the probability density function of x, and let w_c be the corresponding codebook vector that is closest to the x sample in the input space, i.e., the one for which $d(x, w_c)$ is smallest. The VQ procedure tries to minimize the average expected quantization error (*reconstruction error*), which can be concisely expressed by $E = \int f[d(x, w_c)]p(x)d(x)$, where f is some monotonically increasing function of distance d. It should be particularly noted that the index c is a function of x and w_i, whereby the integrand is not continuously differentiable, i.e., c changes abruptly when crossing

a border in the input space where two codebook vectors have the same value for the predefined distance function. After the competition is complete, the algorithm updates the weight coefficients of the winning neuron according to the following prescription: $\Delta w_{ij}^{\text{win}} = \eta \cdot (xj - w_{ij}^{\text{win}})$, where η is a learning step size that usually represents the function of time, $\eta(t)$.

Step 4: The algorithm returns to step 1 or stops the training (e.g., if the value of η is below a critical threshold, or if the predefined number of cycles has passed).

WTA strategy forces the weight vectors of the network to move progressively toward the topological centroids of data distribution, thereby becoming a set of specific prototype (feature) vectors. All the data points located close to the "receptive center" associated with an output neuron will be directly assigned to the same low-dimensional cluster. In other words, the input data, which are much closer to the weight vector of one neuron than to any other weight vector, belong wholly to their specific receptive field. As briefly described above, the main criterion employed to find the winning neuron can be mathematically expressed as the similarity distance between a weight vector, w, and the data vector, x. The most frequently used similarity distances include Euclidian distance: $d = \sqrt{\sum_i (w_i - x_i)^2}$, Manhattan distance: $d = \sum_i |w_i - x_i|$, and the L∞ norm or ultrametric distance that represents the maximum absolute parametric difference: $d_{ij} = \max_{k=1}^{K} |x_{ik} - x_{jk}|$, where x_{ik} is the kth feature of the ith pattern and K is the total number of features. These distances are all the members of the generalized Minkowski metric defined as $d_{ij} = \left[\sum_{k=1}^{K} |x_{ik} - x_{jk}|^r \right]^{1/r}$ and result by substituting r in the Minkowski metric with 1, 2, and ∞, respectively. A complementary approach is to determine the Kohonen element with the maximal formal output: output $= \sum_i (w_i, x_i)$.

Typically, many training epochs are needed to successfully complete the training process. During the completion, each neuron becomes peculiarly sensitive to a particular region of the original input space. The input samples, which fall into the same region, whether they were or were not included in the original training set, are directly mapped onto the same neuron. Due to the obvious simplicity and clarity of their output, SOMs can be remarkably effective in the analysis and visualization of large and complex chemical data sets of different types and structures, particularly when they are used in automatic combination with advanced, interactive graphical tools.

Anzali et al. [29] were the first authors who successfully applied an SOM approach for the analysis and visualization of chemical data. Since then, a huge number of scientific papers dedicated to the application of the self-organizing

methodology in chemoinformatics were subsequently published. For instance, Balakin et al. [30] recently collected a large number of experimental facts and observations as well as several theoretical studies of physicochemical determinants of dimethyl sulfoxide (DMSO) solubility of different organic substances. Initially, the authors compiled a comprehensive reference database following the experimental protocol on compound solubility (55,277 compounds with good DMSO solubility and 10,223 compounds with poor DMSO solubility), then calculated specific physicochemical molecular descriptors (topological, electromagnetic, charge, and lipophilicity parameters), and finally, effectively applied an advanced machine-learning approach for training neural networks to adequately address the solubility. Both supervised (feedforward, backpropagated neural networks) and unsupervised (Fig. 16.4A,B) (self-organizing Kohonen neural networks) learning were used. The resulting neural network models were then externally validated by successfully predicting the DMSO solubility of compounds in an independent test set.

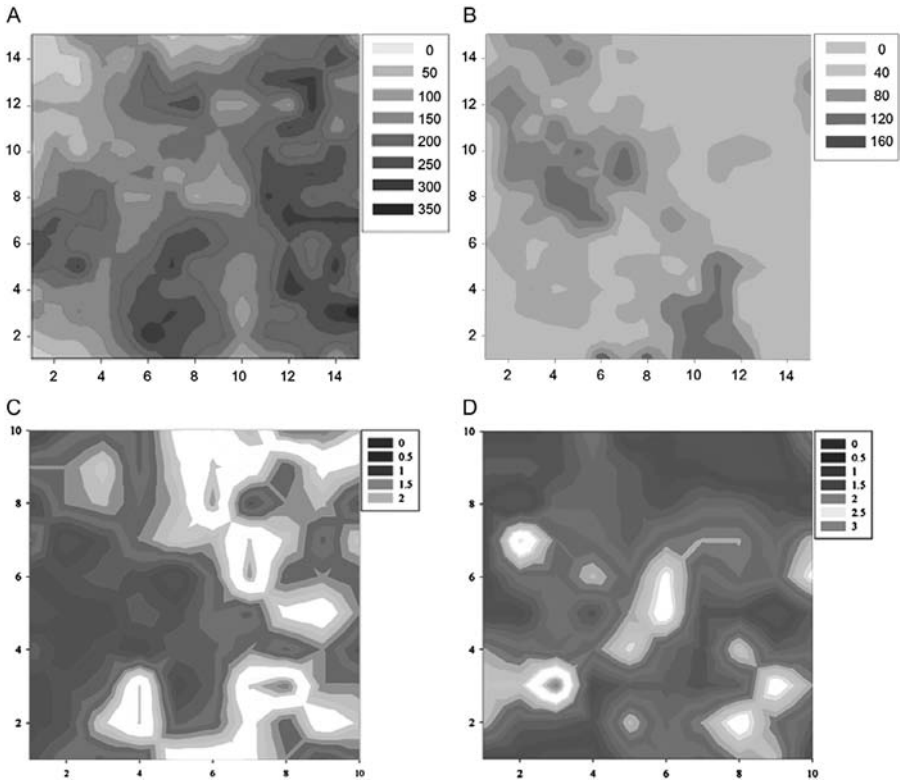


Figure 16.4 Distribution of DMSO(+) (A) and DMSO(–) (B) compounds within the generated Kohonen map; Kohonen network developed for the prediction of cytochrome P450 substrates (C) and products (D) processed within the same map.

Korolev et al. [31] also successfully applied a self-organizing approach for the computational modeling of cytochrome-mediated metabolic reactions (Fig. 16.4C,D). A training database consisted of many known human cytochrome P450 substrates (485 compounds), products (523 compounds), and nonsubstrates for 38 enzyme-specific groups (total of 2200 compounds) was compiled, and most typical cytochrome-mediated metabolic reactions within each group as well as the substrates and products of these reactions were also determined. To assess the probability of P450-related transformations of novel organic compounds, the authors constructed a nonlinear quantitative structure-metabolism relationship (QSMR) model based on the Kohonen self-organizing maps. The developed model incorporated a predefined set of several physicochemical descriptors encoding the key molecular properties that defined the metabolic fate of individual molecules. The result was that the isozyme-specific groups of substrate molecules were conveniently visualized and effectively separated, thereby significantly facilitating the prediction of metabolism. The developed computational model could be successfully applied in the early stages of drug discovery as an efficient tool for the assessment of human metabolism and toxicity of novel compounds, in designing discovery libraries and in lead optimization.

A useful modification of a common WTA strategy originally implicated in the Kohonen algorithm is to use more than just one single winner neuron in the adaptation process repeated during each training cycle [24,32]. As described above, in the Kohonen network, the output layer neurons are neatly arranged in low-dimensional geometry, usually in the two- or three-dimensional plane. Thus, the basic idea of this variant is that during the training not only the winning neuron but also neurons closely located to a winning one within the output layer are also being updated in accordance to the specific neighborhood function. As a result, a new formulation of the average quantization error function E_q (see equation below) has been introduced by Kohonen [33]. Following the modification, the smoothing kernel function h_{ki}, which is a function of the distance between units i and k determined within the whole Kohonen map: $E_q = \int h_{ki} f[d(x, w_i)]p(x)dx$, was subsequently integrated into the key learning role. The minimization of E_q imposes an ordering on the values of w_i as if these vectors were located at the nodes of an elastic net fitted to the density probability distribution $p(x)$ of the input space. Even in the most obvious cases, the minimization of E_q constitutes a complicated nonlinear optimization problem resulting in sustainable solutions that are not explicitly evident immediately. Hence, specific approximation algorithms should be used to provide a successful high-dimensional data representation.

One such algorithm that is commonly based on minimization of E_q is a stochastic approximation method leading to a fairly good approximation of the set of neural weights $\{w_i\}$. Following the mathematical definition of E_q, let $x = x(t)$ at the discrete (iteration) time step t. Let $w_i(t)$ be the approximation of w_i at the time step t and consider the sample function $E_q(t)$ defined as $E_q(t) = \sum_i h_{ki} f[d(x(t)), w_i(t))]$. Then, for the dynamic minimization of E_q, the

approximate optimization approach, which is primarily based on different gradient descent methods, can be effectively applied. Thus, starting from the initial values $w_i(0)$, all the synaptic weight coefficients are updated according to the following rule: $w_i(t+1) = w_i(t) - \left(\frac{1}{2}\right)\lambda(t)\frac{\partial E_q(t)}{\partial w_i(t)}$, where $\lambda(t)$ is a small positive scalar factor that determines the size of the gradient step at the discrete time t. If this function is chosen appropriately, the sequence of $w_i(t)$ will always converge to a set of $w_i^{\text{win}}(t)$ values, which will accurately approximate the solution $w_i(t)$. Although this procedure does not guarantee that a global minimum will always be achieved, a local minimum provides a sufficiently close approximation in many scientific applications. If required, a better local minimum can be readily reached by repeating this procedure with different starting values or using of advanced optimization techniques such as "simulated annealing." Many different variants of the fundamental SOM algorithm can be adequately expressed by the mentioned equation. For example, if d is defined by the Euclidean norm $d(x, w_i) = \|x, w_i\|$ and $f(d) = d^2$, the traditional self-organizing algorithm can be easily obtained and conventionally expressed by the following update rule: $w_i(y + 1) = w_i(t) + h_{ki}(t)[x(t) - w_i(t)]$. The additional rate term $\lambda(t)$ can also be introduced in the neighborhood smoothing kernel function $h_{ki}(t)$ determined around the winner neuron k at the time step t. In accordance with the original Kohonen formulation, the kernel function should proactively be formed by at least two principal parts: by the neighborhood function $h(d, t)$ and by the "learning rate" term $\alpha(t)$ usually defined by $h_{ki}(t) = h(\|r_k - r_i\|, t)\alpha(t)$. A widely used kernel-based function can be conveniently expressed by a Gauss (normal) distribution: $h_{ki}(t) = \alpha(t)\exp\left(-\frac{\|r_k - r_i\|^2}{2\sigma(t)^2}\right)$, where index k corresponds to the neuron for which the Euclidean distance assigned to the input sample is the smallest, r_k and r_i are the respective locations of the kth and ith neurons within the Kohonen lattice ($r_k, r_i \in \mathcal{R}^{2 \text{ or } 3}$), while $\alpha(t)$ is a linear or exponentially decreasing function defined over a finite interval of $0 < \alpha(t) < 1$, and $\sigma(t)$ is the width of the Gaussian function (Fig. 16.5). To ultimately ensure that the convergence of the Kohonen algorithm is successfully achieved, it is extremely important to ascertain that $h_{ki}(t) \to 0$, whereas the learning time $t \to \infty$. The developed approach in which the kernel function is a decreasing function of iteration time, and the distance travelled from the neuron i to the best matching unit k is dynamically changed through the learning time, thereby defining the region of principal influence of the input sample within the SOM.

The simplest definition of $h_{ki}(t)$ corresponds to a "bubble neighborhood function," which is constant over the predefined number of Kohonen neurons located closely to the winning unit while turned to zero elsewhere. In this case, $h_{ki}(t) = \alpha(t)$, if i and k are neighboring units, and $h_{ki}(t) = 0$ otherwise. Let us denote the corresponding set of these neighboring units as $N_k(t)$. Then, the

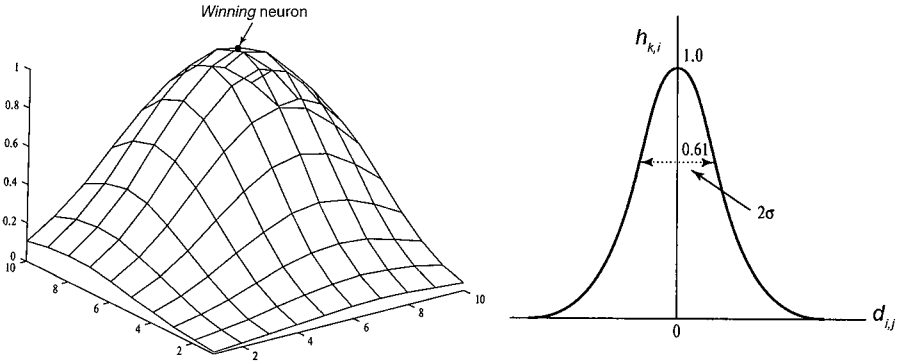


Figure 16.5 A Gauss distribution, one of the most widely applied neighborhood functions integrated in the Kohonen self-organizing algorithm.

fundamental Kohonen learning paradigm can be neatly reformulated in $\begin{cases} w_i(t+1) = w_i(t) + \alpha(t)[x(t) - w_i(t)], i \in N_k(t) \\ w_i(t+1) = w_i(t), i \notin N_k(t) \end{cases}$. During the training process, the algorithm forms the elastic net that folds into a space formed originally by the input data, thereby trying to closely approximate the probability density function of the data examined by the recruitment of additional codebook vectors around the winning neuron.

The method called "dot product map" is another variant of the basic SOM algorithm, which is completely invariant to scale of the input variables. In this approach, the dot product $\eta_i = w_i^T x$ defined between the input vector x and each weight vector w_i can be creatively used as a similarity measure. Consequently, the best matching unit is then selected as soon as the maximum dot product $\eta_k = \max_i\{\eta_i\}$ is found. The updated learning mechanism was designed specifically to normalize the value of the new codebook vector at each time step by the following rule: $w_i(t+1) = \dfrac{w_i(t) + \alpha(t)x}{\|w_i(t) + \alpha(t)x\|}$. This algorithm can be further simplified in accordance to the following equilibrium condition at the convergence limit: $\psi\{h_{ki}(x - m_i^{\text{win}})\} = 0$, where Ψ is the mathematical expectation operator. This expression can be easily reformulated in $w_i^{\text{win}} = \dfrac{\int h_{ki} x p(x)\,dx}{\int h_{ki} p(x)\,dx}$. For instance, if h_{ki} is uniquely defined by the *bubble neighborhood function*, $w_i^{\text{win}} = \dfrac{\int_{y_i} x p(x)\,dx}{\int_{y_i} p(x)\,dx}$, where y_i represents the predefined domain of vector x, the nearest codebook vector belongs wholly to the neighborhood set V_i of unit i, also widely known as the Voronoi region. In turn, this transformation

leads to a more powerful variant of the SOM algorithm also known as "batch map," where the whole data set is directly presented jointly to the Kohonen transducer before any adjustment operations are commenced. The training procedure is simply based on the replacement of the prototype vector by a weighted average value calculated over the total input samples, where the weighting factors are the neighborhood function values h_{ki}. Finally, to continually update the synaptic weights, *batch SOM* effectively uses the following learning rule:

$$w_i(t+1) = \frac{\sum_{j=1}^{n} h_{ki(j)}(t)\, x_j}{\sum_{j=1}^{n} h_{ki(j)}(t)}.$$

16.2.3.1 The Key Variants of SOMs In order to create a beneficial feature map and spatially organize representations of input samples within the Kohonen lattice, the most essential principle seems to be to restrict the learning corrections by subsets of the network units that are grouped in the topological neighborhood of the best-matching unit. It should be noted that a number of methods targeted to define a better matching of an input occurrence with the internal images are currently applied. In addition, the activation neighborhood function, which modulates the sensitivity of each Kohohen element, can be defined in many ways. It is also absolutely necessary in the feature space to clearly and properly define a learning role, for example, significant improvements in matching may be conveniently achieved by using the batch computation or evolutionary tuning approach. Consequently, all such cases will henceforth be regarded as variants of self-organizing map algorithms belonging to the broader category. This category may also include both the supervised and the unsupervised learning methodology. Although a nonparametric regression solution directly assigned to a class of VQ problems is tightly integrated in the SOM algorithm, it does not need any further modifications. There are a significant number of related problems in which the key self-organizing philosophy can be effectively applied by means of various modifications, which include different matching criteria and a nonidentical input strategy. In accordance with the first approach, the matching criteria applied to define the winning neuron can be conceptually generalized in different ways, using either various distance metrics or other definitions of matching. The second approach emphasizes that a straightforward network generalization would be to recruit the input vector x of the initial subsets of samples that is specifically assigned to different, eventually intersecting, areas of the SOM, so that the dimensionalities of x and the m_i may also depend on the topological location. Both strategies include a wide variety of dozens of methods such as hierarchical searching for the "winning neuron" (three-search SOM and Hypermap), dynamic topology of the Kohonen map, and signal space neighborhood function, as well as different methods for the acceleration of learning process.

It seems to be quite interesting to review the key modifications of the self-organizing learning role and topological composition. Thus, there exist a number of unique combinations, alternative modifications, and advanced variants of the basic SOM algorithm reported in the scientific literature. These include the usage of specific neuron learning rate functions and neighborhood size rules, as well as growing map structures. The principal goal of these variations is to enable the SOM to follow the space topology and usually the nested structure of the underlying data set accurately, ultimately achieving more satisfactory results of VQ. The most algorithmically advanced variants and modifications of the fundamental SOM approach are listed briefly below:

- The *tree-structured SOM* [34] is a very fast variant of the basic SOM algorithm, which consists of a set of Kohonen layers that routinely perform a complete quantization of the input data space. The principal differences among these layers lie in the number of codebook vectors, which increases exponentially while the tree is traversed downwards. Thus, data from the upper layers are directly used to train lower layers, reducing the amount of distance calculations needed to find the winning neuron. Consequently, each Kohonen layer provides a more detailed interpretation of the data space.
- The *minimum spanning tree SOM* [35] uses a conventional tree structure as a neighborhood function, which defines the minimal set of key connections needed in order to tightly link together related sets of codebook vectors. From the quantization point of view, the modified algorithm is more stable as compared with the traditional SOM. In contrast, the position of Kohonen elements within the feature space is not rigidly fixed; therefore, in many cases a simple graphical visualization becomes significantly more difficult.
- The *neural gas* [36] is a slightly modified version of the fundamental Kohonen network, which uses a dynamic neighborhood composition that is dramatically changed through the whole training process by analogy with the nature of gas molecules moving. During the execution of the algorithm all Kohonen units are arrayed along the map lattice according to the Euclidean distances toward the input vector x. After ordering is complete, neuron positions are ranked as follows: $d_0 < d_1 < d_2 < \ \dots \ < d_{n-1}$, where $d_k = \|x - w_{m(i)}\|$ indicates the Euclidean distances of ith neuron located in m-position within the Kohonen map from the winning neuron fixed in the position assigned to d_0. In this case, the neighborhood function can be defined narrowly as $G(i, x) = \exp\left(-\frac{m_{(i)}}{\lambda}\right)$, where the term $m_{(i)}$ denotes a queue number determined by sequencing of Kohonen neurons according to $(m_{(i)} = 0, 2, \dots, n - 1)$, parameter λ is an analogue of the neighborhood level used originally in the basic Kohonen algorithm, which decreases linearly or exponentially over the whole learning time. Thus, if λ is zero, the adaptation process is being carried out by

tuning the synaptic weight coefficients of the single winning neuron as such in the classical WTA algorithm, while if λ is not zero, a number of synaptic weight coefficients are being progressively adapted in relation to $G(i, \boldsymbol{x})$. In this context, the neural gas algorithm resembles the "fuzzy" set strategy.

To achieve a beneficial neuron composition, the learning process should be started from a high value of λ that decreases monotonically over the learning time by the linear of exponential functional dependencies. In many cases, a monotonic decrease of the continuous function $\lambda(k)$ can be realized by the following equation: $\lambda(k) = \lambda_{\max} \cdot \left(\frac{\lambda_{\min}}{\lambda_{\max}} \right)^{\frac{k}{k_{\max}}}$, where $\lambda(k)$ is the value of λ at the kth iteration, parameter $k_{\max}$ denotes the maximal number of iterations, while $\lambda_{\min}$ and $\lambda_{\max}$ are manually assigned the maximal and minimal values of λ, respectively. The learning rate coefficient $\eta_i(k)$ can also be changed following the linear or exponential decay laws, for instance, $\eta_i(k) = \eta_i(0) \cdot \left(\frac{\eta_{\min}}{\eta_i(0)} \right)^{\frac{k}{k_{\max}}}$, by analogy to the previous equation used for $\lambda(k)$ calculation. Here, $\eta_i(0)$ is the initial learning rate, while $\eta_{\min}$ takes *a priori* a minimum value corresponding to $k = k_{\max}$ conditions.

- The *convex combination* [28] is an additional method. In the case of nearby similar input objects belonging to the same class, initialization of weight coefficients by a random distribution can potentially lead to the erroneous fusion (or vice versa) to the over-fragmentation of the input data points. To successfully overcome this challenge, the advanced method called *convex combination* based on the key principles of the Kohonen network has recently been developed. *Convex combination* represents the effective learning algorithm used currently for pattern recognition, especially for the successful separation of input samples tightly packed into a multicluster structure. During the training process, weight coefficients are updated iteratively following the classical Kohonen algorithm in an attempt to accurately approximate the tuned components of each input vector. Following the algorithm, normalized input vectors are initially modified by the following equation: $x_i = \alpha(t) \cdot x_i + (1 - \alpha(t)) \cdot \frac{1}{\sqrt{n}}; w_o = \frac{1}{\sqrt{n}}$, where x_i is the ith component of the input vector sample, n is the total number of its components, and $\alpha(t)$ is the coefficient increased from zero to one through the learning time according to the linear or exponential functional dependencies. In summary, each component of the input vector initially takes the same "forced" value (usually close to zero) and henceforth moves gradually toward the original value. At the output of the neural net, all the classes isolated are conveniently located in compact areas identified within the

Kohonen lattice, significantly simplifying the visual analysis of the constructed map.

- The *Duane Desieno method* [37] is based fundamentally on the "conscience" or "memory" mechanism that effectively prevents each Kohonen processing unit from regularly overtaking the others during the competition. In summary, total control for the winning neurons is elegantly achieved to produce a more regular distribution of input samples over the Kohonen map. Following the algorithm, an excitatory neuron that wins the competition more than $\frac{1}{N}$ times, immediately arouses the resident "conscience" keeping the "memory" of the current state during several further iterations. According to this conception, the remaining neurons can also win the competition with equal probability. This approach provides exceptionally smooth and regular maps where the distribution of excitatory neurons is close to normal. The ingenious idea of the algorithm is to trace the time f_j for which the active neuron wins the competition. This parameter can be easily calculated through the total iterations by each neuron using the following equation: $f_j(t + 1) = f_j(t) + \beta(Z_j - f_j(t))$, where Z_j is the key indicator of the winning state of each neuron ($Z_j = 0$ or 1), while β is a manually assigned constant (usually a small positive value, for example, $\beta = 10^{-4}$). The corresponding shift b_j is then calculated immediately after the current winning time f_j is determined: $b_j = \gamma\left(\frac{1}{N} - f_i\right)$, where γ is a manually assigned constant (usually a positive value, for example, $\gamma \sim 10$). After these operations are complete, weight coefficients are tuned iteratively to their own values. In contrast to the classical Kohonen-based learning, the key exception of this method lies in the competition among neurons with the smallest value of the following criterion: $D(W_j, X) - b_j$. The practical role of the shift b_j can be neatly reformulated in: for the too-often-winning neurons, j parameter $f_j > \frac{1}{N}$ and $b_j < 0$, therefore $D(W_j, X) - b_j$ increases uniformly as compared with $D(W_j, X)$, while for the too-rare-winning neurons, $f_j << \frac{1}{N}$, $b_j > 0$, and $D(W_j, X) - b_j$ decreases resulting in the probability to become active to gradually increase.
- The overall goal of the *noise technique* [28] is to promote and facilitate the achievement of focused distribution of the frequency function $\rho(x) > 0$ within the whole range of definition of input vectors x: Ω_x. In this method, a uniformly distributed noise is imposed on each component of the input vector sample that is automatically generated. At the beginning, the noise level should be assigned to a relatively high value so that the "noisy" vector components are significantly different from their original values. By analogy to convex combination, the noise level decreased gradually with the

training epochs according to the linear or exponential functional dependencies. Such tactics are absolutely correct and obviously useful, but "noisy" vector components converge even more slowly toward eigenvalues as compared with convex combination. As a result, a significant excess of weight w_j is finally achieved in the areas occupied by high values of frequency function that completely covers the initial vector space and vice versa.

- The *growing cell structures* [38] method adds or removes map units as needed during the training process instead of working with a predefined number of codebook vectors.
- In an attempt to address the absolute chaos over the input vector space and how a self-organizing engine produces optimal feature image as well as a good convergence of the algorithm, the *two learning stages* [28] method decomposes the total learning process into two principal stages: *self-organization* and *convergence* or *ordering*. During the first training stage, a global, rough topological arrangement of the input samples is generally achieved due to the following machine parameters: number of iterations is close to 1000, the initial learning rate $\eta_0(t)$ should be selected in the range of 0.1–0.3. With great respect to the neighborhood activation function $h_{j,i}(n)$, it must be sufficiently flexible so that a majority of Kohonen neurons initially start to adapt their weight coefficients following the Gauss law. After the first stage is successfully complete, the Kohonen lattice has a rugged structure, which is further fine-adjusted iteratively to produce a smoothly organized map with a tweaking of the neuron weights. As a rule, the basic machine parameters should be assigned as follows: the number of iterations required generally for the convergence stage should be selected based on the role $N_{iter} = 500 \cdot N_{wi}$, where N_{wi} denotes the number of Kohonen neurons, the initial learning rate should be initially assigned to 0.01, and finally, the activation function must be selected so that the nearest processing units, located closely to the "winning neuron," jointly adapt their weight coefficients to produce a highly sensitive state.
- In many cases, the implementation of three-dimensional Kohonen lattice leads to relatively high-resolution maps, resulting in greater completeness and accuracy of a low-dimension representation, as well as an improved classification of input objects can be correctly achieved. Following the *three-dimensional architecture* [28] approach, the activation neighborhood function is commonly presented by a "quasi" normal three-dimensional distribution around the "winning neuron" as against to the two-dimensional Kohonen lattice. Although, the key moments of the postulated learning role are similar to the classical Kohonen algorithm, several advanced training strategies conveniently adapted to the three-dimensional space architecture are also developed to accelerate the learning process.

Being significantly hampered by the available space for the current review, we can only list other methods: bath SOM, operator maps, evolutionary learning SOM, structure-adaptive self-organizing map (SASOM), and adap-

tive-subspace SOM (ASSOM) are related closely with the advanced "episode" and "representative winner" techniques such as hypercube topological and cyclic maps, evolutionary learning filters and functions, feedback-controlled adaptive-subspace SOM (FASSOM), probabilistic extension generative topographic mapping (GTM) [39], and soft-max function.

Although SOMs can be effectively used to solve different classification tasks, there are a number of supervised variants of the basic Kohonen algorithm such as learning vector quantization (LVQ) and the dynamically expanding context (DEC) that, in many cases, may be more appropriate for this task [28]. For example, LVQ-based algorithms are closely related to VQ and SOM. This abbreviation signifies a class of related algorithms, such as LVQ1, LVQ2, LVQ3, and Hypermap-type LVQ, as well as a combined methodology based on SOM–LVQ cooperation and OLVQ1. While VQ and the basic SOMs are in priori unsupervised clustering/learning methods, LVQ uses the supervised principle. On the other hand, unlike SOM, no neighborhoods around the winning neuron are defined during the supervised learning in the basic version of LVQ, whereby also no spatial order of the codebook vectors is expected to ensue. Since LVQ is strictly meant for statistical classification or recognition, its only purpose is to define class regions within the input data space. To this end, a subset of similarly labeled codebook vectors is placed into each class region; even if the class distributions of the input samples would overlap at the class borders, the codebook vectors of each class in these algorithms can be placed in and shown to stay within each class region at all times. The quantization regions, like the Voronoi sets in VQ, are defined by hyperplanes between neighboring codebook vectors. An additional feature in LVQ is that for class borders, one can only take such borders of the Voronoi tessellation that separate Voronoi sets into different classes. The class borders thereby defined are piecewise linear.

At present, several software packages that deal with self-organizing Kohonen mapping are publicly available from the following resources: a computational program for the generation of the Kohonen maps has been developed by the Kohonen group (http://www.cis.hut.fi), KMAP software (Computer-Chemie-Centrum, Germany) (http://www2.chemie.uni-erlangen.de), and Molecular Networks (GmbH, Germany) (http://www.mol-net.de) primarily targeted for chemical applications, and so on. Among these software is InformaGenesis (http://www.InformaGenesis.com), which seems to be one of the most powerful computational programs developed recently for the generation and analysis of self-organizing Kohonen and Sammon maps (this program is discussed in more detail below).

16.2.4 IsoMap

MDS has proven to be hugely successful in many different applications. However, the main principle of MDS is generally based on Euclidean distances; therefore, it does not leave out of account the distribution of the nearest neighboring data points. If high-dimensional data points lie on or near

a curved manifold, MDS might consider two data points as near points, whereas their distance over the manifold is much larger than the typical interpoint distance. IsoMap is a promising new technique, which has resolved this problem [40]. By analogy with Sammon mapping, the algorithm attempts to preserve all pairwise geodesic (*curvilinear*) distances between the input data points within the whole feature space as close as possible. This distance is defined narrowly as the shortest path between a pair of sample points. The path is strictly confined to lie off the low-dimensional manifold, which is not known *a priori*. The geodesic distance can be adequately approximated using the shortest path observed between two sample points obtained by adding all subpaths among the nearest neighboring points. It should be duly noted that this technique produces favorable results only when a representative sampling of input data points across the manifold is presented.

In IsoMap, the geodesic distances among the sample points x_i are directly computed by constructing the neighborhood graph G, in which every data point x_i is intimately connected to its k-nearest neighbor x_{ij} within the studied data set X. In mathematical form, the shortest path observed between two data points can be readily computed using the Dijkstra "shortest-path" algorithm. Using this strategy, the geodesic distances held among the data points x_{ij} in the input space X can be easily computed, thereby instantly forming a complete pairwise geodesic distance matrix, M_{ij}^{G}. In the low-dimensional feature space F, the feature image y_i that corresponds to the data point x_i is then computed by simply applying the fundamental principle of MDS toward the target matrix M_{ij}^{G}. A methodological weakness of the IsoMap algorithm lies in the potential topological instability [41]. Thus, IsoMap may construct significantly erroneous geodesic connections between data points in the neighborhood graph G. Such a local approximation can seriously impair the performance of IsoMap [42]. Fortunately, several advanced approaches were recently proposed to completely overcome this "short-circuiting" problem, for example, by removing data points with large total flows in the shortest-path algorithm [43] or by removing nearest neighbors that violate local linearity of the neighborhood graph [44]. Furthermore, during the training process, IsoMap may be overly sensitive to "holes" (areas of poor/empty population) along the manifold [42]. In addition, if the manifold is not sufficiently convex, the IsoMap can occasionally fail [45].

Despite these drawbacks, IsoMap has been successfully applied for visualization of different types of biomedical data [46]. As recently shown, a graphical visualization of IsoMap models provides a useful tool for exploratory analysis of protein microarray data sets. In most cases, IsoMap planes can adequately explain more of the variance presented in the microarray data in contrast to PCA or MDS [45,47]. Therefore, IsoMap represents a prominent modern algorithm targeted mainly for class discovery and class prediction within high-density oligonucleotide data sets.

In addition, for more detailed analysis of the large protein folding data sets (molecular dynamics trajectories of an Src-Homology 3 protein model), the

algorithm was subsequently modified using landmark points in the geodesic distance calculations. Based on the results, it clearly showed that for an accurate interpretation and analysis of the original data, the approach was far more effective than the linear technique of dimensionality reduction, such as PCA. Whereas the Euclidean metric is based directly on quadratic relationships, IsoMap scales with the third power of the number of data points and then becomes computationally prohibitive for processing and visualizing large data sets. In contrast, the stochastic proximity embedding (SPE) algorithm (see Section 16.2.5) neatly avoids the need to estimate the geodesic distances. It scales linearly with the size of the data set and has been clearly shown to be equivalently effective in dimensionality reduction and NLM as compared with other modern computational algorithms.

To sum up all the arguments described above, the first NLM algorithm introduced by Sammon as well as MDS is applicable solely to relatively small data sets. The Sammon NLM partly alleviates MDS- and PCA-associated problems by introducing a normalization factor in the head error function to give increasing weight to short range distances over long range ones. This scheme, however, is quite arbitrary and fails utterly with highly folded topological structures. To partially overcome this problem, the IsoMap method was originally developed as an alternative approach for the dimensionality reduction and clustering of large data sets of high dimensionality. In stark contrast to MDS, this algorithm effectively uses an estimated geodesic distance instead of the conventional Euclidean one. However, IsoMap requires expensive nearest-neighbor and shortest-path computations, and scales exponentially with the number of examined data points. Running slightly ahead, a similar scaling problem also strikes the SPE algorithm (see Section 16.2.5), a related approach that produces globally ordered maps by constructing locally linear relationships among the input data points.

16.2.5 SPE

This section is entirely devoted to a promising self-organizing algorithm named SPE. This technique was recently developed for embedding a set of related observations into a low-dimensional space that preserves the intrinsic dimensionality and metric structure of the data.

We must remember that extracting the minimum number of independent variables that can describe precisely a set of experimental observations, represents a problem of paramount importance in computational science. With regard to dimensionality reduction, the nonlinear structure of a manifold, which is often embodied in the similarity measure between the input data points, must be transformed ultimately into a low-dimensional space that faithfully preserves these similarities, in the hope that the intrinsic structure of the input data will be reflected adequately in the resulting map [48]. Among a variety of distance metrics, Euclidean geometry is widely used for pairwise comparison of objects represented by the multiparametric data. However, this

conventional similarity measure often tends to grossly underestimate the proximity of data points on a nonlinear manifold and lead to erroneous embeddings. To partially remedy this problem, the IsoMap method (see Section 16.2.4) was recently developed based on the fundamental geodesic principle and classical MDS to find the optimum low-dimensional configuration. Although it has been shown that in the limit of infinite training samples, IsoMap recovers the true dimensionality and geometric structure of the data if it belongs to a certain class of Euclidean manifolds; the proof is of little practical use because the (at least) quadratic complexity of the embedding procedure precludes its use with large data sets. A similar scaling problem plagues locally linear embedding [49], a related approach that produces globally ordered maps by constructing locally linear relationships between the data points.

To avoid these complications, an advanced technique of dimensionality reduction that addresses the key limitations of IsoMap and local linear embedding (LLE) has recently been developed by Agrafiotis and Xu [50]. Although SPE is based on the same geodesic principle first proposed and exploited in IsoMap, it effectively introduces two important algorithmic advances. First, it circumvents the calculation of estimated geodesic distances between the embedded objects, and second, it uses a pairwise refinement strategy that does not require the complete distance or proximity matrix maintaining a minimum separation between distant objects and scales linearly with the size of the data set. In other words, SPE skillfully utilizes the fact that the geodesic distance is always greater than or equal to the input proximity if the latter is an accurate metric. Unlike previous stochastic approaches of nonlinear manifold learning that preferentially preserve local over global distances, the method operates by estimating the proximities between remote objects as lower bounds of their true geodesic distances and uses them as a means to impose global structure. Thus, it can reveal the underlying geometry of the manifold without intensive nearest-neighbor or shortest-path computations. Therefore, it can preserve the local geometry and the global topology of the manifold better than previous approaches. Although SPE does not offer the global optimality guarantees of IsoMap or LLE, it works well in practice, above all due to the linear scaling with the number of points. Therefore, it can be effectively applied to very large data sets that are intractable by conventional embedding procedures.

In mathematical terms, the algorithm is carried out by minimizing the core function $S = \dfrac{\sum_{i<j} (f(d_{ij}, y_{ij})/y_{ij})}{\sum_{i<j} y_{ij}}$, where y_{ij} is the input proximity between the ith and jth points; d_{ij} is their Euclidean distance in the low-dimensional space; $f(d_{ij}, y_{ij})$ is the pairwise stress function defined as $f(d_{ij}, y_{ij}) = (d_{ij} - y_{ij})^2$ if $y_{ij} \le y_c$ or $d_{ij} < y_{ij}$, and $f(d_{ij}, y_{ij}) = 0$ if $y_{ij} > y_c$ and $d_{ij} \ge y_{ij}$; and y_c is the neighborhood radius. The postulated function can be completely minimized by using a self-organizing algorithm that attempts to bring each individual term $f(d_{ij}, y_{ij})$ rapidly to zero. SPE starts with an initial configuration, and iteratively refines

it by repeatedly selecting pairs of objects at random, and adjusting their coordinates so that their distances on the map d_{ij} match more closely their respective proximities y_{ij}. The algorithm proceeds as follows:

1. Initialize the D-dimensional coordinates of the N points, x_i. Select an initial learning rate λ (parameter that decreases linearly or exponentially through the learning time in order to avoid oscillatory behavior, $\lambda > 0$). The tuning is proportional to the disparity $\lambda \frac{\|y_{ij} - d_{ij}\|}{d_{ij}}$.
2. Select a pair of points, i and j, at random and compute their Euclidean distance on the D-dimensional map following the equation $d_{ij} = \|x_i - x_j\|$. If $y_{ij} > y_c$ and $d_{ij} \geq y_{ij}$, where y_c is a predefined neighborhood radius, i.e., if the points are nonlocal and their distance on the map is already greater than their proximity y_{ij}, their coordinates remain unchanged. If $d_{ij} \neq y_{ij}$, i.e., $y_{ij} \leq y_c$, or if $y_{ij} > y_c$ and $d_{ij} < y_{ij}$, update the coordinates x_i and x_j by $\frac{\lambda}{2}\left[(x_i - x_j)\frac{y_{ij} - d_{ij}}{d_{ij} + \gamma}\right] + x_i \rightarrow x_i$ and $\frac{\lambda}{2}\left[(x_j - x_i)\frac{y_{ij} - d_{ij}}{d_{ij} + \gamma}\right] + x_j \rightarrow x_j$, where γ is a small number to avoid division by zero (usually $1.0 \cdot 10^{-10}$).
3. Repeat step 2 for a prescribed number of steps M.
4. Decrease the learning rate λ by prescribed decrement $\delta\lambda$.
5. Repeat steps 2–4 for a prescribed number of cycles C.
6. End of learning.

One potential limitation of SPE is basically related to numerous adjustable and internal parameters. Thus, just like IsoMap and LLE, SPE strongly depends on the choice of the neighborhood radius, y_c. If y_c is too large, the local neighborhoods will include data points from other branches of the manifold, short-cutting them and leading to substantial errors in the final embedding. If it is too small, it may lead to discontinuities, causing the manifold to fragment into a large number of disconnected clusters. The ideal threshold can be determined by examining the stability of the algorithm over a range of neighborhood radii as prescribed by Balasubramanian and Schwartz [41]. Superficially, in addition to the neighborhood radius y_c, SPE also depends on the number of cycles C, the number of steps per cycle M, the initial learning rate λ_0, the annealing schedule for the learning rate, and the initial configuration. It was experimentally found that for most applications, the favorable result can be obtained using $C = 100$, $\lambda_0 = 2.0$, and $\delta\lambda = \frac{\lambda_0 - \lambda_1}{C - 1}$, where λ_1 is the final learning rate that can be any arbitrary small number, typically between 0.01 and 0.1. The initial configuration can be any random distribution of the N points in a D-dimensional hypercube of side length $N^{\frac{1}{D_{yc}}}$. The number of steps per cycle M, or equivalently the total number of pairwise adjustments, therefore

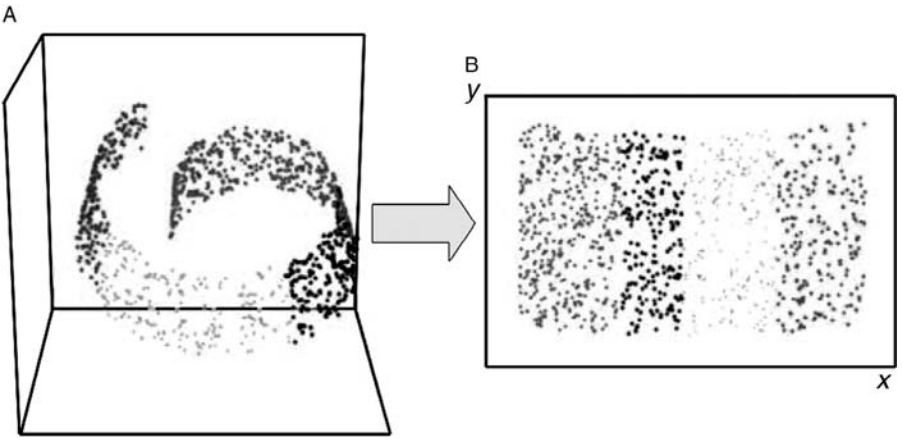


Figure 16.6 (A) Original three-dimensional space of the Swiss roll data set; (B) two-dimensional embedding of Swiss roll data set obtained by SPE.

remains as the only extra free parameter besides the neighborhood radius. *M* should be increased linearly with *N* as SPE's empirical characteristic of linear scaling suggests. Moreover, it was also found that the number of learning steps should be increased for structures with large curvatures.

SPE can be effectively applied in different classification tasks. Thus, this method has been successfully used for the analysis of the Swiss roll data set (Fig. 16.6) [50]. The distances of the points on the two-dimensional map matched the true, analytically derived geodesic distances with a correlation coefficient of 0.9999, indicating a virtually perfect embedding.

SPE can also produce meaningful low-dimensional representations of more complex data sets that do not have a clear manifold geometry. The advantages of this method have been clearly demonstrated using various data sets of different origin, structure, and intrinsic dimensionality, especially in several different chemical data sets. The embedding of a combinatorial chemical library illustrated in Figure 16.7 shows that SPE is able to preserve local neighborhoods of closely related compounds while maintaining a chemically meaningful global structure. For example, amination and Ugi virtual combinatorial libraries have been recently analyzed using the modified SPE algorithm.

The first data set represents a two-component virtual combinatorial library [50] containing 10,000 compounds, derived by combining 100 amines and 100 aldehydes using the reductive amination reaction. Each of the products was described by 117 topological descriptors including molecular connectivity indices, kappa shape indices, subgraph counts, information-theoretic indices, Bonchev–Trinajstic indices, and topological state indices [50]. To eliminate strong linear correlations, which are typical of graph-theoretic descriptors, the data were normalized and decorrelated using PCA. Molecular dissimilarity

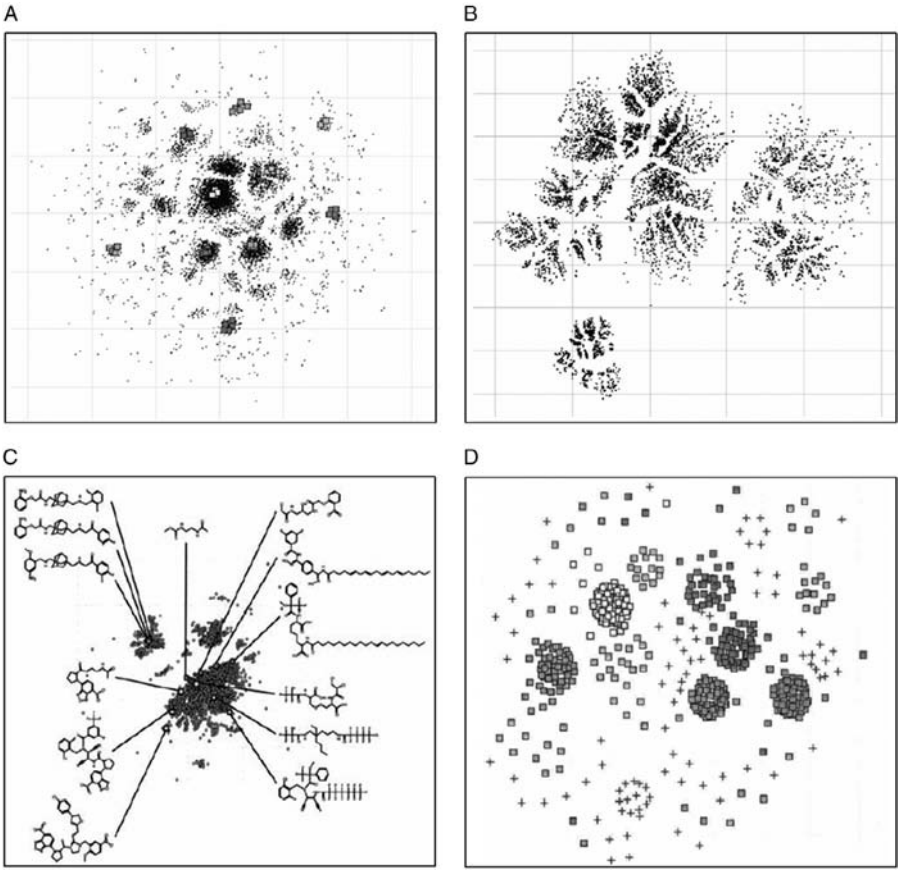


Figure 16.7 (A) Two-dimensional stochastic proximity map of the amination library. Ten representative clusters of closely related compounds that can be seen on the map demonstrate the ability of SPE to preserve close proximities; (B) two-dimensional stochastic proximity map of the generated Ugi library; (C) two-dimensional stochastic proximity map of different types of organic compounds; (D) two-dimensional SPE maps of the Manning kinase domains using a neighborhood radius of 0.89.

was defined as the Euclidean distance in the latent variable space formed by the 23 principal components that accounted for 99% of the total variance in the data.

The second data set represents a four-component virtual combinatorial library containing 10,000 compounds derived by combining 10 carboxylic acids, 10 primary amines, 10 aldehydes and 10 isonitriles using the Ugi reaction. Each of the products was described by an MDL key 166-dimensional binary fingerprint, where each bit encoded the presence or absence of a par-

ticular structural feature in the target molecule, as defined in the ISIS chemical database management system. Molecular dissimilarity was calculated based on the Tanimoto coefficient.

The resulting maps shown in Figure 16.7A,B exhibit the compact clusters that correspond to distinct chemical classes of the tested structures resulting from the discrete nature of the descriptors and the diversity of the chemical fragments employed.

These maps, which are discussed in greater detail in Reference 51, are illustrative of the ability of SPE to identify natural clusters in the data set without prior knowledge or expert guidance. In addition, this method was shown to be highly effective for mapping of different types of organic compounds (Fig. 16.7C). Therefore, this algorithm can be readily applied to other types of binary fingerprint descriptors (that typically consist of a few thousand bits) or to various high-dimensional descriptor spaces that are commonly employed in different QSAR studies.

Although the method is extremely fast, profiling experiments showed that a very significant fraction of the time required for the refinement was spent inside the random number generator (RNG). SPE requires two calls to the RNG for every pairwise refinement step, and for simple proximity measures such as the Euclidean distance or Tanimoto coefficient, this corresponds to a significant fraction of the overall computational work. To avoid this problem, an alternative learning function that reduces the number of RNG calls and thus improves the efficiency of the SPE algorithm was recently developed [52]. Thus, a modified update rule led to a reduction in stress over the course of the refinement. In addition, due to the second term in the modified error function that may be negative, the stress may temporarily increase, allowing the algorithm to escape from the local minima.

SPE can also be applied to an important class of distance geometry problems including conformational analysis [53], nuclear magnetic resonance (NMR) structure determination, and protein-structure prediction [54]. For example, the effective implementation of SPE in combination with the self-organizing superimposition algorithm that has been successfully applied for conformational sampling and conformational search was recently described [55–57]. Basically, the algorithm generates molecular conformations that are consistent with a set of geometric constraints, which include interatomic distance bounds and chiral volumes derived from the molecular connectivity table. SPE has recently been applied to the classification and visualization of protein sequences as well as to reduce the intrinsic dimensionality and metric structure of the data obtained from genomic and proteomic research [58]. Thus, the effectiveness of the algorithm can be illustrated using examples from the protein kinase and nuclear hormone receptor superfamilies (Fig. 16.7D). When used with a distance metric based on a multiple sequence alignment (MSA), the method produced informative maps that preserve the intrinsic structure and clustering of the data. The MSA metric defines dissimilarity as $\xi_{ij} = \sum_{k=1}^{n} \Omega\alpha_{ik}\alpha_{jk}$, where $\Omega\alpha_{ik}\alpha_{jk}$ is the dissimilarity score between amino acids

α_{ik} and α_{jk} as determined by a normalized exchange matrix, and n is the length of the alignment [21].

Finally, it should be noted, that Demartines and Herault [59] recently described a related method for nonlinear dimensionality reduction known as curvilinear component analysis (CCA). Although the key learning function is very different, CCA uses an optimization heuristic that is very similar to the one employed in SPE. The principal difference between these methods is that CCA utterly disregards remote distances, whereas SPE differentiates them from local distances by their intrinsic relationship to the true geodesic distances, and utilizes both types accordingly in order to improve the embedding. In essence, the SPE method views the input distances between remote points as lower bounds of their true geodesic distances, and uses them as a means to impose global structure.

16.3 MAPPING SOFTWARE

Over the past decade, the amount of data arising from various medicinal chemical disciplines, especially from biologic screening and combinatorial chemistry, has literally exploded and continues to grow at a staggering pace. Scientists are constantly being inundated with all types of chemical and biologic data. However, the computational tools for integrating and analysis of the data have largely failed to keep pace with these advances. The majority of computer programs targeted for dimensionality reduction, mapping, and throughput of data analysis have roughly been realized in simple, often inconvenient console format, or as external modules installed under the different pilot platforms, such as Microsoft Excel and MATLAB computing language; these include, but are not limited to, SOM_PAK (*console DOS-based version*) (http://www.cis.hut.fi), Ex-SOM (*Excel-based interface*) (http://www.geocities.com), and SOM Toolbox (*Matlab-based interface*) (http://www.cis.hut.fi). Fortunately, to date, several novel powerful Windows-based programs in the field of dimensionality reduction and multiparametric data analysis are currently available from different commercial and academic sources. Among these software, InformaGenesis (http://www.InformaGenesis.com), NeuroSolutions (http://www.nd.com), and Neurok (http://www.neurok.ru) are typical examples of neural-based computational programs running under Windows and particularly targeted for Kohonen and Sammon mapping. However, there is a relatively small number of such software specialized for chemical data analysis.

InformaGenesis is a powerful software, which is primarily based on the Kohonen SOM algorithm and is enhanced by many advanced modifications and complex-specific modalities. This program has specifically been designed to work under the Windows operating system. In addition to the basic Kohonen settings and learning parameters, it includes significant algorithmic-based improvements, such as "neural gas," "Duane Desieno," "noise technique," and "two learning stages and three-dimensional architecture," as well as

several unique algorithms and specific methods, for instance, "corners," "gradient," and "automatic descriptor selection algorithm (ADSA)". Furthermore, the program is also completely adapted for the analysis of large sets of chemical data of different types and dimensionality. Thus, the specific calculation module SmartMining integrated into the master computing engine of InformaGenesis was also originally developed and scientifically validated in a wide number of drug discovery projects. For example, the SmartMining software calculates more than hundred fundamental molecular descriptors that are generally divided into several logical and functional categories, including the basic specific physicochemical features, such as log *P*, number of H-bond donors, H-bond acceptors and rotatable bonds; and topological and electrotopological descriptors, such as Zagreb, Wiener, and E-state indexes, as well as *quasi*-3D descriptors, such as van der Waals volume and surface. All descriptors are directly calculated by using well-known common models and approximations in the scientific literature [60]. In addition, several algorithms have been progressively modified to obtain more exact feature prediction or/and calculations; for example, van der Waals parameters are calculated fairly accurately considering the overlapped volumes and/or surfaces.

In fact, it becomes increasingly obvious that specific computational programs targeted at the complex analysis of multiparametric/multidimensional data sets are being demanded in modern drug discovery and development at present.

16.4 CONCLUSION

This chapter presents a detailed review specifically focused on advanced computational mapping techniques currently applied to *in silico* drug design and development. Several novel mapping approaches to analysis of the structure–activity relationships within the chemical space of different types and structure were comparatively discussed in order to provide a better understanding of the fundamental principles of dimensionality reduction. However, because of severe space limitations, several mapping techniques and their target applications remain beyond the scope of the current review. We strongly believe this computational mapping area will expand greatly in the future due to the increasing amounts of data being deposited in public databases as well as those that are internally generated in pharmaceutical companies.

REFERENCES

1. Ekins S, Mestres J, Testa B. *In silico* pharmacology for drug discovery: Applications to targets and beyond. *Br J Pharmacol* 2007;152:21–37.
2. Ekins S, Mestres J, Testa B. *In silico* pharmacology for drug discovery: Methods for virtual ligand screening and profiling. *Br J Pharmacol* 2007;152:9–20.
3. Balakin KV, Kozintsev AV, Kiselyov AS, Savchuk NP. Rational design approaches to chemical libraries for hit identification. *Curr Drug Discov Technol* 2006;3:49–65.

4. Bernard P, Golbraikh A, Kireev D, Chretien JR, Rozhkova N. Comparison of chemical databases: Analysis of molecular diversity with self organizing maps (SOM). *Analysis* 1998;26:333–341.
5. Agrafiotis DK. Stochastic algorithms for maximizing molecular diversity. *J Chem Inf Comput Sci* 1997;37:841–851.
6. Bayada DM, Hamersma H, van Geerestein VJ. Molecular diversity and representativity in chemical databases. *J Chem Inf Comput Sci* 1999;39:1–10.
7. Kirew DB, Chretien JR, Bernard P, Ros F. Application of Kohonen neural networks in classification of biologically active compounds. *SAR QSAR Environ Res* 1998;8:93–107.
8. Ros F, Audouze K, Pintore M, Chrétien JR. Hybrid systems for virtual screening: Interest of fuzzy clustering applied to olfaction. *SAR QSAR Environ Res* 2000;11:281–300.
9. Balakin K. Pharma ex machina. *Mod Drug Discov* 2003;8:45–47.
10. Oprea TI. Chemical space navigation in lead discovery. *Curr Opin Chem Biol* 2002; 6:384–389.
11. Oprea TI, Gottfries J. Chemography: The art of navigating in chemical space. *J Comb Chem* 2001;3:157–166.
12. Johnson MA, Maggiora GM. *Concepts and Applications of Molecular Similarity*. New York: Wiley, 1990.
13. Sammon JW. A non-linear mapping for data structure analysis. *IEEE Trans Comput* 1969;C-18:401–409.
14. Chang CL, Lee RCT. A heuristic relation method for nonlinear mapping in the cluster analysis. *IEEE Trans Syst Man Cybern* 1973;SMC-3:197–200.
15. Pykett CE. Improving the efficiency of Sammon's non-linear mapping by using clustering archetypes. *Electron Lett* 1978;14:799–800.
16. Lee RCY, Slagle JR, Blum H. A triangulation method for the sequential mapping of points from N-space to two space. *IEEE Trans Comput* 1977;C-27:288–292.
17. Biswas G, Jain AK, Dubes RC. Evaluation of projection algorithms. *IEEE Trans Pattern Anal Mach Intell* 1981;PAMI-3:701–708.
18. Sadowski J, Wagener M, Gasteiger J. Assessing similarity and diversity of combinatorial libraries by spatial autocorrelation functions and neural networks. *Angew Chem Int Ed Engl* 1995;34:2674–2677.
19. Jiang J-H, Wang J-H, Liang Y-Z, Yu R-Q. A non-linear mapping-based generalized backpropagation network for unsupervised learning. *J Chemom* 1996;10:241–252.
20. Agrafiotis DK, Lobanov VS. Nonlinear mapping networks. *J Chem Inf Comput Sci* 2000;40:1356–1362.
21. Agrafiotis DK. A new method for analyzing protein sequence relationships based on Sammon maps. *Protein Sci* 1997;6:287–293.
22. Balakin KV, Ivanenkov YA, Savchuk NP, Ivashchenko AA, Ekins S. Comprehensive computational assessment of ADME properties using mapping techniques. *Curr Drug Discov Technol* 2005;2:99–113.
23. Willshaw DJ, von der Malsburg C. How patterned neural connections can be set by self-organization. *Proc R Soc Lond B Biol Sci* 1976;194:431–445.

24. Kohonen T. *Self-Organization and Associative Memory*, 3rd edn. New York: Springer-Verlag, 1988.
25. Kohonen T. The self-organizing map. *Proc IEEE* 1990;78:1464–1480.
26. Kohonen T. Physiological interpretation of the self-organizing map algorithm. *Neural Netw* 1993;6:895–905.
27. Linde Y, Buzo A, Gray RM. An algorithm for vector quantizer design. *IEEE Trans Commun* 1980;28:84–95.
28. Kohonen T. *Self-Organizing Maps*. Heidelberg: Springer-Verlag, 1996.
29. Anzali S, Gasteiger J, Holzgrabe U, Polanski J, Sadowski J, Teckentrup A, Wagener A. The use of self-organizing neural networks in drug design. In: *3D QSAR in Drug Design*, edited by Kubinyi H, Folkers G, Martin YC, pp. 273–299. Dordrecht, The Netherlands: Kluwer/ESCOM, 1998.
30. Balakin KV, Ivanenkov YA, Skorenko AV, Nikolsky YV, Savchuk NP, Ivashchenko AA. In silico estimation of DMSO solubility of organic compounds for bioscreening. *J Biomol Screen* 2004;9:22–31.
31. Korolev D, Balakin KV, Nikolsky Y, Kirillov E, Ivanenkov YA, Savchuk NP, Ivashchenko AA, Nikolskaya T. Modeling of human cytochrome P450-mediated drug metabolism using unsupervised machine learning approach. *J Med Chem* 2003;46:3631–3643.
32. Kohonen T. Self-organized formation of topologically correct feature maps. *Biol Cybern* 1982;43:59–69.
33. Kohonen T. Self-organizing maps: Optimization approaches. *Artif Neural Netw* 1991;2:981–990.
34. Koikkalainen P. Tree structured self-organizing maps. In: *Kohonen Maps*, edited by Oja E, Kaski S, pp. 121–130. Amsterdam: Elsevier, 1999.
35. Kangas J, Kohonen T, Laaksonen J. Variants of the self-organizing map. *IEEE Trans Neural Netw* 1990;1:93–99.
36. Martinetz TM, Berkovich SG, Schulten KJ. "Neural gas" network for vector quantization and its application to time series prediction. *IEEE Trans Neural Netw* 1993;4:558–569.
37. DeSieno D. Adding a conscience to competitive learning. In: *Proceedings of the ICNN'88, International Conference on Neural Networks*, pp. 117–124. Piscataway, NJ: IEEE Service Center, 1988.
38. Fritzke B. Growing cell structures—A self-organizing neural network for unsupervised and supervised learning. *Neural Netw* 1994;7:1441–1460.
39. Bishop C, Svensen M, Williams C. GTM: The generative topographic mapping. *Neural Comput* 1998;10:215–234.
40. Tenenbaum JB. Mapping a manifold of perceptual observations. In: *Advances in Neural Information Processing Systems*, edited by Jordan M, Kearns M, Solla S, Vol. 10, pp. 682–688. Cambridge, MA: The MIT Press, 1998.
41. Balasubramanian M, Schwartz EL. The Isomap algorithm and topological stability. *Science* 2002;295:7.
42. Lee JA, Verleysen M. Nonlinear dimensionality reduction of data manifolds with essential loops. *Neurocomputing* 2005;67:29–53.

43. Choi H, Choi S. Robust kernel Isomap. *Pattern Recognit* 2007;40:853–862.
44. Saxena A, Gupta A, Mukerjee A. Non-linear dimensionality reduction by locally linear isomaps. *Lect Notes Comput Sci* 2004;3316:1038–1043.
45. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. *Science* 2000;290:2319–2323.
46. Lim IS, Ciechomski PH, Sarni S, Thalmann D. Planar arrangement of high-dimensional biomedical data sets by Isomap coordinates. In: *Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems*, pp. 50–55. Newton, MA: Butterworth-Heinemann, 2003.
47. Dawson K, Rodriguez RL, Malyj W. Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using Isomap, a nonlinear algorithm. *BMC Bioinformatics* 2005;6:1–17.
48. Borg I, Groenen PJF. *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer, 1997.
49. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science* 2000;290:2323–2326.
50. Agrafiotis DK, Xu H. A self-organizing principle for learning nonlinear manifolds. *Proc Natl Acad Sci USA* 2002;99:15869–15872.
51. Agrafiotis DK, Xu H. A geodesic framework for analyzing. Molecular similarities. *J Chem Inf Comput Sci* 2003;43:475–484.
52. Rassokhin DN, Agrafiotis DK. A modified update rule for stochastic proximity embedding. *J Mol Graph Model* 2003;22:133–140.
53. Spellmeyer DC, Wong AK, Bower MJ, Blaney JM. Conformational analysis using distance geometry methods. *J Mol Graph Model* 1997;15:18–36.
54. Havel TF, Wuthrich K. An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformations in solution. *J Mol Biol* 1985;182:281–294.
55. Zhu F, Agrafiotis DK. Self-organizing superimposition algorithm for conformational sampling. *J Comput Chem* 2007;28:1234–1239.
56. Xu H, Izrailev S, Agrafiotis DK. Conformational sampling by self-organization. *J Chem Inf Comput Sci* 2003;43:1186–1191.
57. Agrafiotis DK, Bandyopadhyay D, Carta G, Knox AJS, Lloyd DG. On the effects of permuted input on conformational sampling of drug-like molecules: An evaluation of stochastic proximity embedding. *Chem Biol Drug Des* 2007;70:123–133.
58. Farnum MA, Xu H, Agrafiotis DK. Exploring the nonlinear geometry of protein homology. *Protein Sci* 2003;12:1604–1612.
59. Demartines P, Herault J. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans Neural Netw* 1997;8: 148–154.
60. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. New York: Wiley-VCH, 2002.

17

DATABASES FOR CHEMICAL AND BIOLOGICAL INFORMATION

Tudor I. Oprea, Liliana Ostopovici-Halip, and Ramona Rad-Curpan

Table of Contents

17.1 INTRODUCTION

Increasing amounts of data and information become available in the biologically related areas of chemistry, which lead to a stringent need to store, archive, organize, and systemize all this information. Such demands are being

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

met by developing comprehensive databases. A database is a collection of permanent data stored in electronic format in a logical and systematic manner, so that software or persons querying for information can easily retrieve them. Despite their increasing size, these collections of data are organized to support a simple mechanism for management and retrieval of information and have become useful in a wide range of applications.

Chemical biology, medicinal chemistry, and molecular biology call for an increased number of databases that index not only chemical information (SciFinder [1], Beilstein [2]) but also biological data (biological and target-specific assay details), chemoinformatics data (calculated and measured properties associated with chemicals), and bioinformatics data (target-related information). The pharmaceutical industry counts on these types of databases for achieving milestones in the drug discovery process, namely, for target (macromolecule) and lead (small-molecule) identification. Such databases, enabled with built-in search engines for appropriately querying chemical and biological information, have become integral to the discovery process and are continuously expanding. The integration process demands hierarchical classification systems to facilitate simultaneous mining after information associated with target-focused chemical libraries and biological families [3].

In this chapter, we present some bioactivity databases having bioinformatics and chemoinformatics content, discuss some aspects of their integration, and describe tools for mapping the edge between chemistry and biology. For a detailed review of bioactivity database assembly, the reader is referred to our earlier work [4].

17.2 DATABASE MANAGEMENT SYSTEMS FOR CHEMISTRY AND BIOLOGY

If biological information can easily be stored in a binary format, as text or graphics, chemical information cannot be captured in a similar manner because a standard database management system [5] (DBMS, software used for designing databases) lacks the appropriate cognizance for handling chemical structures. In order to overcome this issue, special systems or extension modules had to be designed and implemented to enable storage, search, and retrieval of chemical information (data structures). Some representative systems that have been developed until now are presented in the following.

The *Catalyst* [6] software suite from Accelrys provides an integrated environment for three-dimensional (3-D) information management and pharmacophore modeling, valuable in drug discovery research. This integrated environment allows for seamless access to complementary capabilities such as generation of multiple conformations with extensive coverage of conformational space, pharmacophore-based alignment of molecules, shape-based 3-D searching, and automated generation of pharmacophore hypotheses based on structure–activity relationship (SAR) data.

ChemFinder is a small-enterprise DBMS designed by CambridgeSoft Corporation [7] that has been providing free chemical searching to hundreds of thousands of scientists since 1995. It can be used as a stand-alone software or can be connected to Oracle [8] and Microsoft Access. Also, extension modules for Microsoft Word and Microsoft Excel are available. Another product offered by the same company is *ChemOffice WebServer*, an enterprise server for development and also a leading solution platform for scientific data storage and sharing. Accessible through a web browser, the ChemOffice WebServer provides an organized management system for chemical and biological databases. ChemOffice WebServer is included in all of the ChemOffice Enterprise suites (solutions for knowledge management, chemical informatics, biological informatics, and chemical databases). ChemOffice WebServer SDK extends the Microsoft and Oracle platforms, allowing information scientists to use the most powerful development tools. *ChemDraw* is the equivalent of MDL's ISIS/Draw [9] tool for chemical structure drawing [7].

ChemoSoft [10] from Chemical Diversity Labs Inc. is an integrated software environment ensuring chemoinformatics solutions for drug design and for combinatorial and classical chemistry. ChemoSoft offers a low-cost, reliable, and efficient solution due to an interface with standard SQL servers (Oracle, Microsoft SQL Server [11] and Borland Interbase [12]) and *ChemWebServer*. The *SQL Link Library* is designed to connect ChemoSoft to the SQL server, where the user can export, import, browse, and edit data. The ChemWebServer is intended to expose chemical databases via the Internet. ChemoSoft provides useful utilities from the following groups: (1) tools for browsing, editing, and correcting files of certain formats that differ from ChemoSoft ones; (2) tools for correcting errors in database structures, search for tautomers (most probably, multiplicates of the same substance), the refinement of the inconvenient display of structures, and the replacement of a certain moiety of database structures by another fragment; (3) utility for a multiple condition search; and (4) add-in storing and rendering trivial names for predefined structures.

MDL ISIS/Base [9] from Symyx MDL [13] is a flexible desktop database management system for storing, searching, and retrieving chemical structures and associated scientific data. Its form-based searching provides for the end user a customizable and simple exploitation, allowing a combination between chemical structure searches, text, and/or numeric queries. Another ISIS family member, MDL *Isentris* [14] is a desktop environment for efficiently searching data, analyzing results, reporting, and sharing and managing research information in a collaborative manner. Using this product is possible to deliver essential scientific data into scientific workflows; to create solutions that combine proprietary data and commercial information, in-house applications, MDL applications, and specialized software from other vendors; and to search, browse filter, visualize and report warehouse data, and so on.

Isentris and ISIS/Base are enterprise solution chemical DBMSs that can be connected to Oracle [8] using the *MDL cartridge*. These systems require *MDL/Draw* (or *ISIS/Draw*—an older product) for structure drawing. The incorpo-

rated development toolkit allows the creation of dialog boxes, tool bars, buttons, and special forms. Custom modules can be written using the ISIS programming language for automating procedures such as data registration. ISIS/Base/Draw is a good solution for Windows-based front-end systems [9].

Daylight CIS [15] offers a chemical database system that has built-in knowledge about chemical graph theory and it achieves high performance when storing chemical information, Thesaurus Oriented Retrieval (*THOR*) [16]. THOR features extremely fast data-retrieval time, independent of database size or complexity. The primary key used in the database is the molecular structure stored in SMILES (Simplified Molecular Input Line Entry Specification) [17,18] format designed by Daylight—simple and comprehensive chemical language in which molecules and reactions can be specified using ASCII characters representing atom and bond symbols. This feature distinguishes Daylight components from competition, as they require minimal storage, provide space efficiency, and very fast retrieval times; various toolkits are available for high-end customization, tailored to data visualization, query, and storage.

Instant JChem (IJC) is a Java [19] tool from ChemAxon Ltd. [20], based on JChem Base, for the development of applications that allow the searching of mixed chemical structure and nonstructural data and can integrate a variety of database systems (Oracle, SQL, Access, etc.) with web interfaces. By using the JChem Cartridge for Oracle, the user can acquire additional functionalities from within Oracle's SQL. The system includes Marvin, a Java-based chemical editor and viewer. Marvin tool is in the same time a fast substructure, similarity, and exact search engine using two-dimensional (2-D) hashed fingerprints [20]. IJC is a database-centric desktop application that merges MDL-like visual queries and Daylight-like SMARTS (SMiles ARbitrary Target Specification) queries, enabling scientists to perform high-level searches on high-volume chemical databases.

UNITY [21], a DBMS package accessible through SYBYL, the molecular modeling system from Tripos [22], combines database searching with molecular design and analysis tools. UNITY allows one to build structural queries based on molecules, molecular fragments, pharmacophore models, or receptor sites. In addition to atoms and bonds, 3-D queries can include features such as lines, planes, centroids, extension points, hydrogen bond sites, and hydrophobic sites. The UNITY relational database interface within SYBYL provides access to Oracle data associated with structures.

17.3 INFORMATIONAL STRUCTURE OF BIOACTIVITY DATABASES

17.3.1 Chemical Information

Chemical information collected in any database is represented by chemical structures that are encoded into a machine-readable format. In this format,

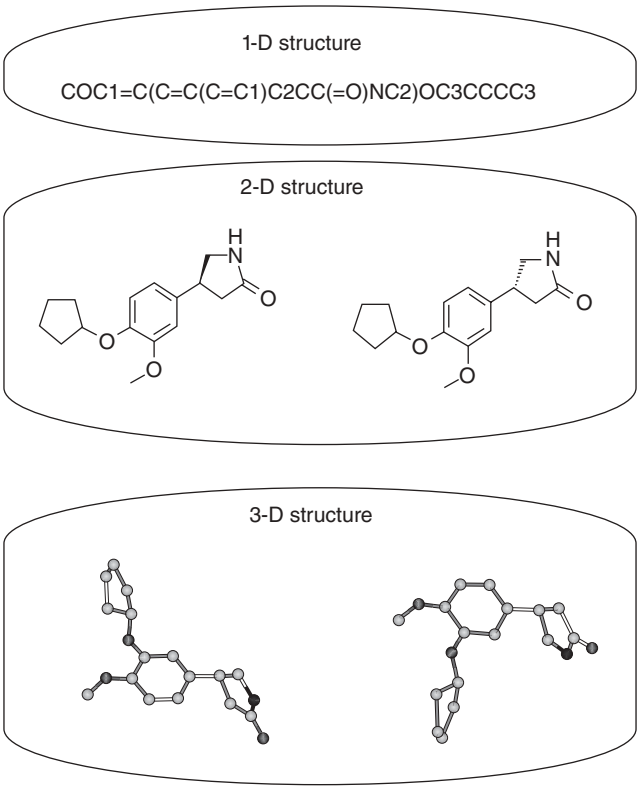


Figure 17.1 The representation of chemical information exemplified for rolipram.

the atomic connectivity, which is the main characteristic for storing chemical structures, is depicted by connection tables that store 2-D and/or 3-D atomic coordinates. The information comprised in a chemical structure can be illustrated hierarchically through one-dimensional (1-D), 2-D, or 3-D representations (Fig. 17.1).

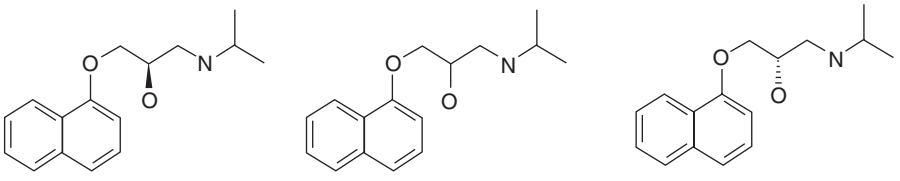
In 1-D representation, chemical structures can be depicted as canonical nonisomeric SMILES [17,18] that do not contain stereochemical information. The SMILES code is a simple comprehensive chemical language that contains the same information as is found in an extended connection table. A molecule is represented as a string using ASCII characters corresponding to atom and bond symbols. The SMILES string is small, taking approximately 1.5 ASCII characters per atom, so it can be manipulated easily for query tasks in chemical databases. In order to simplify a query in chemical databases and to obtain more specific results, one can use SMARTS [23], which is an ASCII character string very similar to SMILES language but different at the semantic level because SMILES describes molecules, whereas SMARTS describes patterns.

Molecular fingerprints are 1-D descriptors used for exhaustive searches on chemical databases, mainly similarity queries. Fingerprints are a very abstract representation of certain structural features of a molecule, being a boolean array, or bitmap, with no assigned meaning to each bit.

An invariant character string representation of chemical structures is the International Chemical Identifier (InChI) [24], a textual identifier developed jointly by IUPAC (International Union of Pure and Applied Chemistry) [25] and NIST (National Institute of Standards and Technology) [26]. InChI was designed to provide a standardized method to encode chemical information and to facilitate the search for this information in databases. InChI enables an automatic conversion of graphical representations for chemical substances into the unique InChI labels, which can be later restored by any chemical structure drawing software. The convention for chemical structure representation in the InChI system uses IUPAC rules as an input structure representation for normalization and canonicalization. Like SMILES technologies from open source projects such as Open Babel [27], InChI is freely available.

2-D representations of chemical structures are simplified graphical models in which atoms are depicted by their atomic symbols and bonds are represented as lines between these symbols. In this type of representation, stereochemical information may be encoded (Fig. 17.2). Many chemical database systems store chemical structures as 2-D representation, perhaps adding isomeric SMILES along with other structural information. Chemical data formats that render chemical information into a machine-readable format are ASCII (text) or binary format. As a consequence of ASCII format redundancy, it is recommendable to deposit data in a compressed format when text format is used. Although data compression is a slow process, it is performed only once when records are registered into the database and a decompression step allows a quick execution. The binary format is less flexible but is comparable in size with the compressed text format. More than a decade ago, Chemical Markup Language [28,29] (CML), a unifier format, was developed as a new approach for managing molecular information using Internet tools like XLM [30,31] and Java [19]. CML can hold very complex information structures and works as an information interchange mechanism.

3-D representations of chemical structures extend the information existent in SMILES or 2-D representations by introducing atomic coordinates,



R-(+)-propanolol (+/–)-propanolol S-(–)-propanolol

Figure 17.2 2-D chemical representation of propanolol and its enantiomers.

information that can be individually unique to conformers. In the 3-D representation, stereochemical information is preserved for both chiral centers and double bonds. The 3-D level is dedicated to structures with *xyz* coordinates (e.g., small-molecule crystallographic data, molecular models, and known bioactive conformers from experimental determinations) and for properties and characteristics that really depend on 3-D structures (volumes, surfaces, VolSurf [32] descriptors, etc.).

In addition to atomic connectivity, annotations using external identifiers and additional attributes regarding, e.g., known protomeric or tautomeric states, as well as structural keywords are important for accurate control of chemical information. Data regarding trivial or generic names, CAS (Chemical Abstracts Service) registry numbers [33,34], or IUPAC nomenclature [35] are very helpful in elucidating some ambiguous records like chemical structure errors, different salt formulations, and tautomers.

17.3.2 Biological Activity Information

Biological activity information refers to bioactivity data, e.g., value and activity type, together with qualitative information, e.g., agonist or antagonist, as well as additional assay information regarding tissues or cells for *in vitro* and whole organism for *in vivo* determinations, assays protocols, and supporting literature.

Biological activity data are expressed as (1) inhibitory concentration at 50% (IC_{50}), which represents the molar concentration of an inhibitor (antagonist) that reduces the biological response (reaction velocity) of a substrate (agonist) by 50%; also other percentage values can be determined—IC_{30}, IC_{90}, and so on; (2) effective concentration at 50% (EC_{50}), which refers to the molar concentration of a substrate (agonist) that produces 50% of the maximal biological effect of that substrate (agonist); (3) inhibition constant (K_i) and direct binding experiment equilibrium dissociation constant (K_d); (4) A_2—the molar concentration of an antagonist that requires double concentration of the agonist to elicit the same submaximal response, obtained in the absence of an antagonist; (5) effective dose at 50% (ED_{50}), which represents the dose of a drug that produces, on average, a specified all-or-none response in 50% of the test population or, if the response is graded, the dose that produces 50% of the maximal response to that drug; and (6) minimum inhibitory concentration (MIC), the lowest concentration of an antimicrobial that will inhibit the visible growth of a microorganism during overnight incubation [36].

To facilitate data mining, activity values should be stored not only in the assay-specific format but also in the same scale, e.g., logarithmic. The standard error of the mean (SEM) got from multiple determinations or the special situations when one compound does not show biological activity or when the exact activity value cannot be the determined can also be stored together with normal activity because it represents useful information for further data mining activity.

Separate but also categorized as bioactivity data are end points relevant to the clinical context: oral bioavailability; metabolic stability; fraction of the drug bound to plasma proteins; renal, hepatic, and systemic clearance; volume of distribution at steady state; fraction of drug excreted unchanged; plasma half-life; maximum recommended daily dose; lethal dose 50%; and so on. These properties, generically classified as absorption, distribution, metabolism, excretion, and toxicity ("ADME/Tox"), are high-level end-point determinations that are more often encountered in late-phase discovery, specifically for potential clinical candidates or drugs. Species-specific animal data, as well as human data, are often available for most drugs. Also, an increasing amount of data is becoming available in the context of allele determinations, whereby specific phenotypes are taken into account for the same biological properties. For example, the influence of "slow" or "fast" metabolizers with respect to drug metabolism is becoming increasingly relevant in the context of drug–drug and food–drug interactions.

17.3.3 Target Information

Besides chemical and biological information, it is increasingly expected that bioactivity databases store target-specific information, namely, target and gene data. Many bioinformatics databases are freely available on the Internet, yet some discrepancies exist, due to different audiences, the major purpose of each resource, and the diversity of classification criteria. Nevertheless, navigation among them is possible due to a set of hyperlinked unique identifiers (the equivalent of chemical names) or uniform resource locators [37] (URLs), which are constructed from the unique identifiers of every entry.

Most bioactivity databases contain a target module where basic biological data and external identifiers are stored. In a simplified case, target information is organized as follows: an internal identifier, used to relate records within the bioactivity database; target description (flat text) that contains the target name, perhaps some synonyms, as well as other information related to its function, classification, and species; and searchable key words and comments related to specific bioassays. Therefore, target fields can cover a large amount of information that may be systematized using different criteria. For example, functional criteria allow targets to be categorized as proteins or as nucleic acids; furthermore, a protein target can be an enzyme, a receptor, an ion channel, a transporter, or perhaps some other (unspecified) protein. Enzymes can further be classified according to the Enzyme Codebook (E.C.) number, according to the six major biochemical classes [38]: oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases; receptors may be categorized as G protein-coupled receptors (GPCRs) [39], nuclear receptors [40] (NRs), integrins, and so on, each with its own subclass hierarchy.

Key words are equally important in database management since they often become an integral part of the query process. Each database should have a predefined dictionary, namely, a set of stored key words that enable the user to easily formulate queries and to navigate across (in particular large) databases. These predefined key words should be descriptive and meaningful; i.e., they have to be significant for a considerable number of entries. Should the key words prove to be too general and apply to the vast majority of the entries, the query outcome would not be relevant. At the opposite end of the spectrum, when the key words are too specific, they will only apply to a very small number of entries and the query will result in incomplete information. Therefore, special care has to be taken when such dictionaries are built. Furthermore, the dictionary is an important tool for the standardization of a database. For example, a significant number of targets can have two or more names that are synonyms; this can be confusing not only to nonbiologists but also to the biologists themselves (see below). A dictionary should keep track of synonyms in order to provide a high success rate for queries. Swiss-Prot [41], a protein-oriented database, solves to some extent the target standardization problem.

17.3.4 Information Drift

Chemical and biological data are subject to temporal drift, i.e., to the rather serious possibility that information changes over time. For example, numbers related to the affinity of propranolol, the first beta-adrenergic antagonist used in the clinic, moved from approximately 50–100 nM in the early 1960s, to a single-digit nanomolar in current literature. This is caused not only by an increased accuracy in assay determination, but also by our deeper understanding of the biology: at least three different beta-adrenergic receptor subtypes have been characterized to date—information that was not available at the time of propranolol's discovery. The racemate has since been separated to its enantiopure components, and we now know that the S-isomer is primarily responsible for the receptor-mediated event, whereas R-propranolol tends to have membrane-specific activities that may relate to its use as antiarrhythmic. Bioactivity values for propranolol and its enantiomers, including affinities to the β_1, β_2, and β_3 adrenoceptors, are summarized in Table 17.1.

This rather simple example illustrates the magnitude of the problem: "old" experiments, while valuable, are rendered obsolete in the light of novel discoveries, and "new" experiments may be required to address specific issues. Chemical structures tend to be resolved into enantiopure components, and perhaps different salt formulations and/or different microenvironment values (such as pH, temperature) can significantly impact the outcome. Furthermore, what initially was considered a single target receptor, e.g., the generically termed "beta" adrenergic receptor from the late 1940s, may prove to be a population of two, or perhaps more, receptors.

TABLE 17.1 Propranolol and Some of Its Bioactivities

Ligand	Receptor	Action	Affinity	Units	Database
S-Propranolol	β_1-Adrenergic (human)	Antagonist	8.9–8.2	pK_i	IUPHAR
Propranolol	β_2-Adrenergic (human)	Antagonist	9.5–9.1	pK_i	IUPHAR
Propranolol	β_3-Adrenergic (human)	Antagonist	7.2–6.3	pK_i	IUPHAR
S-Propranolol	5-HT_{1A} (human)	Antagonist	7.5	pK_i	IUPHAR
S-Propranolol	5-ht_{5a} (human)	Antagonist	5.1	pK_i	IUPHAR
Propranolol	5-HT_{1B} (human)	Antagonist	5.38	pK_i	WOMBAT
Propranolol	5-HT_{1B} (human)	Antagonist	5.38	pK_i	WOMBAT
S-Propranolol	ABCB1 (human)	Inhibitor	3.24	pK_i	WOMBAT
R-Propranolol	ABCB1 (human)	Inhibitor	3.23	pK_i	WOMBAT
S-Propranolol	CYP2D6 (human)	Inhibitor	5.85	pK_i	WOMBAT
R-Propranolol	CYP2D6 (human)	Inhibitor	5.72	pK_i	WOMBAT
S-Propranolol	CYP1A2 (human)	Inhibitor	3.97	pK_i	WOMBAT

5-HT = 5-hydroxytryptamine; ABCB1 = ATP-binding cassette B1 transporter; CYP = cytochrome P450; IUPHAR = International Union of Basic and Clinical Pharmacology; WOMBAT = WOrld of Molecular BioAcTivity database.

17.3.5 Protocol Information

Most bioactivity databases also contain information related to *in vitro* and/or *in vivo* experimental assays, grouped under "protocol information." Such data usually capture information related to the specifics of individual assays: endogenous ligand for a specific target, substrate, radioligand, temperature, pH, buffer, incubation time, method used (spectrophotometric, fluorimetric, etc.), as well as any number of details that may be useful within the context of a data mining experiment.

The protocols are usually standardized, but they are also subject to change, since methods have been constantly improved and technical accuracy continues to evolve. Due to the above, and because molecular biology has a dynamic character and targets are discovered and corrected all the time, updating is a necessity and has to be done on a monthly basis to ensure that the database(s) capture up-to-date, current information.

17.3.6 References

References are formatted bibliographic information fields that usually contain author or inventor names, reference titles, type and name of the publication (journal, patent, book, book chapter, etc.), as well as other specific record locators such as volume, issue or patent number, page numbers, and publisher. Literature data can be stored in full-text databases that capture the whole text of the original published work, or in bibliographic databases that contain only references, thus serving as a way to track specific documents.

Relevant bibliographic databases for chemistry, medicinal chemistry, and biology are the Chemical Abstracts Service [42], Medline [43] (via PubMed), and BIOSIS [44]. These databases index along structures, biological assay, or targets, references related to the captured information. The reference field gives data that describe the content of a document (abstract), being the special feature design to identify an article, patent, book, and so on, related to the specified query.

Bibliographic information may be stored in one or more fields, depending on the type and features of that particular database system. For example, the reference section of each publication captures, typically in text format, bibliographic information that includes authors, title, journal/book name or patent number, page numbers, and so on. Because article and patent data types are different, this reference format is often divergent and sometimes makes reference information handling unproductive and not viable. The simplest solution to this issue is to use a multilevel hierarchical format (Fig. 17.3). A first level should comprise the common elements of any document type like author names, title, or base URLs [37]. Subsequently, the document type should be identified. Based on this information, the following levels index particular elements of the document: journal name, volume and/or issue for a journal, publisher for a book or book chapter, applicants, and patent number for a patent. The more information that is captured, the higher the hit rate for a specific query will be. Additional fields can be assigned for web hyperlinks to the full-text publication, and/or cross-linking identifiers. When online subscription is not available, abstracts can be accessed via digital object identifiers [45,46] (DOIs) or PubMed [43] identifiers. The use of DOIs as identifiers avoids the "information drift" phenomenon (see Section 17.3.4), since Internet-based information is in danger of vanishing due to "link rot" (expired web links). Any referred object can be easily located by accessing the hyperlink

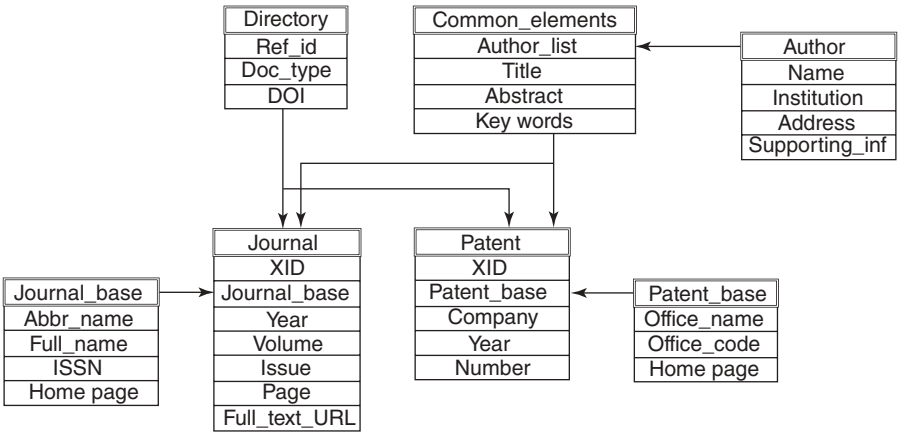


Figure 17.3 A multilevel hierarchical format for the references.

resulted from the concatenation of the base URL http:/dx.doi.org/ with the DOI value [45,46].

17.3.7 Integration with Other Databases

Automated data integration from multiple sources (databases) has become mandatory, since manual multidatabase queries involve large amounts of data transfer, in addition to being time-consuming. As a result of the multidisciplinary characteristics of research, heterogeneous databases no longer fulfill user requirements. For example, in order to get the desired results, the user many need to query a database, find the data of interest by analyzing the results, then use this data to query other databases. It is almost impossible to manage the current information flow with a single DBMS, as different databases have different schemata, data types, and formats. In one data schemata scenario, it is difficult to collect diverse information from different sources, even if similar data sets are used, due to the discrepancy of purpose among various information sources. Another scenario, more suitable for communication diversity, data interchanging between local databases and reinforcing data accessibility without affecting the local database autonomy, is the multidatabase.

The multidatabase supplies full database functionality and works to resolute the discrepancies in data representation and functions between local DBMSs. Federated DBMSs allow a continuous function for existent applications, support controlled integration of existing databases, and make possible incorporation of new applications and new databases [47,48]. Tools like CORBA [49,50], Java [19], XLM [30,31], and HTML [51,52] offer a dominant and flexible technique for integrating data from different databases.

Another extensively used integration technique is the insertion of external identifiers that point to related information from other databases. The frequently used identifiers—links to other web resources—are principally built based on URLs [37], which in turn act as entry points into databases, and associated numeric or alphanumeric identifiers for a specific resource.

17.4 AVAILABLE BIOLOGICAL AND BIOACTIVITY DATABASES

In this section, we present a short overview of some relevant public or commercially available databases that contain information on biological target or small-molecule ligands as well as a number of integrated databases. Due to the ever-changing nature of databases, any numbers and websites discussed here are for informative purpose only, and are likely to become less accurate over time.

17.4.1 Bioactivity Databases

17.4.1.1 Free-Access Databases The number of integrated database resources is increasing [53]. Some of these open-access resources cover a wide range of chemical and biological information, which makes it daunting to

TABLE 17.2 Publicly Available Databases

Database Type	Database Name	Homepage
Target/bioactivity/ chemoinformatics	BIDD (Bioinformatics & Drug Design)	http://bidd.nus.edu.sg/group/bidd.htm
Bioactivity	ChemBank	http://chembank.broad.harvard.edu/
Bioactivity	Drugbank	http://www.drugbank.ca/
Target/bioactivity	KiBank	http://kibank.iis.u-tokyo.ac.jp/
Chemical	LigandInfo	http://ligand.info/
Target/bioactivity	PDSP Ki	http://pdsp.med.unc.edu/pdsp.php
Target/bioactivity	PubChem	http://pubchem.ncbi.nlm.nih.gov/
Chemical	ZINC	http://zinc.docking.org/
Medicinal chemistry	StARlite	Not yet available

summarize. In what follows, we selectively present some of the most relevant and complete databases, as summarized in Table 17.2.

BIDD Database [54] is a collection of databanks that has been developed at the Computational Science Department of the National University of Singapore. These databases are divided in three categories: (1) pharmainformatics databases provide information about therapeutic targets: *Therapeutic Target Database (TTD)* [55] contains 1535 targets and 2107 drugs/ligands; drug adverse reaction, *Drug Adverse Reaction Target (DART)* [56]; *ADME/Tox*, *ADME-Associated Protein* [57] has 321 protein entries and therapeutically relevant multiple pathways contain 11 entries of multiple pathways, 97 entries of individual pathways, 120 targets covering 72 disease conditions along with 120 drugs; (2) bioinformatics databases are represented by the *Computed Ligand Binding Energy (CLiBE)* [58] database that contains 14731 entries (2803 distinctive ligands and 2256 distinctive receptors) and *Kinetic Data of Biomolecular Interactions* (*KDBI)* [59], which currently contains 20,803 records, which involve 2934 protein/protein complexes, 870 nucleic acids, and 6713 small molecules; and the third category is dedicated to (3) herbinformatics databases: *Traditional Chinese Medicine Information Database (TCM-ID)* [60], which includes herbal and chemical composition, molecular structure, functional properties, and therapeutic and toxicity effects. Also, these databases contain chemical structures, associated references, and cross-links to other relevant databases.

ChemBank [61,62], developed at the Broad Institute, and funded in large part by the National Cancer Institute (NCI) [63], is a public, web-based informatics. This database includes freely available data derived from small molecules and small-molecule screens and provides resources for studying the effect that small molecules have on biological systems. This project is intended to help biologists develop new screening methods or biological assays for the identification of chemical probes and to guide chemists to design novel compounds or libraries. ChemBank stores over 1 million compounds, including 150,000 commercially available compounds tagged with vendors' catalog

numbers, varied sets of cell measurements derived from high-throughput screening, and small-molecule microarrays assays. This database is powered by Daylight [15] and has a large pool of analysis tools that allows the relationships between cell states, cell measurements, and small molecules to be determined. ChemBank can be searched using different critera, e.g., molecule name, substructure or similarity search, and assay type. The query result provides information related with the compound (name, SMILES, 2-D representation, molecular weight and formula), vendors, classification, identifiers (CAS numbers [33,34] and ICCB (Institute of Chemistry and Cell Biology) identifier [64]). Also, one can find other important information, namely, characterized activities and observed biological effects, as well as cross references to other databases, such as PubMed [43] supplementary resources relevant to small molecules, chemoinformatics, and high-throughput screening. The search results can be downloaded in two well-known formats, SDF (structure-data file) and XML.

Drugbank [65,66] is a web-based dual-purpose bioinformatics–chemoinformatics initiative developed at the University of Alberta, Canada that combines quantitative, analytic, or molecular-scale information about drugs and drug targets. This database offers drug-related information, e.g., drug's name, chemical structure, experimental and calculated physicochemical properties, pharmacology, mechanism of action, toxicity, and comprehensive data about drug targets, e.g., sequence, structure, and pathways. Drugbank contains detailed information about bibliographic references, interactions with other drugs, and patient information, e.g., clinical indications of a particular drug, dosage forms, and side effects.

Nearly 4800 drug entries in Drugbank are divided into four major categories: (1) Food and Drug Administration (FDA)-approved [67] small-molecule drugs—more than 1480 entries; (2) FDA-approved biotech (protein/peptide) drugs—128 entries; (3) nutraceuticals or micronutrients such as vitamins and metabolites—71 entries; and (4) experimental drugs, which include unapproved drugs, delisted drugs, illicit drugs, enzyme inhibitors, and potential toxins—more than 3200 entries. Also, more than 2500 nonredundant drug target protein sequences are related to the FDA-approved drug entries. Being a web-enabled database, it has many built-in tools and features for query and results display. The database may be queried in different ways: simple or extensive text queries; chemical structure searches using a drawing application (e.g., MarvinSketch [20]) or SMILES strings; sequence search for proteins, which allows user to carry out both simple and multiple sequence queries; and relational query searches permit to the user to select one or more data fields and to search for ranges, occurrences of numbers, words, or string. The results of queries are displayed as an HTML format. Additionally, many data fields are hyperlinked to other databases (ChEBI [68], PubChem [69], KEGG (Kyoto Encyclopedia of Genes and Genomes) [70], PDB (Protein Data Bank) [71], Swiss-Prot [41], GenBank [72]), abstracts, digital images and interactive applets (3-D representations) for viewing molecular structures of drugs and

related targets. Most of the informational contain text, sequence, structure and image data, which can be freely downloaded.

KiBank [73,74] is a free online project that has been developed at the University of Tokyo, Japan [75], and it is designed to support the scientific community, which is interested in structure-based drug design. This database stores information related to the biological activity of chemical compounds, namely, the inhibition constant (Ki) values, species and experimental protocol, as well as chemical information: 3-D structures of target proteins and chemicals. Ki values are gathered from peer-reviewed literature searched via PubMed [43]; bibliographic references are indexed as hyperlinks; and bioactivity data can be downloadable in comma-separated file format. The 3-D structure files of target proteins are originally from Protein Data Bank [71], while the 2-D structure files of the chemicals are collected together with the Ki values and then converted into 3-D ones, being stored in PDB format and MDL MOL, respectively. There are two search methods available: by chemical name—this type of query retrieves a list with all targets that have biological activity data available for that particular compound; and by protein name/function—this target-oriented query provides a list of all compounds that bind to a specified target. These lists are cross referenced and the user can easily toggle between them. KiBank provides structure files of proteins and chemicals ready for use in virtual screening through automated docking methods, while the Ki values can be applied for tests of docking/scoring combinations, program parameter settings, and calibration of empirical scoring functions. Additionally, the chemical structures and corresponding Ki values in KiBank are useful for lead optimization based on quantitative structure–activity relationship (QSAR) techniques. KiBank is updated regularly; the May 2008 version has more than 16,000 entries for Ki values, 50 targets (protein structures), and 5900 chemical structures.

Ligand Info: Small-Molecule Meta-Database [76] is a collection of biologically annotated compounds compiled from various publicly available sources, which are collections of small molecules such as ChemBank [61], ChemPdb [77], KEGG [70], NCI [63], Akos GmbH [78], Asinex Ltd. [79], and TimTec. The current size of this database is 1,159,274 entries. These compounds are collected from different sources and for this reason, different data are included such as FDA approval status and anti-HIV activity; some molecules have predicted biological activity including pharmacological effects, mechanism of action, carcinogenicity, teratogenicity, mutagenicity, embriotoxicity, and druglikeness. The principle behind this project is based on the supposition that small molecules with similar structures have similar binding properties. The developed system allows the end user to search for similar compounds in the database using structural indices that are constructed by averaging indexes of related molecules. Thus, using a Java-based tool, the system can interactively cluster sets of molecules creating index profiles for the user and can automatically download similar molecules from the server in an SDF format. This resource was planned for online virtual screen-

ing, and it enables a rapid and receptive index for compound similarity searching.

PDSP Ki [80] database is free resource initiative developed at the University of North Carolina and is funded, mainly, by the National Institute of Mental Health. This database is a repository of experimental results (Ki), which provides information related with drugs and their binding properties to an outstanding number of molecular targets. It was designed as a data warehouse for in-house and published Ki or affinity values for a large number of drugs and candidate drugs binding to GPCRs, transporters, ion channels. and enzymes. The flexible user interface provides tools for customized data mining (Ki graphs, receptor homology, and ligand selectivity). The queries can be performed on a well-designed interface, which includes the following searchable fields: receptor name, species, tissue source, radiolabeled and tested ligand, bibliographic reference, and Ki values. The end user can search by any field or a combination of them to refine the search criteria, and the system can retrieve the results list cross-linked with corresponding entries in PubChem [69] and PubMed [43]; also, this database can be downloaded as a compressed ASCII file and one can enter his results in the database using the provided input web page. The current size of the database is 46,249 Ki value entries.

PubChem [69]—Public Cheminformatics Database—has been developed as a dedicated chemoinformatics resource by the National Institute of Health being a component of Molecular Libraries Roadmap Initiative [81], and it is hosted by the National Library of Medicine. PubChem is a public database that contains a huge amount of annotated information about the biological activities of small molecules. The chemical and biological information are linked to other databases, like ChemSpider [82], Drugbank [65,66], Sigma-Aldrich, or other Entrez databases allowing the user a direct access to information update on chemical data and biological properties. Also, Pubchem is linked to NCBI's (National Center for Biotechnology Information) 3-D protein structure database and PubMed database, which contains biomedical- and life science-related literature. Powered by Openeye [83] and CACTUS [84] software, PubChem has a complex organization, as three linked databases within the NCBI's Entrez information retrieval system. These are (1) PubChem Substance database, which contains complete descriptions of chemical samples deposited from a variety of sources and links to PubMed citations, 3-D structures for proteins, and biological activity values available from screening, as well as links to the other two databases; (2) PubChem Compound comprises validated chemical depiction information provided to describe substances from PubChem Substance; structures stored are preclustered and cross referenced by identity and similarity groups; furthermore, calculated properties and descriptors are available for searching and filtering of chemical structures; and (3) PubChem BioAssay database enclose bioactivity assays for chemical substances deposited in PubChem Substance; it provides searchable descriptions of each bioassay, including descriptions of the conditions and readouts specific to that screening procedure. The entire information incorporated in

PubChem is free, and it is available and can be downloaded through the PubChem FTP site. The February 2008 release of Pubchem contains more than 40 million substances, 19 million compounds, and 1055 bioassays.

StARlite, a medicinal chemistry database developed for Inpharmatica and Galapagos NV (until 2008) was transferred to the public domain with support of a five-year award from the Wellcome Trust. StARlite, now available from the European Bioinformatics Institute (EBI) is a chemogenomics database that indexes biological and chemical data abstracted from the primary medicinal chemistry literature, for known compounds and their pharmacological effects. It currently contains around 450,000 compounds and covers about 3,000 targets, of which approximately 1,700 are human proteins. StARlite has more than 2 million experimental bioactivities and is updated on a monthly basis, both with new data and also with error-checking. The anticipated growth of StARlite is around 10% per annum, and is intended as a complementary service to Pubchem and ChemBank. Together with other open-source databases, DrugStore (a database of drugs) and CandiStore (a database of approximately 10,000 compounds in clinical development), StARlite forms the ChEMBL chemical database, part of the EMBL (European Molecular Biology Laboratory). Using ChEMBL, one can track the progress of a structure from lead optimization through clinical development and launch phase [85,86]. ZINC is a free database initiative that has been developed at the University of California, San Francisco [86], and it is a curate collection of commercially available chemical compounds, many of them "druglike" or "leadlike," available in 3-D formats ready for docking. Catalogs from more than 40 vendors are uploaded in ZINC, and compounds can be purchased directly from vendors. The user can search this databases using the Java Molecular Editor (JME) for chemical structures [87], which generates SMILES [17,18], or using directly SMILES or SMARTS [23] strings, or compose a query specifying molecular property constraints available in the specific fields. Also, one can search the database entering the vendor's name, catalog number, and/or ZINC code. Additionally, different subsets are available for download, e.g., leadlike compounds, fragment-like, druglike, clean-leads, all-purchasable, etc., and it can be downloaded in different common file formats: mol2, SDF, SMILES, and flexibase. The ZINC8 release contains over 10 million commercially available molecules ready for virtual screening.

17.4.1.2 Commercially Available Databases The trend observed in open-access resources is paralleled by commercial databases as well. The importance of curation and the presence of errors in the literature domain have been discussed elsewhere [88]. Commercial reources can be invaluable for summarizing project-specific information, e.g., bioisosterism, GPCR-active compounds, and kinase inhibitors. Some of the most relevant examples of commercially available databases are summarized in Table 17.3.

BIOSTER [89] is a collection of bioanalogous pairs of molecules (bioisosteres), which contains over 45,000 examples of biologically active molecules

TABLE 17.3 Commercially Available Databases

Database Name	Homepage
AurSCOPE	http://www.aureus-pharma.com/
MediChem	http://www.cambridgesoft.com/
Merck Index	http://www.themerckindex.cambridgesoft.com/
Kinase Knowledgebase	http://www.eidogen-sertanty.com/
MDL Drug Data Report	http://www.mdli.com/
DiscoveryGate	http://www.discoverygate.com/
GVK Biosciences databases	http://http://www.gvkbio.com/
PathArt	http://www.jubilantbiosys.com/
WOMBAT, WOMBAT-PK	http://sunsetmolecular.com/

representing drugs, prodrugs, enzyme inhibitors, peptide mimetics, and agrochemicals selected from existent literature, offered by Accelrys. This database provides key words indicating the mode of action and cross references to reports for each active compound, being a powerful and helpful tool for discovering alternate structures with enhanced efficacy, superior absorption, distribution, metabolism, excretion (ADME), toxicity profiles, and desired physical properties. The *Biotransformations* database offers information about metabolism of drugs, agrochemicals, food additives, and industrial and environmental chemicals in vertebrates. Also, it contains the original literature, test systems, and a set of generic key words. Like BIOSTER, the *Metabolism* database has been designed for use with MDL ISIS. This product supplies biotransformations of organic molecules in a wide variety of species by providing primary information on the metabolic fate of organic molecules [89].

Aureus [90] offers several high-value knowledge databases (*AurSCOPE* [91]) of chemical and biological data, including quantitative activity data on GPCR, kinases, ion channel, and drug–drug interactions (generally Cyp450). These databases contain, besides chemical information, *in vitro* and *in vivo* biological data with complete descriptions of the biological assays. *AurSCOPE GPCR* is a fully annotated structured knowledge database containing chemical and biological information relating to GPCR (around 2300 targets) chemistry, pharmacology, and physiology. This product includes information regarding 500,000 biological activities for more than 100,000 ligands active on GPCR. *AurSCOPE Ion Channel* contains information about ion channel activators, openers, and blockers covering almost all ion channel targets: calcium channels, potassium channels, chloride channels, sodium channels, transmitter-gated channels, and so on. *AurSCOPE Kinase* contains biological activity data mined from journals and patents associated with chemical SARs concerning kinase–ligand interactions. *AurSCOPE ADME/Drug–Drug Interactions* encloses biological and chemical information (97,000 bioactivities for 4250 molecules) related to metabolic properties of drugs (1770 metabolites and 420 targets), which permits the identification of potential drug–drug interaction. *AurSCOPE hERG Channel* contains significant biological and chemical infor-

mation related to the human ether-a-go-go-related gene (hERG) (7750 bioactivities for 1155 ligands).

CambridgeSoft and GVK Biosciences [92] have placed on market *MediChem* database, which is a collection of over 500,000 compounds that have been selected from the top 25 medicinal chemistry journals. Data include chemical information, literature reference (records are also linked to PubMed via a hyperlink), target information (binding information for the target and its mutants), bioactivity information, reaction pathways, chemical properties, availability of reagents, and toxicological information. Records can be queried by target platform, e.g., GPCR, kinase, and ion channels. *Traditional Chinese Medicines* database (also available from Daylight CIS) consists of over 10,000 compounds isolated from 4636 traditional Chinese medicine natural sources, which consisted generally of plants, minerals, and a small number of animals. *Ashgate Drugs Synonyms and Properties* is a database of over 8000 drug substances currently in common use worldwide. The *Merck Index* database is a structure searchable encyclopedia of chemicals, drugs, and biological active compounds. It provides more than 10,000 monographs on single substances and related groups of compounds covering chemical, generic, and brand names. The queries may be performed in different fields like structure and stereochemistry, registry numbers, physical properties, toxicity information, therapeutic uses, and literature [93].

The *Kinase Knowledgebase* [94] (KKB) is an Eidogen–Sertanty [95] database of kinase structure–activity and chemical synthesis data, which provides an overview of published knowledge and patents around kinase targets of therapeutic importance, enabling a detailed understanding of the knowledge space around the target of interest and the relevant antitargets. The presentation of inhibitor structural data consents to group known inhibitors in scaffold groups and outlines a project plan around patentable chemotypes. The overall number of unique small-molecule structures in the KKB is now greater than 440,000 records (with more than 140,000 tested molecules) for over 300 annotated kinase targets, captured from over 3800 journal articles and patents. The curation process captures chemical synthesis steps for those kinase inhibitors with detailed experimental procedures. Chemical information incorporates synthetically feasible reagents that are reported in the context of the claims of a patent being structured in protocols of generic reaction sequences. This generates all specific examples from a patent and also a comprehensive ensemble of structures (the patent space) that can possibly be made by the reported synthetic methodology, which are potentially relevant within the biological activity class. Synthetic pathways leading to these molecules are structurally linked to the biological information. QSAR models based on *in vitro* data and advanced activity models based on cellular activity and toxicity can further be selected.

MDL Discovery Knowledge package from Symyx [13] contains, besides reference works and literature links, one of the most comprehensive collection of bioactivity, chemical sourcing, synthetic methodology, metabolism, toxicology, and environment, health, and safety (EH&S) databases. *MDL Drug Data*

Report (MDDR) [96] is focused on underdevelopment or launched drugs covering the patent literature, journals, meetings, and congresses since 1988. This database counts over 132,000 entries with detailed information regarding chemical structure, biological activity, description of the therapeutic action, patent information (patent number, title, source, and name of invertors), literature references, synonyms (company codes, generic name, trade names, trademark names, etc.), the originating company, and the development phase. The *National Cancer Institute Database 2001. 1* [97] contains more than 213,000 structures gathered in four available NCI databases: (1) the *NCI 127K* database consisting of 127,000 structures with CAS registry numbers; (2) the *Plated Compounds* database containing 140,000 nonproprietary samples, which are offered to the external research community; (3) the *AIDS* database, containing 42,687 entries that have been tested for AIDS antiviral activity; (4) the *Cancer* database containing dose–response data for 37,836 compounds tested for the ability to inhibit the growth of human tumor cell lines. For each record, a 3-D model generated with Corina [98] is available. *MDL Comprehensive Medicinal Chemistry* (CMC), derived from the Drug Compendium in Pergamon's CMC, provides 3-D models and important biochemical properties including drug class, log *P*, and pKa values for over 8400 pharmaceutical compounds. The *MDL Patent Chemistry Database* indexes chemical reactions, substances, and substance-related information from organic chemistry and life sciences patent publications (world, U.S. and European) since 1976. The database contains approximately 3 million reactions, along with at least 3.8 million organic, inorganic, organometallic (and polymeric) compounds, and associated data. All these databases can be searched by compound and property, using MDL ISIS/Base or MDL Database Browser via DiscoveryGate [99]. *DiscoveryGate* is a web-enabled discovery environment that integrates, indexes, and links scientific information to provide direct access to compounds and related data, reactions, original journal articles and patents, and reliable reference works on synthetic methodologies. *CrossFire Beilstein* indexes three primary data domains: substances, reactions, and literature. The substance domain stores structural information with all associated facts and literature references, including chemical, physical, and bioactivity data.

The databases provided by GVK Biosciences [92] are large collections of compounds (more then one million) with further information regarding chemical structures, biological activity, toxicity, and pharmacological data curated from existing literature. More then three million SAR points are indexed in Oracle, XML, and ISIS/Base formats. *MediChem Database*, codistributed by CambridgeSoft, was previosly described. *Target Inhibitor Database* captures information about specific protein families: GPCRs, ion channels, NHRs, transporters, kinases, proteases, and phosphatases. *Natural Product Database* encloses compounds derived from animals, natural plants, marine, and microbial sources. Services also include DNA and protein sequence analysis, protein structure analysis, homology modeling, and visualization tools. *Reaction Database* registers reactions reported in medicinal chemistry journals [100].

The bioinformatics division of Jubilant Biosys offers *PathArt* product as a comprehensive database of biomolecular interactions with tools for searching, analysis, and visualization of data. This product includes a database component and a dynamic pathway articulator component, which build molecular interaction networks from curated databases. The comprised information allows users to upload and to map microarray expression data onto the pathways on over 900 regulatory as well as signaling pathways. *ePathArt* is a single-node locked web-enabled version of PathArt. The *GPCR Annotator* module allows users to classify the GPCR family hierarchy from sequence input and includes a wide range of therapeutically relevant areas related to GPCRs. The main product in the chemoinformatics area is the ChemBioChem suite, which includes curated databases addressed on specific targets. *GPCR ChemBioBase* contains over 400,000 small molecules acting as agonists or antagonists against 60 GPCR receptor classes from 400 journal articles and 2000 patents. *Ion Channel ChemBioBase* contains around 100,000 small molecules that act as ion channel blockers, openers, or activators against ion channels. The *Kinase ChemBioBase* database, produced by Jubilant and codistributed by Accelrys, is a comprehensive collection of over 300,000 small-molecule inhibitors active on more than 700 kinases. Quality-checked SAR points with additional information are collected from about 1500 patents and 500 journal articles. *Nuclear Hormone Receptor ChemBioBase* is a library focused on small ligands published as receptor agonists, antagonists, or modulators against NHRs. *Protease ChemBioBase* is a compilation of 400,000 ligands for proteases active against more than 100 proteases. *Antibacterial and Antifungal Database* contain over 20,000 compounds that possess activity against bacterial and fungal diseases. Another key product offered by this company is *Drug Database*, which captures over 1500 approved drugs related to biological targets [101].

Sunset Molecular Discovery LLC [102] integrates knowledge from target-driven medicinal chemistry with clinical pharmacokinetics data in the WOMBAT Database for Clinical Pharmacokinetics (WOMBAT-PK), and provides up-to-date coverage of the medicinal chemistry literature in WOMBAT, as it appears in peer-reviewed journals [103]. The *WOMBAT* database [104] (current release 2009.1) captures 295,435 (242,485 unique SMILES) chemical structures and associated biological activities against more than 1966 unique targets (GPCRs, ion channels, enzymes, and proteins). Besides exact numeric values (the vast majority), WOMBAT now captures "inactives," "less than," "greater than," as well as percent inhibition values. The The *Target and Biological Information* module provides detailed target information, including biological information (species, tissue, etc.), detailed target and target class information (including hierarchical classification for GPCRs, NHRs, and enzymes), as well as further information regarding the bioassays (e.g., radioligand and assay type). Swiss-Prot reference IDs are stored for most targets (~88%). Additional properties include several experimental and calculated properties for each chemical structure, e.g., counts of miscellaneous atom types, Lipinski's rule-of-five [105] (Ro5) parameters,

including the calculated octanol/water partition coefficient, ClogP and Tetko's calculated water solubility, polar surface area (PSA), and nonpolar surface area (NPSA). The Reference Database contains bibliographic information, including the DOI format with URL links to PDF files for all literature entries, as well as the PubMed ID for each paper. The *WOMBAT-PK* 2009 [104] captures pharmacokinetic data (over 13,000 clinical pharmacokinetic [PK] measurements) in numerical searchable format for 1,230 drugs. Physicochemical characteristics and clinical data are brought together from multiple literature sources. The existent fields allow queries using chemical information (chemical structure, SMILES codes), drug-marketed names, drug target information, multiple PK, and toxicity parameters, which are indexed in both numerical and text format. Both databases are available in the MDL ISIS/Base format, the RDF (Resource Description Framework) format, as well as the Oracle/Daycart [106] (Daylight) format.

17.4.2 Biological Information Databases

The European Bioinformatics Institute [107] (EBI) is a nonprofit academic organization that provides freely available data and bioinformatics services, managing databases of biological data including nucleic acid, protein sequences, and macromolecular structures. The most popular database is the *UniProtKB/Swiss-Prot Protein Knowledgebase Database (Swiss-Prot)* [41], and it is maintained together with the Swiss Institute for Bioinformatics (SIB). Swiss-Prot is a curated protein sequence database that provides a high level of annotation, a minimal level of redundancy, and a high level of integration with other databases. Together with UniProtKB/TrEMBL, it constitutes the Universal Protein Resource (UniProt) Knowledgebase, one component of the UniProt, which allows easy access to all publicly available information about protein sequences. The last release UniProtKB/Swiss-Prot contains 349,480 sequence entries abstracted from 164,703 references, and it is cross referenced with almost 60 different databases. Each sequence entry has captured two types of data: the core data, which correspond to sequence data, taxonomic data, and citation information, and the annotation data, which refer to protein function, secondary and quaternary structure, or similarities to other proteins.

Enzyme Structures Database (EC-PDB) [108] contains the known enzyme structures (22,899 entries) that have been deposited in the Protein Data Bank [71].

The specialized *GPCRDB* [109] and *NucleaRDB* [40] databases collect information about GPCRs and intra-NHRs, respectively. They capture information regarding sequence, structure mutation, and ligand binding data together with data resulting from computational work (phylogenetic trees, multiple sequence alignments, correlated mutation analysis). Each protein has a maximum degree of integration with other biomolecular databases. The two systems are extremely useful since they supply a large amount of the available data from a single source. The enclosed information can be easily accessed

through a hierarchical list of well-known families according to the pharmacological classification of receptors.

The *MEROPS* [110] is a resource for information on peptidases (also termed proteases, proteinases, and proteolytic enzymes) and the proteins that act as their inhibitors. Here are included almost 3000 individual peptidases and inhibitors that can be reached by use of an index under its name, MEROPS identifier, or source organism. The MEROPS database uses a hierarchical, structure-based classification of peptidases. For this, each peptidase is assigned to a family based on statistically significant similarities in amino acid sequence, and families that are thought to be homologous are grouped together in a clan (last release contains 180 families and 49 clans).

Transport Classification Database (TCDB) [111] is a curated database of factual information from over 10,000 published references, containing a comprehensive International Union of Biochemistry and Molecular Biology (IUBMB) approved classification system for membrane transport proteins known as the Transporter Classification (TC) system [112]. The TC system is equivalent to the Enzyme Commission (EC) system [38] for enzyme classification but incorporates phylogenetic information as well. Based on the TC system, the enclosed 3000 protein sequences are classified into over 550 transporter families.

TransportDB [113] is a relational database that describes the predicted cytoplasmic membrane transport protein complement for organisms having available the complete genome sequence. The membrane transport complement is identified and classified into protein families according to the TC classification system [112] for each organism. A regular update of this site is kept with the newly published genomes.

17.5 CONCLUSIONS

The age of informatics-driven pharmaceutical discovery has arrived [53]. Learning how to query disjoint data sources to answer complex knowledge discovery-type questions remains a challenge, but some major hurdles, i.e., data collection and integration, now belong to the past. We are witnessing an unprecedented amount of data integration that enables us, effectively, to consider temporal aspects of the studied systems, an area we term systems chemical biology [114].

The large collection of chemical and biological databases enables discovery scientists to focus on knowledge creation, via data analysis and interpretation. This continues to require familiarity with fundamental principles in both chemistry and biology and in data mining skills. However, we are witnessing an unprecedented integration of bioinformatics and chemoinformatics resources, where data are seamlessly merged into a comprehensive picture. Thus, database systems that seamlessly mine chemical, biological, and target-related data in an integrated manner are as vital as computers.

ACKNOWLEDGMENTS

This work was supported in part by the National Institutes of Health grant U54 MH074425-01 (National Institutes of Health Molecular Libraries Screening Center Network) and by the New Mexico Tobacco Settlement Fund.

REFERENCES

1. American Chemical Society. CAS Online/SciFinder. Available at http://www.cas.org/SCIFINDER/ (accessed April 30, 2008).
2. Elsevier Information System. CrossFire Beilstein Database. Available at http://www.beilstein.com/ (accessed April 30, 2008).
3. Cases M, Garcia-Serna R, Hettne K, Weeber M, van der Lei J, Boyer S, Mestres J. Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family. *Curr Top Med Chem* 2005;5:763–772.
4. Olah M, Oprea TI. Bioactivity databases. In: *Comprehensive Medicinal Chemistry II*, Vol. 3, edited by Taylor JB, Triggle DJ, pp. 293–313. Oxford: Elsevier, 2006.
5. Date CJ. *An Introduction to Database Systems*, 7th edn. New York: Addison Wesley, 2000, pp. 2–57.
6. Accelrys Software Inc. Catalyst. Available at http://www.accelrys.com/products/catalyst/ (accessed April 30, 2008).
7. CambridgeSoft Corporation. Available at http://www.cambridgesoft.com/ (accessed April 30, 2008).
8. Oracle Corporation. Oracle. Available at http://www.oracle.com/ (accessed April 30, 2008).
9. Symyx MDL. ISIS—Integrated Scientific Information System. Available at http://www.mdl.com/products/framework/isis/ (accessed April 30, 2008).
10. Chemical Diversity Labs, Inc. ChemoSoft. Available at http://chemosoft.com/modules/db/ (accessed April 30, 2008).
11. Microsoft Corporation. MS SQL Server. Available at http://www.microsoft.com/sql/ (accessed April 30, 2008).
12. Borland Software Corporation. InterBase 7.5. Available at http://www.borland.com/us/products/interbase/ (accessed April 30, 2008).
13. Symyx MDL. Available at http://www.mdli.com/ (accessed April 30, 2008).
14. Symyx MDL. Isentris. Available at http://www.mdl.com/products/framework/isentris/index.jsp (accessed April 30, 2008).
15. Daylight Chemical Information System, Inc. Available at http://www.daylight.com/ (accessed April 30, 2008).
16. Daylight CIS. THOR v4.8. Available at http://www.daylight.com/products/thor.html (accessed April 30, 2008).
17. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;28:31–36.

18. Weininger D, Weininger A, Weininger JL. Algorithm for generation of unique SMILES notation. *J Chem Inf Comput Sci* 1989;29:97–101.
19. Sun Microsystems, Inc. *Sun Developer Network—Java Technology*. Available at http://java.sun.com/ (accessed April 30, 2008).
20. ChemAxon Ltd. Marvin, JChem Base, JChem Cartridge. Available at http://www.chemaxon.com/ (accessed April 30, 2008).
21. Tripos, Inc. UNITY. Available at http://www.tripos.com/sciTech/inSilicoDisc/chemInfo/unity.html (accessed April 30, 2008).
22. Tripos, Inc. Available at http://www.tripos.com/ (accessed April 30, 2008).
23. Daylight CIS. Available at http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed April 30, 2008).
24. The International Chemical Identifier (InChI). Available at http://www.inchi.info/ (accessed April 30, 2008).
25. International Union of Pure and Applied Chemistry. Available at http://www.iupac.org/ (accessed April 30, 2008).
26. National Institute of Standards and Technology. Available at http://www.nist.gov/ (accessed April 30, 2008).
27. Open Babel, the open source chemistry toolbox. Available at http://openbabel.org/wiki/Main_Page (accessed April 30, 2008).
28. The Chemical Markup Language homepage. Available at http://www.xml-cml.org/, http://cml.sourceforge.net/ (accessed April 30, 2008).
29. Murray-Rust P, Rzepa HS. Chemical markup, XML, and the World Wide Web. 4. CML schema. *J Chem Inf Comput Sci* 2003;43:757–772.
30. World Wide Web Consortium (W3C). Extensible Markup Language (XML). Available at http://www.w3.org/XML/ (accessed April 30, 2008).
31. DuCharme B. *XML: The Annotated Specification*. Upper Saddle River, NJ: Prentice Hall PTR, 1999.
32. Molecular Discovery Ltd. Available at http://moldiscovery.com/ (accessed April 30, 2008).
33. Fisanick W, Shively ER. The CAS information system: Applying scientific knowledge and technology for better information. In: *Handbook of Chemoinformatics*, edited by Gasteiger J, Vol. 2, pp. 556–607. Weinheim, Germany: Wiley-VCH, 2003.
34. American Chemical Society. Chemical Abstracts Service, CAS Registry. Available at http://www.cas.org/EO/regsys.html (accessed April 30, 2008).
35. Wisniewski JL. Chemical nomenclature and structure representation: Algorithmic generation and conversion. In: *Handbook of Chemoinformatics*, edited by Gasteiger J, Vol. 2, pp. 51–79. Weinheim, Germany: Wiley-VCH, 2003.
36. Neubig RR, Spedding M, Kenakin T, Christopoulos A. International Union of Pharmacology Committee on receptor nomenclature and drug classification. XXXVIII. Update on terms and symbols in quantitative pharmacology. *Pharmacol Rev* 2003;55:597–606.
37. Uniform Resource Identifiers (URI). Generic Syntax—Draft Standard RFC 2396. 1998. Available at http://www.ietf.org/rfc/rfc2396.txt (accessed April 30, 2008).

38. International Union of Biochemistry and Molecular Biology (IUBMB). Enzyme Classification. Available at http://www.chem.qmul.ac.uk/iubmb/ (accessed April 30, 2008).
39. Horn F, Weare J, Beukers MW, Hörsch S, Bairoch A, Chen W, Edvardsen O, Campagne F, Vriend G. GPCRDB: An information system for G protein-coupled receptors. *Nucleic Acids Res* 1998;26:277–281.
40. Horn F, Vriend G, Cohen FE. Collecting and harvesting biological data: The GPCRDB and NucleaRDB information systems. *Nucleic Acids Res* 2001;29:346–349.
41. Swiss Institute of Bioinformatics. ExPASy Proteomics Server/Swiss-Prot Protein Knowledgebase. Available at http://www.expasy.org/sprot/ (accessed April 30, 2008).
42. American Chemical Society. Chemical Abstracts Service. Available at http://www.cas.org/ (accessed April 30, 2008).
43. National Center for Biotechnology Information/National Library of Medicine. Entrez PubMed. Available at http://www.ncbi.nlm.nih.gov/entrez/ (accessed April 30, 2008).
44. Thomson Scientific. BIOSIS Bibliographic Database. Available at http://www.biosis.org/, http://www.cas.org/ONLINE/DBSS/biosisss.html (accessed April 30, 2008).
45. The International DOI Foundation. The Digital Object Identifier System. Available at http://www.doi.org/ (accessed April 30, 2008).
46. The International DOI Foundation. DOI Handbook. 2005. Available at http://dx.doi.org/10.1000/186 (accessed April 30, 2008).
47. Bright MW, Hurson AR, Pakzad SH. A taxonomy and current issues in multidatabase systems. *Computer* 1992;25:50–60.
48. Larson JA. *Database Directions: From Relational to Distributed, Multimedia, and Object-Oriented Database Systems*, pp. 45–56. Upper Saddle River, NJ: Prentice-Hall PTR, 1995.
49. Siegel J. *CORBA Fundamentals and Programming*. New York: John Wiley & Sons, Inc., 1996.
50. Object Management Group, Inc. CORBA—Common Object Request Broker Architecture. Available at http://www.corba.org/, http://www.omg.org/ (accessed April 30, 2008).
51. World Wide Web Consortium (W3C). HyperText Markup Language (HTML). Available at http://www.w3.org/MarkUp/ (accessed April 30, 2008).
52. Powell TA. *HTML: The Complete Reference*, 2nd edn. Berkeley, CA: Osborne/McGraw-Hill, 1999.
53. Oprea TI, Tropsha A. Target, chemical and bioactivity databases—Integration is key. *Drug Discov Today Technol* 2006;3:357–365.
54. Bioinformatics & Drug Design Group, Computational Science Department, National University of Singapore. Available at http://bidd.nus.edu.sg/ (accessed April 30, 2008).
55. Chen X, Ji ZL, Chen YZ. TTD: Therapeutic Target Database. *Nucleic Acids Res* 2002;30:412–415.

56. Ji ZL, Han LY, Yap CW, Sun LZ, Chen X, Chen YZ. Drug Adverse Reaction Target Database (DART): Proteins related to adverse drug reactions. *Drug Saf* 2003;26:685–690.
57. Sun LZ, Ji ZL, Chen X, Wang JF, Chen YZ. Absorption, distribution, metabolism, and excretion-associated protein database. *Clin Pharmacol Ther* 2002;71: 405–416.
58. Chen X, Ji ZL, Zhi DG, Chen YZ. CLiBE: A database of computed ligand binding energy for ligand-receptor complexes. *Comput Chem* 2002;26:661–666.
59. Ji ZL, Chen X, Zheng CJ, Yao LX, Han LY, Yeo WK, Chung PC, Puy HS, Tay YT, Muhammad A, Chen YZ. KDBI: Kinetic Data of Bio-molecular Interactions database. *Nucleic Acids Res* 2003;31:255–257.
60. Wang JF, Zhou H, Han LY, Chen X, Chen YZ, Cao ZW. Traditional Chinese Medicine Information Database. *Clinical Pharmacol Therap* 2005;78:92–93.
61. Broad Institute, Cambridge. ChemBank Project. Available at http://chembank.broad.harvard.edu/ (accessed April 30, 2008).
62. Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, Norton S, Brudz S, Sullivan JP, Muhlich J, Serrano M, Ferraiolo P, Tolliday NJ, Schreiber SL, Clemons PA. ChemBank: A small-molecule screening and cheminformatics resource database. *Nucleic Acids Res* 2008;36:D351–D359.
63. National Institutes of Health, National Cancer Institute. Available at http://www.cancer.gov/ (accessed April 30, 2008).
64. Institute of Chemistry and Cell Biology, Harvard Medical School. Available at http://iccb.med.harvard.edu/ (accessed April 30, 2008).
65. Wishart DS, Shart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2007;36:D901–D906.
66. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Zhan Chang Z, Woolsey J. DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;34:D668–D672.
67. U.S. Food and Drug Administration, Department of Health and Human Services. Available at http://www.fda.gov/ (accessed April 30, 2008).
68. EMBL-EBI. Chemical Entities of Biological Interest. Available at http://www.ebi.ac.uk/chebi/ (accessed April 30, 2008).
69. National Center for Biotechnology Information. PubChem. Available at http://pubchem.ncbi.nlm.nih.gov/ (accessed April 30, 2008).
70. Kyoto University. KEGG: Kyoto Encyclopedia of Genes and Genomes. Available at http://www.genome.jp/kegg/ (accessed April 30, 2008).
71. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–242.
72. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank: Update. *Nucleic Acids Res* 2004;32(Database issue):D23–D26.
73. Zhang J-W, Aizawa M, Amari S, Iwasawa Y, Nakano T, Nakata K. Development of KiBank, a database supporting structure-based drug design. *Comput Biol Chem* 2004;28:401–407.

74. Aizawa M, Onodera K, Zhang J-W, Amari S, Iwasawa Y, Nakano T, Nakata K. KiBank: A database for computer-aided drug design based on protein-chemical interaction analysis. *Yakugaku Zasshi* 2004;124:613–619.

75. Quantum Molecular Interaction Analysis Group, Institute of Industrial Science, University of Tokyo. KiBank. Available at http://kibank.iis.u-tokyo.ac.jp/ (accessed April 30, 2008).

76. von Grotthuss M, Koczyk G, Pas J, Wyrwicz LS, Rychlewski L. Ligand.Info small-molecule Meta-Database. *Comb Chem High Throughput Screen* 2004;7: 757–761.

77. Boutselakis H, Dimitropoulos D, Fillon J, Golovin A, Henrick K, Hussain A, Ionides J, John M, Keller PA, Krissinel E, McNeil P, Naim A, Newman R, Oldfield T, Pineda J, Rachedi A, Copeland J, Sitnov A, Sobhany S, Suarez-Uruena A, Swaminathan J, Tagari M, Tate J, Tromm S, Velankar S, Vranken W. E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res* 2003;31:43–50.

78. AKos Consulting & Solutions Deutschland GmbH, Germany. Available at http://www.akosgmbh.de/ (accessed April 30, 2008).

79. ASINEX Ltd., Moscow, Russia. Available at http://www.asinex.com/ (accessed April 30, 2008).

80. PDSP Ki database. Available at http://pdsp.med.unc.edu/pdsp.php (accessed April 30, 2008).

81. Austin CP, Brady LS, Insel TR, Collins FS. NIH molecular libraries initiative. *Science* 2004;306:1138–1139.

82. ChemSpider database. Available at http://www.chemspider.com/Default.aspx (accessed April 30, 2008).

83. OpenEye Scientific Software, Santa Fe, USA. Available at http://www.eyesopen.com/ (accessed April 30, 2008).

84. University of Erlangen-Nürnberg, Erlangen, Germany. The CACTVS System Home Page. Available at http://www2.ccc.uni-erlangen.de/software/cactvs/ (accessed April 30, 2008).

85. Warr, WA. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J Comput Aided Mol Des* 2009;23:195–198.

86. Irwin JJ, Shoichet BK. ZINC—a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 2005;45:177–182.

87. Molinspiration Cheminformatics. Java Molecular Editor. Available at http://www.molinspiration.com/jme/ (accessed April 30, 2008).

88. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M, Oprea TI. *Chemoinformatics in Drug Discovery*, pp. 223–239. New York: Wiley-VCH, 2005.

89. Accelrys Software Inc. Chemical Database Product Listing. Available at http://www.accelrys.com/products/chem_databases/ (accessed April 30, 2008).

90. Aureus Pharma, Paris, France. Available at http://www.aureus-pharma.com/ (accessed April 30, 2008).

91. Aureus Pharma. AurSCOPE. Available at http://www.aureus-pharma.com/Pages/Products/Aurscope.php (accessed April 30, 2008).
92. GVK Biosciences Private Limited, Hyderabad, India. Available at http://www.gvkbio.com/ (accessed April 30, 2008).
93. CambridgeSoft Corporation. Chemical Database. Available at http://www.cambridgesoft.com/databases/ (accessed April 30, 2008).
94. Eidogen–Sertanty. Kinase Knowledgebase. Available at http://www.eidogen-sertanty.com/products_kinasekb.html (accessed April 30, 2008).
95. Eidogen–Sertanty, San Diego, USA. Available at http://www.eidogen-sertanty.com/ (accessed April 30, 2008).
96. Symyx MDL. MDDR—MDL Drug Data Report. Available at http://www.mdli.com/products/knowledge/drug_data_report/ (accessed April 30, 2008).
97. Symyx MDL. MDL Discovery Knowledge Product Listing. Available at http://www.mdli.com/products/knowledge/ (accessed April 30, 2008).
98. Molecular Networks GmbH. CORINA. Available at http://www.mol-net.de/software/corina/ (accessed April 30, 2008).
99. Symyx MDL. DiscoveryGate. Available at https://www.discoverygate.com/ (accessed April 30, 2008).
100. GVK Biosciences. Database products. Available at http://www.gvkbio.com/informatics/dbprod.htm (accessed April 30, 2008).
101. Jubilant Biosys Ltd. Products. Available at http://www.jubilantbiosys.com/products.htm (accessed April 30, 2008).
102. Sunset Molecular Discovery LLC, Santa Fe, USA. Available at http://www.sunsetmolecular.com/ (accessed September 1, 2009).
103. Olah M, Rad R, Ostopovici L, Bora A, Hadaruga N, Hadaruga D, Moldovan R, Fulias A, Mracec M, Oprea TI. *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*, pp. 760–786. New York: Wiley-VCH, 2007.
104. Sunset Molecular Discovery LLC. Products. Available at http://www.sunsetmolecular.com/products/ (accessed September 1, 2009).
105. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 1997;23:3–25.
106. Daylight CIS. DayCart—Daylight Chemistry Cartridge for Oracle. Available at http://www.daylight.com/products/f_daycart.html (accessed April 30, 2008).
107. EBI–European Bioinformatics Institute. Available at http://www.ebi.ac.uk/ (accessed April 30, 2008).
108. European Bioinformatics Institute. Enzyme Structures Database. Available at http://www.ebi.ac.uk/thornton-srv/databases/enzymes/ (accessed April 30, 2008).
109. Horn F, Weare J, Beukers MW, Hörsch S, Bairoch A, Chen W, Edvardsen O, Campagne F, Vriend G. GPCRDB: An information system for G protein-coupled receptors. *Nucleic Acids Res* 1998;26:277–281.
110. Rawlings ND, Tolle DP, Barrett AJ. MEROPS: The peptidase database. *Nucleic Acids Res* 2004;32:D160–D164.
111. Saier Lab Bioinformatics Group, University of California, San Diego. *TCDB—Transport Classification Database*. Available at http://www.tcdb.org/ (accessed April 30, 2008).

112. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). *Membrane Transport Proteins*. Available at http://www.chem.qmul.ac.uk/iubmb/mtp/ (accessed April 30, 2008).

113. Ren Q, Kang KH, Paulsen IT. TransportDB: A relational database of cellular membrane transport systems. *Nucleic Acids Research* 2004;32(Database issue): D284–D288.

114. Oprea TI, Tropsha A, Faulon J-L, Rintoul MD. Systems chemical biology. *Nat Chem Biol* 2007;3:447–450.

18

MINING CHEMICAL STRUCTURAL INFORMATION FROM THE LITERATURE

Debra L. Banville

Table of Contents

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

18.1 INTRODUCTION

18.1.1 Missed Information Costs Time and Money

Both academic and commercial research areas have a growing thirst for chemical knowledge. This thirst is starting to reach massive proportions [1,2]. In economic terms, missed information in the chemical literature costs time, money, and quality, where both the quality of decisions made and the quality of subsequent research output are compromised. In fact, incorrect decisions along the drug pipeline life cycle can cost millions to billions of dollars [3–9]. The cost of finding missed information early in the drug pipeline is relatively small compared to its discovery in the later stages when project teams have to be more reactive than proactive. While cost control is a high priority in the drug industry, there are claims in 2007 that priorities have inverted from (cost > time > quality) to (quality > time > cost) [10]. The thinking is that better information translates into better quality compounds, and obtaining higher-quality drug candidates faster reduces costs.

Similarly, missed information in academic research frequently translates into substantial costs including missed funding opportunities, such as missed deadlines to fund seed projects. Limited access to the necessary information either due to licensing costs, copyright limitations, or technical issues, together with the inability to process the massive amounts of information available, is likely to result in missed information [11]. Access limitations are worse in academia than in industry. Lowering these barriers has been the goal of many individuals such as Paul Ginsparg [12], who, in 1991 developed arXiv, the first free scientific online archive of non-peer-reviewed physics articles that continues today [13]. Many groups have formed to increase the accessibility of academic information such as Scholarly Publishing and Academic Resources Coalition (SPARC), the Science Commons group [14], and the World Wide Web Consortium [15].

There is an acute awareness that the traditional process of scientific publication results in lost information especially chemical structural and biological information, and the traditional business model this process is based on no longer meets the needs of publishers or subscribers. Traditional publication methods puts the onus on the reader to locate the pertinent articles, to find a way to buy or to obtain access to the full-text article, to find the necessary information within each article, and to recreate, in many cases, missing chemical structural information from the text and images provided. Requests by readers to authors for access to the underlying data are voluntary and fraught with many difficulties [11].

18.1.2 Issues of Drug Safety Require Better Information Management Capabilities

As the public outcry increases for safe, novel, less expensive pharmaceuticals, and given that 20% of drug projects are stopped due to drug safety,

researchers are talking about new methodologies for finding and managing relevant information [1,11,16,17]. New methodologies like translational science (the relationship among data/information from different disciplines) and personalized medicine (treating patients based on their unique genetic and physiological profiles) rely on the identification of key entities such as biomarkers [4,18]. This reliance increases the urgency of information and knowledge management [4]. The difficulty of finding possible biomarkers in the published literature is a major challenge. Biomarkers for a disease or a response to a treatment can encompass many types of changes on many different levels (e.g., biochemical, physiological, and cellular). Publications report many of these changes without ever mentioning the word "biomarker." Hence, this information can be easily lost in a sentence or paragraph of an article.

18.2 DIFFERENT NEEDS, DIFFERENT CHALLENGES, AND THE CALL FOR STANDARDIZATION

Barriers to finding information within the literature include restrictive licensing and copyright that can limit access to documents and to the information within the documents, technical restrictions preventing the extraction of key data, and cultural restrictions created from the legacies of a publishing world based on printed media [19]. As scientific publication paradigm shifts to a more open-access electronic environment, publishers, authors, and readers each have different needs and challenges to contend with. But even if information is freely available, free of the various legal/technical/cultural barriers, extraction of chemical structural information still remains a challenge. The primary barrier to chemical structure extraction is the lack of standards in expressing chemical nomenclature [19].

18.2.1 The Ultimate Backend Solution is Universal Standards, Especially for Chemical and Biological Information

The aeronautics industry recognized the importance of timely information gathering and was determined to minimize the frequency of adverse safety-related events [20]: "… By setting standards, maintaining multiple databases to monitor trends, and supporting research to constantly improve systems, the FAA (in collaboration with other agencies such as NASA and NTSB) has made flying safer." The call for standards is an obvious solution for improving chemical information management and for minimizing drug safety issues. The primary barriers to standardization remain cultural and technical.

Publishers of scientific literature, patent filings, and clinical data and researchers from different commercial and academic institutions all have different priorities. For example, authors of published journal articles want wide readership, while journal publishers want a large-paying subscriber base. In another example, inventors of published patent filings need to balance the

coverage of their invention with the risk of revealing "too much," while the patent authorities need to provide easy public access to these inventive publications. Overcoming cultural barriers requires establishing and accepting new processes and a general acceptance of what constitutes a standard. The exponential growth of the Semantic Web and an open-access culture are strongly driving this change.

Overcoming technical barriers requires new processes and tools to facilitate the creation and application of standards to published electronic literature. These processes and tools have to address the needs of the publishers, authors, and readers. Ultimately, these technical processes and tools have to work within the ever-growing translational scientific world that includes a crossover among different disciplines such as chemistry, physics, biology, pharmacology, and medicine.

18.2.2 Main Driver for Standardization with Life Science Literature is Drug Safety

Similar to the aeronautics industry, the healthcare area is looking for ways to improve safety. Innovations in information and knowledge management are one solution, and standardization of the primary literature is necessary to support this solution.

18.3 CURRENT METHODOLOGIES FOR CONVERTING CHEMICAL ENTITIES TO STRUCTURES

18.3.1 Multiple Types of Naming Schemes Require a Diverse Set of Conversion Capabilities

There are many ways to represent chemical compounds in the literature as illustrated in Figure 18.1 for aspirin. Conversion capabilities, as shown in Table 18.1, are required to translate these multiple names into more meaningful structures. Common or trivial compound names require an extensive dictionary or look-up list for conversion to its intended structure. Common abbreviations for compounds, however, can be very ambiguous and require both a look-up list and some understanding of context. For example, the abbreviation, *PCA* has over 60 different expansions or meanings. Index and reference numbers, unlike abbreviations, may have a very specific meaning. The availability of this specific meaning might involve accessing a large variety of databases. Some of these databases, such as the Chemical Abstract Services (CAS), require licensing to capture that relationship between reference numbers and their structure. Misinterpretation and misuse of these reference numbers in the scientific community has frequently resulted in the wrong structures. Although registry/index numbers have been very instrumental in building the chemical databases, they are proprietary and contain no inherent structural information.

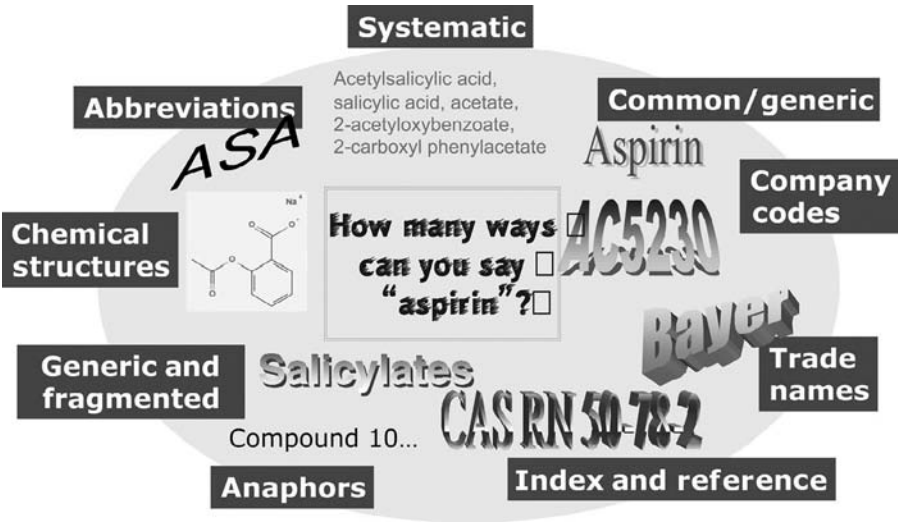


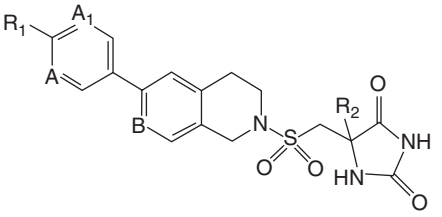
Figure 18.1 How many ways can you say "aspirin"? As shown here, there are at least nine different ways of expressing a chemical compound like aspirin.

TABLE 18.1 Different Ways of Expressing Chemical Structures in Text and of Interpreting Each Group

Chemical Nomenclature Types	Solutions for Interpreting Different Chemical Nomenclatures
Unsystematic names: common, trade names, company codes, index/reference numbers, abbreviations, fragmented names, and generalized structural names	Name lookup/thesaurus or ontologies
Systematic names: IUPAC, CAS, etc.	Name to structure conversion routines
Anaphors, including some abbreviations	Natural language processing
Structural image	Optical image recognition to structure conversion routines

Generic or fragmented names may relate to a family or class of compounds, e.g., salicylates or benzodiazapines. On one level, these can simply appear in a lookup that relates these names to its generalized parent compound. These family names may be available in ontologies that relate the parent, e.g., salicylates and benzodiazepines, to specific compounds in that family, e.g., aspirin and xanax, respectively. Similarly, a family of structures may be drawn in a Markush format that defines the compounds without explicitly stating the naming compounds [21]. Markush representations are typically used in chemical patenting as shown in Figure 18.2.

(54) Title: Novel hydantoin derivatives as metalloproteinase inhibitors



(57) Abstract: The invention provides compounds of formula (I):
wherein R1, R2, A1, and B are as defined in the specification, etc ...

Figure 18.2 Example of generalized/Markush structures typically found in patent documents. These types of structures are named after a Hungarian-born chemist, Dr. Eugene Markush. Dr. Markush claimed generic chemical structures in addition to those actually synthesized in a 1924 patent. The "Markush doctrine" of patent law, which resulted from this patent award, greatly increased flexibility in the preparation of claims for the definition of an invention [21].

Systematic nomenclature is rule based, as discussed below. Several conversion routines that provide a fair amount of success in interpreting the structures from these names are available. Standard rules like those provided through the International Union of Pure and Applied Chemistry (IUPAC) are not applied consistently. This inconsistency results in a huge variation of IUPAC-like names [22]. In the case of CAS nomenclature, the rules generated had a practical purpose related to manual indexing of compounds. These variations in nomenclature are a constant challenge to scientists needing comprehensive conversion of names to structures.

The use of anaphors is generally defined within a document. An author might write the compound name once and put a number or code alongside of it as the anaphor or abbreviation for that compound. In these cases, the ability to identify these associations within the text requires the use of natural language processing (NLP) technologies [19,23].

Finally, converting a picture/image of a structure to a machine-readable form has proven to be difficult to automate as a high-throughput capability. Some of the difficulty is due to large variations in how structures are drawn and in the quality of these images. Chemical image to structure conversion can be broken down into three steps [24]: (1) preprocessing to a identify the image in a document that contains chemical structural information; (2) reconstruction of that chemical structure using vectorized graphical elements and the interpretation of those elements (e.g., dashed lines, wedges, and character recognition) with a compilation of those elements to construct a structure; and finally, (3) post processing to export the structural information and to display them.

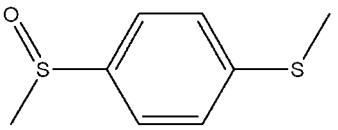
The earliest attempt in this area dates back to 1992 when McDaniel and Balmuth created Kekulé, an optical character recognition (OCR)/chemical recognition capability [25]. One year later, A. Peter Johnson's group published a paper describing their OCR capability called Chemical Literature Data Extraction (CLiDE™), a commercially available application [26–28]. Conversion from a structural image to a computer-readable structure requires extensive error checking and correction. This fact combined with the need to be more comprehensive in the extraction of chemical image information from a multitude of published and internal documents has resulted in a renewed interest by others [24,29,30].

18.3.2 Systematic Chemical Name to Structure Conversion

Chemical nomenclature has evolved from the use of less descriptive, unsystematic, or common/trivial names, like *mandelic acid*, to more descriptive, systematic naming schemes like *2-phenyl-2-hydroxyacetic acid* (IUPAC name) for *mandelic acid*. This need for descriptive systematic naming resulted in the first conference on chemical nomenclature held in Geneva in 1892 [31]. Systematic chemical names follow a linguistic rule set [32–35]. Practitioners with an expertise in generating chemical nomenclature frequently use linguistic terminology to describe chemical naming. For example, the components of the IUPAC name for mandelic acid, "2-," "phenyl," "hydroxyl," and "acetic acid," are called the morphemes, and the arrangement of these morphemes is called the syntax of the name. The meaning behind the syntax is the semantics, and this defines the chemical structure. Two of the most commonly used systematic rule sets are from IUPAC and CAS.

Despite the existence of these two commonly used naming standards, complications in naming and interpreting compounds arise for many reasons. First, the possible lexicons the chemical community draws from can vary considerably. For example, *phthalonitrile* and *o-dicyanobenzene* are the same compound but from a different lexicon. Second, many scientists using chemical nomenclature are not necessarily trained in the nuances of each system. Third, even well-trained chemists using the same lexicon and rules, such as IUPAC, can apply the rules correctly but still create different chemical names as shown in Figure 18.3. Interpreting these names can be a challenge to even a trained chemist. Examples of some subtle and not so subtle naming nuances are shown in Table 18.2.

In 1958, Garfield first recognized that a systematic name could be algorithmically converted into a molecular formula and then to line notation [36,37]. Nine years later, Vander Stouw created an automated approach to convert basic chemical names to structures [38,39]. Another 9 years later, Raynor's automated approach focused on IUPAC nomenclature limited to certain compound classes and some trivial or common names [35,40–43]. Many other approaches were highly focused on specific compound types (e.g., steroids or stereochemical nomenclature) [44,45].



1-(Methylsulfanyl)-4-(methylsulfinyl)benzene
Methyl 4-(methylsulfanyl)phenyl sulfoxide
1-(Methylsulfinyl)-4-(methylthio)benzene
Methyl 4-(methylthio)phenyl sulfoxide
Bis(methylthio)benzene monooxide

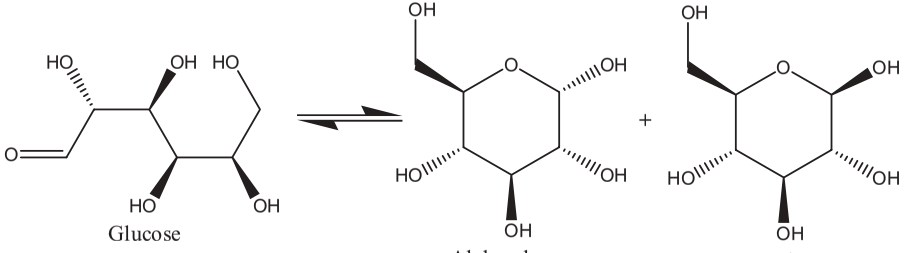
Figure 18.3 As shown in this example, one chemical structure can have multiple systematic names generated by the application of different rule sets and lexicons. Even experienced chemists can easily assign these chemical names to the wrong structures.

These early-automated systems that interpret IUPAC nomenclature generally assumed that chemists have strictly followed the rules. In acknowledging this misassumption, commercial conversion routines allow for some creative and known systematic deviations from the IUPAC rule set [32,46]. As described in Brecher's paper, for *Name = Struct*, the use of spacing and punctuation plays a large role in chemical nomenclature, and subtle difference (intended or unintended) can result in both ambiguity and incorrect structures. Figure 18.4 captures some of these more subtle ambiguities. *Name = Struct* preprocesses using a lexicon perform a series of steps that include appropriate removal of punctuation and correcting common typographical errors. This cleanup of the names also includes recognizing CAS names, inverting their order to be more IUPAC-like, and removing words that do not contribute to the structure such as "glacial" or "95%." *Name = Struct* also divides input names into fragments and attempts to identify the root or parent structure that all the other terms modify. Ambiguities are abundant even at this step. For example, *azine* can refer to aromatic rings as in *tetrasine*, *triasin*, and *diasine* or an acyclic functional group. A list of some known name to structure conversion capabilities are summarized in Table 18.3.

18.3.3 Unsystematic Chemical Name Lookup

The conversion of unsystematic or trivial names to structures requires a thesaurus or synonym look-up database. Unlike chemical research patents that typically provide systematic names, medical journal articles and conference abstracts typically use common drug names to describe compounds of biomedical and clinical interest. The name to structure conversion routines may provide a look-up capability for most common chemical and drug names, but this list is generally limited. Hence, more in-depth look-up databases are required to provide comprehensive coverage for this segment of the literature. While a number of free chemical sources for drug names on the Web abound [47], most are limited in content, search capabilities, quality, and focus.

TABLE 18.2 Several Nuances in Chemical Nomenclature and the Difficulties They Cause in the Interpretation of Chemical Structure

Nuances in Chemical Nomenclature	Examples
More than one lexicon exists that represent generally acceptable morphemes, and these morphemes can have a diverse syntax imposed on them.	Mandelic acid Other unsystematic names Almond acid; amygdalic acid; uromaline Systematic Names Benzeneacetic acid, α-hydroxy- (CAS name); 2-phenyl-2-hydroxyacetic acid (IUPAC name) Other systematic names Phenylglycolic acid; phenylhydroxyacetic acid; (±)-α-hydroxybenzeneacetic acid; (±)-α-hydroxyphenylacetic acid; (±)-2-hydroxy-2-phenylethanoic acid; (±)-mandelic acid; (RS)-mandelic acid; DL-amygdalic acid; DL-hydroxy(phenyl) acetic acid; DL-mandelic acid; paramandelic acid; α-hydroxy-α-toluic acid; α-hydroxyphenylacetic acid ; α-hydroxybenzeneacetic acid; 2-hydroxy-2-phenylacetic acid; 2-phenyl-2-hydroxyacetic acid; 2-phenylglycolic acid
Names are not always used appropriately to define the intended structural form.	*Glucose* is frequently used to mean one or both ring forms and not the open-chain form; the name technically implies  Glucose Alpha-glucose Beta-glucose
Names describe substances but not necessarily their forms or the presence of mixtures or salts.	Copper sulfate versus copper sulfate pentahydrate 5-methoxy-2-[(4-methoxy-3,5-dimethyl-pyridin-2-yl)methylsulfinyl]-3H-benzoimidazole, sodium salt versus magnesium salt
A diverse set of commonly used naming conventions result in a diverse set of expressions for even simple complexes.	$CuSO_4$ Where Cu can be written as cupric = copper = copper(II) = copper (2+) Where SO4 can be written as sulfate = suphate

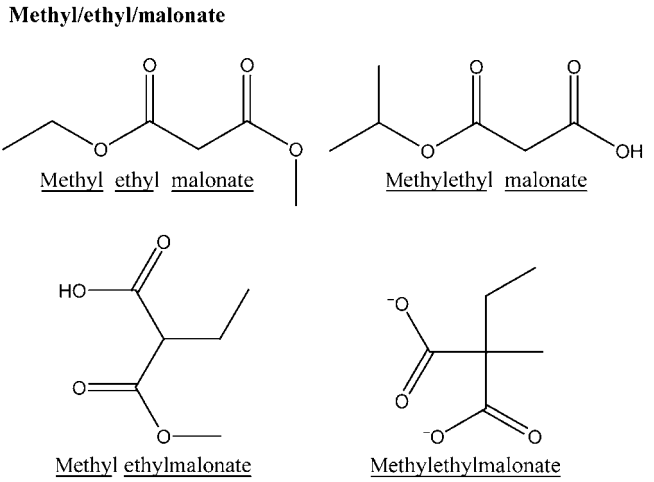


Figure 18.4 An illustration of how variations in spaces can result in different structures. In the case shown here, three components, methyl/ethyl/malonate, can be rewritten with spaces to generate four different chemical structures.

TABLE 18.3 A Summary of Some Commonly Known Chemical to Structure Conversion Capabilities

Application	Application Owner	Reference Links
ACD/Name™	ACD/Labs	http://www.acdlabs.com/
ChemNomParse	University of Manchester	Downloaded OpenSource packages from http://sourceforge.net
nam2mol™	OpenEye Lexichem	http://www.eyesopen.com/
Name to Structure Generation (soon to be released)	ChemAxon	http://www.chemaxon.com/
NamExpert™	ChemInnovation Software	http://www.cheminnovation.com/
OPSIN	Peter Corbett, University of Cambridge	Downloaded as two stand-alone OpenSource packages from http://SourceForge.net: Source Distribution and Jarfile
Name = Struct™	CambridgeSoft	http://www.cambridgesoft.com/

Availability of more definitive sources for this information, such as the CAS Registry file, is limited by the end users' ability to pay the licensing costs. Sources such as PubChem are free to the public with 8 million plus compounds containing synonym lists. However, this is roughly one-fourth the size of CAS's Registry file in number of compounds.

One free source that claims nearly as many compounds as the commercially licensed CAS Registry file with its over 30 million compounds is NIH's Chemical Search Locator Service (CSLS). However, the CSLS does not support an unsystematic name search and is not a database of structures and names, but is a search agent that is dependent on its 80+ host sources for its content and coverage. Some of these host sources are commercial and are only available to licensed users. In contrast, the CAS Registry file allows unsystematic name searches and provides consistent predictable coverage in one database. Ideally in the spirit of open access, a single wiki-style lookup should be available, which supports both unsystematic and structure searching capabilities. Given the volume of chemical information available, it will take the combined efforts of many researchers to create this type of look-up capability. Until then, researchers are reliant on accessing a large variety of sources to gather and collect information-linking structures to a list of unsystematic nomenclature.

18.4 REPRESENTING CHEMICAL STRUCTURES IN MACHINE-READABLE FORMS

The ability to provide machine-readable representations of chemical structural information has been around for many years. The existence of the Web, together with Semantic Web technologies, is starting to provide some consensus on possible standards.

18.4.1 The Language of e-Chem: International Chemical Identifier (InChI), Simplified Molecular Input Line Entry Specification (SMILES), Chemical Markup Language (CML), and More

For anyone routinely using chemical structural information, the simple way to communicate a structure is through some type of stick drawing. It is visual and easily communicates molecular information. For computers designed to read textual codes, these visual images are meaningless. The ability to represent chemical structures in machine-readable form becomes very compelling for researchers wanting to both find key documents containing these structures and find the structures *within* each key document [30]. In the latter case, the structure can be understood in context.

The development of ways to identify and convert molecular information to a machine-readable form started in earnest in the late 1980s and the early 1990s [26,27,48–50]. As summarized in Table 18.4, there are a variety of commonly used identifier types: line notation identifiers (e.g., SMILES and InChI), tabular identifiers (e.g., Molfile and its related SD file type), and portable markup language identifiers (e.g., CML and FlexMol). Each is capable of meeting some of the needs for chemical structural information exchange with computers [30,49,51–58].

TABLE 18.4 A Summary of Some Common Chemical Structure Identifiers

Name	Descriptor	Reference
Simplified Molecular Input Line Entry Specification (SMILES) System	A proprietary line notation for molecules and reactions in a compact linguistic construct. Different types of SMILES can generate nonunique representations for a single structure.	[51]
Molfile (.mol)	A proprietary file format (MDL Elsevier) that uses coordinate and connection information	[53]
International Chemical Identifier (InChI)	A nonproprietary (IUPAC) line notation for representing organic molecules in a compact linguistic construct. Each unique compound has only one unique InChI representation.	[54–57]
Chemical Markup Language (CML)	A nonproprietary domain-specific implementation of XML; capable of capturing a wide range of chemical concepts, e.g., molecules, reactions, and data	[58,59]
FlexMol	A nonproprietary domain-specific implementation of XML. Capable of capturing molecules. This was meant to address specific cases where CML fails to provide unique descriptors, e.g., ferrocene.	[60]

Chemical identifiers found in a document can be packaged inside an extensible markup language (XML) form of the document. As illustrated in Figure 18.5, a name (or image) can be identified and converted to a machine-readable structure such as SMILES and InChI. XML's specialty area of CML allows the structural metadata to be embedded into the document. Ultimately, these chemical tags enrich the document with key structural information that allows researchers to (1) find the documents tagged with their structures of interest and (2) to see how the compound is mentioned in the document. This contextual component can be a very simple and powerful research tool.

18.5 BUILDING CONTEXT WITH NLP TODAY

NLP technology does not understand human speech but dissects language into the parts of speech such as nouns, verbs, and noun phrases. In the mid-1960s, Weizenbaum created a computer program call *Eliza*, which demonstrated the remarkable possibilities of NLP, highlighted the host of complexities

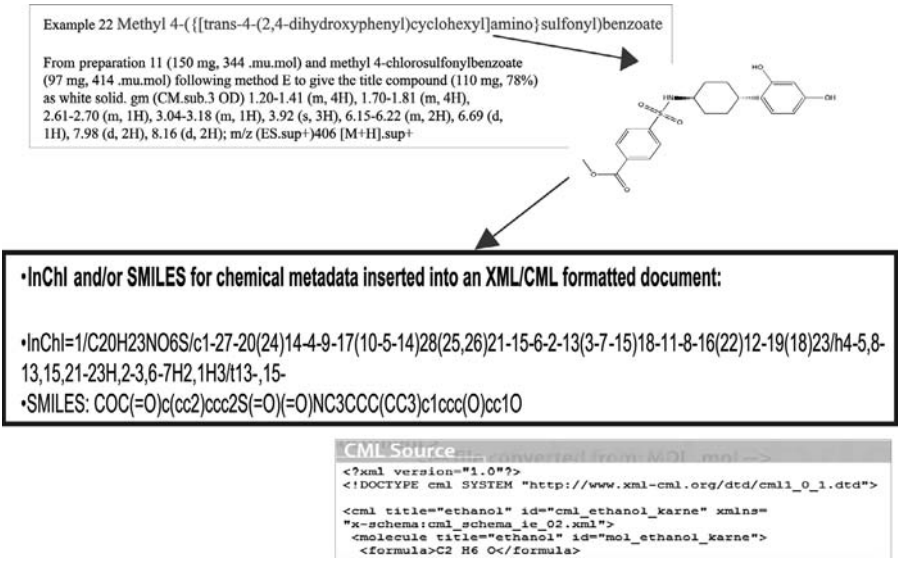


Figure 18.5 An illustration of how an XML-formatted document can be tagged with chemical structural metadata using CML to capture the machine-readable data such as SMILES and InChI textual strings.

prevalent in human speech, and underlined the fact that NLP is not artificial intelligence [59].

Over 30 years later, NLP computer programs are available to provide end users with easier ways to organize and find concepts of interest within unstructured text. In most scenarios, the user submits a query of interest and returns extracted facts. When using NLP queries, the user is faced with a trade-off between precision (the signal to noise or proportion of documents that are actually relevant) and recall (the "hit ratio" or proportion of documents returned). An ideal NLP capability has to provide a good framework for organizing and reviewing the concepts extracted. One possible breakdown of NLP steps is depicted in Figure 18.6 using common NLP terminology. Ultimately, the end user has to determine the accuracy of the information identified especially at the level of *name entity recognition* (NER) (Did it correctly recognize the entities?) and syntactics (Did it correctly build the right associations, e.g., between a drug and its effect?). Semantically, the user has to derive meaning from this text.

Given the lack of standards in life science nomenclature on all levels including biological, chemical, and pharmaceutical terminology, it is not surprising that the NER step shown in Figure 18.6 is the main focus of life science applications. Use of extensive data sources and ontologies to identify all the possible meanings behind a single life science term like *PCA* requires a huge effort. An NLP capability might not understand what molecular anaphor

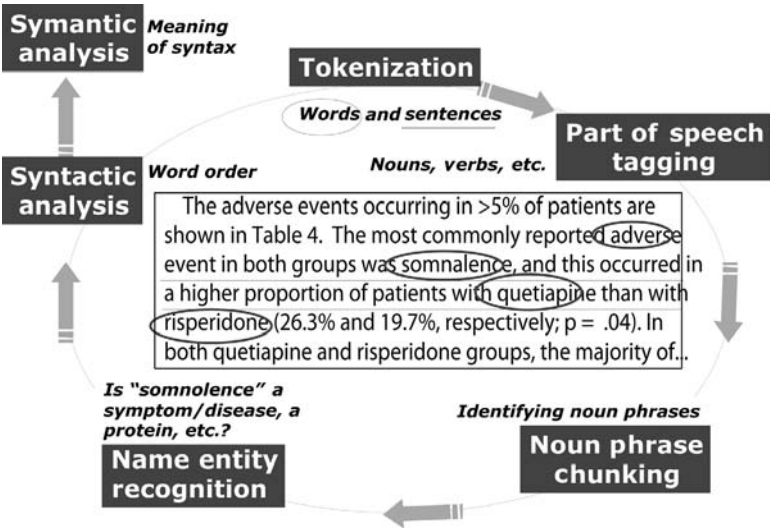


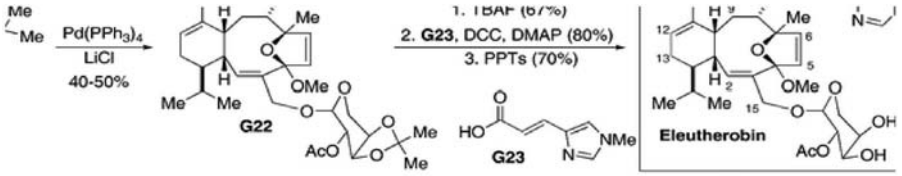
Figure 18.6 One view of how natural language processing (NLP) breaks down scientific literature using linguistic terminology. The process starts with the identification of words and sentences (i.e., tokenization). Name entity recognition (NER) tries to define key terminology (i.e., *risperidone* versus *somnolence* as a *drug* and *symptom*, respectively). In practice, the researcher has to frequently validate many of these steps.

"compound (10)" refers to without additional information found in another part of the document.

In acknowledging the value of NER to capture the authors' science in a computer-readable form, a prominent publisher of chemical information, the Royal Society of Chemistry (RSC), initiated *Project Prospect*. RSC's Project Prospect is based on the premise that text annotation of chemical and other life science entities should be a part of the publication process. As shown in Figure 18.7, chemists can go from reading untagged text to color-annotated text, improving their ability to identify key entities of interest while directly seeing the context in which these entities are being used. Once a chemical entity is identified, links to other articles containing this entity can also be generated.

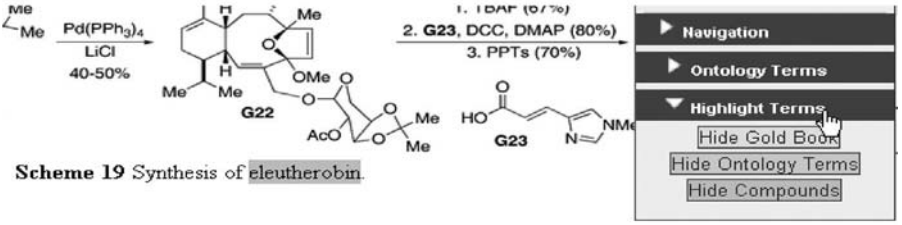
RSC's project is an excellent example of how access to information can be improved using a set of core noncommercial capabilities summarized in Table 18.5 [49,55,56,60–67]. Project Prospect is exploring ways to build this annotation into the natural publishing workflow while providing feedback to continuously improve their NER capabilities. This feedback includes improving chemistry recognition within project *SciBorg* and *Open Biomedical Ontology* (OBO)'s *Gene Ontology* (GO) and *Sequence Ontology* (SO).

In the future, if other publishers provide a similar NER capability, all publishers can benefit by sharing cross-referenced links among their content. With



Scheme 19 Synthesis of eleutherobin.

it from natural sources to only a very sparing extent, we optimized the synthesis, and built up ithetic drug. The compound was then evaluated in *in vivo* screens using xenograft models. As it mely disappointing. It was opined that ester cleavage was apparently a very important liability in eleutherobin in an *in vivo* setting. It seemed that the urocanic acid moiety of eleutherobin was 'et, the agent apparently seemed quite vulnerable to the action of esterases. To deal with this adical modification of eleutherobin; *i.e.* replacement of the ester of the linkage to the urocanic iven the emergence of still another naturally occurring family of tubulin directed anticancer agents, ight to closure. It is well, however, to point out that as part of this process, the actual proof of of the connectivity of the arabinose sector with the terpenoid-like domain was rigorously



Scheme 19 Synthesis of eleutherobin.

it from natural sources to only a very sparing extent, we optimized the synthesis, and built up ithetic drug. The compound was then evaluated in *in vivo* screens using xenograft models. As it mely disappointing. It was opined that ester cleavage was apparently a very important liability in eleutherobin in an *in vivo* setting. It seemed that the urocanic acid moiety of eleutherobin was 'et, the agent apparently seemed quite vulnerable to the action of esterases. To deal with this adical modification of eleutherobin; *i.e.* replacement of the ester of the linkage to the urocanic iven the emergence of still another naturally occurring family of tubulin directed anticancer agents, ight to closure. It is well, however, to point out that as part of this process, the actual proof of of the connectivity of the arabinose sector with the terpenoid-like domain was rigorously

Figure 18.7 The Royal Society of Chemistry (RSC)'s Project Prospect is illustrated here where application of open access/source capabilities can add real value to a literature document. As shown in the top panel, a typical document is an unstructured text that can be made more readable by applying NER annotation to the text. Color coding is used to highlight and distinguish different types of chemical and biological entities.

cross referencing, it is possible to link a compound found in one article automatically to all articles containing that same compound regardless of who the publisher is. Cross referencing among different publishers is becoming a necessity to manage the overflow of chemical information.

Commercial providers of chemical NER tools have focused mainly on patent NER or a broader remit of mining general document types. Unlike

TABLE 18.5 Noncommercial Technologies Used by Project Prospect (See http://www.rsc.org/ for Details on Project Prospect)

Type	Names	References
NLP	SciBorg and Open Source Chemical Analysis Routines 3 (OSCAR3)	[63–66]
Ontologies/terminologies	Gene Ontology (GO), Sequence Ontology (SO), Open Biomedical Ontology (OBO), European Commission—Chemical Entities of Biological Interest (ChEBI), Gold Book (*IUPAC*—chemical terminologies/symbols/units)	[67–69]
Structural information	*IUPAC*—International Chemical Identifier (InChI), Simplified Molecular Input Line Entry Specification (SMILES), Chemical Markup Language (CML)	[51,57,58,59]

TABLE 18.6 Commercially Available Chemical NER Capabilities

Vendor Name	Tool Type
Accelrys	Workflow tool
Fraunhofer Institute and InfoChem	Tools
IBM	Tools and patent database
SureChem	Tools and patent database
TEMIS	Tools
InfoApps/MPERIC	Tools and patent database

publishers, these NER providers do not have control over the quality or format of this content, and the quality issue has a huge impact on their functionality. A summary of some commercial providers is given in Table 18.6. Patents are a logical focus for the NER providers since intellectual property is a rich area of research and business-related information. Since patent documents do *not* require electronic submissions, most are available as image files that are not machine readable.

Any chemical NER capability focused on patents has to deal with OCR-generated text files. As shown in Figure 18.8, OCR documents generally remove the formatting of the original image document (a PDF or TIFF file in general), resulting in many OCR-generated errors including the corruption of systematic chemical names. As discussed earlier, errors in spaces or punctuation within a chemical name could result in failure to convert the name to a structure or to the wrong structure (see Figs. 18.4 and 18.9). Logistically, annotated patent documents have to be stored either at the vendor or at the

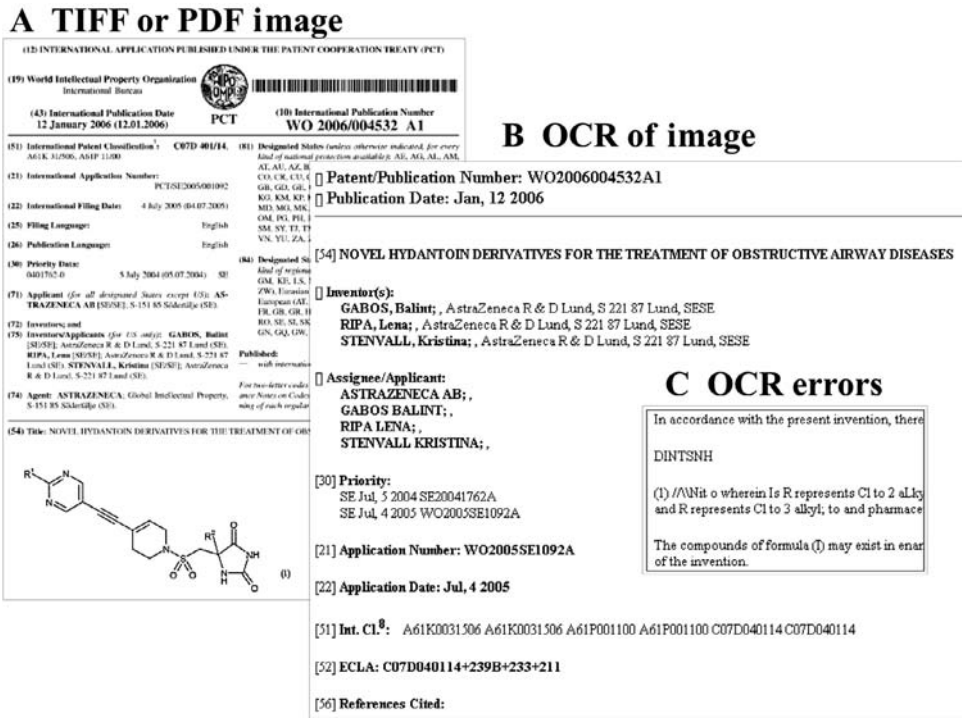


Figure 18.8 (A) Most patent documents start out as images of text (not as text). (B) To apply mining capabilities like chemical NER, these image documents would have to be converted to text by optical character recognition (OCR) tools, but most OCR tools remove the formatting and figures. (C) An example of OCR-generated errors is shown here.

customer's site. The need for customer groups to repeat this process at their individual sites is very costly and time-consuming. Improvements in the patent filing process that encourage the availability of high-quality text documents (e.g., mandating electronic filing) would allow easier application of these chemical NER capabilities.

In summary, the nascent ability to automatically identify both chemical and biological entities and to present this back to the researcher for review is being recognized as a valid and necessary capability within the life science workflow. The ability to identify and to convert chemical names to structures is more dependent on the quality of the document but can be fairly successful as a part of the automated NER process. Image to structure conversion is more difficult. It is the author's hope that efforts like those of RSC's Project Prospect and many others will make this need for image conversion obsolete for future scientific literature [55,65,67,68–70]. However, the need to process older literature will probably remain a continuing driver for improvements in this area [27–29]. Finally, current state of the art in NLP and NER applications requires

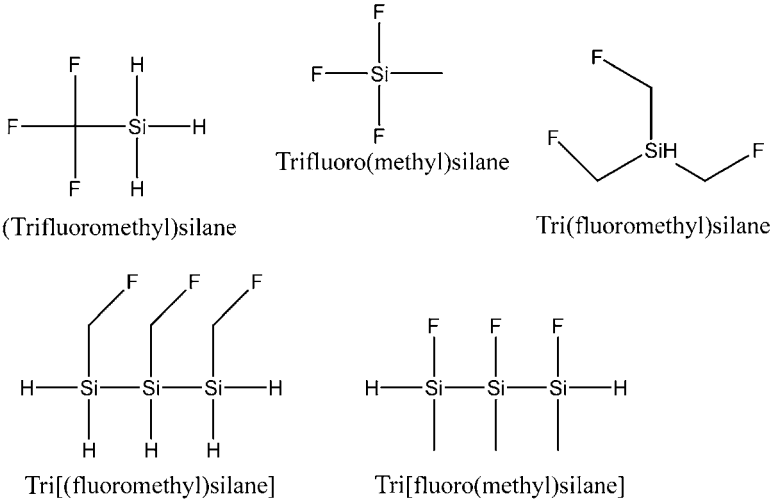


Figure 18.9 Variations in punctuation such as parentheses and brackets, as shown here for *trifluormethylsilane*, are a frequent source of structural misunderstanding.

the end user to provide easy inspection and validation. These validation tools are not necessarily standard in NLP and NER capabilities.

18.6 A VISION FOR THE FUTURE

Building NLP and NER in the life science workflow requires keeping good focus on the end goal, finding key information and building knowledge to drive the best research outcomes possible.

18.6.1 Crossing from Chemistry into Biomedical Space with Chemically Mined Information

Conceptually, a researcher may initiate an NLP/NER query from either chemistry space (i.e., *Can I find the compounds of interest*?) or biology space (i.e., *Can I find protein targets or diseases of interest?*). This is illustrated in Figure 18.10 where a biology query branches into chemistry space to address a key toxicity issue. This is one of many common workflows performed manually in life science areas when time permits.

Ultimately in the life science area, the question of multiple query spaces crosses into areas like drug discovery, drug safety, and trend analysis.

A blurring among multiple disciplines, e.g., chemical/biological/medical/pharmaceutical areas, translates into an essential need to search both literature and other data sources in an integrated manner. A simple text search in any of these disciplines is no longer sufficient to capture the chemical and biological information.

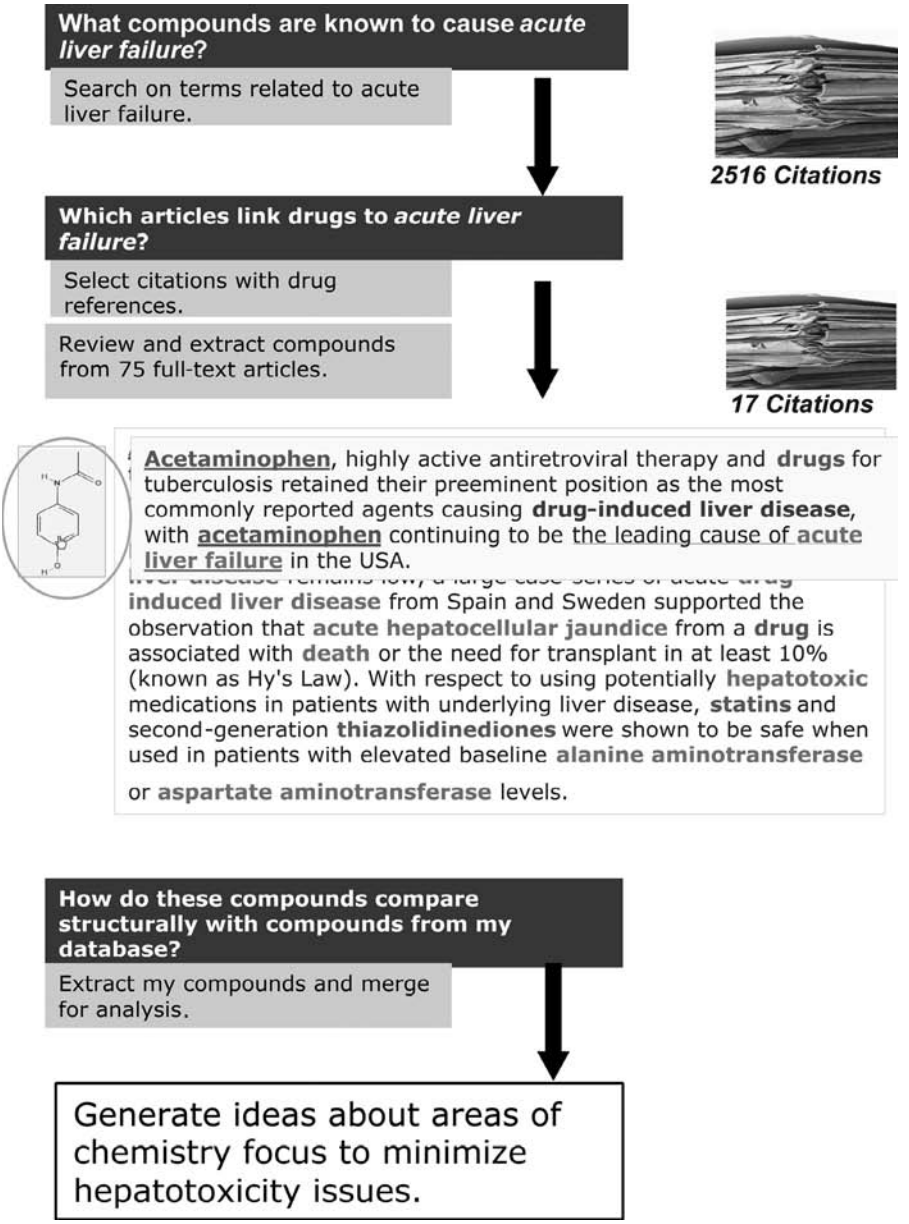


Figure 18.10 Illustrates one possible workflow from biology space (hepatoxicity question) to refocus literature searches and to make key decisions in chemistry space (focusing on compounds with a lower chance of causing hepatoxicity).

The recent proliferation of chemical and biological databases suggests a few additional possibilities for the future: (1) that literature searching will expand to include literature and databases from a single query point [71]; (2) that the ability to easily extract chemical and biological entities from these multiple sources for additional further computational analysis will become more routine; (3) that the barriers to share public information among different research groups and institutions are lowering [13–15]; and (4) as technologies and standards are developed to manage this integrated capability, there will be an increasing demand on authors of journal articles to provide the underlying data along with their report [11].

18.6.2 Text Mining is about the Generation of New Knowledge

The meaning of *text mining* has changed over the last 10 years to include any text analytic capability such as NLP or NER. Text mining was originally intended to mean the discovery of something not known explicitly from text documents but something deduced from textual information. The first example of life science text mining from this perspective occurred in 1986 when a mathematician/information scientist, David R. Swanson examined information on people who suffer from Raynaud's syndrome, an episodic shutting off of blood to fingers and toes. He looked for a syllogism, an unknown connection between symptoms/causes and possible cures. In one case, he asked if there were any agents known to reverse symptomatic blood factors, high viscosity, and cell rigidity/deformity. Literature searches for these blood factors highlighted fish oils that are known to lower blood viscosity and to reduce cell rigidity. Follow-up work in clinical studies successfully demonstrated that fish oils were a treatment for Raynaud's syndrome.

Other examples of text mining can include multiple instances of *drug repurposing*. Most drug repurposing (also known as drug *reprofiling* or *repositioning*) discoveries were the result of researchers connecting key information to generate a valid hypothesis that could be tested in the clinic [69,70,72]. Access to good life science NLP and NER tools can provide easy extraction of key data points. In the hands of creative and experienced researchers, text mining with this extracted data will only serve to increase the opportunities within drug discovery and other areas of life science research.

REFERENCES

1. Hurst, JR. Pharmaceutical industry white paper: Document systems, dimensions for the future. 2007. All Associates Group. Available at http://www.allassociates.com/starproject/pdf/Pharmaceutical_Industry_White_Paper-Nov_2002.pdf (accessed August 31, 2009).
2. Torr-Brown S. Advances in knowledge management for pharmaceutical research and development. *Curr Opin Drug Discov* 2005;8:316–322.

3. Leavitt PM. The role of knowledge mangement in new drug development. 2003. The American Productivity & Quality Center. Available at http://www.providersedge.com/docs/km_articles/Role_of_KM_in_New_Drug_Development.pdf (accessed August 31, 2009).
4. Gaughan A. Bridging the divide: The need for translational informatics. *Future Med* 2006;7:117–122.
5. Banik M, Westgren RE. A wealth of failures: Sensemaking in a pharmaceutical R&D pipeline. *Int J Technol Intelligence Planning* 2004;1:25–38.
6. DiMasi J. The value of improving the productivity of the drug development process: Faster times and better decisions. *Pharmacoeconomics* 2002;20(S3): 1–10.
7. DiMasi J, Hansen R, Grabowski H. The price of innovation: New estimates of drug development costs. *J Health Econ* 2003;22:151–185.
8. Adams C, Brantner V. Estimating the cost of new drug development: Is it really 802 million dollars? *Health Aff* 2006;25:420–428.
9. Myers S, Baker A. Drug discovery an operating model for a new era. *Nat Biotechnol* 2001;19:727–730.
10. Mullin R. Target practice. *Chem Eng News* 2007;85:19–20.
11. Wilbanks J, Boyle J. Introduction to Science Commons. 2006. Available at http://www.sciencecommons.org/about (accessed August 31, 2009).
12. Ginsparg P, Houle P, Joachims T, Sul J-H. Mapping subsets of scholarly information. *Proc Natl Acad Sci U S A* 2004;101:5236–5240.
13. arXiv web site. Available at http://www.arXiv.org (accessed August 31, 2009).
14. Science Commons web site. Available at http://www.sciencecommons.org (accessed August 31, 2009).
15. World Wide Web Consortium web site. Available at http://www.w3c.org (accessed August 31, 2009).
16. Leeson PD, Springthorpe B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat Rev Drug Discov* 2007;6:1–10.
17. Yang C, Richard AM, Cross KP. The art of data mining the minefields of toxicity databases to link chemistry to biology. *Curr Comput Aided Drug Des* 2006;2:135–150.
18. Ogilvie RI. The death of a volunteer research subject: Lessons to be learned. *Can Med Assoc J* 2001;165:1335–1337.
19. Banville DL. Mining chemical structural information from the drug literature. *Drug Discov Today* 2006;11:35–42.
20. Kohn LT, Corrigan JM, Donaldson MS. *To Err is Human: Building a Safer Health System*. Washington, DC: The National Academic Press, 2000.
21. Sibley JF. Too broad generic disclosures: A problem for all. *J Chem Inf Comput Sci* 1991;31:5–9.
22. Panico R, Powell WH, Richer JC. *International Union of Pure and Applied Chemistry (IUPAC). Organic Chemistry Division, Commission on Nomenclature of Organic Chemistry (III.1). A Guide to IUPAC Nomenclature of Organic Compounds: Recommendations 1993*. Oxford: Blackwell Science, 1993.

23. Jackson P, Moulinier I. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*, Vol. 5. Philadelphia: John Benjamin's Publishing Company, 2002.
24. Zimmermann M, Hofmann M. Automated extraction of chemical information from chemical structure depictions. 2007. Touch Briefings. Available at http://www.touchbriefings.com/download.cfm?fileID=12870&action=downloadFile (accessed August 31, 2009).
25. McDaniel JR, Balmuth JR. Kekulé: OCR—Optical chemical (structure) recognition. *J Chem Inf Comput Sci* 1992;32:373–378.
26. Ibison P, Kam F, Simpson RW, Tonnelier C, Venczel T, Johnson AP. Chemical structure recognition and generic text interpretation in the CLiDE project. Proceedings on Online Information 92, London, England, 1992.
27. Ibison P, Jacquot M, Kam F, Neville AG, Simpson RW, Tonnelier C, Venczel T, Johnson WP. Chemical Literature Data Extraction: The CLiDE Project. *J Chem Inf Comput Sci* 1993;33:338–344.
28. Simon A, Johnson AP. Recent advances in the CliDE project: Logical layout analysis of chemical documents. *J Chem Inf Comput Sci* 1997;37:109–116.
29. Zimmermann M, Bui Thi LT, Hofmann M. Combating illiteracy in chemistry: Towards computer-based chemical structure reconstruction. *ERCIM News* 60, January 2005.
30. Degtyarenko K, Ennis M, Garavelli JS. "Good annotation practice" for chemical data in biology. "Storage and Annotation of Reaction Kinetics Data" Workshop, May 2007; Heidelberg, Germany. Available at http://www.bioinfo.de/isb/2007/07/S1/06/ (accessed August 31, 2009).
31. Fennell RW. *History of IUPAC 1919–1987*. Oxford: Blackwell Science, 1994.
32. Brecher J. Name=Struct: A practical approach to the sorry state of real-life chemical nomenclature. *J Chem Inf Comput Sci* 1999;39:943–950.
33. Lugue RI, Cruz Soto JL, Gómez-Nieto MA. Error detection, recovery, and repair in the translation of inorganic nomenclatures. 1. A study of the problem. *J Chem Inf Comput Sci* 1996;36:7–15.
34. Cooke-Fox DI, Kirby GH, Rayner JD. From names to diagrams–By computer. *Chemische Bericht* 1985;21:467–471.
35. Kirby GH, Rayner JD. Computer translation of IUPAC systematic organic chemical nomenclature. 1. Introduction and background to a grammar-based approach. *J Chem Inf Comput Sci* 1989;29:101–105.
36. Garfield E. An algorithm for translating chemical names into molecular formulas. *J Chem Doc* 1962;2:177–179.
37. Garfield E. From laboratory to information explosions…the evolution of chemical information services at ISI. *J Inform Sci* 2001;27:119–125.
38. Vander Stouw GG, Nanitsky I, Rush JE. Procedures for converting systematic names of organic compounds into atom-bond connection tables. *J Chem Doc* 1967;7:165–169.
39. Vander Stouw GG, Elliott PM, Isenbert AZ. Automated conversion of chemical substance names to atom-bond connection tables. *J Chem Doc* 1974;14:185–193.
40. Cooke-Fox DI, Kirby GH, Lord MR, Rayner JD. Computer translation of IUPAC systematic organic chemical nomenclature. 2. Development of a formal grammar. *J Chem Inf Comput Sci* 1989;29:106–112.

41. Cooke-Fox DI, Kirby GH, Lord MR, Rayner JD. Computer translation of IUPAC systematic organic chemical nomenclature. 3. Syntax analysis and semantic processing. *J Chem Inf Comput Sci* 1989;29:112–118.
42. Cooke-Fox DI, Kirby GH, Lord MR, Rayner JD. Computer translation of IUPAC systematic organic chemical nomenclature. 4. Concise connection tables to structure diagrams. *J Chem Inf Comput Sci* 1990;30:122–127.
43. Cooke-Fox DI, Kirby GH, Lord MR, Rayner JD. Computer translation of IUPAC systematic organic chemical nomenclature. 5. Steroid nomenclature. *J Chem Inf Comput Sci* 1990;30:128–132.
44. Stillwell RN. Computer translation of systematic chemical nomenclature to structural formulas—Steroids. *J Chem Doc* 1973;13:107–109.
45. Ihenfeldt WD, Gasteiger J. Augmenting connectivity information by compound name parsing: Automatic assignment of stereochemistry and isotope labeling. *J Chem Inf Comput Sci* 1995;35:663–674.
46. ACD/Labs. ACD/Name to Structure Batch. Available at http://www.acdlabs.com (accessed August 31, 2009).
47. Apodaca R. Thirty-Two Free Chemistry Databases. 2007. Available at http://Depth-First.com/articles/2007/01/24/thirty-two-free-chemistry-databases (accessed August 31, 2009).
48. Borkent JH, Oukes F, Noordik JH. Chemical reaction searching compared in REACCS, SYNLIB and ORAC. *J Chem Inf Comput Sci* 1988;28:148–150.
49. Weininger D. SMILES, a chemical language and information system. *J Chem Inf Comput Sci* 1988;28:31–36.
50. Contreras ML, Allendes C, Alvarez LT, Rozas R. Computational perception and recognition of digitized molecular structures. *J Chem Inf Comput Sci* 1990;30:302–307.
51. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J Chem Inf Comput Sci* 1992;32:244–255.
52. Murray-Rust P, Rzepa HS, Tyrrell SM, Zhang Y. Representation and use of chemistry in the global electronic age. *Org Biomol Chem* 2004;2:3192–3203.
53. Davies TN. XML in chemistry and chemical identifiers. Available at http://www.iupac.org/publications/ci/2002/2404/XML.html (accessed August 31, 2009).
54. Freemantle M. Unique labels for compounds. *Chem Eng News Am Chem Soc* 2002;80(48):33–35.
55. *The IUPAC International Chemical Identifier (InChI™)*. Available at http://www.iupac.org/inchi/ (accessed August 31, 2009).
56. Bachrach SM, Murray-Rust P, Rzepa HS, Whitaker BJ. Publishing Chemistry on the Internet. Network. *Science* 1996;2(3). Available at http://www.netsci.org/Science/Special/feature07.html (accessed August 31, 2009).
57. Murray-Rust P, Rzepa HS. Chemical markup, XML and the Worldwide Web. 1. Basic principles. *J Chem Inf Comput Sci* 1999;39:928–942.
58. Apodaca R. *A Molecular Language for Modern Chemistry: Getting Started with FlexMol*. December 20, 2006. Available at http://depth-first.com/articles/2006/12/20/a-molecular-language-for-modern-chemistry-getting-started-with-flexmol (accessed August 31, 2009).

59. Weizenbaum J. Eliza-A computer program for the study of natural language communication between man and machine. *Commun ACM* 1966;9:36–45.

60. Press Release (2007). Grants and Awards: Meta data prospecting wins award. *The Alchemister Newsletter*, September 26, 2007. Available at http://www.chemweb.com/content/alchemist/alchemist_20070926.html (accessed August 31, 2009).

61. Rupp CJ, Copestake A, Corbett P, Waldron B. Integrating general-purpose and domain-specific components in the analysis of scientific text. Proceedings of the UK e-Science Programme All Hands Meeting (AHM2007), Nottingham, UK. September 10, 2007. Available at http://www.allhands.org.uk/2007/proceedings/papers/860.pdf (accessed September 8, 2009).

62. Corbett P, Batchelor C, Teufel S. Annotation of chemical named entities. In: *BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pp. 57–64. Prague: Association for Computational Linguistics, 2007.

63. Batchelor CR, Corbett PT. Semantic enrichment of journal articles using chimerical named entity recognition. In: *Proceedings of the ACL 2007, Demo and Poster Sessions*, pp. 45–48. Prague: Association for Computational Linguistics, 2007.

64. Corbett PT, Murray-Rust P. *High-Throughput Identification of Chemistry in Life Science Texts. Computational Life Science II Lecture Notes in Computer Science*, Vol. 4216, pp. 107–118. Berlin, Heidelberg: Springer, 2006.

65. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;25:1251–1255.

66. de Matos P, Ennis M, Darsow M, Guedj K, Degtyarenko K, Apweiler R. (2006). ChEBI—Chemical Entities of Biological Interest. *Nucleic Acids Res Database Summary Paper*, 646.

67. Nic M, Jirat J, Kosata B. *IUPAC Gold Book* (also known as the Compendium of Chemical Terminology). Prague: ICT Press. 2002. Available at http://goldbook.iupac.org/index.html (accessed December 20, 2007).

68. Rzepa HS. A history of hyperactive chemistry on the web: From text and images to objects, models and molecular components. *Chim Int J Chem* 1998;52:653–657.

69. Wilkinson N. Semantic Web Use Cases and Case Studies. World Wide Web Consortium (W3C) Use Cases. 2002. Available at http://www.w3c.org/2001/sw/sweo/Public/UseCases/Pfizer/ (accessed September 8, 2009).

70. Lipinski CA. (2006). *Why repurposing works and how to pick a winning drug while avoiding failures*. CBI 2nd Annual Conference on "Drug Repurposing," Philadelphia, USA, January 30, 2006.

71. Oprea TI, Tropsha A. Target, chemical and bioactivity databases – Integration is key. *Drug Discov Today Technol* 2006;3:357–365.

72. Ashburn TT, Thor KB. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 2004;3:673–683.

INDEX

Note: The letters "f" and "t" refer to figures and tables, respectively.

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

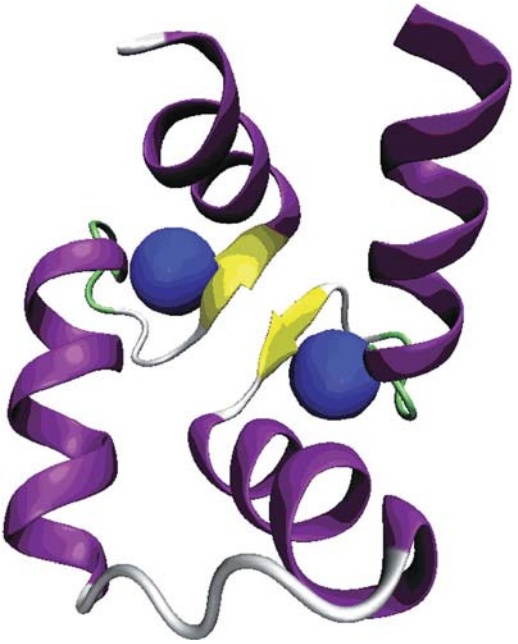


Figure 9.1 C-terminal of calcium-bound calmodulin protein (PDBID 1J7P).

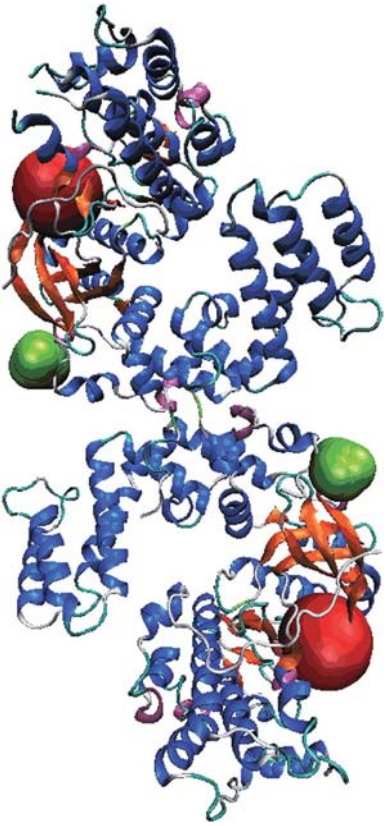


Figure 9.2 G protein-coupled receptor kinase 6 bound to ligands Mg (red) and PO4 (green) (PDBID 2ACX).

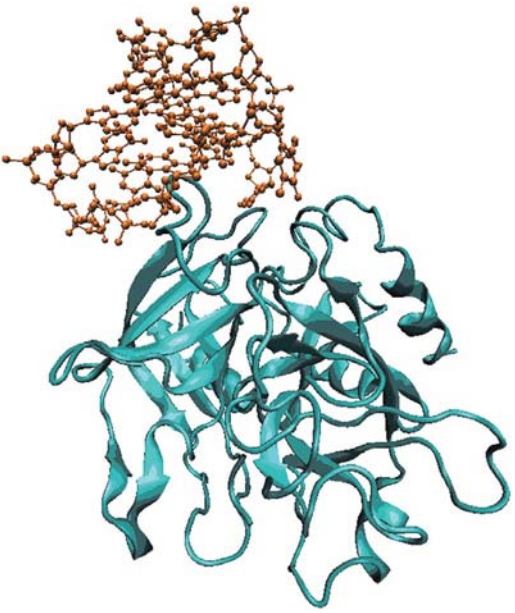


Figure 9.3 Thrombin-binding DNA aptamer (PDBID: 1HAP).

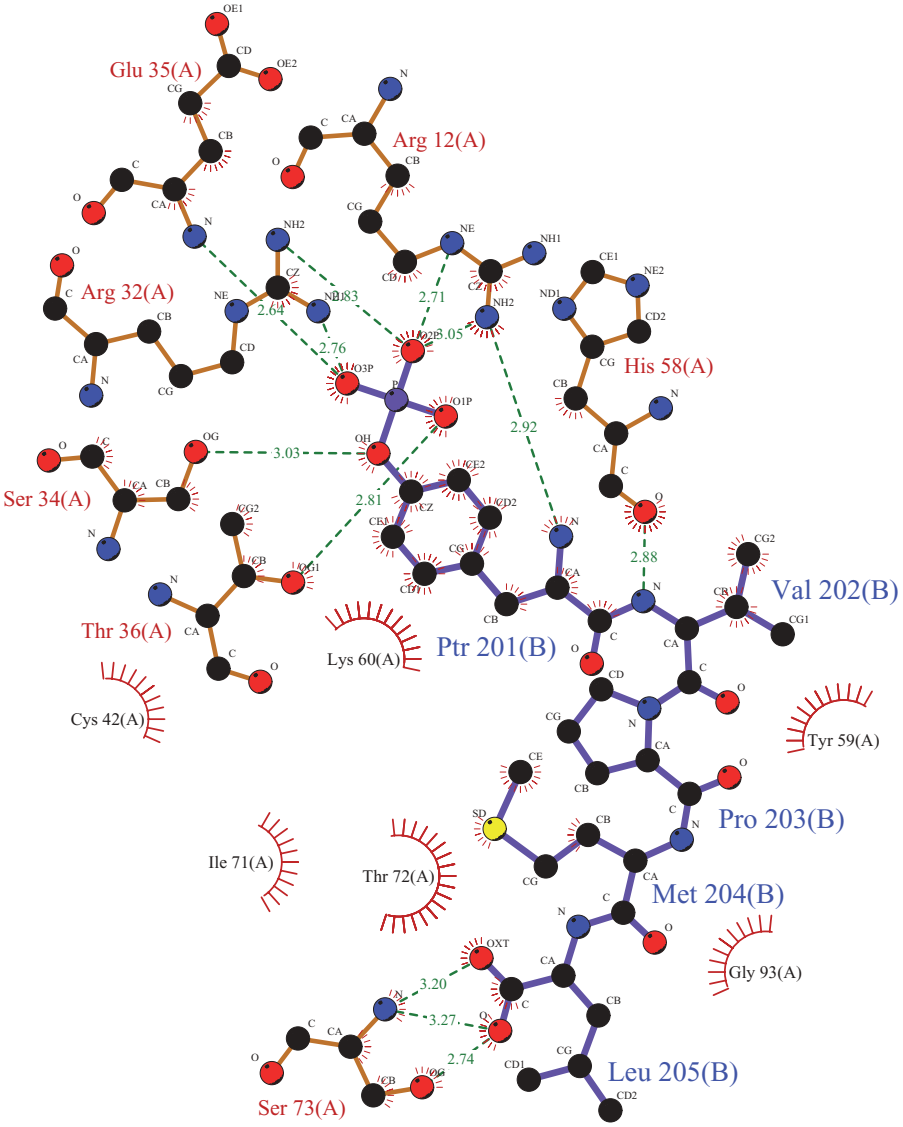


Figure 9.4 SH2 domain complexed with a peptide containing phosphotyrosine (PDBID: 1SHA).

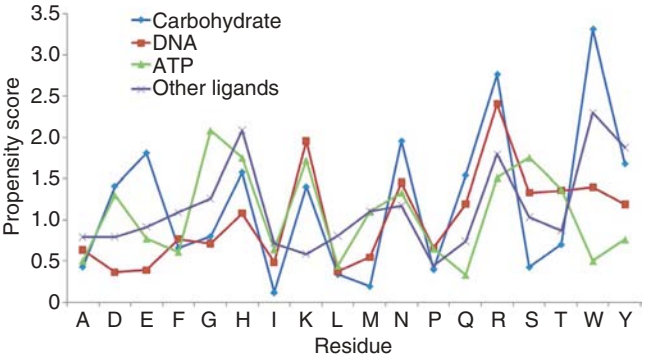


Figure 9.5 Propensity scores of residues in ATP, DNA, carbohydrate, and other ligand binding sites.