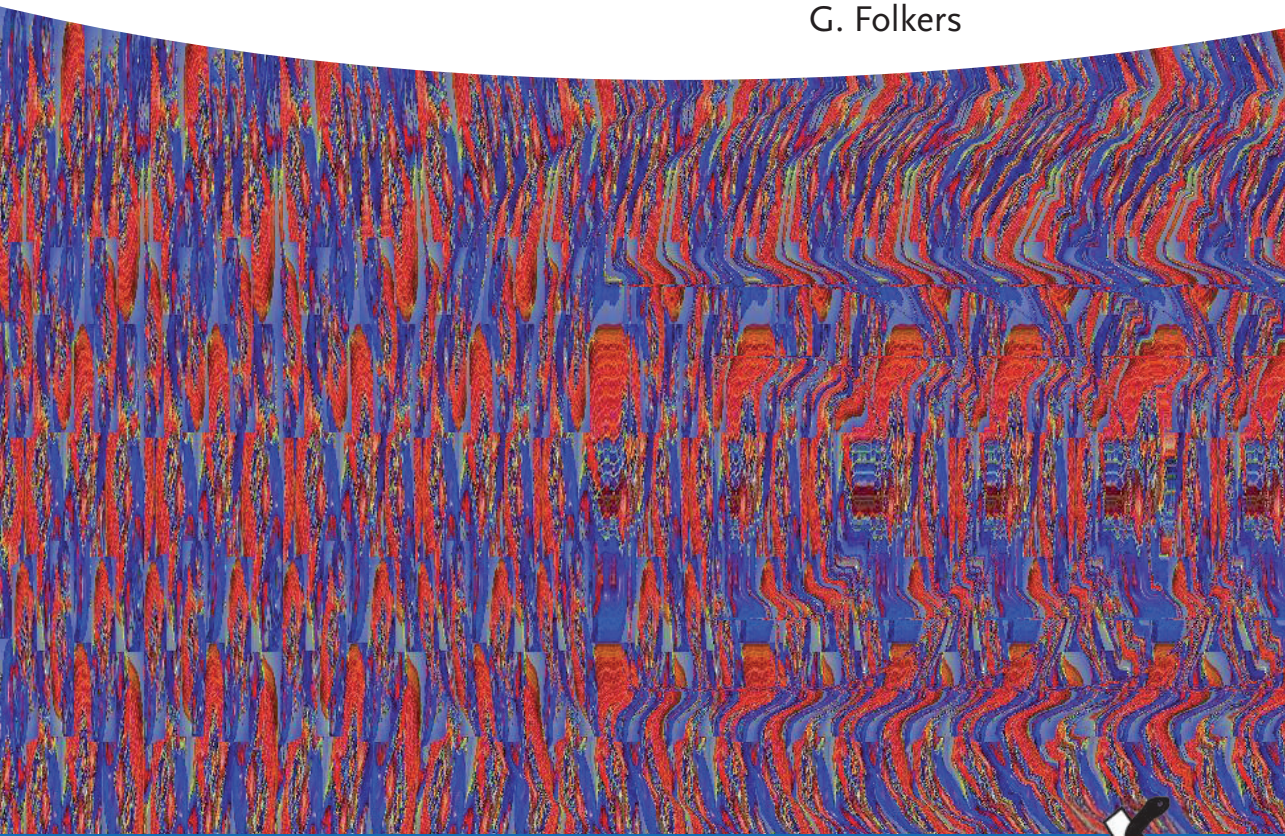Edited by
Rémy D. Hoffmann, Arnaud Gohier, Pavel Pospisil

Data Mining
in Drug Discovery

Volume 57

Series Editors:
R. Mannhold, H. Kubinyi,
G. Folkers

Methods and Principles in Medicinal Chemistry

Edited by
Rémy D. Hoffmann
Arnaud Gohier
Pavel Pospisil

**Data Mining in Drug
Discovery**

Methods and Principles in Medicinal Chemistry

Edited by R. Mannhold, H. Kubinyi, G. Folkers
Editorial Board
H. Buschmann, H. Timmerman, H. van de Waterbeemd, T. Wieland

Previous Volumes of this Series:

Dömling, Alexander (Ed.)

Protein-Protein Interactions in Drug Discovery

2013
ISBN: 978-3-527-33107-9
Vol. 56

Kalgutkar, Amit S./Dalvie, Deepak/Obach, R. Scott/Smith, Dennis A.

Reactive Drug Metabolites

2012
ISBN: 978-3-527-33085-0
Vol. 55

Brown, Nathan (Ed.)

Bioisosteres in Medicinal Chemistry

2012
ISBN: 978-3-527-33015-7
Vol. 54

Gohlke, Holger (Ed.)

Protein-Ligand Interactions

2012
ISBN: 978-3-527-32966-3
Vol. 53

Kappe, C. Oliver/Stadler, Alexander/Dallinger, Doris

Microwaves in Organic and Medicinal Chemistry

Second, Completely Revised and Enlarged Edition

2012
ISBN: 978-3-527-33185-7
Vol. 52

Smith, Dennis A./Allerton, Charlotte/Kalgutkar, Amit S./van de Waterbeemd, Han/Walker, Don K.

Pharmacokinetics and Metabolism in Drug Design

Third, Revised and Updated Edition

2012
ISBN: 978-3-527-32954-0
Vol. 51

De Clercq, Erik (Ed.)

Antiviral Drug Strategies

2011
ISBN: 978-3-527-32696-9
Vol. 50

Klebl, Bert/Müller, Gerhard/Hamacher, Michael (Eds.)

Protein Kinases as Drug Targets

2011
ISBN: 978-3-527-31790-5
Vol. 49

Sotriffer, Christoph (Ed.)

Virtual Screening

Principles, Challenges, and Practical Guidelines

2011
ISBN: 978-3-527-32636-5
Vol. 48

Rautio, Jarkko (Ed.)

Prodrugs and Targeted Delivery

Towards Better ADME Properties

2011
ISBN: 978-3-527-32603-7
Vol. 47

Edited by Rémy D. Hoffmann, Arnaud Gohier,
and Pavel Pospisil

Data Mining in Drug Discovery

WILEY-VCH

Verlag GmbH & Co. KGaA

Series Editors

Prof. Dr. Raimund Mannhold
Rosenweg
740489 Düsseldorf
Germany
mannhold@uni-duesseldorf.de

Prof. Dr. Hugo Kubinyi
Donnersbergstrasse 9
67256 Weisenheim am Sand
Germany
kubinyi@t-online.de

Prof. Dr. Gerd Folkers
Collegium Helveticum
STW/ETH Zurich
8092 Zurich
Switzerland
folkers@collegium.ethz.ch

Volume Editors

Dr. Rémy D. Hoffmann
Prestwick Chemical
Bld. Gonthier d'Andernach
67400 Strasbourg-Illkirch
France

Dr. Arnaud Gohier
Institut de Recherches
Servier
125 Chemin de Ronde
78290 Croissy-sur-Seine
France

Dr. Pavel Pospisil
Philip Morris Int. R&D
Quai Jeanrenaud 5
Biological Systems Res.
2000 NEUCHÂTEL
Switzerland

Cover Description

The cover picture is a 3D stereogram. The pattern is built from a mix of pictures showing complex molecular networks and structures.
The aim of this stereogram is to symbolize the complexity of data to data mine: when looking at them "differently," a shape of a drug pill with a letter D appears!
In order to see it, try parallel or cross-eyed viewing (either you focus your eyes somewhere behind the image or you cross your eyes).

Contents

List of Contributors

Mohammad Afshar
Ariana Pharma
28 rue Docteur Finlay
75015 Paris
France

Kamal Azzaoui
Novartis Institutes for Biomedical
Research (NIBR/CPC/iSLD)
Forum 1 Novartis Campus
4056 Basel
Switzerland

Igor I. Baskin
Strasbourg University
Faculty of Chemistry
UMR 7177 CNRS
1 rue Blaise Pascal
67000 Strasbourg
France

and

MV Lomonosov Moscow State
University
Leninsky Gory
119992 Moscow
Russia

James N.D. Battey
Philip Morris International R&D
Biological Systems Research
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Sven Bergmann
Université de Lausanne
Department of Medical Genetics
Rue du Bugnon 27
1005 Lausanne
Switzerland

Sharon D. Bryant
Inte:Ligand GmbH
Clemens Maria Hofbauer-Gasse 6
2344 Maria Enzersdorf
Austria

Allen Cornett
Novartis Institutes for Biomedical
Research (NIBR/DMP)
220 Massachusetts Avenue
Cambridge, MA 02139
USA

Renée Deehan
Selventa
One Alewife Center
Cambridge, MA 02140
USA

David A. Drubin
Selventa
One Alewife Center
Cambridge, MA 02140
USA

Christof Gaenzler
TIBCO Software Inc.
1235 Westlake Drive, Suite 210
Berwyn, PA 19132
USA

Michael Gilson
University of California San Diego
Skaggs School of Pharmacy and
Pharmaceutical Sciences
9500 Gilman Drive
La Jolla, CA 92093
USA

Janna Hastings
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge CB10 1SD
UK

Julia Hoeng
Philip Morris International R&D
Biological Systems Research
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Nikolai V. Ivanov
Philip Morris International R&D
Biological Systems Research
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Edgar Jacoby
Janssen Research & Development
Turnhoutseweg 30
2340 Beerse
Belgium

Jeremy L. Jenkins
Novartis Institutes for Biomedical
Research (NIBR/DMP)
220 Massachusetts Avenue
Cambridge, MA 02139
USA

Nathalie Jullian
Ariana Pharma
28 rue Docteur Finlay
75015 Paris
France

Esther Kellenberger
UMR 7200 CNRS-UdS
Structural Chemogenomics
74 route du Rhin
67400 Illkirch
France

Thierry Langer
Prestwick Chemical SAS
220, Blvd. Gonthier d'Andernach
67400 Illkirch-Strasbourg
France

Tiging Liu
University of California
San Diego
Skaggs School of Pharmacy and
Pharmaceutical Sciences
9500 Gilman Drive
La Jolla, CA 92093
USA

Gilles Marcou
Strasbourg University
Faculty of Chemistry
UMR 7177 CNRS
1 rue Blaise Pascal
67000 Strasbourg
France

and

MV Lomonosov Moscow State
University
Leninsky Gory
119992 Moscow
Russia

Elyette Martin
Philip Morris International R&D
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Florian Martin
Philip Morris International R&D
Biological Systems Research
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Aurélien Monge
Philip Morris International R&D
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

David Mosenkis
TIBCO Software Inc.
1235 Westlake Drive, Suite 210
Berwyn, PA 19312
USA

George Nicola
University of California San Diego
Skaggs School of Pharmacy and
Pharmaceutical Sciences
9500 Gilman Drive
La Jolla, CA 92093
USA

Florian Nigsch
Novartis Institutes for Biomedical
Research (NIBR)
CPC/LFP/MLI
4002 Basel
Switzerland

Manuel C. Peitsch
Philip Morris International R&D
Biological Systems Research
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Maxim Popov
Novartis Institutes for Biomedical
Research (NIBR/CPC/iSLD)
Forum 1 Novartis Campus
4056 Basel
Switzerland

Pavel Pospisil
Philip Morris International R&D
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

John P. Priestle
Novartis Institutes for Biomedical
Research (NIBR/CPC/iSLD)
Forum 1 Novartis Campus
4056 Basel
Switzerland

Josep Prous Jr.
Prous Institute for Biomedical
Research
Research and Development
Rambla Catalunya 135
08008 Barcelona
Spain

Jordi Quintana
Parc Científic Barcelona (PCB)
Drug Discovery Platform
Baldiri Reixac 4
08028 Barcelona
Spain

Didier Rognan
UMR 7200 CNRS-UdS
Structural Chemogenomics
74 route du Rhin
67400 Illkirch
France

Ansgar Schuffenhauer
Novartis Institutes for Biomedical
Research (NIBR/CPC/iSLD)
Forum 1 Novartis Campus
4056 Basel
Switzerland

Alain Sewer
Philip Morris International R&D
Biological Systems Research
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Christoph Steinbeck
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SD
UK

Ty M. Thomson
Selventa
Cambridge, MA 02140
USA

Yannic Tognetti
Ariana Pharma
28 rue Docteur Finlay
75015 Paris
France

Antoni Valencia
Prous Institute for Biomedical
Research, SA
Computational Modeling
Rambla Catalunya 135
08008 Barcelona
Spain

Thibault Varin
Eli Lilly and Company
Lilly Research Laboratories
Lilly Corporate Center
Indianapolis, IN 46285
USA

Jurjen W. Westra
Selventa
Cambridge, MA 02140
USA

Preface

In general, the extraction of information from databases is called data mining. A database is a data collection that is organized in a way that allows easy accessing, managing, and updating its contents. Data mining comprises numerical and statistical techniques that can be applied to data in many fields, including drug discovery. A functional definition of data mining is the use of numerical analysis, visualization, or statistical techniques to identify nontrivial numerical relationships within a data set to derive a better understanding of the data and to predict future results. Through data mining, one derives a model that relates a set of molecular descriptors to biological key attributes such as efficacy or ADMET properties. The resulting model can be used to predict key property values of new compounds, to prioritize them for follow-up screening, and to gain insight into the compounds' structure–activity relationship. Data mining models range from simple, parametric equations derived from linear techniques to complex, nonlinear models derived from nonlinear techniques. More detailed information is available in literature [1–7].

This book is organized into four parts. Part One deals with different sources of data used in drug discovery, for example, protein structural databases and the main small-molecule bioactivity databases.

Part Two focuses on different ways for data analysis and data enrichment. Here, an industrial insight into mining HTS data and identifying hits for different targets is presented. Another chapter demonstrates the strength of powerful data visualization tools for simplification of these data, which in turn facilitates their interpretation.

Part Three comprises some applications to polypharmacology. For instance, the positive outcomes are described that data mining can produce for ligand profiling and target fishing in the chemogenomics era.

Finally, in Part Four, systems biology approaches are considered. For example, the reader is introduced to integrative and modular analysis approaches to mine large molecular and phenotypical data. It is shown how the presented approaches can reduce the complexity of the rising amount of high-dimensional data and provide a means for integrating different types of omics data. In another chapter, a set of novel methods are established that quantitatively measure the biological impact of chemicals on biological systems.

The series editors are grateful to Remy Hoffmann, Arnaud Gohier, and Pavel Pospisil for organizing this book and to work with such excellent authors. Last but not least, we thank Frank Weinreich and Heike Nöthe from Wiley-VCH for their valuable contributions to this project and to the entire book series.

<table>
<tr><td>Düsseldorf</td><td><i>Raimund Mannhold</i></td></tr>
<tr><td>Weisenheim am Sand</td><td><i>Hugo Kubinyi</i></td></tr>
<tr><td>Zürich</td><td><i>Gerd Folkers</i></td></tr>
<tr><td>May 2013</td><td></td></tr>
</table>

References

1 Cruciani, G., Pastor, M., and Mannhold, R. (2002) Suitability of molecular descriptors for database mining: a comparative analysis. *Journal of Medicinal Chemistry*, **45**, 2685–2694.

2 Obenshain, M.K. (2004) Application of data mining techniques to healthcare data. *Infection Control and Hospital Epidemiology*, **25**, 690–695.

3 Weaver, D.C. (2004) Applying data mining techniques to library design, lead generation and lead optimization. *Current Opinion in Chemical Biology*, **8**, 264–270.

4 Yang, Y., Adelstein, S.J., and Kassis, A.I. (2009) Target discovery from data mining approaches. *Drug Discovery Today*, **14**, 147–154.

5 Campbell, S.J., Gaulton, A., Marshall, J., Bichko, D., Martin, S., Brouwer, C., and Harland, L. (2010) Visualizing the drug target landscape. *Drug Discovery Today*, **15**, 3–15.

6 Geppert, H., Vogt, M., and Bajorath, J. (2010) Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *Journal of Chemical Information and Modeling*, **50**, 205–216.

7 Hasan, S., Bonde, B.K., Buchan, N.S., and Hall, M.D. (2012) Network analysis has diverse roles in drug discovery. *Drug Discovery Today*, **17**, 869–874.

A Personal Foreword

The term data mining is well recognized by many scientists and is often used when referring to techniques for advanced data retrieval and analysis. However, since there have been recent advances in techniques for data mining applied to the discovery of drugs and bioactive molecules, assembling these chapters from experts in the field has led to a realization that depending upon the field of interest (biochemistry, computational chemistry, and biology), data mining has a variety of aspects and objectives.

Coming from the ligand molecule world, one can state that the understanding of chemical data is more complete because, in principle, chemistry is governed by physicochemical properties of small molecules and our "microscopic" knowledge in this domain has advanced considerably over the past decades. Moreover, chemical data management has become relatively well established and is now widely used. In this respect, data mining consists in a thorough retrieval and analysis of data coming from different sources (but mainly from literature), followed by a thorough cleaning of data and its organization into compound databases. These methods have helped the scientific community for several decades to address pathological effects related to simple (single target) biological problems. Today, however, it is widely accepted that many diseases can only be tackled by modulating the ligand biological/pharmacological profile, that is, its "molecular phenotype." These approaches require novel methodologies and, due to increased accessibility to high computational power, data mining is definitely one of them.

Coming from the biology world, the perception of data mining differs slightly. It is not just a matter of literature text mining anymore, since the disease itself, as well as the clinical or phenotypical observations, may be used as a starting point. Due to the complexity of human biology, biologists start with hypotheses based upon empirical observations, create plausible disease models, and search for possible biological targets. For successful drug discovery, these targets need to be druggable. Moreover, modern systems biology approaches take into account the full set of genes and proteins expressed in the drug environment (omics), which can be used to generate biological network information. Data mining these data, when structured into such networks, will provide interpretable information that

leads to an increased knowledge of the biological phenomenon. Logically, such novel data mining methods require new and more sophisticated algorithms.

This book aims to cover (in a nonexhaustive manner) the data mining aspects for these two parallel but meant-to-be-convergent fields, which should not only give the reader an idea of the existence of different data mining approaches, algorithms, and methods used but also highlight some elements to assess the importance of linking ligand molecules to diseases. However, there is awareness that there is still a long way to go in terms of gathering, normalizing, and integrating relevant biological and pharmacological data, which is an essential prerequisite for making more accurate simulations of compound therapeutic effects.

This book is structured into four parts: Part One, Data Sources, introduces the reader to the different sources of data used in drug discovery. In Chapter 1, Kellenberger *et al.* present the Protein Data Bank and related databases for exploring ligand–protein recognition and its application in drug design. Chapter 2 by Nicola *et al.* is a reprint of a recently published article in *Journal of Medicinal Chemistry* (2012, 55 (16): 6987–7002) that nicely presents the main small-molecule bioactivity databases currently used in medicinal chemistry and the modern trends for their exploitation. In Chapter 3, Hastings *et al.* point out the importance of chemical ontologies for the standardization of chemical libraries in order to extract and organize chemical knowledge in a way similar to biological ontologies. Chapter 4 by Martin *et al.* presents the importance of a corporate chemical registry system as a central repository for uniform chemical entities (including their spectrometric data) and as an important point of entry for exploring public compound activity databases for systems biology data.

Part Two, Analysis and Enrichment, describes different ways for data analysis and data enrichment. In Chapter 5, Battey *et al.* didactically present the basics of plant pathway construction, the potential for their use in data mining, and the prediction of pathways using information from an enzymatic structure. Even though this chapter deals with plant pathways, the information can be readily interpreted and applied directly to metabolic pathways in humans. In Chapter 6, Azzaoui *et al.* present an industrial insight into mining HTS data and identifying hits for different targets and the associated challenges and pitfalls. In Chapter 7, Mosenkis *et al.* clearly demonstrate, using different examples, how powerful data visualization tools are key to the simplification of complex results, making them readily intelligible to the human brain and eye. We also welcome Chapter 8 by Marcou *et al.* that provides a concrete example of the increasingly frequent need for powerful statistical processing tools. This is exemplified by the use of R in the chemoinformatics process. Readers will note that this chapter is built like a tutorial for the R language in order to process, cluster, and visualize molecules, which is demonstrated by its application to a concrete example. For programmers, this may serve as an initiation to the use of this well-known bioinformatics tool for processing chemical information.

Part Three, Applications to Polypharmacology, contains chapters detailing tools and methods to mine data with the aim to elucidate preclinical profiles of small

molecules and select potential new drug targets. In Chapter 9, Prous *et al.* nicely present three examples of knowledge bases that attempt to relate, in a comprehensive manner, the interactions between chemical compounds, biological entities (molecules and pathways), and their assays. The second part of this chapter presents the challenges that these knowledge-based data mining methodologies face when searching for potential mechanisms of action of compounds. In Chapter 10, Jullian *et al.* introduce the reader to the advantages of using rule-based methods when exploring polypharmacological data sets, compared to standard numerical approaches, and their application in the development of novel ligands. Finally, in Chapter 11, Bryant *et al.* familiarize us with the positive outcomes that data mining can produce for ligand profiling and target fishing in the chemogenomics era. The authors expose how searching through ligand and target pharmacophoric structural and descriptor spaces can help to design or extend libraries of ligands with desired pharmacological, yet lowered toxicological, properties.

In Part Four, Systems Biology Approaches, we are pleased to include two exciting chapters coming from the biological world. In Chapter 12, Bergmann introduces us to integrative and modular analysis approaches to mine large molecular and phenotypical data. The author argues how the presented approaches can reduce the complexity of the rising amount of high-dimensional data and provide a means to integrating different types of omics data. Moreover, astute integration is required for the understanding of causative links and the generation of more predictive models. Finally, in the very robust Chapter 13, Sewer *et al.* present systems biology-based approaches and establish a set of novel methods that quantitatively measure the biological impact of the chemicals on biological systems. These approaches incorporate methods that use mechanistic causal biological network models, built on systems-wide omics data, to identify any compound's mechanism of action and assess its biological impact at the pharmacological and toxicological level. Using a five-step strategy, the authors clearly provide a framework for the identification of biological networks that are perturbed by short-term exposure to chemicals. The quantification of such perturbation using their newly introduced impact factor "BIF" then provides an immediately interpretable assessment of such impact and enables observations of early effects to be linked with long-term health impacts.

We are pleased that you have selected this book and hope that you find the content both enjoyable and educational. As many authors have accompanied their chapters with clear concise pictures, and as someone once said "one figure can bear thousand words," this Personal Foreword also contains a figure (see below). We believe that the novel applications of data mining presented in these pages by authors coming from both chemical and biological communities will provide the reader with more insight into how to reshape this pyramid into a trapezoidal form, with the enlarged knowledge area. Thus, improved data processing techniques leading to the generation of readily interpretable information, together with an increased understanding of the therapeutic processes, will enable scientists

to take wiser decisions regarding what to do next in their efforts to develop new drugs.

We wish you a happy and inspiring reading.

Strasbourg, March 14, 2013

Remy Hoffmann, Arnaud Gohier,
and Pavel Pospisil
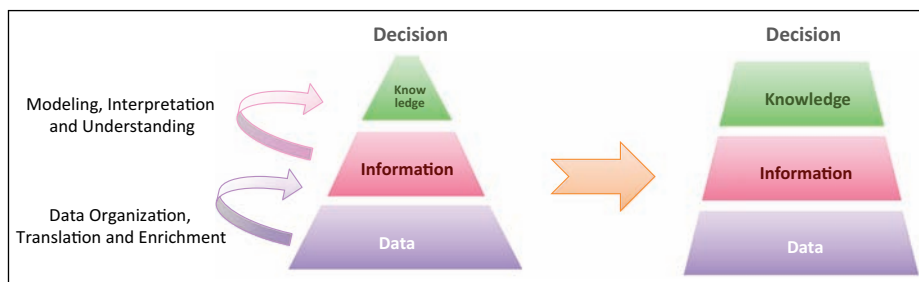
Part One
Data Sources

1
Protein Structural Databases in Drug Discovery

Esther Kellenberger and Didier Rognan

1.1
The Protein Data Bank: The Unique Public Archive of Protein Structures

1.1.1
History and Background: A Wealthy Resource for Structure-Based Computer-Aided Drug Design

The Protein Data Bank (PDB) was founded in the early 1970s to provide a repository of three-dimensional (3D) structures of biological macromolecules. Since then, scientists from around the world submit coordinates and information to mirror sites in the Unites States, Europe, and Asia. In 2003, the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB, USA), the Protein Data Bank in Europe (PDBe) – the Macromolecular Structure Database at the European Bioinformatics Institute (MSD-EBI) before 2009, and the Protein Data Bank Japan (PDBj) at the Osaka University formally merged into a single standardized archive, named the worldwide PDB (wwPDB, http://www .wwpdb.org/) [1]. At its creation in 1971 at the Brookhaven National Laboratory, the PDB registered seven structures. With more than 75 000 entries in 2011, the number of structures being deposited each year in PDB has been constantly increasing (Figure 1.1).

The growth rate was especially boosted in the 2000s by structural genomics initiatives [2,3]. Research centers from around the globe made joint efforts to overexpress, crystallize, and solve the protein structures at a high throughput for a reduced cost. Particular attention was paid to the quality and the utility of the structures, thereby resulting in supplementation of the PDB with new folds (i.e., three-dimensional organization of secondary structures) and new functional families [4,5].

The TargetTrack archive (http://sbkb.org) registers the status of macromolecules currently under investigation by all contributing centers (Table 1.1) and illustrates the difficulty in getting high-resolution crystal structures, since only 5% targets undergo the multistep process from cloning to deposition in the PDB.
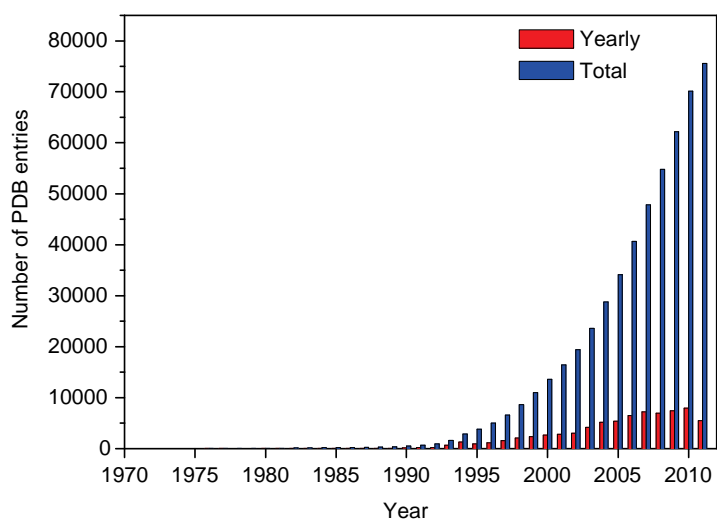
Figure 1.1 Yearly growth of deposited structures in the Protein Data Bank (accessed August 2011).

If only 450 complexes between an FDA-approved drug and a relevant target are available according to the DrugBank [6], the PDB provides structural information for a wealth of potential druggable proteins, with more than 40 000 different sequences that cover about 18 000 clusters of similar sequences (more than 30% identity).

Table 1.1 TargetTrack status statistics.

Status	Total number of targets	Relative to "cloned" targets (%)	Relative to "expressed" targets (%)	Relative to "purified" targets (%)	Relative to "crystallized" targets (%)
Cloned	192 735	100.0	—	—	—
Expressed	120 526	62.5	100.0	—	—
Soluble	35 436	18.4	29.4	—	—
Purified	45 105	23.4	37.4	100.0	—
Crystallized	14 472	7.5	12.0	32.1	100.0
Diffraction-quality crystals	7059	3.7	5.9	15.7	48.8
Diffraction	7522	3.9	6.2	16.7	52.0
NMR assigned	2262	1.2	1.9	5.0	—
HSQC	3409	1.8	2.8	7.6	—
Crystal structure	4953	2.6	4.1	11.0	34.2
NMR structure	2136	1.1	1.8	4.7	—
In PDB	8618	4.5	7.2	19.1	45

Accessed August 2011.

1.1.2
**Content, Format, and Quality of Data: Pitfalls and Challenges
When Using PDB Files**

1.1.2.1 **The Content**
The PDB stores 3D structures of biological macromolecules, mainly proteins (about 92% of the database), nucleic acids, or complexes between proteins and nucleic acids. The PDB depositions are restricted to coordinates that are obtained using experimental data. More than 87% of PDB entries are determined by X-ray diffraction. About 12% of the structures have been computed from nuclear magnetic resonance (NMR) measurements. Few hundreds of structures were built from electron microscopy data. The purely theoretical models, such as *ab initio* or homology models, are no more accepted since 2006. For most entries, the PDB provides access to the original biophysical data, structure factors and restraints files for X-ray and NMR structures, respectively. During the past two decades, advances in experimental devices and computational methods have considerably improved the quality of acquired data and have allowed characterization of large and complex biological specimens [7,8]. As an example, the largest set of coordinates in the PDB describes a bacterial ribosomal termination complex (Figure 1.2) [9]. Its structure determined by electron microscopy includes 45 chains of proteins and nucleic acids for a total molecular weight exceeding 2 million Da.
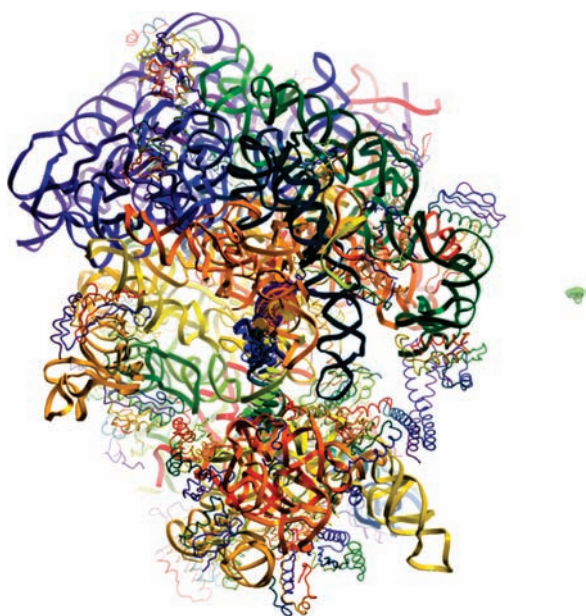
Figure 1.2 Comparative display of the largest macromolecule in the PDB (*Escherichia coli* ribosomal termination complex, PDB code 1ml5, left) and of a prototypical drug (aspirin, PDB code 2qqt, right).

To stress the quality issue, one can note the recent increase in the number of crystal structures solved at very high resolution: 90% of the 438 structures with a resolution better than 1 Å was deposited after year 2000. More generally, the enhancement in the structure accuracy translates into a more precise representation of the biopolymer details (e.g., alternative conformations of an amino acid side chain) and into the enlarged description of the molecular environment of the biopolymer, that is, of the nonbiopolymer molecules, also named ligands. Ligands can be any component of the crystallization solution (ions, buffers, detergents, crystallization agents, etc.), but it can also be biologically relevant molecules (cofactors and prosthetic groups, inhibitors, allosteric modulators, and drugs). Approximately 11 000 different free ligands are spread across 70% of the PDB files.

1.1.2.2 The Format
The conception of a standardized representation of structural data was a requisite of the database creation. The PDB format was thus born in the 1970s and was designed as a human-readable format. Initially based on the 80 columns of a punch card, it has not much evolved over time and still consists in a flat file divided into two sections organized into labeled fields (see the latest PDB file format definition at http://www .wwpdb.org/docs.html). The first section, or *header*, is dedicated to the technical description and the annotation (e.g., authors, citation, biopolymer name, and sequence). The second one contains the coordinates of biopolymer atoms (ATOM records), the coordinates of ligand atoms (HETATM records), and the bonds within atoms (CONECT records). The PDB format is roughly similar to the connection table of MOL and SD files [10], but with an incomplete description of the molecular structure. In practice, no information is provided in the CONECT records for atomic bonds within biopolymer residues. Bond orders in ligands (simple, double, triple, and aromatic) are not specified and the connectivity data may be missing or wrong. In the HETATM records, each atom is defined by an arbitrary name and an atomic element (as in the periodic table). Because the hydrogen atoms are usually not represented in crystal structures, there are often atomic valence ambiguities in the structure of ligands.

To overcome limits in data handling and storage capacity for very large biological molecules, two new formats were introduced in 1997 (the macromolecular crystallographic information file or mmCIF) and 2005 [the PDB markup language (PDBML), an XML format derivative] [11,12]. They better suit the description of ligands, but are however not widely used by the scientific community. There are actually few programs able to read mmCIF and PDBML formats, whereas almost all programs can display molecules from PDB input coordinates.

1.1.2.3 The Quality and Uniformity of Data
Errors and inconsistencies are still frequent in PDB data (see examples in Table 1.2). Some of them are due to evolution in time of collection, curation, and processing of the data [13]. Others are directly introduced by the depositors because of the limits in experimental methods or because of an incomplete knowledge of the chemistry and/ or biology of the studied sample. In 2007, the wwPDB released a complete

Table 1.2 Common errors in PDB files and effect of the wwPDB remediation.

Description of errors	Impacted data	Status upon remediation
Invalid source organism	Annotation	Fixed
Invalid reference to protein sequence databases	Annotation	Fixed
Inconsistencies in protein sequences[a]	Annotation	Fixed
Violation of nomenclature in protein[b]	Structure	Fixed
Incomplete CONECT record for ligand residues	Structure	Partly solved
Wrong chemistry in ligand residues	Structure	Partly solved
Violation of nomenclature in ligand[c]	Structure	Unfixed
Wrong coordinates[d]	Structure	Unfixed

a) In HEADER and ATOM records.
b) For example, residue or atom names.
c) Discrepancy between the structure described in the PDB file and the definition in the Chemical Component Dictionary.
d) For example, wrong side chain rotamers in proteins.

remediated archive [14]. In practice, sequence database references and taxonomies were updated and primary citations were verified. Significant efforts have also been devoted to chemical description and nomenclature of the biopolymers and ligands. The PDB file format was upgraded (v3.0) to integrate uniformity and remediation data and a reference dictionary called the Chemical Component Dictionary has been established to provide an accurate description of all the molecular entities found in the database. To date, however, only a few modeling programs (e.g., MOE[1] and SYBYL[2]) make use of the dictionary to complement the ligand information encoded in PDB files.

The remediation by the wwPDB yielded in March 2009 to the version 3.2 of the PDB archive, with a focus on detailed chemistry of biopolymers and bound ligands. Remediation is still ongoing and the last remediated archive was released in July 2011. There are nevertheless still structural errors in the database. Some are easily detectable, for example, erroneous bond lengths and bond angles, steric clashes, or missing atoms. These errors are very frequent (e.g., the number of atomic clashes in the PDB was estimated to be 13 million in 2010), but in principle can be fixed by recomputing coordinates from structure factors or NMR restraints using a proper force field [15]. Other structural errors are not obvious. For example, a wrong protein topology is identified only if new coordinates supersede the obsolete structure or if the structure is retracted [16]. Hopefully, these errors are rare. More common and yet undisclosed structural ambiguities concern the ionization and the tautomerization of biopolymers and ligands (e.g., three different protonation states are possible for histidine residues).

1) Chemical Computing Group, Montreal, Quebec, Canada H3A 2R7.
2) Tripos, St. Louis, MO 63144-2319, USA.

To evaluate the accuracy of a PDB structure, querying the PDB-related databases PDBREPORT and PDB_REDO is a good start [15]. PDBREPORT (http://swift.cmbi .ru.nl/gv/pdbreport/) registers, for each PDB entry, all structural anomalies in biopolymers. PDB_REDO (http://www.cmbi.ru.nl/pdb_redo/) holds rerefined copies of the PDB structures solved by X-ray crystallography (Figure 1.3).

PDB_REDO

New PDB code

PDB entry 3rte

Structure

Spacegroup	I 4 2 2		
Cell dimensions	a: 122.441 Å	b: 122.441 Å	c: 155.050 Å
	α: 90.00°	β: 90.00°	γ: 90.00°
Resolution	2.10 Å		

Experimental data

Reflections	All: 32084	Test set: 1625 (5.1%)
Resolution range	38.72 Å	2.10 Å

R-values etc.

	From PDB header	Calculated from data	After conservative optimisation	After full optimisation
R	0.1620	0.1612	0.1636	0.1624
R-free	0.2070	0.2058	0.1975	0.1962
σR-free		0.0036	0.0035	0.0034
R-free Z-score		-0.89	2.31	2.32

WHAT_CHECK validation

	Original PDB entry	Conservatively optimised	Fully optimised
1st generation packing quality[1]	0.439	0.465	0.497
2nd generation packing quality[1]	-0.509	-0.443	-0.431
Ramachandran plot appearance[1]	-0.275	-0.036	-0.032
Chi-1/Chi-2 rotamer normality[1]	-0.900	0.110	0.555
Backbone conformation[1]	-0.345	-0.239	-0.271
Bond length RMS Z-score[2]	0.843	0.356	0.351
Bond angle RMS Z-score[2]	0.843	0.530	0.521
Total number of bumps[3]	32	26	23
Unsatisfied H-bond donors/acceptors[3]	16	18	20
Full WHAT_CHECK reports	Link	Link	Link

[1] Higher is better
[2] Should be lower than 1.000
[3] Fewer is better

Download

- Conservatively optimised structure (PDB | MTZ)
- Fully optimised structure (PDB | MTZ)
- YASARA scenes (for visualisation of the results)
- All files (compressed)
- PDB structure
- Structure factors

Figure 1.3 PDB_REDO characteristics of the 3rte PDB entry.

The quality issue was recently discussed in a drug design perspective with benchmarks for structure-based computer-aided methods [17–19]. A consensual conclusion is that the PDB is an invaluable resource of structural information provided that data quality is not overstated.

1.2
PDB-Related Databases for Exploring Ligand–Protein Recognition

The bioactive structure of ligands in complex with relevant target is of special interest for drug design. During the last decade, many databases of ligand/protein information have been derived from the PDB. Their creation was always motivated by the ever-growing amount of structural data. Each database however has its own focus, which can be a large-scale analysis of ligands and/or proteins in PDB complexes, or training and/or testing affinity prediction, or other structure-based drug design methods (e.g., docking). Accordingly, ligands are either thoroughly collected across all PDB complexes or only retained if satisfying predefined requirements. As a consequence, the number of entries in PDB-related databases ranges from a few thousands to over 50 000 entries. These databases also differ greatly in their content. This section does not intend to establish an exhaustive list. We have chosen to discuss only the recent or widely used databases and to group them according to their main purposes (Table 1.3).

1.2.1
Databases in Parallel to the PDB

The wwPDB contributors have developed free Web-based tools to match chemical structures in the PDB files to entities in the Chemical Component Dictionary; the Ligand Expo and PDBeChem resources are linked to the RCSB PDB and PDBe, respectively, and provide the chemical structure of all ligands of every PDB file [20,21]. A few other databases also hold one entry for each PDB entry. The Het-PDB database was designed in 2003 at the Nagahama Institute of Bio-Science and Technology to survey the nonbiopolymer molecules in the PDB and to draw statistics about their frequency and interaction mode [22]. It is still monthly updated and covers 12 000 ligands in the PDB. It revealed that the most repeated ligands in the PDB were metal ions, sugars, and nucleotides, all of which can be considered as part of the functional protein as a result of a posttraductional modification or as cofactors. Another important database was developed at Uppsala University to provide structural biologists with topology and parameters file for ligands [23]. This database named HIC-Up was maintained until 2008 by G. Kleywegt, who now leads the PDBe. Another useful service has been offered by the Structural Bioinformatics group in Berlin: the Web interface of the SuperLigands database allows the search for 2D and 3D similar ligands in the PDB [24]. The last update of SuperLigands was made in December 2009. Other PDB ligand warehouses have been developed during the last decade, but, like HIC-Up and SuperLigands, are not actively

Table 1.3 Representative examples of PDB-related databases useful for drug design.

Databases	Dates[a)]	Content	Web site
Repository of PDB ligands			
Ligand Expo	2004-	>13 000 different ligands. Experimental and ideal coordinates of ligands (PDB, SD, mmCIF formats)	ligand-expo.rcsb.org
PDBeChem	2005-	>13 000 different ligands. Experimental and ideal coordinates of ligands (PDB, SD, mmCIF formats)	www.ebi.ac.uk/pdbe/
HET-PDB	2004-	12 262 different ligands in 74 732 PDB files (August 2011). Navigator only, no download	hetpdbnavi.nagahama-i-bio.ac.jp
HiC-Up	1997–2008	7870 different ligands (March 2008). Experimental and ideal coordinates of ligands in PDB format. Dictionary files (X-PLOR/CNS, O, TNT)	xray.bmc.uu.se/hiccup
SuperLigands	2005–2009	10 085 different ligands in 401 300 complexes. Experimental coordinates of ligands in PDB and MOL formats	bioinformatics.charite.de/superligands/
Experimental binding affinities			
PDBBind	2004-	Affinity data for 7986 PDB complexes	http://www.pdbbind.org.cn
Binding MOAD	2005-	Affinity data for 4782 PDB complexes	www.bindingmoad.org
BindingDB	2001-	721 721 affinity data for 60 179 proteins and 316 172 ligands, including PDB complexes	www.bindingdb.org/bind
ChEMBL	2008-	>5 million affinity data for 8603 proteins and >1 million ligands, including PDB complexes	www.ebi.ac.uk/chembl
Structural description of protein-ligand complexes			
Relibase	2003-	Experimental coordinates of the complex (in PDB and MOL2 format) or of the isolated ligand (in SD and MOL2 format)	relibase.ccdc.cam.ac.uk
sc-PDB	2006-	9891 protein–ligand complexes with refined hydrogen atom positions. Separate coordinates for ligands (SD and MOL2 format), protein (PDB and MOL2 format), and active site (MOL2 format)	bioinfo-pharma.u-strasbg.fr/scPDB/
PSMDB		5266 nonredundant protein–ligand complexes. Separate coordinates for ligands (SD format) and proteins (PDB format)	compbio.cs.toronto.edu/psmdb

a) The year of database creation is that of relative primary publication. It is followed by the year of the database last updated (- indicates that the database is still updated).

maintained, since the RCSB PDB and the PDBe directly integrate most of their data or services.

1.2.2
Collection of Binding Affinity Data

A few databases collect binding affinities such as experimentally determined inhibition (IC_{50}, K_i) or dissociation (K_d) constant for PDB complexes. The larger ones are Binding MOAD, PDBbind, and BindingDB [25–27]. Both Binding MOAD and PDBbind were developed at the University of Michigan, and have in common the separation of biologically relevant PDB ligands from invalid ones, such as salts and buffers. Their focuses are however different. For example, PDBbind disregards any complex without binding data, whereas Binding MOAD groups proteins into functional families and chooses the highest affinity complex as a representative. BindingDB considers only potential drug targets in the PDB, but collects data for many ligands that are not represented in the PDB.

In all cases, data gathering implies the manual review of the reference publications in PDB files and, more generally, expert parsing of scientific literature. BindingDB also contains data extracted from two other Web resources, PubChem BioAssay and ChEMBL. PubChem BioAssay database at the National Center for Biotechnology Information (NIH) contains biological screening results. ChEMBL is the chemogenomics data resource at the European Molecular Biology Laboratory. It contains binding data and other bioactivities extracted from scientific literature for more than a million bioactive small molecules, including many PDB ligands.

Affinity databases were recently made available from two of the wwPDB mirror sites. The RCSB PDB Web site now includes hyperlinks to the actively maintained ones, BindingDB and BindingMOAD. The PDBe Web site communicates with ChEMBL.

1.2.3
Focus on Protein–Ligand Binding Sites

As already described, RCSB PDB and PDBe resources currently provide chemical description and 3D coordinates for all ligands in the PDB. They also provide tools for inspection of protein–ligand binding (Ligand Explorer at RCSB PDB and PDBe-Motifs at PDBe). But as already discussed in this chapter, PDB data are prone to chemical ambiguities and not directly suitable to finely describe nonbonded intermolecular interactions. Several initiatives aimed at the structural characterization of protein–ligand interactions at the PDB scale. Among the oldest one is Relibase that automatically analyzes all PDB entries, identifies all complexes involving non-biopolymer groups, and supplies the structural data with additional information, such as atom and bond types [28]. Relibase allows various types of queries (text searching, 2D substructure searching, 3D protein–ligand interaction searching, and ligand similarity searching) and complex analyses, such as automatic superposition

of related binding sites to compare ligand binding modes. The Web version of Relibase is freely available to academic users, but does not include all possibilities for exploration of PDB complexes.

If Relibase holds as many entries as PDB holds ligand–protein complexes, other databases were built using only a subset of the PDB information. For example, the sc-PDB is a nonredundant assembly of 3D structures for "druggable" PDB complexes [29]. The druggability here does not imply the existence of a drug–protein complex, but that both the binding site and the bound ligand obey topological and physicochemical rules typical of pharmaceutical targets and drug candidates, respectively. Strict selection rules and extensive manual verifications ensure the selection in the PDB of binary complexes between a small biologically relevant ligand and a druggable protein binding site. The preparation, content, and applications of the sc-PDB are detailed in Section 1.3.

Along the same lines, the PSMDB database endeavors to set up a smaller and yet most diverse data set of PDB ligand–protein complexes [30]. Full PDB entries are parsed to select structures determined by X-ray diffraction with a resolution lower than 2 Å, with at least one protein chain longer than 50 amino acids, and a noncovalently bound small ligand. The PDB file of each selected complex was split into free protein structure and bound ligand(s). The added value of PSMDB does not consists in these output structure files that contain the original PDB coordinates, but in the handling of redundancy at both the protein and ligand levels.

With the growing interest of the pharmaceutical industry for fragment-based approach to drug design [31], several applications focusing on individual fragments derived from PDB ligands have recently emerged. Algorithms for molecule fragmentation were applied to a selection of PDB ligands defining a library of fragment binding sites [32] to map the amino acid preference of such fragments [33] or to extract possible bioisosteres [34].

1.3
The sc-PDB, a Collection of Pharmacologically Relevant Protein–Ligand Complexes

We decided in 2002 to set up a collection of protein–ligand binding sites called sc-PDB, originally designed for reverse docking applications [35]. While docking a set of ligands to a single protein was already a well-established computational technique for identifying potentially interesting novel ligands, the reverse paradigm (docking a single ligand to a set of protein active sites) was still a marginal approach. The main difficulty was indeed to automate the setup of protein–ligand binding sites with appropriate attributes, such as physicochemical (e.g., ionization and tautomerization states) and pharmacological properties of the ligand. It was not our intention to cover all ligand–protein complexes in the PDB, but rather to compile a large and yet not redundant set of experimental structures for known or potential therapeutic targets that had been cocrystallized with a known drug/inhibitor/activator or with a small endogenous ligand that could be replaced by a drug/inhibitor/activator (e.g., sildenafil in phosphodiesterase-5 is an adenosine mimic).
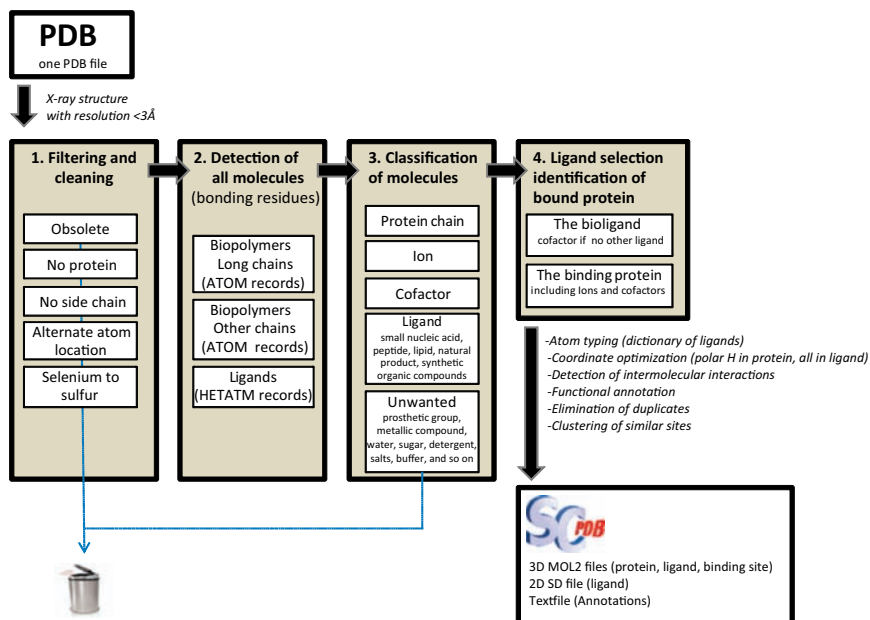
Figure 1.4 Flowchart to select sc-PDB entries from the PDB. Unwanted molecules at step 3 are identified using a dictionary or simple filters (based on ligand molecular weight, ligand surface area buried into the protein, number of amino acids close to the ligand, number of rings, and number of rotatable bonds of ligand). The bioligand in step 4 is the ligand that passes step 3 and maximizes the product of ligand molecular weight and surface area buried into the protein.

Selection rules as well as the applicability domain of the database have considerably evolved over time and are reviewed in the following sections.

1.3.1
Database Setup and Content

In brief, the selection scheme is made of simple and intelligible selection rules for the function and properties of the protein, the physicochemical properties of its ligand, and its binding mode (Figure 1.4).

The first publicly available version of the database was released in 2004 [35]. The database was named sc-PDB (acronym for *sc*reening the *P*rotein *D*ata *B*ank) (Table 1.4). At that time, it contained the atomic coordinates of proteins and their "druggable" binding sites. The protein was defined as all biopolymer chains, ions, and cofactors in the vicinity of the ligand. The binding site includes only the protein residues less than 6.5 Å away from the ligand. Noteworthy, all atoms were represented, including the hydrogen atoms not described in crystal structures. From 2005 onward, the sc-PDB has also provided the atomic coordinates of ligands. The ligand chemistry has been validated using an in-house dictionary, manually built from

Table 1.4 Annotation and available search options in the Web interface to the sc-PDB.

Object	Properties
PDB X-ray structure	PDB identifier
	Resolution
	Deposition date
Ligand	HET code
	Chemical structure
	Formula
	Molecular weight
	LogP
	LogS
	Polar surface area
	H-Bond donor count
	H-Bond acceptor count
	Number of rotatable bonds
	Number of rings
	Rule-of-five number of violations
Protein	Name
	EC number
	Uniprot accession number
	Uniprot name
	Source organism name
	Source organism taxonomy
	Source organism kingdom
	Mutant/wild type
Ligand binding site	Ion/cofactor
	Number of residues
	Number of nonstandard amino acids
	Number of chains
	Average B-factor
	Center of mass
Protein–ligand interactions	Number of hydrophobic interactions
	Aromatic face-to-face interactions
	Aromatic face-to-edge interactions
	H-Bond (donor in protein or ligand)
	Ionic interaction (cation in protein or ligand)
	Metal coordination
	Affinity data (K_i, K_d, IC_{50}, or pK_d)
	Ligand buried surface area

scratch then supplemented since 2007 by manually checked entries of the PDB Chemical Component Dictionary. The all-atoms representation of both partners of sc-PDB complexes have allowed us to refine the position of polar hydrogen atoms in the protein binding site and to compute an optimized pose of the bound ligand [29].

Figure 1.5 sc-PDB output for PDB protein–ligand complexes (3 hits) between an indole-containing ligand (blue substructure) of molecular weight <350 and a human kinase to which the ligand donates at least one hydrogen bond.

The sc-PDB is annually updated and regularly enriched with new information (ligand descriptors, binding mode encoded into an interaction fingerprint (IFP) [36], and cavity volume) and new functionalities (classification of similar binding sites [37]). A Web interface enables querying the database by combining requests about ligand chemical structures and properties, protein function and source organism, binding site properties, and ligand/protein binding properties (Figure 1.5).

The current version of the database contains 9891 entries corresponding to 3039 different proteins (according to protein sc-PDB name [37]) and 5505 different ligands (according to canonical SMILES strings). The sc-PDB protein space is redundant. There are 395 different proteins with more than 5 copies and single-copy proteins represent 55% of the database entries. Noteworthy is the complex nature of many proteins: a cofactor is bound to 219 proteins; calcium, magnesium, manganese, cobalt, zinc, or iron ions are found in 981 different proteins. No sc-PDB ligands are located at the interface of a protein–protein complex. The functional and species distribution of sc-PDB proteins reflects the bias in protein function space of the PDB itself, yet the sc-PDB is enriched in enzymes. The sc-PDB ligands space is also redundant and most prevalent ligands are cofactors and other nucleotides, which are also the most promiscuous ligands (e.g., more than 100 different protein targets for adenoside 5′-diphosphate or nicotinamide adenine dinucleotide). About 75% of the sc-PDB ligands is not primary bioorganic metabolites (nucleic acids, peptides, amino acids, sugars, or lipids) or their derivatives. Most of them pass the Lipinksi's rule of five (69%

with no violations and 20% with a single violation). The sc-PDB ligand space does not match that of commercial drugs because of a bias toward polar and flexible ligands. Finally, the sc-PDB ligand ensemble is not very diverse: for more than half of sc-PDB ligands, the ligand molecule is highly similar to at least one molecule in the pool of nonidentical ligands (with similarity evaluated by the Tanimoto coefficient, computed on feature-based circular 2D FCFP4 fingerprints, higher than 0.6).

1.3.2
Applications to Drug Design

1.3.2.1 **Protein–Ligand Docking**
The sc-PDB database has been developed for reverse docking applications [35] and is therefore an invaluable source for establishing large-scale docking benchmarks. Most validation studies, which flourished in the literature in the last decade, have been applied to a restricted set of a few hundred PDB targets [38–41] and in the best cases to a "clean" set of high-resolution protein structures in which erroneous PDB data (Table 1.2) have been removed [42]. In daily drug discovery programs, many targets under investigation do not obey such strict rules. Assessing the robustness of docking algorithms against a larger and more representative set of protein 3D structures is therefore of interest. The sc-PDB provides a unique source for such benchmarks since ligand, protein, and active site coordinates have been preprocessed and are ready for automated docking. When applied to a collection of 5681 complexes, Tietze and Apostoklasis reported with the GlamDock software [43] an accuracy (RMSD to the X-ray structure below 2.0 Å) significantly lower than that obtained with restricted protein sets with only 77% of sampling accuracy (RMSD of the best pose <2.0 Å) and 47% of scoring accuracy (RMSD of the top-ranked pose <2 Å). Along the same lines, we reported the accuracy of four docking algorithms in posing low molecular weight fragments into druggable sc-PDB binding sites and observed that ranking poses by a pure topological scoring function based on protein–ligand interaction fingerprints were much superior to poses by classical energy-based scoring functions [36].

Coming back to the seminal application for which the sc-PDB archive was initially developed (reverse docking), it appeared quite soon that the concept could be easily applied to a large and heterogeneous set of binding sites with a naïve target ranking scheme consisting of simple docking scores. Serial docking of four test ligands (biotin, methotrexate, 4-hydroxytamoxifen, and 6-hydroxy-1,6-dihydropurine ribonucleoside) to a collection of 2148 binding sites enabled recovering the known target(s) of the later ligands within the top 1% scoring entries, using the GOLD docking algorithm. These results were quite encouraging since these validated *per se* the reverse docking concept and notably the automated binding site setup protocol despite well-known insufficiencies regarding, for example, ionization/tautomerization of binding site residues as well as water-mediated ligand binding effects. These initial trials were applied to high-affinity ligands, which were relatively selective for very few targets. When applied to smaller and more permissive compounds

(e.g., AMP), a larger list of potential targets (top 5 to 10%) had to be selected to fish the correct protein targets [35]. The main reason was an inaccurate scoring of the "good" binding sites, which was not a real surprise with regard to the abundant literature about the limitations of fast scoring functions utilized in docking algorithms [19,44]. In order to overcome these severe limitations, alternative target ranking schemes independent of any energy calculation have been developed. One particular problem in docking-based target fishing is that the distribution of docking scores may be quite heterogeneous across different binding sites with diverse physicochemical properties. Therefore, score normalization according to either ligand and/or target properties is necessary to get rid of frequent target hitters [45–47]. Another promising approach consists in the conversion of protein–ligand coordinates (docking poses) into simple 1D IFPs [36]. Assuming that a virtual hit is more likely to be a true hit if it shares a similar target–ligand interaction profile with a known ligand, docking poses can be ranked by decreasing similarity of the IFP to that of the reference compound(s). Combining docking scores with IFP similarities allows removing many false positives (wrong targets with high docking scores), while still selecting the true targets in the final hit list [48].

1.3.2.2 Binding Site Detection and Comparisons

The sc-PDB provides, for each entry, all-atom Cartesian coordinates for the ligand, the target, and the binding site. By "binding site" we mean any monomer (amino acid, ion, cofactor, or prosthetic group) within 6.5 Å of any ligand heavy atom. Although the definition is conservative and excludes many potentially interesting pockets, it presents the advantage to favor cavities with well-described ligand occupancy. sc-PDB entries, therefore, can be used by cavity detection algorithms [49] to predict the most likely ligand binding sites and whether they are druggable or not, in other words, if the pocket could accommodate an orally available rule of five compliant drug-like molecule. When applied to 4915 sc-PDB protein structures, Volkamer *et al.* reported that the ligand is present in one of the three largest pockets in 90% of cases [50]. We used a grid-based cavity detection method (VolSite) to map cavity points with pharmacophoric properties of the closest protein atom, thus defining an ideal virtual ligand for each binding site (Figure 1.6).

Predicting the druggability of a given target from its three-dimensional structure is an intense field of research in order to reduce attrition rates in pharmaceutical discovery [51]. As druggability is by far more complex than the simple propensity of a particular protein cavity to accommodate high-affinity drug-like compounds, other terms, such as "bindability" [52] or "ligandability" [51] have been proposed recently, since they better capture target property ranges (cavity volume, polarity, and buriedness) known to be important for druggable targets [52–56]. Since these important properties are theoretically encoded in the aforementioned cavity site points, we investigated whether the present cavity descriptors might be suitable for predicting the ligandability of cavities from their 3D structures. A training set of 62 cavities (50% druggable and 50% undruggable) was assembled from literature [53,57] and the distribution of site point properties was given as input for a support vector machine (SVM) classifier. The best cross-validated classification model
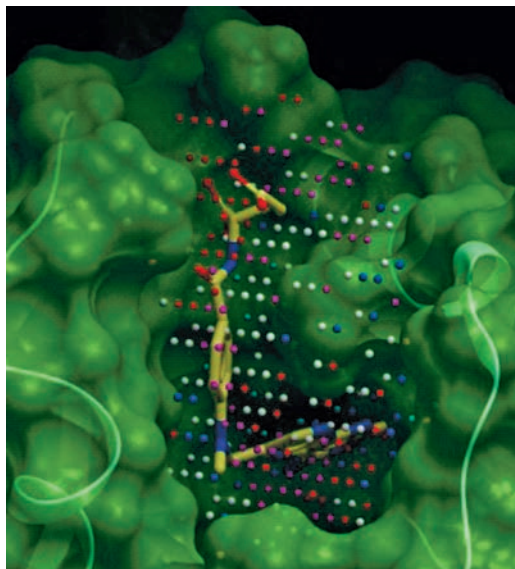
Figure 1.6 Detection and pharmacophoric annotation of VolSite cavity points in the X-ray structure of *Lactobacillus* dihydrofolate reductase (PDB code 4dfr). The cognate ligand (methotrexate, sticks) is shown in the binding site of the protein (green transparent surface). Cavity points are colored by pharmacophoric properties (H-bond acceptor and negative ionizable, red: H-bond donor and positive ionizable, blue: hydrophobe, white: aromatic, cyan: null, magenta).

achieves a very good accuracy of 80% and a Matthews correlation coefficient (MCC) of 0.62. Of course, larger sets of proteins of known (non)druggability are necessary to draw general conclusions, but the observed trend is quite promising and suggests that druggable target triage may be considered at an early level of drug discovery programs on condition that a high-resolution X-ray structure is available.

A second interesting application of the sc-PDB is the quantitative measure of its binding sites. Assuming that similar binding sites recognize similar ligands, comparing binding sites notably in the absence of 3D structure conservation permits identifying unexpected secondary targets for bioactive ligands. Several alignment-dependent or alignment-independent binding site comparison methods have been benchmarked on diverse collections of sc-PDB ligand binding sites [58–61] and have enabled the definition of global and local similarity thresholds for defining two sites as similar. Screening a library of binding sites for similarity to any given query is, therefore, possible and has already yielded the identification of an unexpected off-target (Synapsin I) for some but not all serine/threonine protein kinase inhibitors (Figure 1.7) [62].

Interestingly, only inhibitors of binding sites (cyclin-dependent kinase type 2, pim-1, and casein kinase II) predicted similar to that of Synapsin I were indeed found to bind to Synapsin I, sometimes with nanomolar affinities, whereas inhibitors of binding sites distant to that of Synapsin I (e.g., checkpoint kinase 1, protein kinase A, HSP-90α, DAG kinase, and DNA topoisomerase II) were not recognized by the enzyme [62].
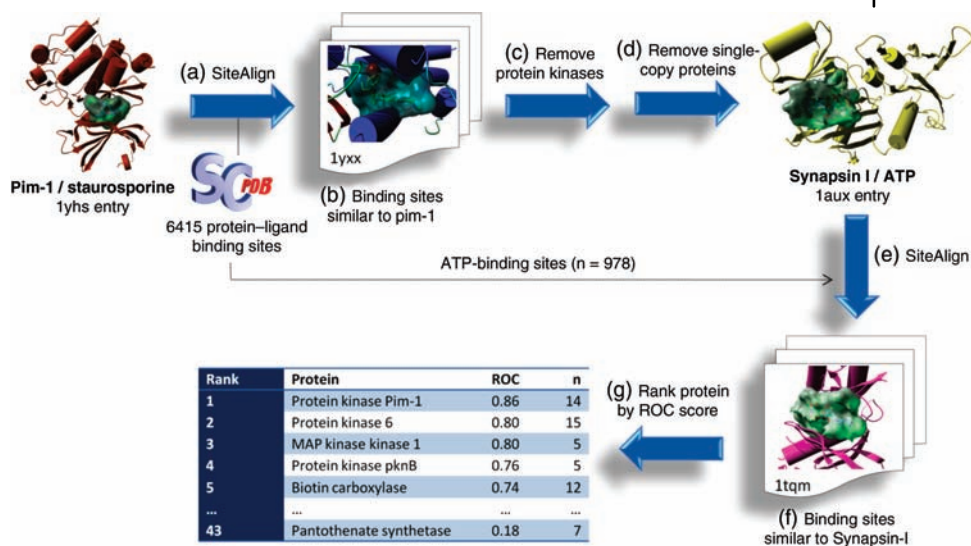
Figure 1.7 Computational protocol used to detect local similarities between ATP-binding sites in pim-1 kinase and Synapsin I. The ATP-binding site in pim-1 kinase (occupied by the ligand staurosporine) is compared with SiteAlign [58] (step a) to 6415 binding sites stored in the sc-PDB database. Among the top scoring entries (step b), Synapsin I is the only protein not belonging to the protein kinase target family (step c) and present in numerous copies (step d). A systematic SiteAlign comparison (step e) of the ATP-binding site in Synapsin I with 978 other ATP-binding sites (from the sc-PDB) suggests that some but not all ATP-binding sites of protein kinases (steps f and g) resemble that of Synapsin I [62].

1.3.2.3 Prediction of Protein Hot Spots

The structural knowledge encoded by 3500 protein–ligand complexes in the sc-PDB has been used to derive a model able to discriminate, from simple 1D cavity fingerprints, 120 000 ligands interacting from 500 000 ligand-noninteracting protein atoms [63]. When applied to a novel complex, the model was able to predict with 70% accuracy the protein atoms that are likely to interact with a ligand and, therefore, prioritize protein structure-based pharmacophore queries specifically targeting these hot spots.

1.3.2.4 Relationships between Ligands and Their Targets

The sc-PDB data set offers the opportunity to delineate evolutionary relationships between ligands and their targets or binding sites. By examining the distribution patterns of sc-PDB ligands in the protein universe, Ji *et al.* reported that synthetic compounds (e.g., enzyme inhibitors) tend to bind to a single protein fold, whereas "superligands" (metabolites) are much more permissive and can be accommodated by more than 10 different protein folds [64]. Target fold promiscuity was almost found for ancestral ligands (e.g., nucleotide-containing metabolites) that appeared quite early in the evolution and behave as hubs of metabolic networks. Interestingly,

these ligands share common physicochemical properties (high flexibility and polarity) responsible for their promiscuity. Likewise, the analysis of cofactor usage (organic molecules and transition metal ions) by primitive redox proteins in the sc-PDB clearly shows that organic cofactors (NAD and NADP) are much more used than metals, probably because of the abundance of neutral residues at the border of the corresponding binding sites [65]. Finally, a survey of known interactions between phenolic ligands and their sc-PDB targets provides some explanations for the classically observed discrepancy between potent *in vitro* and moderate *in vivo* antioxidant properties of phenols [66]. A tight hydrogen bonding of phenolic moieties to many sc-PDB proteins suggests that reactive oxidative species (ROS) cannot be scavenged by phenols if they are already engaged in interactions with surrounding proteins.

Relationships between ligands and their targets could also be integrated in rational drug discovery programs. For example, retrieving from the sc-PDB, 171 diverse protein kinases cocrystallized with ATP competitors and aligning their binding sites led to the observation that crystal water patterns (position, hydrogen bond network to the kinase, and known inhibitor) were not necessarily conserved despite very high binding site similarities, thus suggesting novel avenues for optimizing the fine selectivity of kinases inhibitors [67]. By comparing the structure of unrelated targets binding to the same natural flavonoids, Quinn and coworkers introduced the concept of protein fold topology (PFT) [68] characterized by short stretches of not necessarily conserved secondary structures providing shared anchoring points to a common ligand. The concept was demonstrated for natural products binding to both biosynthetic enzymes and therapeutic targets and may explain why natural compounds are abundant among existing drugs [69].

1.3.2.5 Chemogenomic Screening for Protein–Ligand Fingerprints

In a recent report, Meslamani and Rognan describe a novel protein cavity kernel able to quantitatively measure the 3D similarity between two sc-PDB binding sites. A novel chemogenomic screening method based on a SVM was designed to browse the sc-PDB protein–ligand space and predict binary protein–ligand interactions from separate ligand and cavity fingerprints. The best SVM model was able to predict with a high recall (70%) and exquisite specificity (99%) and precision (99%) the binding of 14 117 external ligands to a set of 531 sc-PDB targets [70].

1.4
Conclusions

Exploiting structural knowledge on known protein–ligand complexes is a key step in the rational design of bioactive compounds. This knowledge has gained considerable value in the recent years, thanks to parallel endeavors of structural biologists and computational biologists/chemists to release an ever-increasing number of high-quality data. Many smart algorithms to parse and analyze the PDB have been described in the last couple of years with a large spectrum of applications ranging

from hit identification and optimization to massive ligand profiling against a large array of possible targets. With the expected better coverage of the therapeutic target space by the PDB in the coming years, we anticipate a significant boost of rational drug discovery and notably a better interplay between protein structure-based and ligand-centric methods.

References

1 Berman, H., Henrick, K., Nakamura, H., and Markley, J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, **35**, D301–D303.

2 Dessailly, B.H., Nair, R., Jaroszewski, L., Fajardo, J.E., Kouranov, A., Lee, D., Fiser, A., Godzik, A., Rost, B., and Orengo, C. (2009) PSI-2: structural genomics to cover protein domain family space. *Structure*, **17**, 869–881.

3 Nair, R., Liu, J., Soong, T.T., Acton, T.B., Everett, J.K., Kouranov, A., Fiser, A., Godzik, A., Jaroszewski, L., Orengo, C., Montelione, G.T., and Rost, B. (2009) Structural genomics is the largest contributor of novel structural leverage. *Journal of Structural and Functional Genomics*, **10**, 181–191.

4 Chandonia, J.M. and Brenner, S.E. (2006) The impact of structural genomics: expectations and outcomes. *Science*, **311**, 347–351.

5 Brown, E.N. and Ramaswamy, S. (2007) Quality of protein crystal structures. *Acta Crystallographica Section D*, **63**, 941–950.

6 Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A.C., and Wishart, D.S. (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research*, **39**, D1035–D1041.

7 Joachimiak, A. (2009) High-throughput crystallography for structural genomics. *Current Opinion in Structural Biology*, **19**, 573–584.

8 Montelione, G.T. and Szyperski, T. (2010) Advances in protein NMR provided by the NIGMS Protein Structure Initiative: impact on drug discovery. *Current Opinion in Drug Discovery & Development*, **13**, 335–349.

9 Klaholz, B.P., Pape, T., Zavialov, A.V., Myasnikov, A.G., Orlova, E.V., Vestergaard, B., Ehrenberg, M., and van Heel, M. (2003) Structure of the *Escherichia coli* ribosomal termination complex with release factor 2. *Nature*, **421**, 90–94.

10 Dalby, A., Nourse, J.G., Hounshell, W.D., Gushurst, A.K.I., Grier, D.L., Leland, B.A., and Laufer, J. (1992) Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences*, **32**, 244–255.

11 Bourne, P.E., Berman, H.M., McMahon, B., Watenpaugh, K.D., Westbrook, J.D., and Fitzgerald, P.M.D. (1997) Macromolecular crystallographic information file. *Methods in Enzymology*, **277**, 571–590.

12 Westbrook, J., Ito, N., Nakamura, H., Henrick, K., and Berman, H.M. (2005) PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics*, **21**, 988–992.

13 Dutta, S., Burkhardt, K., Swaminathan, G.J., Kosada, T., Henrick, K., Nakamura, H., and Berman, H.M. (2008) Data deposition and annotation at the worldwide protein data bank. *Methods in Molecular Biology*, **426**, 81–101.

14 Henrick, K., Feng, Z., Bluhm, W.F., Dimitropoulos, D., Doreleijers, J.F., Dutta, S., Flippen-Anderson, J.L., Ionides, J., Kamada, C., Krissinel, E., Lawson, C.L., Markley, J.L., Nakamura, H., Newman, R., Shimizu, Y., Swaminathan, J., Velankar, S., Ory, J., Ulrich, E.L., Vranken, W., Westbrook, J., Yamashita, R., Yang, H., Young, J., Yousufuddin, M., and Berman, H.M. (2008) Remediation of the protein data bank archive. *Nucleic Acids Research*, **36**D426–D433.

15 Joosten, R.P., te Beek, T.A., Krieger, E., Hekkelman, M.L., Hooft, R.W., Schneider, R., Sander, C., and Vriend, G. (2011) A series of PDB related databases for everyday needs. *Nucleic Acids Research*, **39**, D411–D419.

16 Joosten, R.P. and Vriend, G. (2007) PDB improvement starts with data deposition. *Science*, **317**, 195–196.

17 Hartshorn, M.J., Verdonk, M.L., Chessari, G., Brewerton, S.C., Mooij, W.T.M., Mortenson, P.N., and Murray, C.W. (2007) Diverse, high-quality test set for the validation of protein–ligand docking performance. *Journal of Medicinal Chemistry*, **50**, 726–741.

18 Hawkins, P., Warren, G., Skillman, A., and Nicholls, A. (2008) How to do an evaluation: pitfalls and traps. *Journal of Computer-Aided Molecular Design*, **22**, 179–190.

19 Dunbar, J.B., Smith, R.D., Yang, C.-Y., Ung, P.M.-U., Lexa, K.W., Khazanov, N.A., Stuckey, J.A., Wang, S., and Carlson, H.A. (2011) CSAR benchmark exercise of 2010: selection of the protein–ligand complexes. *Journal of Chemical Information and Modeling*, **51**, 2036–2046.

20 Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H.M., and Westbrook, J. (2004) Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, **20**, 2153–2155.

21 Golovin, A. and Henrick, K. (2008) MSDmotif: exploring protein sites and motifs. *BMC Bioinformatics*, **9**, 312.

22 Yamaguchi, A., Iida, K., Matsui, N., Tomoda, S., Yura, K., and Go, M. (2004) Het-PDB Navi.: a database for protein–small molecule interactions. *Journal of Biochemistry*, **135**, 79–84 [Erratum: *Journal of Biochemistry (Tokyo)*, 2004, 135 (5), 651.].

23 Kleywegt, G.J. and Jones, T.A. (1998) Databases in protein crystallography. *Acta Crystallographica Section D*, **54**, 1119–1131.

24 Michalsky, E., Dunkel, M., Goede, A., and Preissner, R. (2005) SuperLigands: a database of ligand structures derived from the Protein Data Bank. *BMC Bioinformatics*, **6**, 122.

25 Benson, M.L., Smith, R.D., Khazanov, N.A., Dimcheff, B., Beaver, J., Dresslar, P., Nerothin, J., and Carlson, H.A. (2008) Binding MOAD, a high-quality protein–ligand database. *Nucleic Acids Research*, **36**, D674–D678.

26 Wang, R., Fang, X., Lu, Y., Yang, C.Y., and Wang, S. (2005) The PDBbind database: methodologies and updates. *Journal of Medicinal Chemistry*, **48**, 4111–4119.

27 Liu, T., Lin, Y., Wen, X., Jorissen, R.N., and Gilson, M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research*, **35**, D198–D201.

28 Hendlich, M., Bergner, A., Gunther, J., and Klebe, G. (2003) Relibase: design and development of a database for comprehensive analysis of protein–ligand interactions. *Journal of Molecular Biology*, **326**, 607–620.

29 Kellenberger, E., Muller, P., Schalon, C., Bret, G., Foata, N., and Rognan, D. (2006) sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *Journal of Chemical Information and Modeling*, **46**, 717–727.

30 Wallach, I. and Lilien, R. (2009) The protein–small-molecule database, a non-redundant structural resource for the analysis of protein–ligand binding. *Bioinformatics*, **25**, 615–620.

31 Rognan, D. (2012) Fragment-based approaches and computer-aided drug discovery. *Topics in Current Chemistry*, **317**, 201–222.

32 Moriaud, F., Doppelt-Azeroual, O., Martin, L., Oguievetskaia, K., Koch, K., Vorotyntsev, A., Adcock, S.A., and Delfaud, F. (2009) Computational fragment-based approach at PDB scale by protein local similarity. *Journal of Chemical Information and Modeling*, **49**, 280–294.

33 Wang, L., Xie, Z., Wipf, P., and Xie, X.-Q. (2011) Residue preference mapping of ligand fragments in the Protein Data Bank. *Journal of Chemical Information and Modeling*, **51**, 807–815.

34 Wood, D.J., de Vlieg, J., Wagener, M., and Ritschel, T. (2012) Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. *Journal of Chemical Information and Modeling*, **52**, 2031–2043.

35 Paul, N., Kellenberger, E., Bret, G., Muller, P., and Rognan, D. (2004) Recovering the true targets of specific ligands by virtual screening of the Protein Data Bank. *Proteins: Structure, Function, and Bioinformatics*, **54**, 671–680.

36 Marcou, G. and Rognan, D. (2007) Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *Journal of Chemical Information and Modeling*, **47**, 195–207.

37 Meslamani, J., Rognan, D., and Kellenberger, E. (2011) sc-PDB: a database for identifying variations and multiplicity of "druggable" binding sites in proteins. *Bioinformatics*, **27**, 1324–1326.

38 Verdonk, M.L., Berdini, V., Hartshorn, M.J., Mooij, W.T., Murray, C.W., Taylor, R. D., and Watson, P. (2004) Virtual screening using protein–ligand docking: avoiding artificial enrichment. *Journal of Chemical Information and Computer Sciences*, **44**, 793–806.

39 Kellenberger, E., Rodrigo, J., Muller, P., and Rognan, D. (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins*, **57**, 225–242.

40 Kontoyianni, M., McClellan, L.M., and Sokol, G.S. (2004) Evaluation of docking performance: comparative data on docking algorithms. *Journal of Medicinal Chemistry*, **47**, 558–565.

41 Perola, E., Walters, W.P., and Charifson, P.S. (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins*, **56**, 235–249.

42 Hartshorn, M.J., Verdonk, M.L., Chessari, G., Brewerton, S.C., Mooij, W.T., Mortenson, P.N., and Murray, C.W. (2007) Diverse, high-quality test set for the validation of protein–ligand docking performance. *Journal of Medicinal Chemistry*, **50**, 726–741.

43 Tietze, S. and Apostolakis, J. (2007) GlamDock: development and validation of a new docking tool on several thousand protein–ligand complexes. *Journal of Chemical Information and Modeling*, **47**, 1657–1672.

44 Ferrara, P., Gohlke, H., Price, D.J., Klebe, G., and Brooks, C.L., 3rd (2004) Assessing scoring functions for protein–ligand interactions. *Journal of Medicinal Chemistry*, **47**, 3032–3047.

45 Yang, L., Wang, K., Chen, J., Jegga, A.G., Luo, H., Shi, L., Wan, C., Guo, X., Qin, S., He, G., Feng, G., and He, L. (2011) Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome: clozapine-induced agranulocytosis as a case study. *PLoS Computational Biology*, **7**, e1002016.

46 Yang, L., Chen, J., and He, L. (2009) Harvesting candidate genes responsible for serious adverse drug reactions from a chemical-protein interactome. *PLoS Computational Biology*, **5**, e1000441.

47 Vigers, G.P. and Rizzi, J.P. (2004) Multiple active site corrections for docking and virtual screening. *Journal of Medicinal Chemistry*, **47**, 80–89.

48 Kellenberger, E., Foata, N., and Rognan, D. (2008) Ranking targets in structure-based virtual screening of 3-D protein libraries: methods and problems. *Journal of Chemical Information and Modeling*, **48**, 1014–1025.

49 Perot, S., Sperandio, O., Miteva, M.A., Camproux, A.C., and Villoutreix, B.O. (2010) Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today*, **15**, 656–667.

50 Volkamer, A., Griewel, A., Grombacher, T., and Rarey, M. (2010) Analyzing the topology of active sites: on the prediction of pockets and subpockets. *Journal of Chemical Information and Modeling*, **50**, 2041–2052.

51 Edfeldt, F.N., Folmer, R.H., and Breeze, A. L. (2011) Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discovery Today*, **16**, 284–287.

52 Sheridan, R.P., Maiorov, V.N., Holloway, M.K., Cornell, W.D., and Gao, Y.D. (2010) Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *Journal of Chemical Information and Modeling*, **50**, 2029–2040.

53 Cheng, A.C., Coleman, R.G., Smyth, K.T., Cao, Q., Soulard, P., Caffrey, D.R., Salzberg, A.C., and Huang, E.S. (2007)

Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnology*, **25**, 71–75.

54 Hajduk, P.J., Huth, J.R., and Fesik, S.W. (2005) Druggability indices for protein targets derived from NMR-based screening data. *Journal of Medicinal Chemistry*, **48**, 2518–2525.

55 Halgren, T.A. (2009) Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling*, **49**, 377–389.

56 Schmidtke, P. and Barril, X. (2010) Understanding and predicting druggability: a high-throughput method for detection of drug binding sites. *Journal of Medicinal Chemistry*, **53**, 5858–5867.

57 Huang, N. and Jacobson, M.P. (2010) Binding-site assessment by virtual fragment screening. *PLoS One*, **5**, e10109.

58 Schalon, C., Surgand, J.S., Kellenberger, E., and Rognan, D. (2008) A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins*, **71**, 1755–1778.

59 Weill, N. and Rognan, D. (2010) Alignment-free ultra-high-throughput comparison of druggable protein–ligand binding sites. *Journal of Chemical Information and Modeling*, **50**, 123–135.

60 Totrov, M. (2011) Ligand binding site superposition and comparison based on Atomic Property Fields: identification of distant homologues, convergent evolution and PDB-wide clustering of binding sites. *BMC Bioinformatics*, **12** (Suppl. 1), S35.

61 Kasahara, K., Kinoshita, K., and Takagi, T. (2010) Ligand-binding site prediction of proteins based on known fragment–fragment interactions. *Bioinformatics*, **26**, 1493–1499.

62 Defranchi, E., Schalon, C., Messa, M., Onofri, F., Benfenati, F., and Rognan, D. (2010) Binding of protein kinase inhibitors to Synapsin I inferred from pair-wise binding site similarity measurements. *PLoS One*, **5**, e12214.

63 Barillari, C., Marcou, G., and Rognan, D. (2008) Hot-spots-guided receptor-based pharmacophores (HS-Pharm): a knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores. *Journal of Chemical Information and Modeling*, **48**, 1396–1410.

64 Ji, H.F., Kong, D.X., Shen, L., Chen, L.L., Ma, B.G., and Zhang, H.Y. (2007) Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. *Genome Biology*, **8**, R176.

65 Ji, H.F., Chen, L., and Zhang, H.Y. (2008) Organic cofactors participated more frequently than transition metals in redox reactions of primitive proteins. *Bioessays*, **30**, 766–771.

66 Shen, L., Ji, H.F., and Zhang, H.Y. (2007) How to understand the dichotomy of antioxidants. *Biochemical and Biophysical Research Communications*, **362**, 543–545.

67 Barillari, C., Duncan, A.L., Westwood, I. M., Blagg, J., and van Montfort, R.L.M. (2011) Analysis of water patterns in protein kinase binding sites. *Proteins: Structure, Function, and Bioinformatics*, **79**, 2109–2121.

68 McArdle, B.M., Campitelli, M.R., and Quinn, R.J. (2006) A common protein fold topology shared by flavonoid biosynthetic enzymes and therapeutic targets. *Journal of Natural Products*, **69**, 14–17.

69 Kellenberger, E., Hofmann, A., and Quinn, R.J. (2011) Similar interactions of natural products with biosynthetic enzymes and therapeutic targets could explain why nature produces such a large proportion of existing drugs. *Natural Product Reports*, **28**, 1483–1492.

70 Meslamani, J. and Rognan, D. (2011) Enhancing the accuracy of chemogenomic models with a three-dimensional binding site kernel. *Journal of Chemical Information and Modeling*, **51**, 1593–1603.

2
Public Domain Databases for Medicinal Chemistry

George Nicola, Tiqing Liu, and Michael Gilson

2.1
Introduction

Medicinal chemists today find themselves in an increasingly information-rich environment. An abundance of compound activity and affinity data is being published, and medicinal chemistry data are increasingly connected with a broader world of data from the realms of bioinformatics and systems biology. In recent years, a number of publicly accessible, chemistry-oriented databases of interest to medicinal chemists have been established to facilitate access to medicinal chemistry data and their biological links, with the aim of accelerating the discovery of new medications. In order to maximize their usefulness, it is important that researchers in pertinent fields be fully aware of these resources and exploit their full potential.

Decades of growth worldwide in the pharmaceutical industry and of academic drug discovery efforts, along with technological advances that speed up compound synthesis and assays [1], and the advent and growth of the related fields of chemical biology and chemical genomics have led to an ongoing flood of publications with valuable data regarding new compounds and their biological activities. About 20 000–30 000 new compounds are now published per year in some of the main medicinal chemistry journals, and this rate has accelerated in recent years (as detailed later). However, publication in conventional journals traps data in a form where they are inaccessible to computer search and retrieval. For example, it is not possible to search standard scientific articles for compounds of interest or to reliably extract machine-readable representations of compounds from chemical drawings in articles. As a consequence, the conventional publishing paradigm can severely restrict the discoverability and usability of medicinal chemistry data.

The parallel growth of information technology and the emergence of the World Wide Web in the 1990s have created important new opportunities for dissemination of data. Biologists – especially structural and molecular biologists – seized these opportunities, establishing central data resources such as the Protein Data Bank (PDB) [2] and GenBank [3] and laying the foundations for the field of bioinformatics. The first public protein-ligand database aimed at serving the drug discovery

community, BindingDB, came on line in late 2000. This resource has grown substantially and has since been joined by other important databases with related scope and goals. According to Pathguide, a Web resource for online databases, at least 43 protein–compound interaction databases [4,5] and many other useful, yet free, chemical databases are now available [6]. Such resources are of increasing value not only for basic uses like finding and downloading structure–activity relationship (SAR) data for a protein target of interest but also for emergent applications that become possible as the medicinal chemistry data set grows to provide a comprehensive picture of small molecules in the larger biological context. For example, if a cell-based screen reveals that a new compound inhibits apoptosis, then one might seek similar compounds that bind apoptosis-related proteins and thus hypothesize that the new compound also binds one of these targets. Similarly, if one is prioritizing several lead compounds for further development, the observation that one lead is similar to a published compound known to bind a different target might lead one to reduce its priority to minimize off-target effects. In another scenario, marking all the proteins in a defined signaling pathway according to which ones are already targeted by FDA-approved drugs might lead to suggestions for a multidrug therapy to maximally suppress signaling.

Here, we aim first to help medicinal chemists take advantage of the growing array of freely accessible medicinal chemistry-oriented databases by discussing three central resources focused on small molecule binding and bioactivity, BindingDB, ChEMBL, and PubChem, and noting as well several other small molecule databases that are also of great value. (Readers interested in additional perspectives will enjoy other recent reviews [7–12].) In particular, Section 2.2 seeks to help users over the initial barriers encountered when one starts to use these rather complex resources by summarizing information on their organization and methods of accessing key types of data, information that is not always easy to glean from their respective Web sites. Subsequent sections then offer broader discussions of the field, and some readers may wish to jump directly to Section 2.3 that uses the available medicinal chemistry data to derive interesting overviews of the available medicinal chemistry data or to Section 2.4 that offers views on the future of online compound databases and their applications, including the possibility of integrating related databases to minimize overlapping efforts, addressing the challenge of getting data into databases where they can be most useful, and the role of medicinal chemistry databases in systems biology and systems pharmacology.

2.2
Databases of Small Molecule Binding and Bioactivity

This section is intended to help medicinal chemists understand and start using BindingDB, ChEMBL, and/or PubChem. It provides an overview of what each database contains and how the information is structured, since this is important for effective use, explains how to perform basic tasks, and notes special capabilities. We envision the new user accessing these Web sites with the present chapter as a guide.

This section also includes thumbnails of a number of other medicinal chemistry-related databases that readers are likely to find useful.

2.2.1
BindingDB

2.2.1.1 History, Focus, and Content

BindingDB (www.BindingDB.org) began in the late 1990s at the University of Maryland as apparently the first publicly accessible affinity database. Since its inception, BindingDB has collected primarily protein–small molecule binding affinity data. In particular, BindingDB focuses on quantitative data, such as Ki, Kd, and IC_{50} measurements where there is a well-defined protein target. As of April 2012, BindingDB contains 793 068 binding data from 5583 protein targets and 349 917 small molecules. These holdings include about 60 000 data that have been manually extracted from journals by curators at BindingDB, including some sets that have been submitted by authors. The entries collected by BindingDB curators directly from the literature contain a particularly high level of details on assay conditions such as pH, temperature, and buffer composition. A large fraction of the data in BindingDB are merged in from other open databases listed below, notably ChEMBL [13,14] and PubChem [12,15–17], as well as PDSP Ki [18,19]. In each case, BindingDB carries out additional processing to ensure that all imported data meet current BindingDB criteria. For example, BindingDB imports only those measurement data from ChEMBL that include a well-defined protein target (TARGET_TYPE = 'PROTEIN'). For PubChem, BindingDB imports only quantitative affinity data (i.e., Confirmatory Assays – described later). In the case of PDSP, it is sometimes necessary to supplement the existing data, such as with a manually curated protein sequence or a machine-readable representation of the ligand. It is worth noting that few, if any, public database projects have the internal resources to systematically check all incoming data for possible errors. BindingDB therefore sends emails to authors inviting them to check their own data as presented on the BindingDB Web site and report any errors for correction. Indeed, readers of this chapter are also invited to find their data in BindingDB at the Author page www.bindingdb.org/bind/ByAuthor.jsp and to send in any corrections that may be needed.

2.2.1.2 Browsing, Querying, and Downloading Capabilities

BindingDB offers a range of methods to find and access data; some of the most broadly useful ones are described in video tutorials available through the BindingDB home page. One of the simplest ways to find data in BindingDB is to type any text of interest into the Full Search box at the top of the home page. This generates a powerful Google-type search for related data on compound names, protein names, article titles, assay descriptions, and author names. Wild cards are allowed here; for example, adeny* yields hits to any word starting with "adeny." Following the links to data in the resulting hit list leads to a comprehensive Results Table (bit.ly/ws4vLt) [20], where each row contains one target–ligand pair along with a rich set of links to further data on the target, the ligand, and the target–ligand combination (described

later), as well as connections to further details, compound availability, and information on the origins of the data. The links on the left-hand side of the main Web page provide more specific access to data, according to targets, compounds, citations, and protein sequence and structure. Highlights of these capabilities are as follows.

Targets: The Name link under Targets provides an alphabetical list of protein targets with direct links to data in the Results Table and to Articles. The Target list makes it easy to download an SDfile with all the compounds and affinity data for any protein target, with either 2D or computed 3D coordinates. One can, moreover, search by target name in conjunction with various conditions, such as IC_{50} range (bit.ly/AyOWyq) [21], molecular weight, etc. (bit.ly/AAUiVz) [22]. Finally, BLAST sequence search [23] can be used to find data for targets of interest (bit.ly/xuN2IY) [24].

Compounds: Users may draw a compound or paste it in a SMILES string with the ChemAxon plug-in and then search for data in BindingDB by compound, substructure, and chemical similarity. These searches may include filters by affinity range, molecular weight, target name, etc. (bit.ly/zL842y) [25]. One may query BindingDB with multiple compounds simultaneously, via the batch search page (bit.ly/w0A1G5) [26]. BindingDB also provides access to binding data based on the names of 3431 FDA-approved drugs, through cross-referencing of the Drugs@FDA database [27]. For example, BindingDB has about 60 measurements for nifedipine, the active ingredient of the calcium channel blocker Adalat (bit.ly/wXIziD) [28].

Citations: BindingDB allows users to view all the data associated with a particular author (bit.ly/wHLXDl) [29], article (bit.ly/ydpChT) [30], or institution (bit.ly/yMYvx2) [31]. In addition, the pull-down menus on the Journal/Citation page provide immediate links to SDfiles with all the compounds and affinity data for each available article. Users of Web-based reference managers may directly import citations for data of interest into BindingDB Web pages. As detailed on the BindingDB home page, Zotero uses a Firefox extension, Cite-U-Like uses a Bookmarklet browser plug-in, and Mendeley uses a Web Importer plug-in.

Protein structure: The PDB [2] contains three-dimensional structures of a number of protein–ligand complexes for which binding data are available in BindingDB, and one may search BindingDB by PDB ID or HET ID, allowing matches for either 85 or 100% BLAST [23] sequence identity (bit.ly/zVd6z2 [32] and bit.ly/x9f9Yd [33], respectively).

Users may download the entire BindingDB database as an Oracle data dump or as an SD file that includes not only the compounds but also the activity data, such as targets and affinities (bindingdb.org/bind/chemsearch/marvin/SDFdownload.jsp). Also available on this download page are proteins in FASTA format and other specialized data sets, and opportunities are provided on various Web pages within the site to download subsets of data, such as all data for a given target protein, or all data from a given article, again in the form of data-rich SDfiles. These can be imported directly into chemical viewers and spreadsheets. The data are provided under the nonrestrictive Creative Commons Attribution-ShareAlike 3.0 Unported license [34].

2.2.1.3 Linking with Other Databases

The BindingDB Results Table provides an array of links to further information about each binding measurement and the molecules involved. In each row, links for the Target, the Ligand, or the Target and Ligand together are presented in separate columns (e.g., bit.ly/zq9oW3 [35]) (see Figure 2.1). For example, all proteins for which structural information is available are linked to the appropriate entries in the PDB. The biological role of each Target may be explored by following links from Targets to corresponding pathways in systems biology databases, including Reactome [36], KEGG [37], and NCI's Pathway Interaction Database [38]; the broader concept of linking compound databases with systems biology to support systems pharmacology is discussed later. Links are also provided from BindingDB data to the corresponding articles in PubMed, as well as to the related databases, including many of those discussed in this chapter. Finally, in 2011, BindingDB began providing links from ligands to matching compounds in the ZINC database of commercially available compounds, zinc.docking.org [39], in order to help users obtain physical samples of compounds for further experimental study.

In addition, a number of databases provide links from their data to relevant data in BindingDB. For example, PDB users will find links to BindingDB from structure
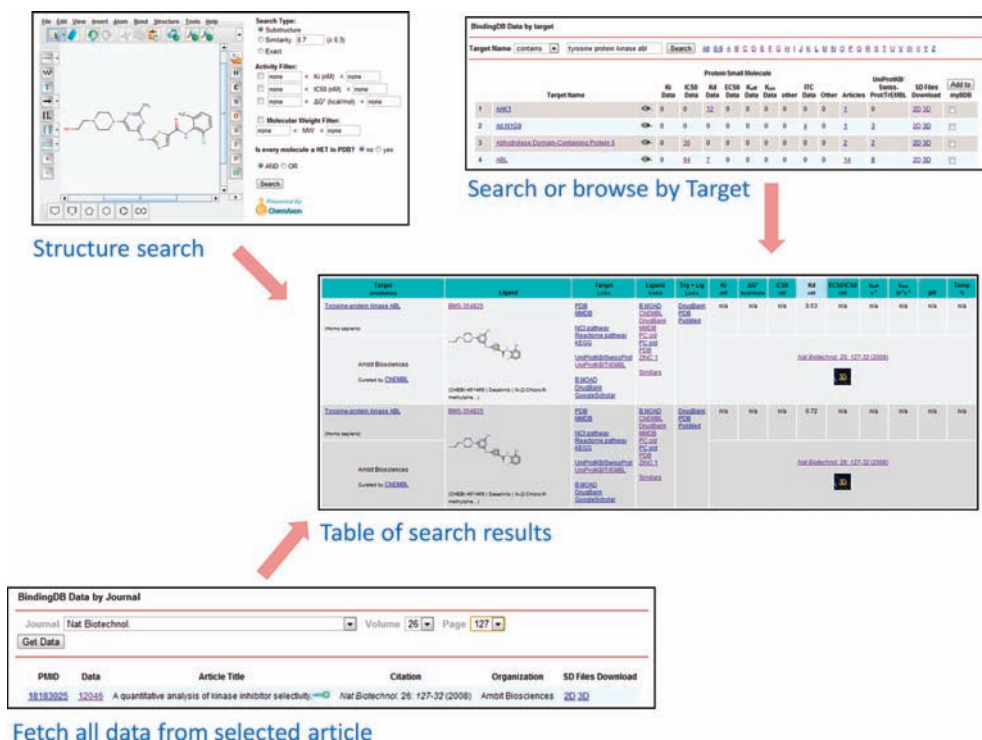
Figure 2.1 Collage of selected tools for finding data in BindingDB, along with sample search results. See text for further details.

entries for which affinity data are available, such as PDB entry 2GQG. Similarly, one may navigate from articles in PubMed to the corresponding data in BindingDB, for viewing and downloading, by expanding PubMed's LinkOut options and following the one to BindingDB, such as on the following page: www.ncbi.nlm.nih.gov/pubmed/17718712.

2.2.1.4 Special Tools and Data Sets

BindingDB also provides a number of Web-based tools and data subsets to help users take advantage of this large data collection. For example, the Find my Compound's Target page (bit.ly/zX0SfQ) [40] allows one to identify possible targets of new compounds. One draws a compound or uploads a file with multiple compounds, and BindingDB reports all protein targets known to bind similar compounds. This capability can be used to predict off-target binding, and hence side effects, of a new compound. It can also be used to generate hypotheses regarding the mechanistic targets of compounds found to be active in an empirical bioassay, such as a cell-based screen.

BindingDB also provides several online virtual screening methods allowing one to select a group of compounds in BindingDB that are known to be active against a given target protein and use them as a basis for discovering other potential actives in an uploaded compound library. The simplest and faster method, Maximum Similarity, ranks the uploaded compounds according to their maximum similarity to any of the known actives. This method uses Tanimoto similarity based upon JChem [41,42] fingerprints. A second method, binary kernel discrimination (BKD), uses a training set of compounds to produce a model that can then be applied to the structures of other compounds in order to predict their likely activity [43]. Here, the actives are divided into reference and training sets of equal size. Each set is then supplemented with 500 other drug-like compounds presumed to be inactive, and JChem binary fingerprints are computed for all compounds. The BKD comparison is used to rank the test-set compounds based on the reference set and the enrichment of actives at the top of the ranked list is reported in order to provide the user with information on the predictivity of the BKD model. If the user wishes to proceed, based on these results, then the reference and training sets are combined into one large reference set and used to rank a large set of compounds uploaded by the user. A third method [44] uses the Support Vector Machine (SVM) machine-learning approach [45]. This divides the first 100 actives into training and test sets, and again supplements these with 500 other compounds presumed to be inactive. Here, however, numerical descriptors, rather than binary fingerprints, are computed for each compound. The training set is used to set up an SVM model that will discriminate actives from inactives, and this model is evaluated with the test set. The results are reported, and if the user wishes to proceed, then descriptors are computed for the user's uploaded compounds and the SVM model is applied to rank them. It is worth noting that each of these methods has both strengths and weaknesses, and users are free to download the data from BindingDB and apply their own approaches.

In order to support the parameterization and validation of algorithms for computer-aided ligand discovery, BindingDB provides a series of validation sets manually curated from the larger data collection (bit.ly/yTctqN) [46]. Each validation

set comprises a series of congeneric compounds with measured affinities for one protein target, where the crystal structure of at least one compound in the series has been solved in a complex with the target. To support more basic studies of molecular recognition, BindingDB also houses a small collection of affinity data for small, nonprotein receptors and their ligands (bindingdb.org/bind/HostGuest.jsp).

Finally, BindingDB has also begun an initial implementation of a personalization aspect of the database, named myBDB (bindingdb.org/mybdb/login.jsp). This feature allows registered users to save searches for subsequent visits to the resource.

2.2.2
ChEMBL

2.2.2.1 History, Focus, and Content
The ChEMBL database (http://www.ebi.ac.uk/chembl/) began as a set of commercial products known as StARlite, CandiStore, and DrugStore (chembl.blogspot.com/) [14]. With funding from The Wellcome Trust, these were essentially moved to the public domain (described later) under the aegis of the European Bioinformatics Institute, an outpost of the European Molecular Biology Laboratory near Cambridge in the United Kingdom. ChEMBL's outsourced curation effort captures a broad range of medicinal chemistry data from the scientific literature. These include biological activities, such as cell-based assay data and protein–ligand affinities, although ChEMBL's curation of binding data does not include details like buffer composition and experimental conditions. About 40% of ChEMBL data are imported from PubChem, and the database also includes several large screening data sets (described later). As of April 2012, the ChEMBL database contains about 7 million measurements for 1.1 million compounds and 8900 protein targets.

2.2.2.2 Browsing, Querying, and Downloading Capabilities
A search bar on the ChEMBL home page (www.ebi.ac.uk/chembl/) provides direct access to searches by name and certain database IDs for Compounds, Targets, or Assays. Here, a Target may be not only a protein but also an entire organism, such as the yeast *Candida albicans* in the case of an antifungal bioassay. A series of tabs along the top of the home page provide access to a range of more detailed search and browsing options, organized primarily by Compounds and Targets. Highlights of these capabilities are as follows:

Targets: The Protein Target Search tab allows sequence-based searches with BLAST. These yield a table of Targets with their BLAST scores, with links to the UniProtKB protein database and further information in ChEMBL. The Browse Targets tab enables intuitive browsing of protein targets through a hierarchy of protein types (e.g., enzymes and ion channels) or a taxonomy of organisms, where, again, a Target may be an organism or a protein from an organism. The results of a Target search are presented in a table with UniProtKB IDs, gene names, and information on how many compounds and activity data are associated with each Target. A pull-down menu at the top right of the table allows

one to access the bioactivity data, optionally filtered according to parameters such as IC_{50} range. Alternatively, one may click on the name of a Target of interest in order to view a richly informative Target Report Card, as described later.

Compounds: The Compound Search tab allows one to draw a compound or fragment with a choice of JME [47], Marvin [48], or JDraw [49] sketcher and search ChEMBL by identity, similarity, or substructure. Alternatively, one may search ChEMBL for a list of compounds by pasting multiple SMILES [50] strings into a text window. Any of these searches leads to a compound table, where clicking on a compound leads to an informative Compound Report Card (described later). The compound table is also equipped with a pull-down menu allowing all or selected compounds to be downloaded as an SDfile containing the molecular structures or as a table of compound IDs with various computed descriptors, such as molecular weight and computed logP estimates. The pull-down menu also provides access to the bioactivity data for the selected compounds, as already described for Targets. An appealing alternative to the compound table display is provided in the form of scatter plots of computed compound properties, with color-coded data points linked back to compound data. An additional Browse Drugs tab on the ChEMBL home page focuses on the subset of ChEMBL compounds that are marketed drugs and provides commercial and pharmaceutical information such as a compound's approved drug name and its route of administration.

The Report Card format is a distinctive feature of the ChEMBL site (Figure 2.2). Thus, clicking on a Compound in a search result table leads to a Compound Report Card, which provides a range of additional information such as names and database identifiers, links to clinical trial information, computed properties, and links to the same compound in other databases such as DrugBank, PubChem, and the Protein Data Bank in Europe, PDBe [51]. Importantly, a set of pie charts and associated links at the bottom of the Compound Report Card provide direct access to bioactivity and other data for this compound in ChEMBL. Similarly, clicking on a Target in a ChEMBL result table leads to a Target Report Card, which contains not only further Target identifiers and links but also histograms of molecular weight, AlogP, and polar surface area for the compounds tested against this Target. One may navigate to a result table for all compounds tested against this Target, or else choose the range of a compound parameter by clicking on histogram bars and then generating a table of results for only compounds within this range. Analogous Assay Report Cards and Document Report Cards provide details of assay techniques and the documents from which ChEMBL data are drawn.

Each release of ChEMBL is freely available from an FTP server in a variety of formats, including Oracle 9i, 10g, 11g; MySQL; an SD file of compound structures; and a FASTA file of the target sequences. The data are provided under the nonrestrictive Creative Commons Attribution-ShareAlike 3.0 Unported license.

2.2.2.3 Linking with Other Databases
ChEMBL data are cross-linked with a number of other molecular databases, primarily through the various ChEMBL Report Cards (already described). For

Figure 2.2 Sample of a ChEMBL Target Report Card (a), Compound Report Card (b), and Document Report Card (c). Only the top portion of each card is shown, due to space limitations. See text for further details.

example, Target Proteins are linked to three-dimensional structure data in PDBe and sequence data and annotations in Ensembl [52] and UniProtKB [53]; Compounds are linked to ChemSpider [11,54], DrugBank [3,55], PDBe, PubChem, Wikipedia, and ChEBI [56], EBI's compound database. Articles described in Document Report Cards are linked primarily to EBI's publicly accessible journal database CiteXplore. In turn, CiteXplore's listing of each medicinal chemistry article includes a list of compounds in the article, each with a link to a ChEMBL Compound Report Card. Similarly, compounds in protein crystal structures are linked from PDBe to Compound Report Cards in ChEMBL.

2.2.2.4 Special Tools and Data Sets

ChEMBL is attuned to applications in drug discovery and pharmaceutics. For example, as already noted, it provides tabular and graphical displays of a variety of computed compound properties relevant for drug design. Another unique tool is the DrugEBIlity service, which uses structural data to evaluate whether a protein can be targeted with small molecules (www.ebi.ac.uk/chembl/drugebility/structure). One may upload a PDB format structure file, choose an existing PDB ID, or use BLAST to find similar proteins of known structure as a basis for this evaluation. In addition to druggability ratings, the server also provides a graphical display of the

protein's potential binding sites. ChEMBL also includes a Drug Approvals tab with information on new FDA drug approvals 2009–2011 ("Orange Book" data), and Compound Report Cards include links to clinical trials data (clinicaltrials.gov), when available. Finally, the Kinase SARfari and GPCR SARfari tools provide alternative access portals to Target, Compound, and activity data for two key families of therapeutic targets that are well represented in the ChEMBL database.

ChEMBL hosts a series of special data sets related to tropical pathogens in its ChEMBL-NTD (neglected tropical diseases) pages (www.ebi.ac.uk/chemblntd/). The data sets, which comprise thousands of compounds, are the result of compound screening campaigns, typically against whole *Plasmodium* and *Trypanosoma* organisms, from GlaxoSmithKline (GSK) [57], Novartis-GNF [58], St. Jude Children's Research Hospital [59], and Drugs for Neglected Diseases Initiative (www.dndi.org/).

2.2.3
PubChem

2.2.3.1 History, Focus, and Content
The PubChem database (http://pubchem.ncbi.nlm.nih.gov/) [16,17,60] is a U.S. government initiative started in 2004 by the National Institutes of Health within the National Center for Biotechnology Information (NCBI). Its broad goal is to collect and disseminate information on the biological activities of small molecules. PubChem focused initially on assay data from the high-throughput compound screening programs supported by NIH's Molecular Libraries Roadmap Initiative. However, it also accepts chemical structures and assay data from other sources, and such depositions have substantially expanded PubChem's data collection. For example, although the PubChem initiative does not include the extraction of activity data from journal articles, PubChem's incorporation of the BindingDB and ChEMBL data sets allows it to provide access to a large body of literature data. PubChem currently houses about 33 million distinct chemical entries (pubchem.ncbi.nlm.nih.gov/help.html#faq), activity data drawn from about 4800 NIH Molecular Libraries assays, 45 000 journal articles, and several hundred other sources, such as pharmaceutical companies and individual research groups.

In order to make effective use of PubChem, it is helpful for one to have a basic knowledge of its conceptual framework. First, the information in PubChem is organized into Compounds, Substances, and BioAssays. A given chemical can be listed as both a Compound and a Substance, where the Compound listing is its single standard representation, while the Substance listing corresponds to the specific material used in a given BioAssay. Thus, a given Compound can correspond to multiple Substances, and there are about three times as many Substances as Compounds in PubChem. It is the Compound listings that will generally be most meaningful to PubChem users. It is also worth noting that PubChem includes many Compounds for which there are no BioAssay data. All activity and binding data, including those drawn from the literature, are represented in terms of BioAssays. There are three types of BioAssay record: Summary, Primary, and Confirmatory. A Summary record contains an overview of a given experiment. A Primary record

contains results of a primary screen in which each compound is listed simply as Active or Inactive at a given concentration (e.g., 10 μM). A Confirmatory record reports the effective concentrations (e.g., IC_{50} values) of compounds found to be Active in a Primary screen, based on a multiconcentration dose–response study. For BioAssays with well-defined protein targets, target information is provided through seamless links to the NCBI protein database.

2.2.3.2 Browsing, Querying, and Downloading Capabilities

The PubChem home page provides immediate access to text-based searches within BioAssay, Compound, and Substance listings. (As already noted, the Compound listings are in general more useful than Substances.) One may also click through to a comprehensive tool for chemical structure searches (described later). An Advanced Search link on the home page allows more fine-tuned searching capabilities of each category (Figure 2.3). Clicking on the BioActivity Analysis or Bioactivity Summary links leads to a particularly useful BioActivity Services page (pubchem.ncbi.nlm.nih. gov/assay/), offering Compound-centric, Target-centric, and Assay-centric query tools. Several useful paths into this rich data set are now described.

Structure search, with filters

Structure-activity matrix
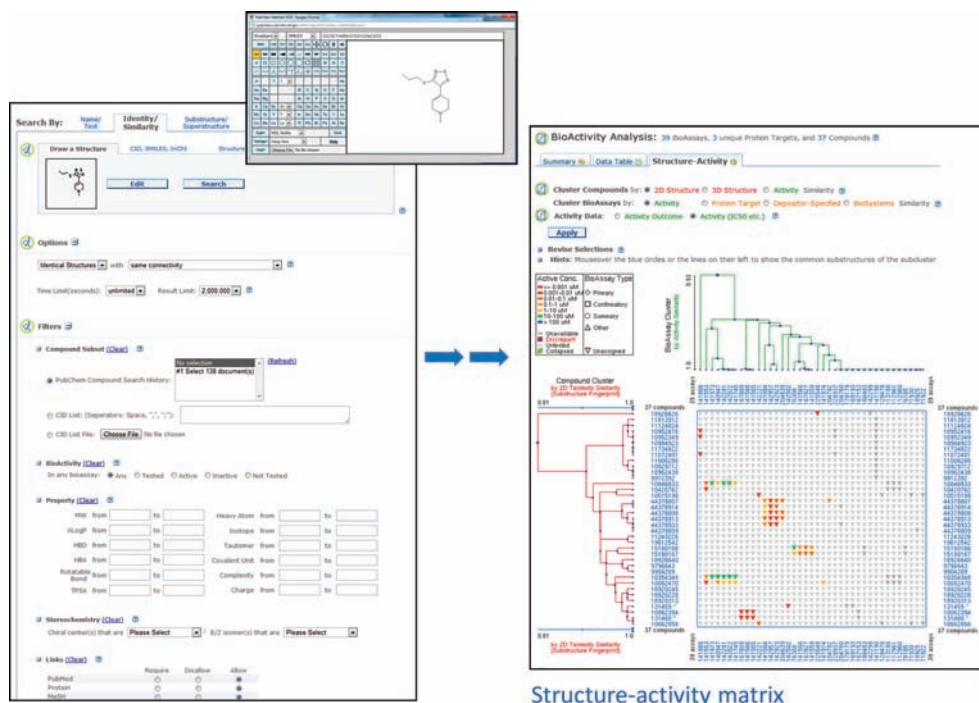
Figure 2.3 Tools for searching PubChem by chemical structure along with a variety of filters (left), and an interactive structure–activity relationship matrix, with Compounds listed along the left and BioAssays along the top.

Target: There are at least three ways of accessing compound and activity data for a given protein target. One is to choose the Target-centric option on the BioActivity Services page and enter the name of one's protein of interest, such as Rin1, into the Search by protein family text box. This search leads to a list of BioAssays for this protein target with an array of information, including the number of Active compounds by various criteria. Clicking on these numbers leads to data tables showing compounds and their activities, along with tools for downloading the data in various formats, such as comma-separated value (CSV) with a database ID for each compound. The compound activity table also provides tools for plotting the data and gaining an overview of the compounds and their activities through clustering and dendrogram displays.

A second way of accessing data for a given protein target is to enter through the NCBI Protein database. For example, one may search the NCBI Protein database for Rin1. This leads to a list of hits, where one may check the box for the *Homo sapiens* variant and then, on the right, choose PubChem BioAssay from the Find Related Data pull-down menu on the right. This reveals a further Option pull-down menu, where one may choose Bioassay by Target (identical proteins). This in turn leads to a list of BioAssays involving Rin1 from which one may choose the Confirmatory BioAssay and thus access a BioAssay Summary page for this quantitative data set. Clicking on either Show Data Active or All leads again to a table of compounds and their activities, as already described.

Compound: Chemically oriented compound searches are available at the Structure Search page (pubchem.ncbi.nlm.nih.gov/search/), which allows queries by chemical identity, similarity, and substructure; molecular formula; and three-dimensional structure similarity. Tools are also provided to filter Compounds by many criteria, such as computed chemical properties and depositor. Results are displayed as a list of compounds with a range of navigation options. Simply clicking on a compound leads to a Compound Summary page, described later. Alternatively, one may search more directly for BioAssay data for compounds of interest via the BioActivity Services page. From here, hits go directly to a data table of compounds and activities (see above).

Bioassay: PubChem contains a wealth of information on high-throughput assay methods for various protein targets and bioactivities. For example, if one wishes to learn about assays for protein Rin1, one may type this protein name into the first text box under the Assay-centric tab on the BioActivity Services page. This leads directly to a list of high-throughput assays involving the protein of interest.

PubChem provides an informative Summary page for each Compound and BioAssay, similar in spirit to the ChEMBL Report Cards (already described). A Compound Summary page (e.g., 1.usa.gov/x3eREX) [61] provides 2D and sometimes 3D representations of the Compound, along with a wealth of additional information and links. These include alternative identifiers such as synonyms, InChI Identifier, and SMILES; computed characteristics such as molecular weight, XlogP, and number of H-bonding groups; links to BioAssays results for this Compound, toxicity information from the National Library of Medicine

ChemIDplus resource, and representations of the Compound in vendor catalogs and other databases; and links to similar Compounds within PubChem. The precise content of a BioAssay Summary page depends upon the assay type. In general, a BioAssay Summary (e.g., 1.usa.gov/x2nfRP) [62] provides a direct link to the assay data from a Show Data link near the top of the page, followed potentially by information on the protein Target and on Compounds tested and found active, and information on the assay itself, often including a detailed protocol. An array of links leads to further information, such as related BioAssays and Targets.

Many PubChem pages offer downloads of data subsets, while an FTP server (http://ftp.ncbi.nlm.nih.gov/pubchem) allows users to download complete listings of Compounds, Substances, BioAssays, and associated information. Users are referred to the original submitters of the various data set for any possible license terms (http://ftp://ftp.ncbi.nlm.nih.gov/pubchem/README).

2.2.3.3 Linking with Other Databases

As a component of the NCBI, PubChem is tightly integrated with the other bioinformatics databases available at the NCBI Web site, such as those for gene sequence, protein sequence and structure, gene expression, and the scientific literature, via bidirectional links that allow seamless navigation across NCBI resources. The relatively new BioSystems component of NCBI [63] (www.ncbi.nlm.nih.gov/biosystems) places protein Targets into the context of biomolecular pathways and other functional groupings, such as structural complexes, and includes links to external resources such as KEGG [64,65] and Reactome [36,66]. Protein targets are also linked to the curated NCBI Conserved Domain Database (CDD) (www.ncbi.nlm.nih.gov/cdd) [67] and to the three-dimensional structures of closely related proteins contained in the NCBI Molecular Modeling Database (MMDB) (http://www.ncbi.nlm.nih.gov/structure) [68]. Such links help to identify and characterize conserved binding sites in proteins. Many other external links are also provided. For example, Compound Summary pages provide links to external information in categories like Use and Manufacturing, Safety and Handling, Chemical Vendors, and so on, when available. For data imported to PubChem from other databases, such as ChEMBL and BindingDB, PubChem includes links to the corresponding information in those resources.

2.2.3.4 Special Tools and Data Sets

PubChem offers a unique set of tools for analyzing groups of Compounds. For example, a Compound search (e.g., by similarity to a drawn structure and optionally with filters according to activities and computed properties; pubchem.ncbi.nlm.nih.gov/search/) leads to a page with a list of Compounds meeting the search criteria. One may then use check boxes to select any or all of these compounds, and then, on the right-hand side of the page, choose BioActivity Analysis, Structure Clustering, or a link to biomolecular pathways involving the selected Compounds. Choosing BioActivity Analysis, and then the Structure–Activity tab, leads to an interactive heat plot showing the activities of the Compounds against multiple BioAssays, along with a hierarchical clustering of Compounds by chemical similarity and of BioAssays by Compound activity profiles (Figure 2.3) [17].

As already noted, PubChem focuses in particular on high-throughput screening data from the Molecular Libraries Screening Centers Network (MLSCN), 10 centers with a diverse set of screening platform technologies. The MLSCN is a component of the NIH Molecular Libraries Roadmap, and along with the Molecular Libraries Probe Production Centers Network (MLPCN), with nine centers, offers biomedical researchers access to their large-scale screening capabilities, along with medicinal chemistry and informatics aimed at discovering chemical probes to explore the functions of genes and signaling pathways in health and disease [69]. The molecular library centers are NIH's New Pathways to Discovery initiative, which aims to advance the understanding of biological systems. The unique high-throughput assay data in PubChem obtained directly from these screening centers are not typically present in the published literature.

2.2.4
Other Small Molecule Databases of Interest

There are dozens more chemically oriented databases of potential interest to medicinal chemists. Several noteworthy ones have been summarized alphabetically.

Binding MOAD (bindingmoad.org) gathers high-quality protein–ligand structures from the PDB (about 17 000 currently) and annotates as many as possible (about 5600 currently) with measured binding affinity data collected from the scientific literature [70–72]. Binding MOAD is thus particularly relevant to structure-based drug discovery. One may browse and search structure and affinity data via a protein classification, PDB ID, enzyme classification number, keyword, or author.

ChemSpider (www.chemspider.com) is a freely accessible chemical database [54] containing more than 26 million distinct molecules with links to information about properties and availability in over 400 data sources, such as compound catalogs and databases, including many of those listed in this article. The Web interface uses a crowdsourcing approach to expand and improve the data set, by allowing users to enter or correct entries. One may query by, for example, compound name, structure, database identifier, and computed properties; available information for each compound includes names, properties, spectra, vendors, data sources, and patents.

DrugBank (drugbank.ca) [3,55,73] is a smaller but richly annotated public database of approved and experimental drugs, including a total of about 6700 small molecules and biopharmaceuticals. One may browse and query by, for example, structure, pathway, protein sequence, and drug interactions. The data set includes pharmacological and pharmacokinetic data, dosage forms, solubilities, drug–drug interactions, metabolism information, target, and pathway data. An extensive set of downloads is provided.

GRAC and *IUPHAR-DB* (www.guidetopharmacology.org [74] and www.iuphar-db.org [75,76]) are two complementary and integrated databases that collect a range of pharmacological information on GPCRs, ion channels, and nuclear receptors from the primary literature. These data, which include small molecule

activities and affinities, are reviewed by expert international subcommittees and consultants and are linked to related information in other online resources. These databases currently house data for about 1800 small molecules and 600 different proteins spanning the targets of about half of all current licensed drugs.

PDBbind (pdbbind.org.cn and pdbbind.org) [77–79], like BindingMOAD, collects measured affinities for many types of complexes in the PDB, including protein–small molecule, protein–protein, and nucleic acid–small molecule systems. The current version at pdbbind-org.cn provides about 8000 data, of which about 6000 are for protein–small molecule complexes, and is free for academic and commercial use, on acceptance of a license agreement.

PDSP Ki (pdsp.med.unc.edu), the database of the Psychoactive Drug Screening Program at the University of North Carolina [19], contains about 55 000 binding measurements for 7500 drugs and other compounds with 740 receptors, neurotransmitter transporters, ion channels, and enzymes. The query interface is based primarily on pull-down menus and Ki limits. At no cost to academics engaged in mental health research, the same group provides experimental compound screening services with a variety of assays, including bioavailability predictions (e.g., $CaCo_2$) and cardiotoxicity (e.g., human ether-related gene (HERG)).

SuperTarget (insilico.charite.de/supertarget/) provides various views of over 330 000 interactions involving about 6000 targets and 200 000 compounds, along with annotated pathway diagrams and the ability to browse for targets categorically, such as by function and cellular location [80,81].

Therapeutic Targets Database (bidd.nus.edu.sg/group/cjttd/TTD_HOME.asp) [82] focuses on proven and prospective drug targets and their associated drugs and candidate drugs, providing extensive links to related information, such as sequence and pathway data. Both text-based and chemical similarity searches are supported, and many data sets may be downloaded.

ZINC (zinc.docking.org), based at the University of California, San Francisco, is a free database of over 21 million commercially available compounds [39]. Compounds are organized into various subsets, such as target-focused, natural products, metabolites, lead-like, and fragment-like, and are annotated with the time frame for their availability. Small arbitrary subsets may also be assembled by the user. Compounds are downloadable in popular molecular docking formats with precomputed three-dimensional conformations, in order to facilitate virtual structure-based screening.

2.3
Trends in Medicinal Chemistry Data

The combined holdings of BindingDB, ChEMBL, and PubChem enable a broad overview of trends in published medicinal chemistry data. Here, we examine rates of data production overall and by journal and institution, as well as statistical distributions of, for example, compound molecular weight and compounds per target protein. Clearly, many other analyses are also enabled by these resources.
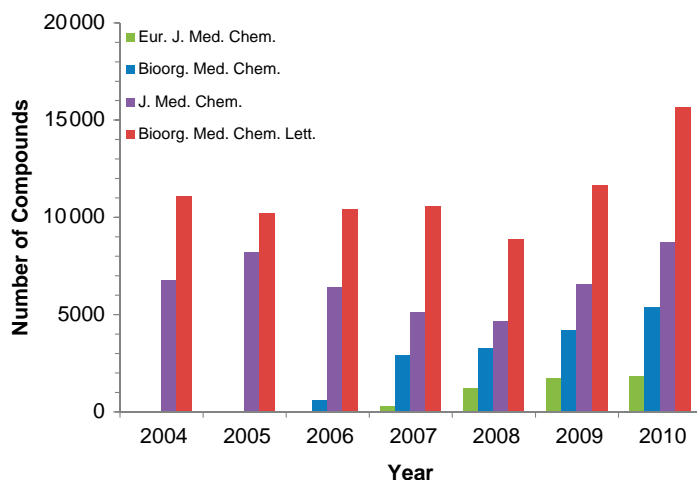
Figure 2.4 Trends in published unique small molecules by year. The data are the union (JChem 5.2 full structure search) of the holdings of BindingDB and ChEMBL for the four medicinal chemistry journals with the most data.

The number of unique small molecules published annually has increased year on year since 2008 (Figure 2.4), while the number of protein–small molecule binding measurements has followed a similar trend but at a higher level (Figure 2.5). (Note, however, that although ChEMBL has sought to exhaustively curate the core
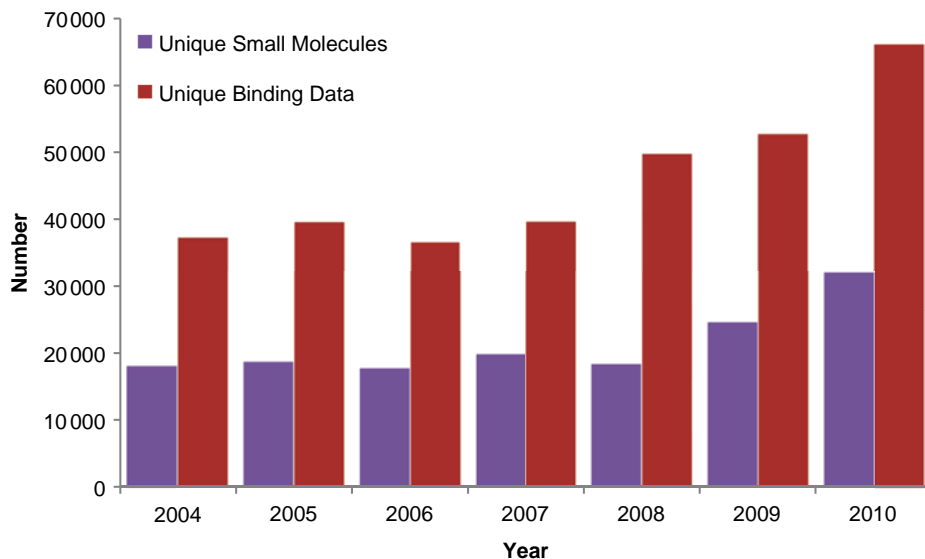
Figure 2.5 Trends in published unique small molecules and associated binding data by year. The data are the union (JChem 5.2 full structure search) of the holdings of BindingDB and ChEMBL across 34 curated journals.
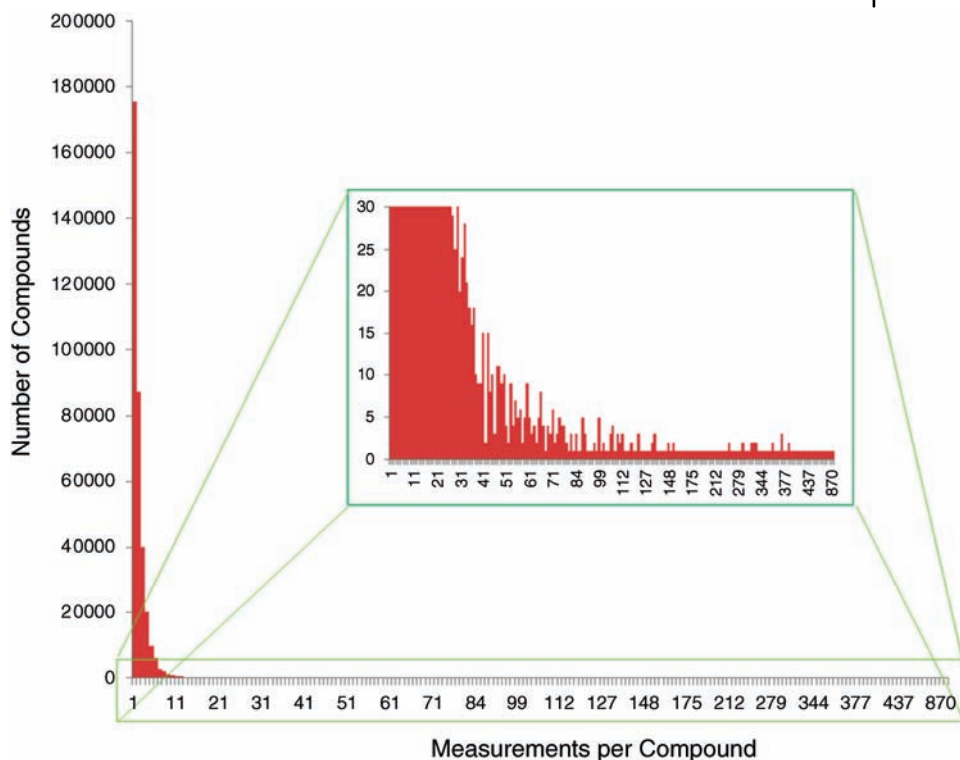
Figure 2.6 Number of binding measurements per compound in BindingDB. For example, there are nearly 180 000 compounds with one binding measurement. Inset shows the long tail of the distribution, which contains a few compounds with hundreds of measurements each.

medicinal chemistry journals, it is not guaranteed that all articles in the targeted journals were captured every year.) The difference between these two quantities implies that multiple measurements are available for some compounds, and this relationship is depicted in Figure 2.6 for the data in BindingDB. Although nearly 180 000 compounds have only one measurement, about 80 000 have two, 40 000 have three, and so on. In fact, as shown in the inset, there is a long tail in this distribution, due to a small number of compounds with tens or hundreds of measurements apiece. These outliers are mainly kinase inhibitors that have been tested against many mutants of many kinases, but several other classes are also represented there. The distribution of the number of compounds studied per target is depicted in Figure 2.7. Not surprisingly, there are many targets, such as neurotransmitter receptors, clotting factors, and kinases, against which hundreds and even thousands of compounds have been tested. The bump in the distribution at about 40 compounds per target appears to result from the reuse of several compound panels in various assays. Further details of these distributions are available at the BindingDB Web site (bit.ly/uu6ZNn [83] and bit.ly/uz9HeV [84]). Finally, it is interesting to observe that, since 2004, the distribution of compound molecular weights has
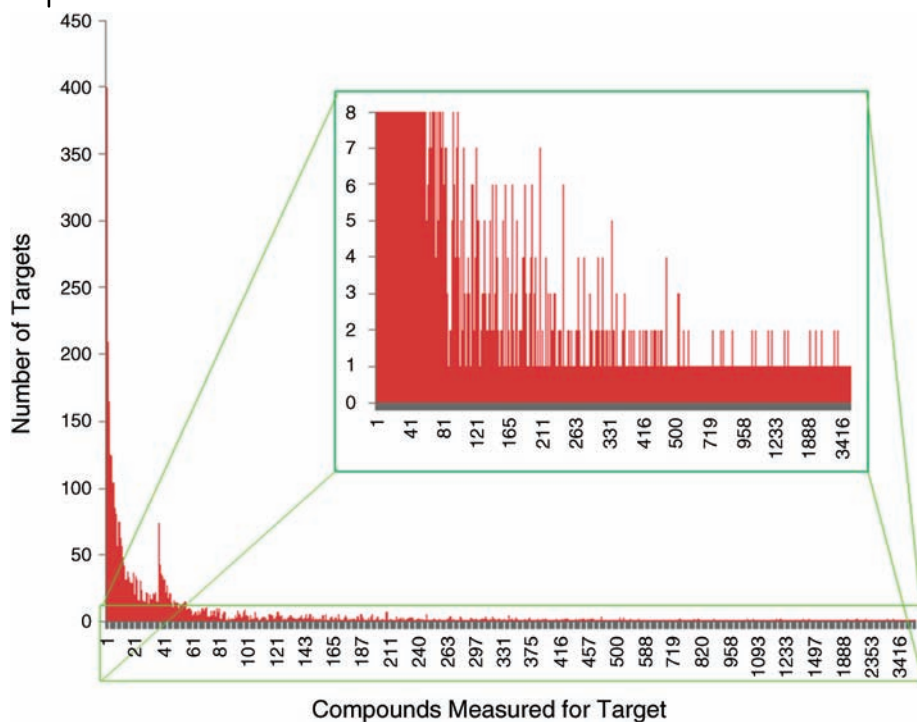
Figure 2.7 Number of protein targets in BindingDB having a given number of compounds for which affinities were measured. For example, there are about 400 targets for which one compound's affinity was tested. Inset shows the long tail of the distribution, which contains targets for which hundreds or thousands of compounds have been measured.

sharpened dramatically, with more between 200 and 600 Da, and mostly in the 200–400 Da range (Figure 2.8).

The number of new compounds in BindingDB and ChEMBL from the most highly represented journals each year is examined in Figure 2.4. (The analogous breakdown of new measurements per year parallels new compounds closely, although at a higher level, and is therefore not shown.) There is a rather consistent laddering of journals by the numbers of new compounds they publish, with most in *Bioorganic and Medicinal Chemistry Letters*, followed by *Journal of Medicinal Chemistry* and *Bioorganic & Medicinal Chemistry*. Interestingly, although the total number of new compounds per year was rather level from 2004 to 2008 (Figure 2.4), this overall trend masked a drop in new compounds in *Journal of Medicinal Chemistry* and a rise in compounds in *Bioorganic & Medicinal Chemistry* and *European Journal of Medicinal Chemistry*. However, since 2008, the number of new compounds in all of these journals has risen together, and their relative shares have not changed appreciably.

Perhaps surprisingly, academia generates nearly half of the medicinal chemistry data in the combined holdings of BindingDB, ChEMBL, and PubChem BioAssays
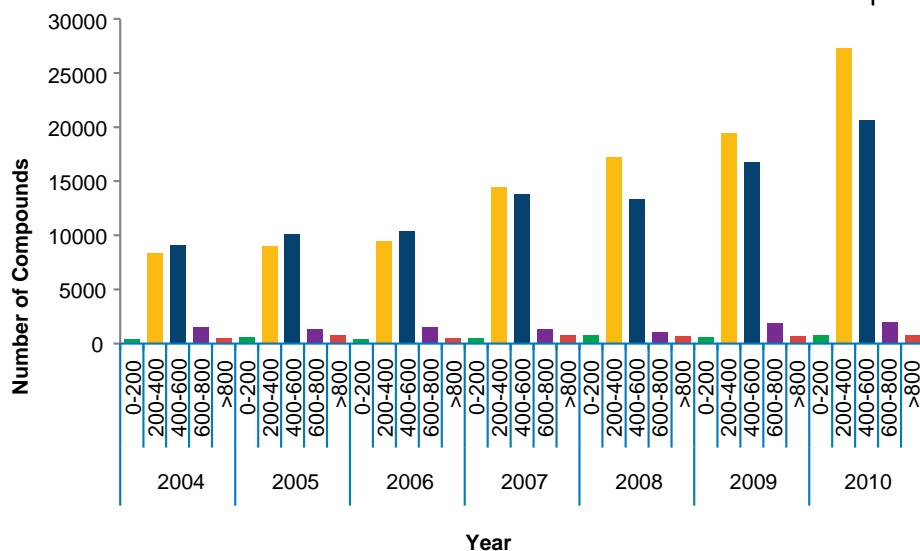
Figure 2.8 Molecular weights of new compounds in ChEMBL, BindingDB, and PubChem BioAssays, by year.

(Figure 2.9). It is unlikely, however, that this distribution reflects the volume of data actually generated in these two sectors, as many corporate data are not published. Also note that about one third of the academic data derive from screening centers such as the Scripps Research Institute Molecular Screening Center and the New Mexico Molecular Libraries Screening Center.
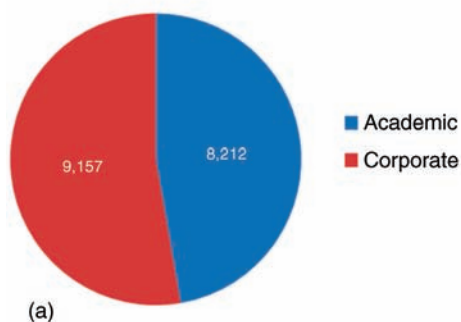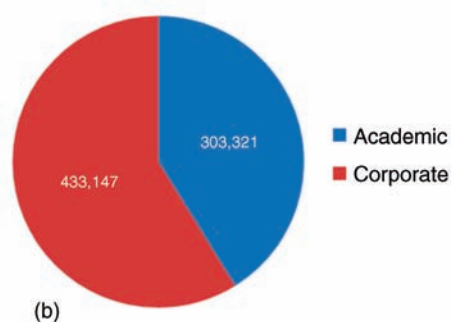
Figure 2.9 Institutional sources of articles (a) and compound activity data (b) in BindingDB, PubChem BioAssay, and ChEMBL. These data include only measurements with a defined protein target. Each confirmatory PubChem BioAssay is counted as an article. Institutional sources were obtained based on keywords (e.g., "university" and "institute") in the Affiliation information in PubMed article entries and were spot-checked by hand.

2.4
Directions

2.4.1
Strengthening the Databases

2.4.1.1 Coordination among Databases

The existence of several publicly accessible medicinal chemistry databases provides substantial benefits, while defining a need for coordination to minimize the duplication of effort and maximize the value to users. One current benefit is the availability of a diversity of user interfaces and capabilities to support a range of applications and preferences. Another benefit is the high-level sustainability and stability in the face of potential data losses and uncertainties of continued scientific funding for any single project. There is also a valuable opportunity to distribute the workload of journal curation across projects. Indeed, BindingDB, ChEMBL, and PubChem are increasingly sharing data and curation efforts. For example, while ChEMBL's outsourced curation focuses on core medicinal chemistry journals such as *Journal of Medicinal Chemistry* and *Bioorganic & Medicinal Chemistry Letters*, BindingDB is now engaged in curation of chemical biology journals and others not covered by ChEMBL, such as *Chemistry & Biology*, *Nature Chemical Biology*, and *ACS Chemical Biology*. The protein–ligand data sets in the latter journals often are particularly interesting, because they involve proteins that are currently in the process of being identified as candidate drug targets, or compounds that explore innovative chemistries. To further increase efficiencies, there is now a collaborative effort between BindingDB and ChEMBL to compare each other's existing data holdings for discrepancies, and thus potential errors. Ultimately, the greatest efficiency and service to users may be achieved by following the models of other large database endeavors. For example, the Worldwide Protein Data Bank (wwPDB, www.wwpdb.org) [85] comprises four different projects, two in the United States [2,86], one in Japan [87], and one in Europe [88], which share a core data set, as well as annotation and validation strategies, while presenting the data differently and with emphasis on different user communities.

2.4.1.2 Data Quality

Data quality is of fundamental importance, and it is of interest to consider the origins and nature of errors in the public medicinal chemistry databases. Data errors may be separated into three classes: scientific errors, errors of transcription, and data handling errors. Scientific errors result from problems with an experiment or its technical analysis. Transcription errors arise during the writing and publication of the data or during the extraction of the data from the publication and its subsequent entry into the database. Data handling errors result from problems in the database itself, such as the introduction of a mismatch between a table of compounds and a table of targets during a database update. A systematic evaluation of the quality of data in these public databases would be of interest, and one could in principle use statistical sampling to characterize overall data quality without having

to examine every entry. It would be even more valuable to identify and correct errors throughout these massive data sets, but this would be a much larger challenge. Some of the issues in error checking are now discussed.

Although there is no perfect way to detect scientific errors, it is possible for an expert to judge the suitability of the method reported in the paper, as done, for example, in NIST's evaluated database of thermodynamics of enzyme-catalyzed reactions [89]. Concerns that might be identified in this way could include failure to ascertain the active enzyme concentration [90], or reported enzyme inhibition by a compound that is a known aggregator [91]. Perhaps only the authors of an article can identify transcription errors that are enshrined in their publication, but errors introduced during the extraction of data from an article and their entry into a database can be detected by painstaking comparisons between database entries and associated articles. The same is true for data handling errors, but the latter, once detected, can often be corrected *en masse* by undoing the database manipulation that generated them. It is worth noting that meaningfully categorizing errors can also be challenging. For example, an error in stereochemistry may not be considered equally severe as an incorrect chemical structure. However, if these two types of error are put into different categories, rather than being lumped together, then more articles will need to be surveyed in order to gather meaningful statistics in both categories. There can also be ambiguities that are difficult to resolve, such as when a paper provides data for a protein target without specifying its subtype; e.g., beta-adrenergic receptor, as opposed to beta-1- or beta-2-adrenergic receptor. Other errors, such as in the name of an author, are significant, but do affect the scientific content of the database.

Evaluating and ultimately correcting the data extracted from tens of thousands of papers will be an enormous undertaking [92,93]. Given the limited resources available to these projects, a community effort may be the only way to make inroads. It is in this spirit that the BindingDB project routinely emails article authors inviting them to correct any errors they may find in their BindingDB entries. Perhaps 1–2% of these messages receive a reply, and of these, about one third report an error. Users who notice errors in BindingDB, ChEMBL, and PubChem are also invited to submit corrections at http://www.bindingdb.org/bind/sendmail.jsp, chembl-help@ebi.ac.uk, info@ncbi.nlm.nih.gov, respectively. However, a more systematic approach would be for experts to adopt specific protein targets, overseeing the crowdsourcing of corrections to the associated data [10]. Similar approaches are already being used by Wikipedia, ChemSpider, and the IUPHAR databases.

2.4.1.3 Linking Journals and Databases

All of the literature data in these databases are entered by employees or contractors who read each article, extract the pertinent data, and enter it into one of the databases. This labor-intensive curation process is time consuming and costly and inevitably introduces errors. The magnitude of these parallel curation efforts is highlighted by the graphs in Figures 2.4 and 2.5 and the data production rate will only grow in the coming years, as research in emerging economies accelerates and technological advances yield a wealth of new candidate drug targets [94]. The challenge of keeping up with this data flow was the topic of a panel discussion

at a recent database conference, which included the leaders of most of the largest databases already discussed, as well as many participants from industry, publishing, and government.[1]

The consensus that emerged is that a new mechanism is needed, in which authors and/or journals make the data in their new articles available in a simple, machine-readable format. For example, authors might provide a file with a list of protein targets, SMILES strings, and affinity data. This could reside in the online supplementary information, or might be uploaded directly to a central Web portal from which any database team could draw those data that fall within the scope of their project. The field of structural biology offers two interesting models. In the case of macromolecular structures, authors routinely deposit their machine-readable structure data into one of the PDB portals so that they may be incorporated into the global wwPDB databases, and journals do not accept papers that report new structures without a PDB ID. Small molecule structure data are typically published via *Acta Crystallographica Section E*, in which each online article is associated with a short crystallographic information file (cif), which users may freely download and use.

In the case of medicinal chemistry data, electronic submission should be quite straightforward, as most authors already have their data in machine-readable format when they are preparing their articles, for example, in the form of spreadsheets and ChemDraw files. The chief challenge for our community might be defining the precise set of data to be uploaded. For example, although it is clear that each compound should be defined, it may not be so clear how much information should be provided about the experimental method and conditions. Regardless of the details, it is clear that joining machine-readable data to every medicinal chemistry article will lead to medicinal chemistry databases that are dramatically more sustainable, accurate, and complete.

2.4.2
Next-Generation Capabilities

The public compound activity databases now provide an informatics foundation on which many new research capabilities can be built. For example, the fact that researchers increasingly read articles on computer screens rather than paper provides an opportunity for tighter integration between journals and databases [95]. Articles then become live, interactive media, which provide seamless access to a world of related information, while also serving as documentation for database entries (Phillip Bourne, personal communication). Building tighter, interactive connections between medicinal chemistry and pathway databases [36–38,63,96–99] also has enormous potential to strengthen research. For example, the ability to display pathways while highlighting proteins already known to have small molecule binders will help medicinal chemists view their work in a broad biological context. It will also draw the attention of systems biologists to compounds that may be useful biological

probes and to potential new avenues for drug discovery. The SuperTarget database [80] is one significant effort along these lines, while the Reactome pathway database [36,66] and the Cytoscape software [100] provide related network viewing capabilities by using the PSICUQIC Web service [101,102], which has links to multiple molecular interaction databases.

Data will also be more smoothly linked to computational analysis and prediction tools. For example, one might collect a set of active compounds at BindingDB, transfer them to another online resource that does machine learning, and then use the result set of rules to computationally filter a compound catalog in search of new actives. The candidate actives could furthermore be piped through a database integrating pathway and medicinal chemistry data, in order to flag potentially unanticipated on- or off-target effects. Each step of such a process might be carried out on a different computer somewhere on the Web, with the user directing the flow of data and collecting the output. Many other capabilities could be used in an online informatics network, for example, methods of predicting druggable protein binding sites [103–107] or of estimating the physical properties of compounds. Software is already available that allows users to direct data flows involving multiple data and computational resources in a flexible manner [108–114], and the continued development of such technologies will enable many new informatics tools to speed up drug discovery.

2.5
Summary

Historically, medicinal chemistry data were not well connected to the informatics world, but this situation has now changed decisively. Here, we have focused on three prominent, publicly accessible chemical activity databases: BindingDB, ChEMBL, and PubChem, each with its own unique user interface and scientific focus. These resources allow users to browse, query, and download hundreds of thousands of data extracted from the medicinal chemistry literature, along with additional data from other sources such as the NIH screening centers. We also more briefly reviewed seven complementary chemical databases of interest to many medicinal chemists. Analyses of the holdings of BindingDB and ChEMBL indicate that the rate of publication of medicinal chemistry data has grown by about 50% since 2007 and appears to continue on an upward trend. This is exciting scientifically, but it also means the work of extracting and managing the data is growing. We therefore discussed potential approaches to strengthening the database system, including further coordination among the various projects, the community quality control efforts, and the development of a simple mechanism for authors to make their data available in electronic format concurrently with publication. Finally, we discussed future research capabilities that will grow from integration of the medicinal chemistry databases with more biologically oriented databases, as well as with Web-based tools for computational analysis and prediction. In sum, the emerging system of publicly accessible medicinal chemistry databases is rapidly becoming a

critical infrastructure component for drug discovery efforts worldwide and is opening doors to valuable, new applications at the interfaces of chemistry and biology.

References

1 Portoghese, P.S. (2011) My farewell to the journal of *Medicinal Chemistry*. *Journal of Medicinal Chemistry*, **54**, 8235.

2 Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**, 235–242.

3 Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006) DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Research*, **34**, D668–D672.

4 Bader, G. and Donaldson, S. (2012) Pathguide: the pathway resource list. www.pathguide.org/ (accessed May 16, 2012).

5 Bader, G.D., Cary, M.P., and Sander, C. (2006) Pathguide: a pathway resource list. *Nucleic Acids Research*, **34**, D504–D506.

6 Apodaca, R. (2012) Depth-First. depth-first.com/articles/2011/10/12/sixty-four-free-chemistry-databases/ (accessed May 29, 2012).

7 Gaulton, A. and Overington, J.P. (2010) Role of open chemical data in aiding drug discovery and design. *Future Medicinal Chemistry*, **2**, 903–907.

8 Wassermann, A.M. and Bajorath, J. (2011) BindingDB and ChEMBL: online compound databases for drug discovery. *Expert Opinion on Drug Discovery*, **6**, 683–687.

9 Li, Q., Cheng, T., Wang, Y., and Bryant, S.H. (2010) PubChem as a public resource for drug discovery. *Drug Discovery Today*, **15**, 1052–1057.

10 Williams, A.J. (2008) Public chemical compound databases. *Current Opinion in Drug Discovery & Development*, **11**, 393–404.

11 Williams, A.J. (2008) Internet-based tools for communication and collaboration in chemistry. *Drug Discovery Today*, **13**, 502–506.

12 Gozalbes, R. and Pineda-Lucena, A. (2011) Small molecule databases and chemical descriptors useful in chemoinformatics: an overview. *Combinatorial Chemistry & High Throughput Screening*, **14**, 548–558.

13 Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J.P. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, **40**, D1100–D1107.

14 ChEMBL (2012) www.ebi.ac.uk/chembl/ (accessed May 16, 2012).

15 Bryant, S. (2006) PubChem: an information resource linking chemistry and biology. Abstracts of Papers of the American Chemical Society, 231.

16 Bolton, E.E., Wang, Y., Thiessen, P.A., and Bryant, S.H. (2008) PubChem: integrated platform of small molecules and

biological activities, in *Annual Reports in Computational Chemistry*, vol. 4 (eds A.W. Ralph and C.S. David), Elsevier, Bethesda, MD, pp. 217–241.

17 Wang, Y, Xiao, J., Suzek, T.O., Zhang, J., Wang, J., and Bryant, S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, **37**, W623–W633.

18 Evans, J.M., Lopez, E., and Roth, B.L. (2001) The NIMH Psychoactive Drug Screening Program's Ki database: an on-line, searchable database of receptor-ligand affinity values. *Society for Neuroscience Abstracts*, **27**, 2076.

19 Roth, B.; Driscoll, J. PDSP. http://pdsp.med.unc.edu/indexR.html (accessed May 16, 2012).

20 The Binding Database: http://bit.ly/ws4vLt or http://www.bindingdb.org/jsp/dbsearch/PrimarySearch_pubmed.jsp?pubmed=50006488&pubmed_submit=TBD (accessed May 16, 2012).

21 The Binding Database. http://bit.ly/AyOWyq or http://www.bindingdb.org/bind/ByKI.jsp?specified=IC50 (accessed May 16, 2012).

22 The Binding Database. http://bit.ly/AAUiVz or http://www.bindingdb.org/bind/ByMolWeight.jsp (accessed May 16, 2012).

23 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

24 The Binding Database: http://bit.ly/ws4vLt or www.bindingdb.org/bind/BySequence.jsp (accessed May 16, 2012).

25 The Binding Database: http://bit.ly/zL842y or www.bindingdb.org/bind/chemsearch/marvin/index.jsp (accessed May 16, 2012).

26 The Binding Database. http://bit.ly/w0A1G5 or www.bindingdb.org/bind/BatchStructures.jsp (accessed May 16, 2012).

27 Drugs@FDA, www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm (accessed May 16, 2012).

28 The Binding Database. http://bit.ly/wXIziD or www.bindingdb.org/bind/ByFDAdrugs.jsp (accessed May 16, 2012).

29 The Binding Database. http://bit.ly/wHLXDl or www.bindingdb.org/bind/ByAuthor.jsp (accessed May 16, 2012).

30 The Binding Database. http://bit.ly/wHLXDl or www.bindingdb.org/bind/ByJournal.jsp (accessed May 16, 2012).

31 The Binding Database. http://bit.ly/yMYvx2 or www.bindingdb.org/bind/ByInstitution.jsp (accessed May 16, 2012).

32 The Binding Database. http://bit.ly/zVd6z2 or www.bindingdb.org/bind/ByPDBids.jsp (accessed May 16, 2012).

33 The Binding Database. http://bit.ly/x9f9Yd or www.bindingdb.org/bind/ByPDBids_100.jsp (accessed May 16, 2012).

34 Creative Commons (2011) Attribution-ShareAlike 3.0 Unported.

35 The Binding Database. http://bit.ly/zq9oW3 or www.bindingdb.org/jsp/dbsearch/PrimarySearch_ki.jsp?polymerid=50000007&target=ABL1&tag=polkd&column=Kd&energyterm=kcal/mole&startPg=0&Increment=50&submit=Search (accessed May 16, 2012).

36 Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E., and Stein, L. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, **33**, D428–D432.

37 Ogata, H., Goto, S., Fujibuchi, W., and Kanehisa, M. (1998) Computation with the KEGG pathway database. *Biosystems*, **47**, 119–128.

38 Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Research*, **37**, D674–D679.

39 Irwin, J.J. and Shoichet, B.K. (2005) ZINC: a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, **45**, 177–182.

40 The Binding Database. http://bit.ly/zX0SfQ or www.bindingdb.org/bind/chemsearch/marvin/BatchStructures.jsp (accessed May16, 2012).

41 Csizmadia, F. (2000) JChem: Java applets and modules supporting chemical

database handling from web browsers. *Journal of Chemical Information and Computer Sciences*, **40**, 323–324.

42 ChemAxon (2011) JChem, 5.6, Budapest, Hungary.

43 Harper, G., Bradshaw, J., Gittins, J.C., Green, D.V.S., and Leach, A.R. (2001) Prediction of biological activity for high-throughput screening using binary kernel discrimination. *Journal of Chemical Information and Computer Sciences*, **41**, 1295–1300.

44 Jorissen, R.N. and Gilson, M.K. (2005) Virtual screening of molecular databases using a support vector machine. *Journal of Chemical Information and Modeling*, **45**, 549–561.

45 Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Machine Learning*, **20**, 273–297.

46 The Binding Database. http://bit.ly/ yTctqN or http://bindingdb.org/ validation_sets/index.jsp (accessed May 16, 2012).

47 Ertl, P. (2012) JME Molecular Editor, Novartis.

48 ChemAxon (2012) Marvin Sketch, 5.7.

49 Accelrys (2012) JDraw.

50 Weininger, D. (1988) SMILES, a chemical language and information system: 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, **28**, 31–36.

51 Velankar, S., Best, C., Beuth, B., Boutselakis, C.H., Cobley, N., Sousa Da Silva, A.W., Dimitropoulos, D., Golovin, A., Hirshberg, M., John, M., Krissinel, E.B., Newman, R., Oldfield, T., Pajon, A., Penkett, C.J., Pineda-Castillo, J., Sahni, G., Sen, S., Slowley, R., Suarez-Uruena, A., Swaminathan, J., van Ginkel, G., Vranken, W.F., Henrick, K., and Kleywegt, G.J. (2010) PDBe: Protein Data Bank in Europe. *Nucleic Acids Research*, **38**, D308–D317.

52 Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H.S., Rios, D., Ritchie, G.R., Ruffier,

M., Schuster, M., Sobral, D., Spudich, G., Tang, Y.A., Trevanion, S., Vandrovcova, J., Vilella, A.J., White, S., Wilder, S.P., Zadissa, A., Zamora, J., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Hubbard, T.J., Parker, A., Proctor, G., Vogel, J., and Searle, S.M. (2011) Ensembl 2011. *Nucleic Acids Research*, **39**, D800–D806.

53 Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., and Yeh, L. S. (2005) The universal protein resource (UniProt). *Nucleic Acids Research*, **33**, D154–D159.

54 ChemSpider: the free chemical database (2012) www.chemspider.com/ (accessed May 16, 2012).

55 Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, **36**, D901–D906.

56 Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, **36**, D344–D350.

57 Gamo, F.J., Sanz, L.M., Vidal, J., de Cozar, C., Alvarez, E., Lavandera, J.L., Vanderwall, D.E., Green, D.V., Kumar, V., Hasan, S., Brown, J.R., Peishoff, C.E., Cardon, L.R., and Garcia-Bustos, J.F. (2010) Thousands of chemical starting points for antimalarial lead identification. *Nature*, **465**, 305–310.

58 Meister, S., Plouffe, D.M., Kuhen, K.L., Bonamy, G.M., Wu, T., Barnes, S.W., Bopp, S.E., Borboa, R., Bright, A.T., Che, J., Cohen, S., Dharia, N.V., Gagaring, K., Gettayacamin, M., Gordon, P., Groessl, T., Kato, N., Lee, M.C., McNamara, C.W., Fidock, D.A., Nagle, A., Nam, T.G., Richmond, W., Roland, J., Rottmann, M., Zhou, B., Froissard, P., Glynne, R.J., Mazier, D., Sattabongkot, J., Schultz, P.G., Tuntland, T., Walker, J.R., Zhou, Y.,

Chatterjee, A., Diagana, T.T., and Winzeler, E.A. (2011) Imaging of Plasmodium liver stages to drive next-generation antimalarial drug discovery. *Science*, **334**, 1372–1377.

59 Guiguemde, W.A., Shelat, A.A., Bouck, D., Duffy, S., Crowther, G.J., Davis, P. H., Smithson, D.C., Connelly, M., Clark, J., Zhu, F., Jimenez-Diaz, M.B., Martinez, M.S., Wilson, E.B., Tripathi, A.K., Gut, J., Sharlow, E.R., Bathurst, I., El Mazouni, F., Fowble, J.W., Forquer, I., McGinley, P.L., Castro, S., Angulo-Barturen, I., Ferrer, S., Rosenthal, P.J., Derisi, J.L., Sullivan, D. J., Lazo, J.S., Roos, D.S., Riscoe, M.K., Phillips, M.A., Rathod, P.K., Van Voorhis, W.C., Avery, V.M., and Guy, R. K. (2010) Chemical genetics of *Plasmodium falciparum*. *Nature*, **465**, 311–315.

60 Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., Zhou, Z., Han, L., Karapetyan, K., Dracheva, S., Shoemaker, B.A., Bolton, E., Gindulyte, A., and Bryant, S.H. (2012) PubChem's BioAssay Database. *Nucleic Acids Research*, **40**, D400–D412.

61 PubChem Compound. http://1.usa.gov/ x3eREX or http://pubchem.ncbi.nlm.nih. gov/summary/summary.cgi?cid=184691 (accessed May 16, 2012).

62 PubChem BioAssay. http://1.usa.gov/ x2nfRP or http://pubchem.ncbi.nlm.nih. gov/assay/assay.cgi?aid=33734&loc=ea_ ras (accessed May 16, 2012).

63 Geer, L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., He, J., He, S., Liu, C., Shi, W., and Bryant, S.H. (2010) The NCBI BioSystems database. *Nucleic Acids Research*, **38**, D492–D496.

64 Minoru, K. (1997) A database for post-genome analysis. *Trends in Genetics*, **13**, 375–376.

65 Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **27**, 29–34.

66 Reactome (2012) www.reactome.org/ ReactomeGWT/entrypoint.html (accessed May 16, 2012).

67 Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R. C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Lu, F., Marchler, G.H., Mullokandov, M., Omelchenko, M.V., Robertson, C.L., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D., Zhang, N., Zheng, C., and Bryant, S.H. (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research*, **39**, D225–D229.

68 Madej, T., Addess, K.J., Fong, J.H., Geer, L.Y., Geer, R.C., Lanczycki, C.J., Liu, C., Lu, S., Marchler-Bauer, A., Panchenko, A. R., Chen, J., Thiessen, P.A., Wang, Y., Zhang, D., and Bryant, S.H. (2012) MMDB: 3D structures and macromolecular interactions. *Nucleic Acids Research*, **40**, D461–D464.

69 Molecular Libraries Probe Production Centers Network (MLPCN)(2012) (accessed May 16, 2012).

70 Carlson, H.A. (2012) Binding MOAD. (accessed May 16, 2012).

71 Hu, L., Benson, M.L., Smith, R.D., Lerner, M.G., and Carlson, H.A. (2005) Binding MOAD (Mother Of All Databases). *Proteins*, **60**, 333–340.

72 Benson, M.L., Smith, R.D., Khazanov, N. A., Dimcheff, B., Beaver, J., Dresslar, P., Nerothin, J., and Carlson, H.A. (2008) Binding MOAD, a high-quality protein–ligand database. *Nucleic Acids Research*, **36**, D674–D678.

73 Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A.C., and Wishart, D.S. (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research*, **39**, D1035–D1041.

74 Alexander, S.P., Mathie, A., and Peters, J. A. (2011) Guide to Receptors and Channels (GRAC), 5th edition. *British Journal of Pharmacology*, **164** (Suppl. 1), S1–S324.

75 Sharman, J.L., Mpamhanga, C.P., Spedding, M., Germain, P., Staels, B., Dacquet, C., Laudet, V., Harmar, A.J., and Nc, I. (2011) IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data. *Nucleic Acids Research*, **39**, D534–D538.

76 Harmar, A.J., Hills, R.A., Rosser, E.M., Jones, M., Buneman, O.P., Dunbar, D.R., Greenhill, S.D., Hale, V.A., Sharman, J.L., Bonner, T.I., Catterall, W.A., Davenport, A.P., Delagrange, P., Dollery, C.T., Foord, S.M., Gutman, G.A., Laudet, V., Neubig, R.R., Ohlstein, E.H., Olsen, R.W., Peters, J., Pin, J.P., Ruffolo, R.R., Searls, D.B., Wright, M.W., and Spedding, M. (2009) IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic Acids Research*, **37**, D680–D685.

77 Wang, R., Fang, X., Lu, Y., and Wang, S. (2004) The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, **47**, 2977–2980.

78 Cheng, T., Li, X., Li, Y., Liu, Z., and Wang, R. (2009) Comparative assessment of scoring functions on a diverse test set. *Journal of Chemical Information and Modeling*, **49**, 1079–1093.

79 Wang, R., Fang, X., Lu, Y., Yang, C.Y., and Wang, S. (2005) The PDBbind database: methodologies and updates. *Journal of Medicinal Chemistry*, **48**, 4111–4119.

80 Gunther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E.G., Gewiess, A., Jensen, L.J., Schneider, R., Skoblo, R., Russell, R.B., Bourne, P.E., Bork, P., and Preissner, R. (2008) SuperTarget and Matador: resources for exploring drug–target relationships. *Nucleic Acids Research*, **36**, D919–D922.

81 Hecker, N., Ahmed, J., von Eichborn, J., Dunkel, M., Macha, K., Eckert, A., Gilson, M.K., Bourne, P.E., and Preissner, R. (2012) SuperTarget goes quantitative: update on drug–target interactions. *Nucleic Acids Research*, **40**, D1113–D1117.

82 Zhu, F., Shi, Z., Qin, C., Tao, L., Liu, X., Xu, F., Zhang, L., Song, Y., Liu, X., Zhang, J., Han, B., Zhang, P., and Chen, Y. (2012) Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Research*, **40**, D1128–D1136.

83 The Binding Database. http://bit.ly/uu6ZNn or www.bindingdb.org/bind/ByDataLigand.jsp (accessed May 16, 2012).

84 The Binding Database. http://bit.ly/uz9HeV or www.bindingdb.org/bind/ByMonomersTarget.jsp (accessed May 16, 2012).

85 Berman, H., Henrick, K., and Nakamura, H. (2003) Announcing the worldwide Protein Data Bank. *Nature Structural Biology*, **10**, 980.

86 (1995) Battle for BMRB Biological Magnetic Resonance Data Bank. *Nature Structural Biology*, **2**, 811–812.

87 Kinjo, A.R., Suzuki, H., Yamashita, R., Ikegawa, Y., Kudou, T., Igarashi, R., Kengaku, Y., Cho, H., Standley, D.M., Nakagawa, A., and Nakamura, H. (2012) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Research*, **40**, D453–D460.

88 Velankar, S., Alhroub, Y., Best, C., Caboche, S., Conroy, M.J., Dana, J.M., Fernandez Montecelo, M.A., van Ginkel, G., Golovin, A., Gore, S.P., Gutmanas, A., Haslam, P., Hendrickx, P.M., Heuson, E., Hirshberg, M., John, M., Lagerstedt, I., Mir, S., Newman, L.E., Oldfield, T.J., Patwardhan, A., Rinaldi, L., Sahni, G., Sanz-Garcia, E., Sen, S., Slowley, R., Suarez-Uruena, A., Swaminathan, G.J., Symmons, M.F., Vranken, W.F., Wainwright, M., and Kleywegt, G.J. (2012) PDBe: Protein Data Bank in Europe. *Nucleic Acids Research*, **40**, D445–D452.

89 Goldberg, R.N., Tewari, Y.B., and Bhat, T.N. (2004) Thermodynamics of enzyme-catalyzed reactions: a database for quantitative biochemistry. *Bioinformatics*, **20**, 2874–2877.

90 Kuzmic, P., Elrod, K.C., Cregar, L.M., Sideris, S., Rai, R., and Janc, J.W. (2000) High-throughput screening of enzyme inhibitors: simultaneous determination of tight-binding inhibition constants and enzyme concentration. *Analytical Biochemistry*, **286**, 45–50.

91 McGovern, S.L., Caselli, E., Grigorieff, N., and Shoichet, B.K. (2002) A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *Journal of Medicinal Chemistry*, **45**, 1712–1722.

92 Williams, A.J. and Ekins, S. (2011) A quality alert and call for improved

curation of public chemistry databases. *Drug Discovery Today*, **16**, 747–750.

93 Williams, A.J., Ekins, S., and Tkachenko, V. (2012) Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug Discovery Today*, **17**, 685–701.

94 Wang, S. and Georg, G.I. (2012) Transition in leadership: opportunities and challenges. *Journal of Medicinal Chemistry*, **55**, 1.

95 Bourne, P. (2005) Will a biological database be different from a biological journal? *PLoS Computational Biology*, **1**, e34.

96 Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2009) ConsensusPathDB: a database for integrating human functional interaction networks. *Nucleic Acids Research*, **37**, D623–D628.

97 Pico, A., Kelder, T., van Iersel, M., Hanspers, K., Conklin, B., and Evelo, C. (2008) WikiPathways: pathway editing for the people. *PLoS Biology*, **6**, e184.

98 Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D.D., Liu, P., Gautam, B., Ly, S., Guo, A.C., Xia, J., Liang, Y., Shrivastava, S., and Wishart, D.S. (2010) SMPDB: the Small Molecule Pathway Database. *Nucleic Acids Research*, **38**, D480–D487.

99 Sreenivasaiah, P.K., Rani, S., Cayetano, J., Arul, N., and Kim, D.H. (2012) IPAVS: Integrated Pathway Resources, Analysis and Visualization System. *Nucleic Acids Research*, **40**, D803–D808.

100 Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**, 2498–2504.

101 Kerrien, S. and Aranda, B. (2012) PSICQUIC View. www.ebi.ac.uk/Tools/webservices/psicquic/view/main.xhtml (accessed May 16, 2012).

102 Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F.S., Ceol, A., Chautard, E., Dana, J.M., De Las Rivas, J., Dumousseau, M., Galeota, E., Gaulton, A., Goll, J., Hancock, R.E., Isserlin, R., Jimenez, R.C., Kerssemakers, J., Khadake, J., Lynn, D.J., Michaut, M., O'Kelly, G., Ono, K., Orchard, S., Prieto, C., Razick, S., Rigina, O., Salwinski, L., Simonovic, M., Velankar, S., Winter, A., Wu, G., Bader, G.D., Cesareni, G., Donaldson, I. M., Eisenberg, D., Kleywegt, G.J., Overington, J., Ricard-Blum, S., Tyers, M., Albrecht, M., and Hermjakob, H. (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nature Methods*, **8**, 528–529.

103 Nicola, G., Smith, C.A., and Abagyan, R. (2008) New method for the assessment of all drug-like pockets across a structural genome. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, **15**, 231–240.

104 An, J., Totrov, M., and Abagyan, R. (2004) Comprehensive identification of "druggable" protein ligand binding sites. *Genome Information*, **15**, 31–41.

105 Keller, T.H., Pichota, A., and Yin, Z. (2006) A practical view of 'druggability'. *Current Opinion in Chemical Biology*, **10**, 357–361.

106 Halgren, T. (2007) New method for fast and accurate binding-site identification and analysis. *Chemical Biology and Drug Design*, **69**, 146–148.

107 Henrich, S., Salo-Ahen, O.M., Huang, B., Rippmann, F.F., Cruciani, G., and Wade, R.C. (2010) Computational approaches to identifying and characterizing protein binding sites for ligand design. *Journal of Molecular Recognition*, **23**, 209–219.

108 Stevenson, J.M. and Mulready, P.D. (2003) Pipeline pilot 2.1. *Journal of the American Chemical Society*, **125**, 1437–1438.

109 Accelrys (2012) Pipeline Pilot, 8.5.

110 Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kotter, T., Meinl, T., Ohl, P., Thiel, K., and Wiswedel, B. (2009) KNIME: The Konstanz Information Miner. *SIGKDD Explorations*, **11**, 26–31.

111 Berthold, M.H., Cebron, N., Dill, F., Di Fatta, G., Gabriel, T.R., Georg, F., Moinl, T., Ohl, P., Sieb, C., and Wiswedel, B. (2006) KNIME: the Konstanz Information Miner, Industrial Simulation Conference, University of Palermo, Palermo, Italy, June 5–7, 2006.

112 Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P., and Oinn, T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, **34**, W729–W732.

113 Oinn, T., Greenwood, M., Addis, M., Alpdemir, M.N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M.R., Senger, M., Stevens, R., Wipat, A., and Wroe, C. (2006) Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, **18**, 1067–1100.

114 Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B., and Mock, S. (2004) Kepler: an extensible system for design and execution of scientific workflows, 16th International Conference on Scientific and Statistical Database Management, Petros Nomikos Conference Center, Santorini Island, Greece, June 21–23.

3
Chemical Ontologies for Standardization, Knowledge Discovery, and Data Mining

Janna Hastings and Christoph Steinbeck

3.1
Introduction

Scientific research is increasingly benefitting from technological advances. New platforms enable novel investigations for insight into previously opaque areas of biology. The advent of high-throughput technology is enabling such investigations to generate data at an unprecedented rate throughout the life sciences. These data have the potential to steer solutions to some of the longest-standing mysteries of biological functioning. However, analyzing the data and retrieving *relevant* information becomes much more difficult with the increase in volume and diversity [1].

Sophisticated computational processing is essential to filter, organize, and search for patterns in biological data. In the biomedical sciences, large-scale data management is increasingly being facilitated by *ontologies*: computable logic-based representations of human domain knowledge, designed to serve many different data management analysis purposes [2,3].

First, ontologies provide a standardization of terminology and reference identifiers for a domain in order that different sources of data can be aggregated through shared annotations. Second, they provide a hierarchical organization of entities in the domain to enable flexible aggregation. Such aggregation is an essential component in several approaches to data-driven scientific discovery. Third, they facilitate browsing and searching, driving the interfaces of several bioinformatics databases. Further, their underlying representation in logical languages allows intelligent applications to perform complex reasoning tasks such as checking for errors and inconsistencies in the represented knowledge.

This chapter is devoted to the use of chemical ontologies in support of data mining for drug discovery. Section 3.2 gives a general introduction to ontologies, their structure, and underlying technology. Section 3.3 introduces the state of the art in chemical ontology technology. Sections 3.4–3.6 deal with the uses of chemical ontologies in standardization, knowledge discovery, and data mining. Finally, Section 3.7 gives an overview and highlights some exciting current research areas.

3.2
Background

Ontology has long been a discipline of philosophy concerned with the study of existence or simply *what is*. In the late twentieth century, ontology re-emerged as a subdiscipline of knowledge representation in artificial intelligence research within computer science circles, where the focus was on describing domain knowledge in a fashion that enabled logic-based reasoning processes to derive inferences in a way that simulate human reasoning [4]. Finally, in the bioinformatics genomic revolution, ontologies of a more terminological bent emerged to solve challenges in the standardization of database annotations relating to genes and gene products [5,6]. These different intellectual paradigms have been synthesized into the discipline of modern biomedical ontology, or *bio-ontology* [7].

Where biomedical data are described and labeled using unconstrained text, different terminology is often used for similar or identical things. Such terminological variance is normal and reflects natural language; usually humans have no difficulty in resolving ambiguous usages of terminology and discrepant labels. However, due to the sheer volumes of research data being generated, it is necessary to develop computational methods of aggregating and aligning like with like. One approach to addressing this issue is to adopt shared standards for the categorization of data. Agreement in annotation across different databases increases the value of a standardized terminology, allowing for easier cross-domain integration and querying. Ontologies formalize the meaning of terminology used in a domain and provide a reference for disambiguation of terminology usage. Increasingly, research in the life sciences needs to integrate knowledge and results from multiple disparate fields and methodological approaches in order to gain insight into underlying biological mechanisms. This is the case, for example, when studying the genetic and epigenetic factors in understanding behavioral phenotypes, or in the development of predictive models to enable personalized and translational medicine. Research results from diverse disciplines, such as genetics, molecular biology, physiology, chemistry, psychology, and medicine, have to be integrated in order to build a coherent picture of what is known in order to address key research gaps, and ontologies are meeting this need.

Bio-ontologies represent biological or medical entities of scientific interest together with their properties and the relationships between them in a computationally processable language, enhanced with metadata such as definitions, synonyms, and references to databases. They have enjoyed increasing success in addressing the large-scale data integration requirements emerging from the recent increase in data volume [3]. The longest-standing example of a successful bio-ontology is Gene Ontology (GO) [5], which is used, *inter alia*, to unify annotations between disparate biological databases and for the statistical analysis of large-scale genetic data, to identify genes that are significantly enriched for specific functions. The Gene Ontology describes the functions, processes, and cellular components of genes and gene products, such as proteins. It is actively maintained and as of 2012 consists of more than 30 000 terms [8].

3.2.1
The OBO Foundry: Ontologies in Biology and Medicine

The Open Biomedical Ontologies (OBO) Foundry [9] is an organization that is coordinating the development of a suite of interoperable reference ontologies for scientific application domains such as biology and medicine, centered around the first-of-its-kind Gene Ontology [5]. As part of this coordination effort, the OBO Foundry requests that prospective member ontologies strive to follow a set of shared, community-agreed principles that facilitate reuse of ontologies in multiple projects and support orthogonality between the ontologies. Ontologies are orthogonal when they cover domains that do not overlap, or where a small overlap exists (such as between chemistry and biology). This is dealt with by reuse of shared identifiers between the two ontologies. Ontologies that are submitted to the OBO Foundry are first admitted to the OBO Library. They then undergo a peer review process, and if the outcome of this review process is that they display a substantial level of compliance with these guidelines, they are then included as OBO Foundry ontologies. The full list of current OBO Foundry and OBO Library ontologies is available at http://www.obofoundry.org/. At the time of writing, there were eight Foundry ontologies and just over a hundred ontologies in the Library. All ontologies may be downloaded in full from the link on the Foundry Web site, or they may be browsed in the BioPortal interface (http://bioportal.bioontology.org/) [10].

Most of the OBO Foundry principles are designed to facilitate the use of ontologies in the standardization of database annotations, which is the first use case that motivated the development of the Gene Ontology. In particular, the principles emphasize the use of stably maintained semantics-free (numeric) unambiguous identifiers for entities contained in the ontology. Such identifiers are an essential requirement if annotations to an ontology are to be created across multiple databases, because different update and release cycles will certainly result in dead links within downstream databases if IDs are allowed to disappear from the source ontology. Also, using semantics-free identifiers means that the identifiers can remain stable, while the underlying ontology, for example, labels and hierarchical organization, changes. Having clearly delineated scope for a particular ontology and not overlapping with other ontologies, another OBO principle, is also very helpful for database annotation standardization, since a plethora of similar-sounding options from different ontologies bewilders and deters curators who need to use ontologies for annotation.

Another benefit of the adoption of ontologies for data annotation is that they facilitate flexible data aggregation. Science searches for generalities and patterns in the world. Such generalities allow predictions to be made and contribute to our understanding of underlying mechanisms. Grouping individuals together in a hierarchical structure allows discovery of commonalities at different levels. Ontologies provide a very generic and flexible structure that can be organized hierarchically to arbitrary depth.

One of the most pressing challenges in data-driven biology is the proliferation of different databases. While the growth in open data is *prima facie* a good thing as

more data become available for data-driven research, a substantial portion of researcher's time is lost trying to unify the available data from multiple different resources. Unifying data based on names and other metadata is a hard, patchy work. But if the data are annotated with a shared ontology, the integration has already been done at the time of the annotation. Of course, this annotation may also be incorrect, but at least performing it in a centralized manner shifts the effort to the producer, rather than the consumer, of the data, enabling quality control to be applied at the source. Thus, to address this need for data integration, many bio-ontologies are developed as community efforts and used in annotation across multiple databases.

3.2.2
Ontology Languages and Logical Expressivity

Bio-ontologies are developed and exchanged in a shared ontology language such as the Web Ontology Language (OWL), version 2 [11] or the OBO format [12]. The OBO language was developed by the Gene Ontology project [5] to provide an ontology language that was more human-readable than the then available technical alternatives that were the precursors of the OWL language. Many of the tools that will be discussed in the following sections still rely on the OBO language with its graph-based ontology structure, but the OBO and OWL languages are now inter-convertible to a large extent [13], and increasingly the community is moving toward OWL for ontology maintenance in order to harness an expanding set of available ontology development tools. The OBO Ontologies Release Tool [14] is a software library that provides support for OBO Foundry and OBO Library ontologies to produce release versions of ontologies in both OBO and OWL format.

OWL ontologies consist of several distinct components. These include classes, which are entities in the domain, known as "terms" in the OBO format, and properties, which are known as "relations" in OBO format. Metadata such as names and synonyms are considered annotations in OWL and are ignored from the perspective of the logical properties of the language. In OWL, properties representing relationships between classes are more complex than the corresponding OBO relations, since in OWL it is necessary to capture the type of restriction that is represented by the property axiom. This may be *existential*, in that the relation expresses the knowledge that all members of the first class are related to *some* members of the other class, or *value*, in that the relation may express the knowledge that members of the first class are *only* related to members of that other class (for the specified type of property).

OWL is based on Description Logics [15], a family of decidable logical languages optimized for the expression of large-scale terminological knowledge such as is found within large biomedical vocabularies. Logical languages are defined by the types of logical axioms that they support. OWL supports many different types of logical axioms, and the combination of axiom types that is used in a given ontology defines the expressivity of that ontology. There is a trade-off between expressive power and the tractability of a logic-based language. Increasing the expressivity of the language usually means that it takes longer to perform reasoning tasks.

The axiom types available in the OWL language include the following:

- Declaring and naming atomic classes
- owl:Thing, the "top" class that is a superclass of every other class in an ontology
- owl:Nothing, the "bottom" class that is a subclass of every other class in an ontology, and is used to highlight inconsistent classes when there are errors in the ontology
- Intersection (AND)
- Union (OR)
- Negation (complement, NOT)
- One of (enumeration)
- Restriction on object property: some values
- Restriction on object property: all values
- Restriction on object property: minimum cardinality
- Restriction on object property: maximum cardinality
- Restriction on data property: value or value range

A powerful feature of OWL is the ability to perform *automatic classification* using highly optimized OWL reasoners. Reasoners are software toolkits that are able to compute inferences on the underlying logic used by the ontology – that is, to logically deduce the consequences of the knowledge that is expressed, including inconsistencies. Examples of reasoners that work with OWL ontologies are Fact++ [16], Pellet [17], and HermiT [18], each of which supports different ranges of operators and may have different performance profiles on different reasoning tasks. As an example of OWL reasoning, the following axioms are given:

$$\text{ZincAtom } \texttt{subclassOf } \text{MetalAtom} \tag{3.1}$$

$$\text{MetallicCompound } \texttt{equivalentTo } \text{Compound } \texttt{and } \text{hasAtom } \texttt{some } \text{MetalAtom} \tag{3.2}$$

$$\text{ZincOxide } \texttt{subclassOf } \text{Compound } \texttt{and } \text{hasAtom } \texttt{some } \text{ZincAtom} \tag{3.3}$$

An OWL reasoner can automatically infer from Eqs. (3.1)–(3.3) that ZincOxide is a subclass of MetallicCompound.

OWL has *open world* semantics, which means that inferences can only be drawn based on what is explicitly captured in the knowledge base, and that absence of additional information has no implication that the information doesn't exist. For example, given a statement that *Square* and *Circle* are subclasses of *Shape* in a particular knowledge base, together with the knowledge that some entity *A has_shape C*, we nevertheless cannot infer that *C* is either a *Square* or a *Circle*, as we do not know anything about which other shapes may exist that have not been specified (i.e., in the open world). If this type of inference is desired, it is necessary to add an axiom specifying that all shapes are either squares or circles. This is called a *closure axiom*.

The primary editor that is used to maintain OWL ontologies is Protégé[1]. Protégé provides an extensive framework for ontology development, allowing editing of

---

1) The Protégé ontology editing tool. http://protege.stanford.edu/ (accessed November 2012).

classes and properties, reasoner integration, and ontology visualization, as well as being supported by an extensive set of custom plug-ins.

3.2.3
Ontology Interoperability and Upper-Level Ontologies

There are a growing number of bio-ontologies, and these ontologies are increasingly used in combination with each other in support of database annotation and other objectives. When multiple ontologies are used in combination, there is a need to provide a shared view across the content of the ontologies. This generates challenges such as the need to anchor different ontologies within a common context and to remove redundancies between them. It is also important to coordinate a set of shared ontology relationships across multiple ontologies, since application logic often depends on the nature of the relationships used to encode such knowledge.

To address these challenges, upper-level ontologies provide core foundational entities and relationships that are intended to be shared and reused across multiple different domain-specific ontologies [19]. Widely used in the biomedical domain, the Basic Formal Ontology (BFO) [20,21] is one such upper-level ontology. BFO offers a small ontology of core entities that encode foundational distinctions relevant in any domain. For example, BFO distinguishes objects that endure through time and exist independently, such as humans and trees; properties that endure through time but need a bearer in order to exist, such as color; and processes that unfold in time, such as cell division. BFO also includes relationships that are relevant to multiple ontologies, such as those for parthood and participation.

The object–property–process distinction mirrors the distinction of different subontologies of the Gene Ontology [5]: the cellular component branch is concerned with the physical objects and their parts that make up the cellular environment, the molecular function branch with the properties and activities of gene products within their environment, and the biological process branch with the processes that take place within biological organisms, such as cell division and reproduction. A similar distinction can also be drawn in chemistry, between the molecules that are the objects in the domain, their properties such as mass and aromaticity, and processes that they are involved in, such as chemical reactions.

3.3
Chemical Ontologies

With the large-scale availability of chemical data through projects such as PubChem [22], making sense of the data has become one of the most pressing challenges facing researchers. Traditional large-scale data management methods in chemistry include chemical structure-based algorithmic and statistical methods for the construction of hierarchies and similarity landscapes. These techniques are essential not only for human consumption of data in the form of effective browsing and searching, but also in scientific methods for interpreting underlying

biological mechanisms and detecting bioactivity patterns associated with chemical structure [23].

For the domain of biologically interesting chemistry, the Chemical Entities of Biological Interest (ChEBI) ontology [24] is an ontology of chemical entities such as atoms, molecules, and chemical substances. ChEBI classifies chemical entities according to shared structural features, for example, carboxylic acids are all molecular entities that possess the characteristic carboxy group, and according to their activities in biological and chemical contexts, for example, acting to kill or inhibit the growth of bacterial infections. As of December 2012, ChEBI contained just over 30 000 entities. ChEBI has been harnessed in diverse use cases, including the annotation of chemicals in biological databases [25–27], the automatic identification of chemicals in natural language text [28], formalizing the chemistry underlying the GO [8], and large-scale metabolome prediction [29].

In chemistry, algorithmic and statistical methods for chemical classification and data mining are in widespread use. One advantage of logic-based methods such as ontologies is that they allow the knowledge to be explicitly expressed *as knowledge*, that is, as statements that are comprehensible, true and self-contained, and available for modification by persons without a computational background. This is in contrast to statistical methods that operate as black boxes and to procedural methods that require a programmer in order to manipulate or extend them. Algorithmic cheminformatics methods are often based on the features encoded in chemical structure representations [23]. By contrast, chemical ontologies also allow classification by nonstructure-based features. For example, many use cases demand the identification of a knowledge base of chemical entities that share functional activity in order to do primary research in a particular domain. In research into odor perception, it may be necessary to identify a knowledge base of all odorant molecules. Where the primary purpose of a research project is not primarily chemical in nature, the implementation of a targeted chemical database is a costly overhead.

ChEBI provides both a structure-based and a functional subontology, with roots "chemical entity" and "role," respectively. The "chemical entity" subontology of ChEBI is concerned with the classification of molecular entities into structure-based classes by virtue of shared structural features. Included in the chemical entity ontology are fully specified molecules such as and erythromycin A, and structure-based classes such as steroid and macrolide. These molecules may be furthermore interrelated with structural relationships such as "is enantiomer of" and "is tautomer of" where appropriate to capture closely related chemical structures. Included in the "role" subontology are drug usage classes such as antidepressant and antifungal; chemical reactivity classes such as solvent, acid, and base; and biological activity classes such as hormone [30]. Where specific targets are known, the role class is specified down to the molecular level of granularity, such as "cyclooxygenase inhibitor" for paracetamol. Chemicals are linked to roles using the "has role" relationship. ChEBI is mapped to BFO and interrelated with the Gene Ontology [31]. An overview of the content of ChEBI showcasing the important ontology classifications for the hormone oxytocin is illustrated in Figure 3.1.
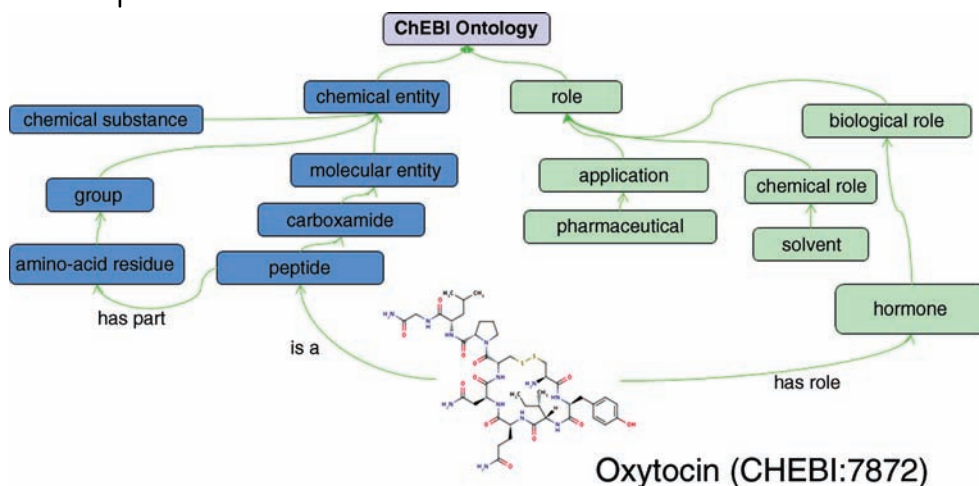
Figure 3.1 Core divisions in content in the ChEBI ontology between chemical entities and roles. Fully specified chemical entities such as the hormone molecule oxytocin are assigned role categories with the "has role" relationship and structure-based classes with the "is a" relationship. The "has part" relationship may be used to specify important parts of structure-based classes.

An area of active ongoing research at present is that of structure-based classification in chemical ontologies [32–34]. At present, ChEBI is manually maintained by a team of chemists, but efforts are underway to harness the semantics of the OWL language in combination with features of the chemical structures to provide partial automation of error detection and classification. The chemical domain benefits from a great deal of underlying regularity in the entities in the domain, that is, it is constrained by the regularities observed in real chemicals. Logical axioms expressing these underlying regularities can thus be harnessed to help manage the overhead of maintaining a large ontology. Explicitly formalizing class definitions also enables disambiguation of different class definitions that are used by different communities in reference to the same named entities. For example, some communities may use the term "hydrocarbons" as encompassing derivatives such as chlorohydrocarbons, while other communities may use the term in a stricter sense as molecules only composed of hydrogen and carbon. Making these definitions explicit enables different chemical hierarchies to be computed according to the preferred definition for a given community by classification tools, avoiding the maintenance of entirely different ontologies to accommodate such differences.

There is a fundamental challenge in the *full* integration of chemical structure-based algorithms with logic-based OWL reasoning, namely that OWL is fundamentally unable to correctly represent cyclic structures, such as molecular entities containing rings [35]. Given this limitation, progress has been made to partially represent aspects of chemical structure and expose those to OWL-based reasoning. One of the first applications of OWL for chemical classification was made by

Villanueva-Rosales and Dumontier [36], who encoded functional groups into an OWL ontology. The Dumontier representation allows powerful use to be made of the relatively compact knowledge base of functional groups in terms of the classification of arbitrary molecules in a defined hierarchy. Care, however, had to be taken to only encode the structures of the functional groups insofar as they were not cyclic. The same team extended this work more recently to include axiomatic definitions of lipid classes in the Lipid Ontology [32]. In both the cases, the classification is dependent on the detection of specific functional groups in molecular structures, which is done using normal algorithmic approaches [23]. A prototype of a self-classifying ontology for chemicals in ChEBI was presented in Ref. [33], which includes an algorithm for the discovery of shared features among groups of chemical structures, the representation of such features in an OWL ontology, and the automatic classification of that ontology using OWL reasoning. The features that are detected include common functional groups and, additionally, the presence of charges and cycles.

We have identified the following types of structural features used in chemical class definitions in Ref. [34]:

- *Presence or absence of specific parts.* As ChEBI currently includes molecular parts beneath a "group" hierarchy and uses the "has part" relationship to relate the full molecules to their parts, this feature can already be accommodated in an OWL-based chemical ontology such as ChEBI. However, the representation of such parts is one part of the problem, and the detection of matches between parts and whole molecules to enable automated classification is another. In the general case, including also cyclic molecules, in order to do something similar to substructure matching to classify molecules based on interesting parts of their structures, a different formalism to OWL is needed for the chemical ontology. Description Graphs have been proposed as one such formalism [35].
- *The number or count of specific parts.* While ChEBI does not capture this information at present, it is within the scope of the OWL formalism through capturing *cardinality* constraints on the "has part" relationships. For example, it could be expressed that a tricarboxylic acid is a class of chemical entities that "has part" minimum three carboxy groups. The algorithms reported in Ref. [33] were capable of discovering such constraints on classes. However, within that approach it is not possible to exclude the inference that a tricarboxylic acid – with three carboxy groups – is also a dicarboxylic acid – with *at least two* carboxy groups. It is trivially true that a chain of n methylene groups is also a chain of $(n-1)$ methylene groups. However, it would be misleading to describe a molecule with an attached dodecyl group as a methylated compound simply because it contains a substructure with the formula CH_3 at the end of the alkyl chain.
- *Calculated properties of the chemical such as overall charge or molecular weight.* OWL is capable of handling data properties such as integers or strings. These allow the definition of classes referring to particular values or value ranges. For example, it is possible to define *small* molecules as molecules whose molecular weight is less than 800 Da. These data restrictions are supported by OWL reasoners just as are object property restrictions.

- *Topological features such as polycyclic cages or molecular knots.* Topological features such as overall regularity or cage structure fall outside the expressivity of OWL, although efforts have investigated the use of higher-order logics for that purpose [37].
- *Structural formulas, such as hydrocarbons (strictly defined, excluding heteroatoms), in which atoms of types other than hydrogens and carbons are absent.* Due to the open world semantics of OWL, everything that is not explicitly stated in the ontology is assumed to be *not known to hold* rather than *known not to hold*. For example, ChEBI contains "organic molecule" and "inorganic molecule" as two classes. However, if it is not explicitly stated in the ontology that *all* molecules are either organic or inorganic, the ontology cannot infer that a molecule is inorganic simply from a statement to the effect that it is not organic. This open world property of the semantics is a challenge for strict and exclusive class definitions. This is an area where the closed world semantics of alternative formalisms, such as the description graphs formalism described in Ref. [35], may be more practical. A particularly challenging class to define within a logic-based formalism is those captured by a parameterized molecular formula, such as alkenes which are described by the formula C_nH_{2n}. Constraints on number of atoms of particular sorts can be expressed using OWL cardinality restrictions, but this facility does not allow the relationship between the number of carbons and the number of hydrogens to be expressed.

3.4
Standardization

Bio-ontologies following the tradition of the Gene Ontology [5], including ChEBI [24], were primarily developed to enable standardization across bioinformatics databases. Ontologies enable standardization through the use of semantics-free stable identifiers, the annotation of extensive synonyms and cross-references, and through striving to represent community agreement about the entities in a given domain such that the ontology can be widespread among different subgroups of the wider community. In systems biology, for example, the use of unambiguous, publicly available, and perennial identifiers for model components is becoming increasingly recognized as being essential for sharing and reusing models [38]. Ontology-based identifiers are required if such models are to be used in computational pipelines [2]. The same applies to many scientific domains, including chemistry. The ChEBI ontology is already widely used in model and pathway annotation. ChEBI is the primary chemical annotation and identification standard in the BioModels [26] and Reactome [25] databases and in the Gene Ontology [8]. ChEBI is also listed as a secondary identifier for chemical information in many additional databases such as KEGG [39], DrugBank [40], and HMDB [41], enabling ontology-based data integration. To further facilitate data integration, ChEBI maintains an extensive set of database cross-references as metadata associated with classes in the ontology.

The semantic Web is a worldwide effort in which data are being brought online from heterogeneous databases in the readily integratable "triple" format (subject,

predicate, object). Cheminformatics data are being brought onto the semantic Web in great volumes [42]. Bringing data onto the semantic Web allows it to be used remotely, without the data having to be downloaded and stored locally on the researcher's machine. Semantic Web-enabled software fetches desired data from distributed repositories that support cross-resource query answering over multiple data sources. A key challenge that arises in this context over and above the data integration challenges that ordinary data warehouse applications face is managing the provenance of the information coming online in order to track and deal appropriately with different levels of data quality. Different vocabularies may refer to the same sorts of data with different labels or identifiers. For example, molecular weight as a property of a molecule may be referred to in one database as "MWEIGHT", in another as simply "WEIGHT" or even "MASS," in another as "MOLWEIGHT," and so on. These different labels obscure the fact that the data are comparable and should be integrated, thus "hiding" portions of the data from algorithmic processes of extraction. On the other hand, multiple implementations of an algorithm may use the same terminology, while they can produce different outputs due to heuristics, optimizations, errors, or outright differences in the interpretation of the terms. This can lead to incorrect deductions when the results of calculations are made available under the same terminological label without further provenance as to which implementation was used to generate the data. To address these challenges in the domain of calculated or measured chemical data being brought online, the Chemical Information (CHEMINF) ontology was developed [43].

Both measurement and prediction of property values are ways to derive information about chemical or biological properties and represent them so that they can be accessible for further research. Properly reproduced on the semantic Web, such values can be reused in multiple scientific analyses. This highlights the importance of maintaining the *provenance* of the information – detailing the algorithms that were used to generate calculated property values and/or the experimental conditions under which data were generated. To address this need, the CHEMINF ontology includes entities for different types of chemical property, algorithm, toolkit, and descriptors, with definitions and axioms describing what they are specifically about.

The use of standard identifiers and ontologies is of particular importance in the context of the semantic Web to enable cross-resource integration and querying. An example of chemical structures and properties being brought onto the semantic Web is described in Ref. [44].

3.5
Knowledge Discovery

Newly generated knowledge across many different research areas is reported in the primary scientific literature. However, the body of literature is growing at such a rate that it is not possible to stay on top of all developments in a given field.

Computational processing of the primary literature in order to identify publications of interest or to amass all the publications in which specific entities are mentioned is increasingly important. Such literature reports are often phrased in terms of classes of chemical entities rather than individual fully specified molecules. Biological knowledge such as the actions of enzymes in biological pathways is also described in terms of classes rather than individual molecules. For this reason, chemical ontologies are very important for the identification of relevant entities in natural language text, and ChEBI has been used for "chemical text mining" applications [28].

A closely related task is that of computing systematic names for chemical structures and reverse-engineering structures from specified names. IUPAC naming rules for compounds such as described in Ref. [45] and implemented in tools, including the open source Opsin [46], provide a method for obtaining a systematic name from a given chemical structure and for interpreting a name to determine the intended underlying structure. Importantly, rules for chemical naming in IUPAC confer similar information to the classification of molecular entities into hierarchies in the sense that parts of a chemical name correspond to parts of the molecule, and the same parts of the molecule are also used for parts-based classification. Thus, there could be a close integration between software that computes names and software that computes classification. However, it should be noted that IUPAC rules generate systematic names, which can be unwieldy and lengthy, and that chemists in many cases prefer to use shorter trivial names such as "caffeine". Such trivial names cannot be automatically computed and need to be stored in a knowledge base such as ChEBI.

Text processing and knowledge discovery has been enhanced by implementation of ontology-based similarity for improved classification and text mining [47]. Similarity has many important applications in life science research. Most importantly, similar entities can be expected to behave similarly in similar contexts [23]. We can learn something about unknown entities, and make predictions about their behavior, by examining knowledge about similar entities. But for this endeavor to yield the best results, the measures used for similarity must have real biological relevance.

Semantic similarity is named as such because it encodes similarity measures that harness deeper features than only the superficial structure of an entity. For example, the two words "cell" and "cell" are structurally identical, but may be semantically very different in meaning – if one means "cell" as the word is used in biology, and the other "cell" meaning the place where prisoners sleep. Ontologies aim to encode many relevant portions of information about the meaning of the entities they represent, through the topological graph structure around a particular node, the different relationships used, and in the case of more expressive OWL ontologies, the features of the logical axioms. In chemistry, the most commonly used measure for calculating similarity compares features of the chemical structure to quantify a score. Chemical structural representations depict the atoms and bonds from which the chemical is composed, with information about their types and local configuration. Similarity algorithms pick out structural features from such representations and convert them into a bit array (a string of 1s and 0s) or "fingerprint", which is then able to be rapidly and numerically compared for similarity with the fingerprint from another structure using (usually) something
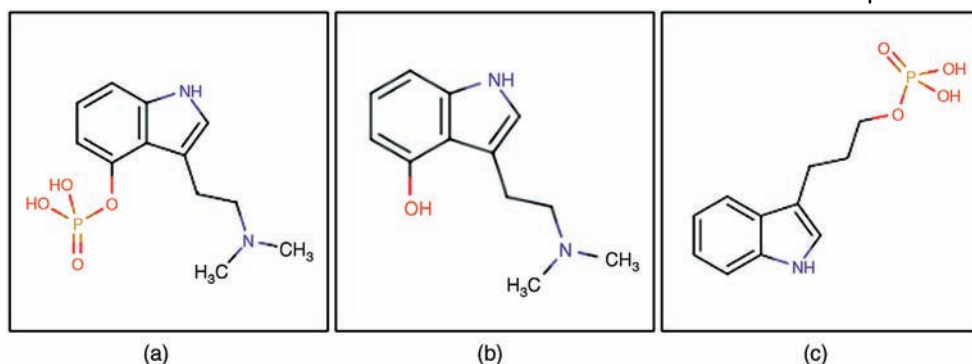
Figure 3.2 The chemical structure for (a) psilocybin (CHEBI:8614), (b) psilocin (CHEBI:8613), and (c) 3-(indol-3-yl)propyl phosphate (CHEBI:28162).

like the Tanimoto score. Psilocybin (CHEBI:8614), psilocin (CHEBI:8613), and 3-(indol-3-yl)propyl phosphate (CHEBI:28162) are structurally very similar, as illustrated in Figure 3.2.

ChEBI gives the similarity score between psilocybin and psilocin at 76%, and between psilocybin and 3-(indol-3-yl)propyl phosphate at 75%. Different databases might give different pairwise scorings depending on the underlying features used in the comparison. Human chemists may also give different judgments about the relative similarity of the two pairs of compounds depending on the application they have in mind. However, the point to note is that it is clearly not possible to say much of significance based on a mere distinction of 1% in similarity. While 3D shape-based similarity measures may yield better results in some cases, many organic molecules are flexible and thus may adopt multiple conformers, hindering computation of shape-based similarities. In terms of a semantic similarity using the ontology relationships captured in ChEBI, however, we can immediately observe rather strong differences between these two pairs. While both psilocybin and psilocin are classified as tryptamine alkaloids (CHEBI:48274) and as having the role hallucinogen (CHEBI:34599), 3-(indol-3-yl)propyl phosphate has immediate structural parent monoalkyl phosphate (CHEBI:25381) and no role annotated. In terms of having strong biological relevance, the first pair is much closer in similarity than the second pair. This is not reflected in the structural similarity measure, but would be reflected in a semantic similarity measure using the relationships asserted in the ChEBI ontology. A hybrid similarity metric for chemical entities, called Chym, has been developed that combines a structure-based similarity measure with a semantic similarity measure based on ChEBI [47]. In application of Chym to several classification problems, the authors have shown that the hybrid measure yields better (more biologically meaningful) results than a "straight" structural similarity measure.

CMPSim [48] is another Web tool that also uses the information contained in the ChEBI ontology in order to calculate similarity. This time, the tool is used to quantify the similarity between metabolic pathways. This tool can be used to quickly find the

semantic similarity between KEGG pathways. To do so, a pathway is mapped into the ChEBI compounds that participate in it, and then the two pathways are compared by comparing the ChEBI similarities.

3.6
Data Mining

Ontology-annotated data are organized by the structure of the ontologies into categories that can serve as the framework around several statistical data mining techniques. One technique that makes particularly prominent use of the structure of ontologies is *ontology-based enrichment analysis*. In an ontology-based enrichment analysis, an annotated data set is compared to a background set of annotations to find whether some of the ontology categories are statistically overrepresented in the annotated data as compared to the background annotations. This technique has found widespread use in functional genomics research, in which sets of genes annotated in the Gene Ontology are compared to the full set of GO annotations to determine which functional ontology categories are over- or underrepresented in the sample in question. There are many variations on this theme in different implementations. For example, the genes of interest may be selected as those overexpressed in a particular microarray experiment and the background reference set may be the full set of genes present in the microarray experiment, rather than the full set of genes annotated for a given species [49].

An example of a Gene Ontology based enrichment tool is the Biological Networks Gene Ontology tool (BiNGO) [50], which is available as a plug-in to the Cytoscape network analysis toolkit [51]. BiNGO assesses overrepresentation or underrepresentation of GO categories for a set of genes as compared to the GO annotations for a particular species. It is fully interactive with the Cytoscape network visualization software, and can take as input genes from selected subsets of networks in the main Cytoscape network view. Of particular relevance here is that BiNGO is also able to work with custom ontologies and annotation sets. For example, BiNGO is able to load ChEBI annotations to the Gene Ontology to do GO enrichment for a set of chemicals, or ChEBI annotations to ChEBI role classes to perform role enrichment for a set of chemicals. Figure 3.3 shows an example output of the BiNGO tool running against the ChEBI role ontology with input a set of metabolites that were implicated in a study as having some involvement in bipolar disorder. It is interesting to note that the neurotransmitter role is enriched in this set of metabolites, as is the osmolyte role. This result, while not unexpected in this case, provides support for the neurochemical mechanisms believed to underlie bipolar disorder.

A Web-based tool that performs ChEBI role enrichment for metabolic data is MBRole [52]. MBRole performs enrichment analysis from ChEBI role annotations that are grouped into biological roles, chemical roles, and applications. MBRole is also able to perform pathway enrichment analysis using KEGG pathways. Data mining approaches have also been used to establish links between drugs and
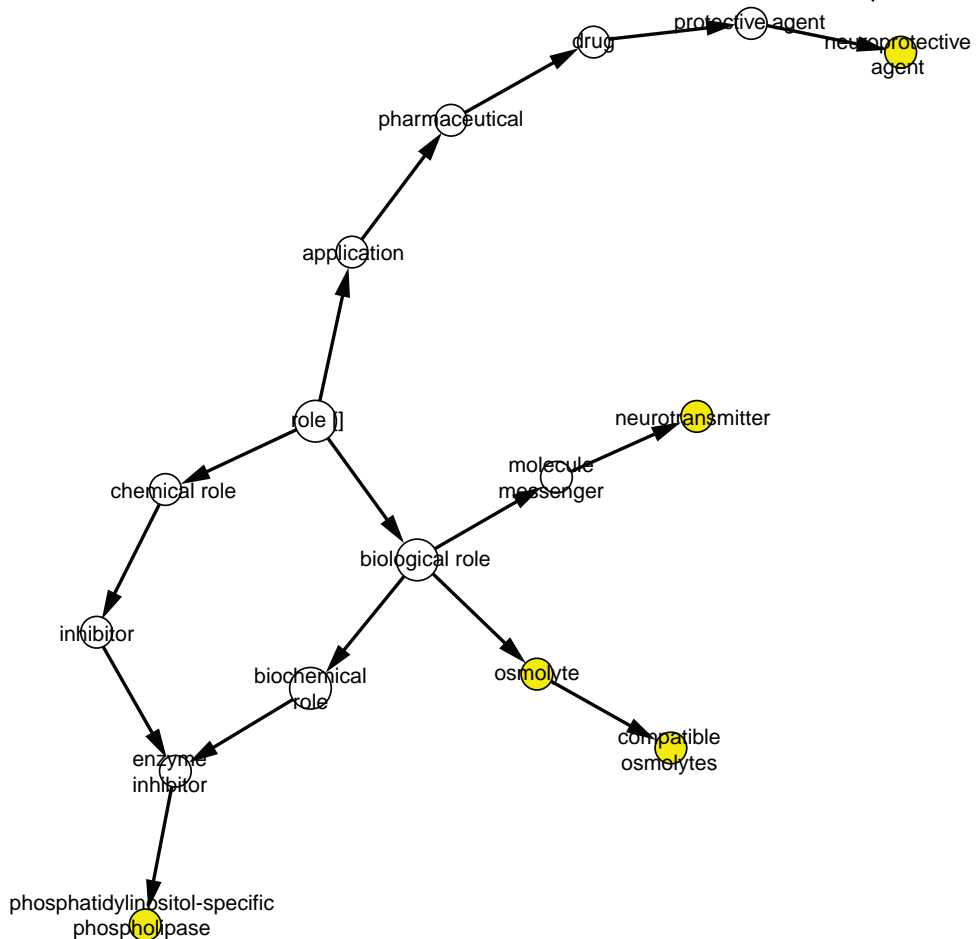
Figure 3.3 A ChEBI role ontology enrichment result for a set of metabolites implicated as having an involvement in bipolar disorder, using the BiNGO tool. The color of the nodes indicates the p-value of the enrichment, with white being not significant (those nodes are included just to build the graph output) and orange very significant. The visualization is customizable within the Cytoscape framework.

pathways [53]. In this utility, a large-scale mapping between disease pathways and chemicals that may be used to perturb them is provided through the integration of information about drugs, genes, diseases, and pathways. The approach uses a multiontology enrichment analysis with a human disease ontology and a chemical ontology in combination.

An interesting data mining challenge is the prediction of ontology annotations for as-yet unannotated data. The Gene Ontology annotation team uses a set of rules for automatically predicting ontology annotations for as-yet unannotated data, but these are mainly based on knowledge about the orthology of genes

rather than data mining existing annotations or examining the structure of the ontology [54]. Ontologies provide categories for data to be aggregated into; such categories can then form the input to machine learning algorithms and statistical models of various sorts, which in turn can provide predictions for novel data. Supervised methods, such as Bayesian classifiers, decision trees, and support vector machines, can be used to classify compounds for a particular functional activity class. However, these approaches result in binary output for a particular class membership rather than allocation of compounds to the ontology data set. Supervised machine learning for prediction of chemical class membership based on an existing hierarchy would require large training sets of chemicals that are already classified into such a hierarchy. Although ChEBI could in principle act as such a training set, the size of the classified data is still a tiny fraction of the enormous chemical space, and the problem is further complicated by the fact that the leaf nodes thus far contain fairly few structures. Our research on methods to automatically extend the ChEBI classification aims to address this parcity of leaf nodes. Manually constructed classifications may furthermore be far from complete in the sense that an arbitrary compound belongs to a vast number of classes, yet will only have been classified under one or two – those deemed to be the most relevant. Automated reasoning will also serve to address this shortcoming, although the determination of the most relevant subset of classifications for a given chemical may remain peculiarly a human ability for some time in the future. Another interesting problem is matching the categorization inferred by purely data-driven approaches, such as clustering across a given research data set, with that created by human annotators and encoded in the ontology.

3.7
Conclusions

We have given some background to ontologies as they are in use in the biomedical sciences, highlighted the state of the art in chemical ontologies, and described several application areas in which ontologies are being used to support knowledge discovery, standardization, and data mining. In general, the standardization effect of ontology use makes ontology-annotated data highly suitable for all other aspects of data mining, including specifically meta-analysis, as it reduces the overhead of data integration in making use of the data.

An exciting direction that life sciences ontology development is currently expanding into is that of increasing interrelationships between the ontologies themselves. For example, relationships links from the Gene Ontology to the ChEBI ontology are being created to explicitly represent chemical participation in biological processes. This shows vast potential for enabling the sort of whole-systems scientific modeling that is needed to transform basic knowledge about biology into predictive models and simulations that allow scientists to design and investigate perturbations for explicit therapeutic endpoints *in silico*.

Acknowledgments

References

1 Lambrix, P. (2004) Ontologies in bioinformatics and systems biology, in *Artificial Intelligence Methods and Tools for Systems Biology (Computational Biology)*, vol. 5 (eds W. Dubitzky, F. Azuaje, A. Dress, M. Vingron, G. Myers, R. Giegerich, W. Fitch, and P.A. Pevzner), Springer, The Netherlands, pp. 129–145.

2 Courtot, M., Juty, N., Knüpfer, C., Waltemath, D., Zhukova, A., Dräger, A., Dumontier, M., Finney, A., Golebiewski, M., Hastings, J., Hoops, S., Keating, S., Kell, D.B., Kerrien, S., Lawson, J., Lister, A., Lu, J., Machne, R., Mendes, P., Pocock, M., Rodriguez, N., Villeger, A., Wilkinson, D.J., Wimalaratne, S., Laibe, C., Hucka, M., and Novère, N.L. (2011) Controlled vocabularies and semantics in systems biology. *Molecular Systems Biology*, **7**, 543.

3 Harland, L., Larminie, C., Sansone, S.A., Popa, S., Marshall, M.S., Braxenthaler, M., Cantor, M., Filsell, W., Forster, M.J., Huang, E., Matern, A., Musen, M., Saric, J., Slater, T., Wilson, J., Lynch, N., Wise, J., and Dix, I. (2011) Empowering industrial research with shared biomedical vocabularies. *Drug Discovery Today*, **16** (21–22), 940–947. doi: 10.1016/j.drudis.2011.09.013.

4 Gruber, T.R. (2009) Ontology, in *Encyclopedia of Database Systems* (eds L. Liu and M.T. Özsu), Springer.

5 The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

6 Hunter, L. (2002) Ontologies for programs, not people. *Genome Biology*, **3**, 1002.1–1002.2.

7 Smith, B. (2003) Ontology, in *Blackwell Guide to the Philosophy of Computing and Information* (ed. L. Floridi), Blackwell, Oxford, pp. 155–166.

8 The GO Consortium (2011) The Gene Ontology: enhancements for 2011. *Nucleic Acids Research*, **40** (DI), D559–D664.

9 Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., and Lewis, S. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, **25** (11), 1251–1255. doi: 10.1038/nbt1346.

10 Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., and Musen, M.A. (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, **37** (Web Server issue), W170–W173. doi: 10.1093/nar/gkp440.

11 Grau, B.C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., and Sattler, U. (2008) OWL 2: The next step for OWL. *Web Semantics*, **6**, 309–322. doi: 10.1016/j.websem.2008.05.001.

12 The Gene Ontology Consortium (2012) The OBO language, version 1.2., http://www.geneontology.org/GO.format.obo-1_2.shtml (accessed October 2012).

13 Hoehndorf, R., Oellrich, A., Dumontier, M., Kelso, J., Rebholz-Schuhmann, D., and Herre, H. (2010) Relations as patterns: bridging the gap between OBO and OWL. *BMC Bioinformatics*, **11** (1), 441. doi: 10.1186/1471-2105-11-441.

14 Mungall, C., Dietze, H., Carbon, S., Ireland, A., Bauer, S., and Lewis, S. (2012) Continuous integration of open biological ontology libraries, http://bio-ontologies.knowledgeblog.org/405 (accessed October 4, 2012).

15 Baader, F., Calvanese, D., McGuiness, D., Nardi, D., and Patel-Schneider, P. (2003) *Description Logic Handbook*, 2nd edn, Cambridge University Press.

16 Tsarkov, D. and Horrocks, I. (2006) FaCT++ description logic reasoner: system description. *Proceedings of the International Joint Conference on Automated Reasoning (IJCAR 2006)*, Springer, pp. 292–297.

17 Sirin, E., Parsia, B., Cuenca Grau, B., Kalyanpur, A., and Katz, Y. (2007) Pellet: a practical OWL-DL reasoner. *Journal of Web Semantics*, **5**, 51–53.

18 Shearer, R., Motik, B., and Horrocks, I. (2008) HermiT: a highly-efficient OWL reasoner, in *Proceedings of the 5th Workshop on OWL: Experiences and Directions*, vol 432 (eds C. Dolbear, A. Ruttenberg, and U. Sattler), CEUR-WS, Karlsruhe, Germany. Available online at http://ceur-ws.org/Vol-432/.

19 Smith, B. and Ceusters, W. (2010) Ontological realism as a methodology for coordinated evolution of scientific ontologies. *Applied Ontology*, **5**, 139–188.

20 Smith, B. and Grenon, P. (2004) The cornucopia of formal ontological relations. *Dialectica*, **58**, 279–296.

21 Smith, B. (1998) The basic tools of formal ontology, in *Formal Ontology in Information Systems*, IOS Press.

22 Bolton, E., Wang, Y., Thiessen, P.A., and Bryant, S.H. (2008) PubChem: integrated platform of small molecules and biological activities, *Annual Reports in Computational Chemistry* (eds A.W. Ralph and C.S. David), vol. 4, Elsevier.

23 Wegner, J.K., Sterling, A., Guha, R., Bender, A., Faulon, J.L., Hastings, J., O'Boyle, N., Overington, J., van Vlijmen, H., and Willighagen, E. (2012) Cheminformatics. *Communications of the ACM*, **55** (11), 65–75.

24 de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., and Steinbeck, C. (2010) Chemical Entities of Biological Interest: an update. *Nucleic Acids Research*, **38**, D249–D254.

25 Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, E. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, **37**, D619–D622.

26 Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M., Snoep, J., Hucka, M., Le Novère, N., and Laibe, C. (2010) BioModels database: an enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, **4**, 92.

27 Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R., and Hermjakob, H. (2007) Intact: open source resource for molecular interaction data. *Nucleic Acids Research*, **35**, D561–D565.

28 Corbett, P. and Murray-Rust, P. (2006) High-throughput identification of chemistry in life science texts, in *Computational Life Sciences II* (eds M. Berthold, R. Glen, and I. Fischer), Springer, Berlin, pp. 107–118.

29 Swainston, N., Jameson, D., Li, P., Spasic, I., Mendes, P., and Paton, N.W. (2010) Integrative information management for systems biology, in *Proceedings of the 7th International Conference on Data Integration in the Life Sciences*, Springer, Berlin, pp. 164–178.

30 Batchelor, C., Hastings, J., and Steinbeck, C. (2010) Ontological dependence, dispositions and institutional reality in chemistry, in *Proceedings of the 6th Formal Ontology in Information Systems Conference* (eds A. Galton, and R. Mizoguchi), IOS Press, Toronto.

31 Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., and Steinbeck, C. (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research*, **41** (Database issue) D456–D463, doi: 10.1093/nar/gks1146.

32 Chepelev, L., Riazanov, A., Kouznetsov, A., Low, H.S., Dumontier, M., and Baker, C. (2011) Prototype semantic infrastructure for automated small molecule classification and annotation in lipidomics. *BMC Bioinformatics*, **12** (1), 303.

33 Chepelev, L.L., Hastings, J., Ennis, M., Steinbeck, C., and Dumontier, M. (2012) Self-organizing ontology of biochemically relevant small molecules. *BMC Bioinformatics*, **13**, 3.

34 Hastings, J., Magka, D., Batchelor, C., Duan, L., Stevens, R., Ennis, M., and Steinbeck, C. (2012) Structure-based classification and ontology in chemistry. *Journal of Cheminformatics*, **4** (1), 8, http://www.jcheminf.com/content/4/1/8. doi: 10.1186/1758-2946-4-8.

35 Magka, D., Motik, B., and Horrocks, I. (2011) Modelling structured domains using description graphs and logic programming, Technical Report, Department of Computer Science, University of Oxford.

36 Villanueva-Rosales, N. and Dumontier, M. (2007) Describing chemical functional groups in OWL-DL for the classification of chemical compounds. Proceedings of OWL: Experiences and Directions (OWLED 2007), CEUR, Austria.

37 Kutz, O., Hastings, J., and Mossakowski, T. (2012) Modelling highly symmetrical molecules: linking ontologies and graphs, in *Artificial Intelligence: Methodology, Systems, and Applications* (eds A. Ramsay and G. Agre), Springer, Berlin, pp. 103–111. doi: 10.1007/978-3-642-33185-5_11.

38 Swainston, N., Smallbone, K., Mendes, P., Kell, D.B., and Paton, N.W. (2011) The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. *Journal of Integrative Bioinformatics*, **8**, 186.

39 Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, **34**, D354–D357. doi: 10.1093/nar/gkj102.

40 Wishart, D., Knox, C., Guo, A., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, **34**, D668–D672. doi: 10.1093/nar/gkj067.

41 Wishart, D.S., Knox, C., Guo, A.C.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J.A., Lim, E., Sobsey, C.A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H.J., and Forsythe, I. (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research*, **37** (Database issue), D603–D610. doi: 10.1093/nar/gkn810.

42 Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., and Wild, D. (2010) Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics*, **11** (1), 255. doi: 10.1186/1471-2105-11-255.

43 Hastings, J., Chepelev, L., Willighagen, E., Adams, N., Steinbeck, C., and Dumontier, M. (2011) The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. *PLoS ONE*, **6** (10). e25513. doi: 10.1371/journal.pone.0025513.

44 Willighagen, E.L. and Wikberg, J.E.S. (2010) Linking open drug data to cheminformatics and proteochemometrics, in *SWAT4LS-2009: Semantic Web Applications and Tools for Life Sciences, CEUR – Workshop Proceedings*, vol. 559 (eds M.S. Marshall, A. Burger, P. Romano, A. Paschke, and A. Splendiani), CEUR.

45 McNaught, A.D. and Wilkinson, A. (1997) *IUPAC Compendium of Chemical Terminology (the "Gold Book")*, 2nd edn, Blackwell Scientific Publications, Oxford. doi: 10.1351/goldbook.

46 Lowe, D.M., Corbett, P.T., Murray-Rust, P., and Glen, R.C. (2011) Chemical name to structure: Opsin, an open source solution. *Journal of Chemical Information and Modeling*, **51** (3), 739–753. doi: 10.1021/ci100384d.

47 Ferreira, J.a.D. and Couto, F.M. (2010) Semantic similarity for automatic classification of chemical compounds. *PLoS Computational Biology*, **6** (9), e1000937. doi: 10.1371/journal.pcbi.1000937.

48 Grego, T., Ferreira, J.D., Pesquita, C., Bastos, H., Vicosa, D.V., Freire, J., and Couto, F.M. (2010) Chemical and

metabolic pathway semantic similarity, Technical Report, University of Lisbon, Faculty of Sciences, LASIGE.

49 da Huang, W, Sherman, B.T., and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, **37** (1), 1–13.

50 Maere, S., Heymans, K., and Kuiper, M. (2005) BiNGO: a cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics (Oxford, England)*, **21** (16), 3448–3449. doi: 10.1093/bioinformatics/bti551.

51 Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks.

Genome Research, **13** (11), 2498–2504. doi: 10.1101/gr.1239303.

52 Chagoyen, M. and Pazos, F. (2011) MBRole: enrichment analysis of metabolomic data. *Bioinformatics (Oxford, England)*, **27** (5), 730–731. doi: 10.1093/bioinformatics/btr001.

53 Hoehndorf, R., Dumontier, M., and Gkoutos, G.V. (2012) Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. *Bioinformatics (Oxford, England)*, **28** (16), 2169–2175. doi: 10.1093/bioinformatics/bts350.

54 Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. (2004) The Gene Ontology annotation (GOA) database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, **32** (Suppl. 1), D262–D266. doi: 10.1093/nar/gkh021.

4
Building a Corporate Chemical Database
Toward Systems Biology

Elyette Martin, Aurélien Monge, Manuel C. Peitsch, and Pavel Pospisil

4.1
Introduction

Effective data mining of databases populated with chemicals and their known bioactivity data requires there to be a structured description of the compounds that are represented by accurate structural representations and names, with clearly defined terms used to describe biological activities associated with individual compounds. Even though many compound names used in publicly available databases are considered to be "standard" (IUPAC naming convention, CAS Registry Number®, PubChem IDs, or SMILES codes), when building a robust corporate in-house database, the quality and the accuracy of chemical representations and nomenclature should be of a higher standard. In other words, common identifiers are not sufficient. For example, a CAS number might not exist for a newly synthesized molecule, or the IUPAC name is not sufficient to describe a molecule with complex stereochemistry.

Here, we present our concept for building a corporate chemical registration system that we refer to as the Unique Compound and Spectra Database (UCSD), which allows chemists to register small molecules in a unique, nonambiguous way. The system can accurately handle complex chemical structures, small molecular mixtures, stereoisomers, and molecules with nondetermined structures and register them as unique records. In Section 4.2, the process for associating molecules with their analytical chemistry data is presented. The addition of analytical spectra from techniques such as NMR, chromatography, or mass spectrometry serves as supporting evidence for compound identification and purity and allows scientists to make comparisons versus known reference spectra. In this way, the UCSD increases user confidence in the accuracy of registered records. In addition, the concept of UCSD suggests what database features should be considered when building a database for chemicals present in complex chemical matrices. Examples from our experience of building such a system within R&D at Philip Morris International, Inc. (PMI) are presented.

Before any kind of data mining of a chemical and biological database is undertaken, it is crucial to have a clearly defined and unambiguous way to extract and

manage data from different sources such as external databases and literature and record the requested information in a coherent in-house system. Unlike in-house databases that are fully controllable, public databases use standard or common chemical names and therefore may contain ambiguous names and synonyms. In Section 4.7, our method for linking records from the UCSD to popular compound activity databases is briefly presented. Further integration of such organized chemical data with pharmacological, toxicological, and biological databases together with advanced computational data mining approaches (as presented in this book) becomes an important component for the expanding systems biology-based drug discovery efforts.

4.2
Setting the Scene

Small molecule chemistry is of central importance for R&D in many companies in diverse areas such as pharmaceutical, nutraceutical, food flavoring and processing, tobacco, and cosmetic industries. These institutions all face similar problems, including the question of how best to register and store small molecule information in their corporate collections. The complexity of compound registration is also significantly increased when two or more compounds are required to be registered together as a mixture, which has particular mixture-specific properties. In general, all scientists working in this arena are faced with the same questions, namely, which technology should be used, which type of data should be stored, how will physical samples of molecules be managed, how will the uniqueness of chemical structures be defined, and how will the correctness of chemical structures entered by chemists be ensured. Surprisingly, this topic is rarely covered in scientific publications, and few insights can be gained from chemoinformatics books [1–5]. This is partly because the development of such registration systems is a very technical challenge for developers and partly because it is a rapidly evolving field. As such, this chapter is based upon the work performed within R&D at PMI, which was recently published in the Ref. [6].

In the absence of a common chemistry registration platform, chemical information is generally retained in a variety of locations, as illustrated in Figure 4.1. Lists are kept at the team or scientist level in diverse formats such as Excel files or ISIS/Base entries[1]. Transferring the data from several locations into a single registration system poses a challenge because in many cases the information related to a molecule is incomplete (e.g., some data are not provided or are not provided in a uniform way) and there is often no molecular structure available, just a name. Building a centralized unique compound database is the first step to tackle this issue.

---

1) ISIS/Base is an information management framework that provides extensive chemical representation features and capabilities for searching chemical structures. It is now part of Accelrys.com.
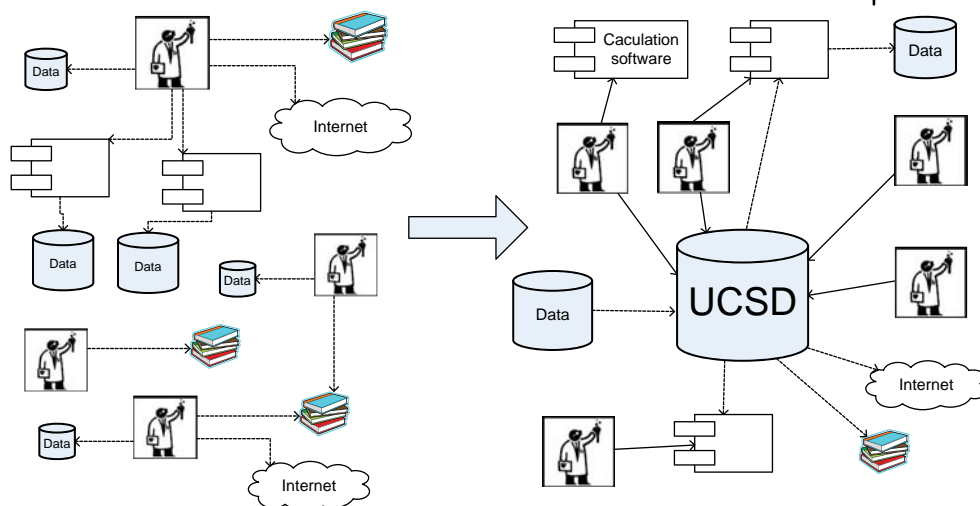
Figure 4.1 Centralizing chemical data. Without a designated data registry, scientists consult publications and Web sites, create or explore databases, and simulate data in processes (boxes) independent from each other (left side). The challenge of connecting different data formats, activities, and scientific approaches can be met by creating a unique chemical registration system (right side).

4.2.1
Concept of Molecule, Substance, and Batch

Chemical data in the UCSD are organized into different interlinked categories. The three main entities are "molecule," "substance," and "batch," the definitions for which may differ slightly from other chemical registration systems. In the UCSD, these terms are defined as follows:

- *Molecule* is the neutral form of a chemical structure without any charge, counterion, or hydrate. If a molecule is charged, the system converts it to its neutral equivalent and records its salt form at the *substance* level.
- *Substance* is equivalent to a molecule with additional information for the salts (its counterion or hydrate) associated with the molecule at the *substance* level. For neutral molecules, *substance* is recorded as *molecule* with the note "No Salt." For charged molecules, each *substance* is annotated with its type of counterion(s) or hydrate(s) and its coefficient (e.g., Na^+, 2). An exception is made for substances containing quaternary ammonium cations. Because they remain permanently charged, independent of solution pH, it is not logical to create a neutral form for registration. In this case, the system does not neutralize the molecule.
- *Batch* is defined as an occurrence of a compound. In many companies, this is usually a physical sample, for example, a compound synthesized in the laboratory. In our company it can be a compound identified by mass spectrometric methods
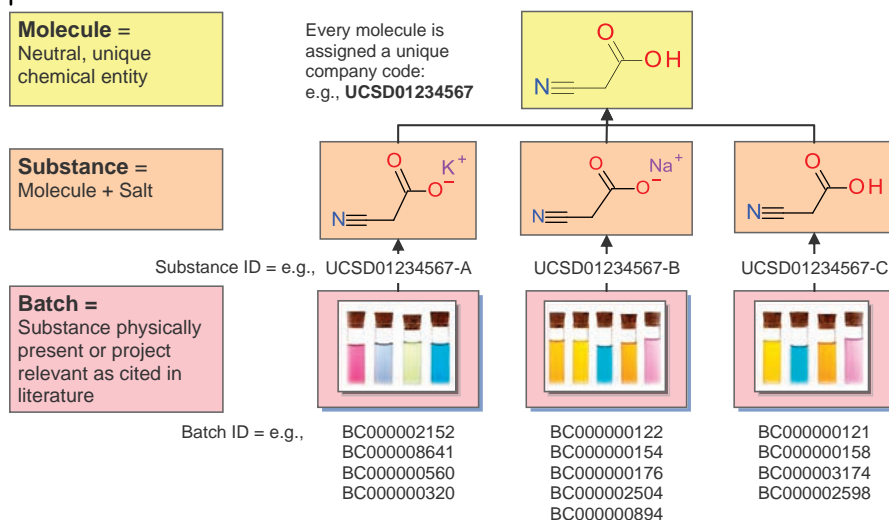
Figure 4.2 Hierarchy of molecule, substance, and batch entities in the UCSD. An example of the hierarchy is shown for 2-cyanoacetic acid. Three substances are associated with the molecule: two salts and one neutral substance. Each substance gets a letter after the molecule code (e.g., A in UCSD01234567-A). Several batches are registered for each substance. Batch codes are generated incrementally upon registration and also serve as the submission ID.

from the tobacco plant, a material purchased for analysis, or even a nonphysical sample represented as a compound cited from the relevant literature.

In the UCSD, one molecule can be linked to several substances (e.g., salt, hydrate) and each substance may have several batches (Figure 4.2). This allows users from different teams to enter batches (e.g., different laboratory procedures) related to the same substance and molecule, which can be a new one or a molecule already registered within the system. Each batch can only be linked to a single substance that in turn can only be linked to a single molecule. Batches are stored with experimentally relevant information entered manually by submitters. At the *molecule* level, a single molecule is represented by a unique, property-independent code generated by the registration system (e.g., "UCSD code" or "PMI code" at PMI). The substance codes are the same as the molecule codes, but a unique letter is added to distinguish each individual salt/hydrate. Batches are assigned their own unique codes (which eases the process of batch reassignment, see Section 4.4.4), which are generated incrementally during the submission process. Batch codes also serve as the submission ID for submitters and registrars.

4.2.2
Challenge of Registering Diverse Data

The source of chemical compounds to be stored in a chemical registration system greatly depends upon the business of the company. For instance, in pharmaceutical

companies, compounds are synthesized internally or purchased from external suppliers and can represent millions of individually identified chemical structures. In the flavor and fragrance industry, compounds are often natural products extracted from plants.

For the tobacco industry in general, and for PMI R&D in particular, chemical constituent sources are relatively limited in comparison with traditional pharmaceutical inventories. Approximately, 8400 compounds have been identified from tobacco plants and tobacco smoke [7]. It is important, however, that the system be as universal as possible and have the capacity to hold millions of compounds. The challenge posed by the implementation of a registration system in this context is not the number of compounds to be registered, but rather the wide range of chemistries represented (peptides, natural products, sugars, and complex products resulting from tobacco combustion). As such, a critical step in the project was to identify what was to be registered and how to ensure that the quality criteria were met effectively.

4.3
Dealing with Chemical Structures

4.3.1
Chemical Cartridges

The core technical functionality of the registration system is to handle chemical structures. Basically, this means that each chemical structure must be stored as a unique representation, and that the drawing of the structure must include all structural information such as stereochemistry, so that users have the possibility to search by exact structure, substructure, or similarity. Although it is possible to generate unique codes and perform searches using classical chemoinformatics tools such as Accelrys Pipeline Pilot [8] or Chemistry Development Kit [9], we believe the most suitable approach is to use a chemical cartridge as the central core of the registration system.

A chemical cartridge is basically a database plug-in that gives chemical handling functionalities to the database. The quality of chemical representation depends on several factors, one of which is the type of chemical cartridge used. There is a wide range of chemical cartridges available on the market, using different underlying database technologies (Table 4.1). Cartridges usually differ in their performance, searching mechanism (exact match, substructure, similarity, etc.), and ability to store large quantities of structures in the most efficient way. Chemical cartridges have the advantage of offering very good performance (for compound searches) because chemical structures are indexed in the database.

While these are critical elements, it should be stressed that there is another element that is usually underestimated: the concordance of the input system used by the end user (sketcher) and the underlying cartridge. These usually share the same chemical representation rules, but full alignment between them is not always guaranteed. The UCSD overcomes this problem by ensuring that molecules are

Table 4.1 Main chemical cartridges systems.

Name; Web source	Database type	Available for free
Accelrys Accord; accelrys.com/products/informatics	Oracle	No
Accelrys Direct; accelrys.com/products/informatics	Oracle	No
ChemAxon JChem; chemaxon.com/jchem/intro	Oracle	No
GGA Software Services Bingo; ggasoftware.com/opensource/bingo	Oracle and SQL	Yes
IDBS ActivityBase; idbs.com/products-and-services/activitybase-suite	Oracle	No
Molsoft MolCart; molsoft.com/molcart.html	MySQL	No
Mychem; mychem.sourceforge.net	MySQL	Yes
OrChem; orchem.sourceforge.net	Oracle	Yes
PerkinElmer CambridgeSoft Oracle Cartridge; chembionews.cambridgesoft.com/WhitePapers/Default.aspx?whitepaperID=18	Oracle	No
Pgchem::tigress; pgfoundry.org/projects/pgchem	PostgreSQL	Yes

drawn and stored in the exact same manner (see later in this chapter). In order to match our design and concept, the chemical cartridge of the UCSD is the Accelrys Direct (formerly Symyx Direct) chemical cartridge (Table 4.1).

4.3.2
Uniqueness of Records

The concept of the UCSD, in order to ensure data uniformity and uniqueness, requires that structures be standardized prior to registration in the system, and that this uniqueness be defined at the level of neutral molecules. As a consequence, when a new molecule is submitted or if a duplicate of the corresponding compound is found in the database, the system will create a new batch for the compound. It is also important for users to have the provision to register compounds for which structures are not known. Such cases are annotated as "no structure," denoting that no particular chemical structure is defined. This is useful, for example, for analytical chemists working with mass spectrometry who might encounter the same peak in several gas chromatography and liquid chromatography mass spectra, without being able to identify the compound (Sections 4.4 and 4.6).

For salts, the neutral form of the molecule is drawn and associated with the appropriate counterions (selected from a predefined list of ions) and ratios. In the same manner, hydrates are not drawn, but are associated with the chemical structure. Therefore, the system must be able to verify the uniqueness of the molecule, regardless of whether it is in the form of a salt or a hydrate. The canonical representation for each structure is generated by the system's chemical cartridge, and any tautomers of the same molecule are given the same canonical representation.

Each level (*molecule*, *substance*, or *batch*) requires specific information that is either entered manually by scientists during the submission step of the registration or calculated automatically by the system. For example, information related to a project, scientist, laboratory notebook reference, or analytical results is stored at the *batch* level (manual entry). Information related to the chemical substance, that is, IUPAC name, codes (InChI and SMILES), and physicochemical properties, such as molecular weight and logP, are calculated automatically and stored at the *substance* level. Similar information is stored at the *molecule* level of the neutral chemical structure.

The IUPAC names, SMILES, and InChI codes are generated by the software attached to the UCSD platform (e.g., IUPAC name is generated by ACD/Labs [10], SMILES, and InChI by the chemical cartridge). However, these names and codes can only partially represent the stereochemistry. When a molecule contains more than a single group of relative stereocenters, chemical line notations using SMILES and InChI are not sufficient to accurately represent the stereochemistry. For example, the "either" bond (drawn as wavy line) linked to a stereocenter cannot be encoded in InChI or SMILES.

Nevertheless, SMILES and InChI representations are still generated in the UCSD because these popular chemical line notations are useful to query external databases and data mine them. In order to omit any ambiguity and ensure that scientists are working with a single unique structure, the UCSD generates a unique internal company code, which in our case starts with letters PMI. In other words, no record can have two different codes and there cannot be two or more molecular entities sharing the same code. This is the important prerequisite for unambiguous data assembly and data mining. It should be noted, however, that considerable progress to guarantee uniqueness of structural descriptions using line notations has been recently made in the field of chemoinformatics. Examples are given in two recent publications [11,12], introducing and discussing yaInChI and CTISMILES codes, respectively.

4.3.3
Use of Enhanced Stereochemistry

One of the main difficulties with chemical registration systems is the representation of uncertainties of stereoconfigurations and mixtures of stereoisomers. To represent this as precisely as possible, even when the absolute configuration is not known, the UCSD platform uses a leading system of stereocenter descriptions developed by *Accelrys* (formerly *Symyx*). The system is called *Accelrys enhanced stereochemistry labeling* (V3000 format) and uses directly embedded labels in the drawings of the structure to allow precise configuration of the molecule for each possibility. Hence, this detailed 2D drawing of the structure with enhanced labeling minimizes ambiguity and guarantees the uniqueness of the molecule.

For example, the configuration for the two centers of 4-chloropentan-2-ol in the mixture of stereoisomers can be known or partially known (Figure 4.3). Embedded stereochemical labeling allows us to represent the relative stereoconfiguration of each stereogenic center. The six stereoisomers and stereoisomeric mixtures would be registered as six different molecules in the UCSD, each being a different entity; thus, removing the uncertainty of known or unknown stereoconfiguration.
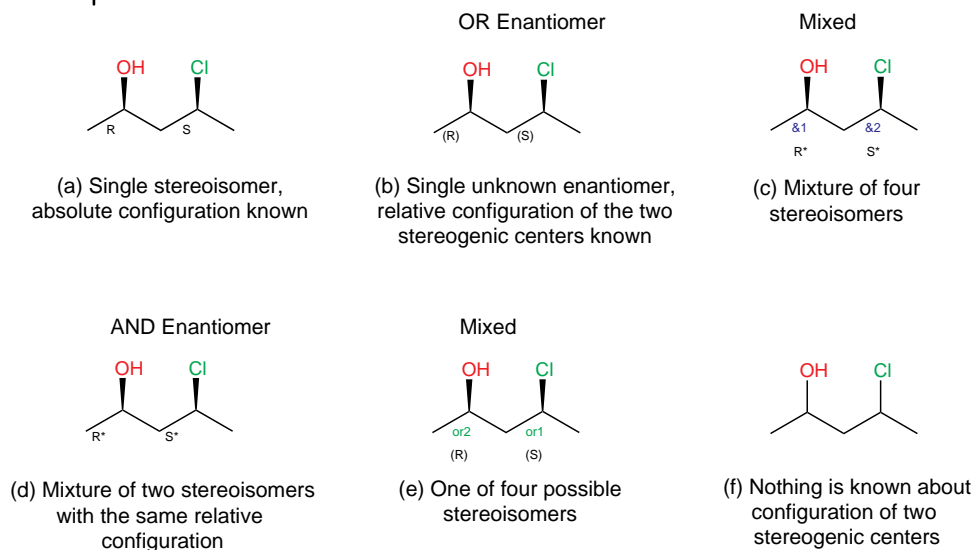
OR Enantiomer Mixed

(a) Single stereoisomer, absolute configuration known

(b) Single unknown enantiomer, relative configuration of the two stereogenic centers known

(c) Mixture of four stereoisomers

AND Enantiomer Mixed

(d) Mixture of two stereoisomers with the same relative configuration

(e) One of four possible stereoisomers

(f) Nothing is known about configuration of two stereogenic centers

Figure 4.3 Different cases of stereochemistry differentiated by embedded labels. All embedded annotations have a defined meaning: stereocenter configuration is either absolutely known [R and S labels in part (a)] or stereocenters have known relative configurations [R and S labels in parts (b–e)]. Larger labels above the structure indicate if the enantiomer is considered (b and d) or it is a mixture of stereoisomers (c–e). When nothing is known about the stereochemistry of the molecule, the annotations are absent (f). Annotations are managed by Accelrys Draw and Accelrys Direct chemical cartridge.

4.4
Increased Accuracy of the Registration of Data

4.4.1
Establishing Drawing Rules for Scientists

Since some molecules can be drawn in more than one way, it is important to define both a set of rules that chemists must follow when drawing molecules and the representations that are not allowed. The seamless translation of structure from the chemist to the chemical cartridge is then ensured.

In order for correct recognition by the chemical cartridge, structures must be drawn in a nonperspective way (Figure 4.4). When a structure contains bridging atoms, the bonds that are attached to the bridging atoms should not be marked as stereo bonds; instead, explicit hydrogens should be used. The correct drawing of *cis–trans* isomerism is also important.

Drawing sugars encompasses its own known challenges. Sugars can be represented in different ways, but not all of them are fully interpretable by the chemical cartridge. For linear sugars, the preferred drawing should use the line–angle structure rather than the Fischer projection [13]. Cyclic structures of monosaccharides should be

(a) Bridged ring

Perspective drawing   Nonperspective drawing

(b) Isomerism *cis-trans*

Disallowed

Allowed
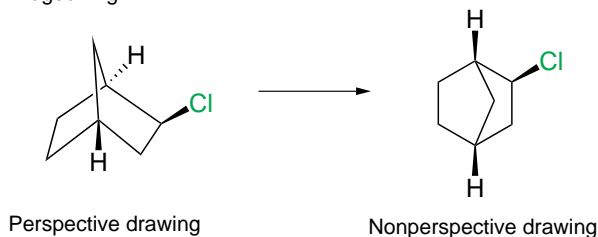
(c) Cyclic sugar

Disallowed drawing   Allowed drawing
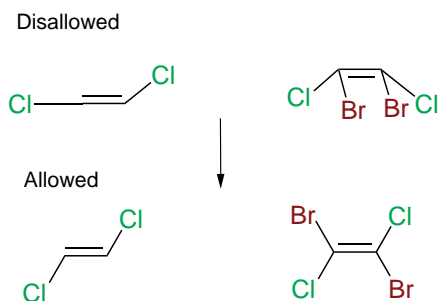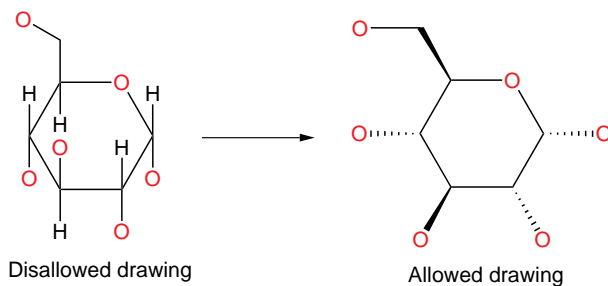
Figure 4.4   Allowed and disallowed representation of molecules. To be correctly understood by the chemical cartridge, drawings must be nonperspective (a) and in the allowed form for *cis–trans* isomerism (b). Allowed drawings for sugars are based on the Haworth projection of cyclic forms as shown in the example of α-D-Glucose (c).

represented using the Haworth projection[2] with a nonperspective drawing that is easy to translate into an acceptable drawing (Figure 4.4). Bonds above the plane of the carbon ring are marked "up," and bonds beneath the plane of the carbon ring are marked "down." Hydrogen atoms, which are explicit in Haworth projections, are also implicit in structures that are drawn for registration.

2) Haworth projection, http://goldbook.iupac.org/H02749.html.

4.4.2
Standardization of Compound Representation

Drawing rules provide the support necessary for scientists to represent chemical structures in a manner that will be correctly understood by the chemical cartridge. However, since all structures must be checked before they are entered into the database, some standardization rules must be automatically applied. For this purpose, Accelrys Cheshire (formerly Symyx Cheshire [14]), a chemistry-oriented scripting platform, was chosen. With this tool, it is possible to apply corporate standards to check, adjust, and neutralize chemical structures. In some cases, structures can be standardized or checked automatically. Some examples of standardization rules and error checks are presented in Figure 4.5.

In addition to an automatic check of the structure drawing by Accelrys Cheshire, well-defined drawing rules and validation by an expert established in the company

Standardization
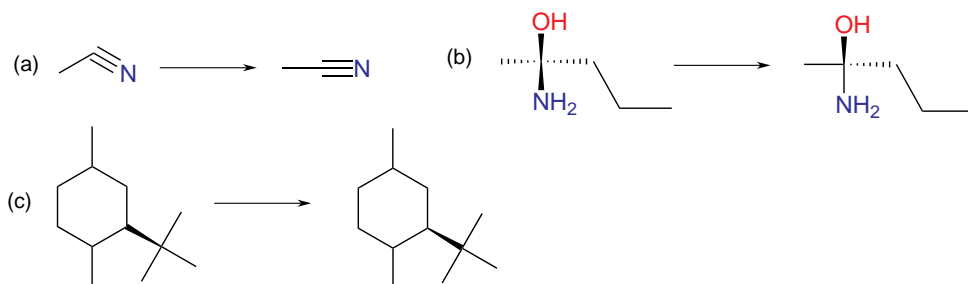
Flagged as erroneous

Technical limitations
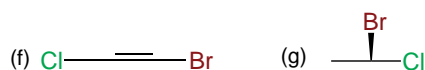
Figure 4.5 Examples of automatic standardization rules and error checks. Some standardization rules are applied automatically by Accelrys Cheshire: the nitrile group can be automatically redrawn linearly (a). For a stereocenter it is not necessary to have two double up bonds and/or two double down bonds (b), and up and/or down bonds must be oriented to the stereocenter (c). Some compounds cannot be corrected, but can be detected as erroneous (d and e) because it is not possible to determine the stereochemistry (d) or the valence is not correct (e). In some cases, it is not possible to correct or automatically detect the error. Some drawings meet technical limitations; the configuration of the stereobond (f) or of the stereocenter (g) cannot be determined.

Figure 4.6 Example of a standardization of the structure in the UCSD registration system. The nitrile group of the 2-cyanoacetate is redrawn linearly and the acid group is put into the neutral form. The output of the normalization script (right part: *Normalized structure*) is presented to the user, who can accept the changes prior to the submission of the molecule (tick the check box *Accept Normalization*). Salt can be selected using the *Molecule & Substance* button.

help to ensure that only correct structures are stored in the database. An example of the validation of chemical structures with Accelrys Cheshire in our platform is presented in Figure 4.6.

4.4.3
Three Roles and Two Staging Areas

The workflow for the registration process is another important element that helps to ensure the highest quality of registration and minimize incorrect entries due to human error. In the UCSD, three different roles are defined in the system (Figure 4.7):

- The role of *Viewer* includes all users that can search and view registered data via a Web interface (except for some "restricted fields" reserved for specific teams).
- The role of *Submitter* is reserved for scientists who can create new information. Submitters have the same privileges as Viewers, but can also access a submission form to insert new records in the database.
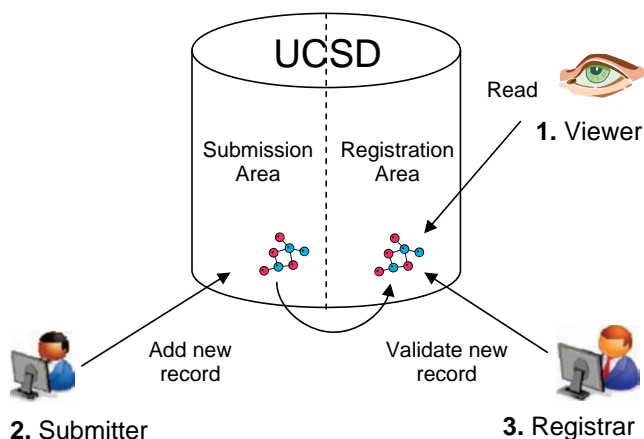
Figure 4.7 Three user roles and two staging areas of the database.

- The role of *Registrar* is restricted to chemoinformaticians. Registrars have the same privileges as Submitters, but are also responsible for checking and validating data in the submission area (source, project ID, etc.) and giving approval for registering an entry.

Authentication to access the database is managed by Lightweight Directory Access Protocol (LDAP) accounts, with three groups corresponding to the three roles.

Nevertheless, when submitting a molecule for registration, the submitters have to follow a registration workflow. Namely, chemists wishing to add a new batch to the UCSD should consider whether the chemical substance fulfills the criteria (agreed within the company) for the registration.

The submission form contains various fields of data associated with the batch (e.g., common name, source, internal identifier), some of which are mandatory. An automated control ensures that a molecule can be submitted only if all mandatory data and validation fields are completed correctly (e.g., experimental molecular weight field accepts only numbers). When the user validates the drawing of the molecular structure (or for the unknown compound, the label "no structure" is written), a normalization process is automatically executed, and the normalized structure is displayed next to the original drawing (Figure 4.6). The submitter can then decide whether or not to approve this normalization.

When the form is completed, the *Submitter* clicks the Submit button to store all the data in a transitory area called the *submission area*. This area is the "waiting room" for compounds to be reviewed and validated by the database *Registrar*. In this area, any potential user errors are checked by the *Registrar* and corrected. Approved molecules (and all related information) are then copied into the *registration area*, which is the final database. The two staging areas allow scientists to be part of the registration process and the *Registrar* to check discrepancies and errors. This two-stage process brings an additional layer of security for the UCSD and increases confidence in the

accuracy of registered data. When a submitted molecule is not approved, the *Submitter* receives an e-mail notification with the reason why the batch has been rejected and has the option to modify and resubmit it.

4.4.4
Batch Reassignment

The system also allows records to be updated for the inclusion of newly discovered properties for individual structures. A *batch* can be reassigned to a different *substance* if the user realizes that the structure was entered incorrectly or determines stereogenic centers during later experiments. If a *substance* no longer has a *batch*, it is archived as inactive. If after this process a molecule no longer has a *substance*, it is archived. There is no "delete" process; all new entries are assigned chronologically with new codes. This procedure allows the correction of errors without losing any information related to the archiving process.

4.4.4.1 Unknown Compounds Management
Another important point concerning batch reassignment is the registering of compounds with undetermined structure. As mentioned earlier, there is a provision to register a compound with unknown structure with the "no structure" label. However, it can happen that scientists, especially analytical chemists, identify mass spectra peak corresponding to a specific mass of the compound that they see repetitively in their analyses. Such an entry is referred to as a "known unknown." In the UCSD, upon registering such a compound with a "no structure" label, a special temporary code "UNK" (as for Unknown) is generated (e.g., UNK1, UNK2) and not the UCSD code. The batch of this analysis is then assigned to the molecule with the UNK code. Several batches of the same unknown can be assigned to the same molecule; hence, the system handles identical unknowns under one unique UNK code (e.g., UNK1). When two or more unknowns reveal themselves to be the same molecule, the batch can be reassigned to a single code (e.g., from UNK2 to UNK1). If the chemical structure of a "known unknown" compound is determined at a later date, the molecule can be drawn and the batches of this originally unknown molecule can be assigned to a known molecule with a UCSD code.

4.4.5
Automatic Processes

In a chemical database, to further minimize possible errors, it is preferable to have names and certain aspects, such as ADMET properties, calculated automatically. Even though our system generates corporate compound IDs (UCSD codes) upon the registration of each entry, it is important that chemical names and other identifiers such as IUPAC names or CAS numbers be recorded in order to provide links to molecules in external databases to facilitate any data mining process.

ACD/Labs Name Batch tool [10] is used to automatically and accurately generate most names according to IUPAC guidelines from the molecular structure. However,

naming structures with enhanced stereochemistry is still an issue: see example of (2*R*,4*S*)-4-chloropentan-2-ol in Figure 4.3 for which IUPAC names with such detailed description of stereochemistry cannot be generated. The software Accelrys Pipeline Pilot is used to predict ADMET properties and calculate lead-like and Lipinski indicators. ACD/Labs PhysChem is used to automatically calculate water solubility values of the molecules.

In addition, the system automatically calculates the theoretical molecular mass of both molecule and substance. The system neutralizes the charged atoms of the molecule by adding hydrogen(s) to these atoms and the submitter selects the corresponding counterion from the predefined dictionary of salts for substance. As the charged form of the molecule is not represented in the database, the molecular mass of the salt is decreased by the theoretical molecular mass of the one (or several) added hydrogen(s).

4.5
Implementation of the Platform

4.5.1
Database

The database is hosted on Oracle 11g with Accelrys Direct 8.0 cartridge. The major effort in the design phase of the project was to conceive the database model (Figure 4.8). The two areas (*submission* and *registration*) are clearly separated in the data model. The submission part of the scheme is a buffer area for the data containing all the information entered by the user. The registration tables contain the validated information. Molecule, substance, and batch tables reflect the organization of the chemical data presented before. Properties are stored in dedicated
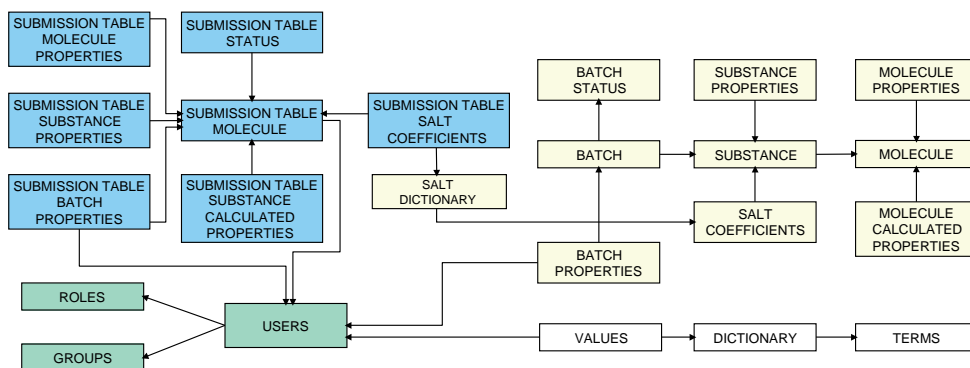
Figure 4.8 Database model for the chemical part of UCSD. Database is composed of tables related to the submission (blue), registration (yellow), security (green), and additional properties (white) areas.
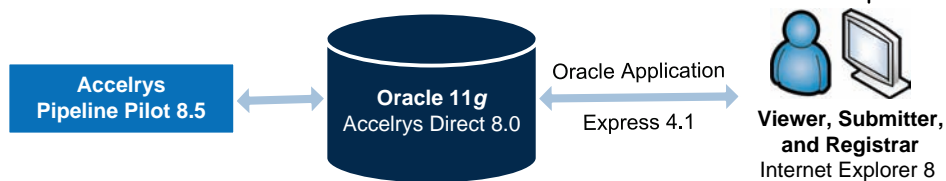
Figure 4.9 Software architecture of the UCSD. The database part is based on Oracle and Accelrys Direct. The software to input data and the visualization interface was developed using Oracle Application Express, and it is accessible with a Web browser. Pipeline Pilot is used to compute physicochemical properties and chemical names.

tables for each level of information. Information related to security and additional properties is stored in separate tables.

4.5.2
Software

Several programs were used to build the UCSD platform. Accelrys Pipeline Pilot 8.5 was used to compute the physicochemical properties (e.g., molecular weight, logP) and IUPAC names (generated by ACD/Labs Name) for any new molecule entry. In order to normalize the structures, Accelrys Cheshire is called directly from Oracle. Oracle Application Express 4.1 (Apex) was used to develop the main part of the platform (Figure 4.9). Apex is a tool integrated by default in Oracle 11*g* and dedicated for building Web interfaces for Oracle databases in a very efficient way. The Web interface built in Apex is then the point of entry to the database for Viewers, Submitters, and Registrars. The only specific software that is required to be installed on the user's desktop PC is the software used to draw structures, for example, Accelrys Draw.

4.5.3
Data Migration and Transformation of Names into Structures

It is widely acknowledged that the implementation of a chemical registration system can be a lengthy and difficult process. In a business environment, it is critical to deliver the system in the shortest timeframe possible with maximum efficiency. In the case of the UCSD, the team at PMI comprised three chemoinformaticians, one project manager, and support from the former Symyx consulting team. This team was empowered by management to make all decisions regarding the chemical representation and standardization rules, the data structure, and the technical implementation strategy. A production system was available 4 months after the project kickoff.

Database population is perhaps one of the most underestimated and neglected processes in terms of time and resources. In our case, once the UCSD was ready to be populated with compound entries, the migration of all existing data was recognized as being the critical step because data and names (or CAS numbers) for molecules are often stored locally in diverse file formats.
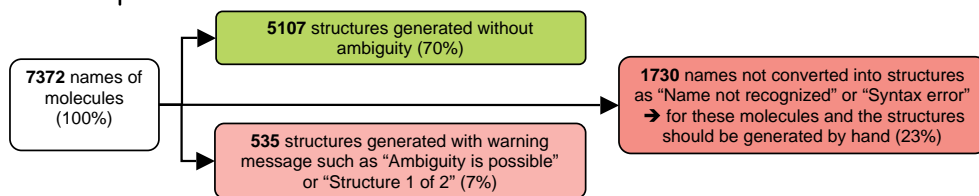
Figure 4.10 Success rates for converting chemical names to structures. In this example, around 70% of the names were transformed correctly into structures, 7% of the generated structures were ambiguous (generated warning messages), and 23% of the names were not recognized by the software.

When building the UCSD without any existing infrastructure, two major sources of compounds were available: (i) names of molecules cited in the literature and (ii) internal lists of compounds, often annotated by their IUPAC names, common names (e.g., harmane), or CAS numbers. As these names do not always follow IUPAC recommendations, their transformation into structures and subsequent importation into the UCSD was difficult. The issue was addressed using the standard software module "ACD/Name to Structure Batch" from ACD/Labs. This software generates accurate structures for entire libraries of compound names. Even though this software is capable of transforming large numbers of names into structures, a considerable amount of time must be spent by the submitters and registrars in checking and correcting the structural representation of molecules. The overall yield for generating correct structures is presented in Figure 4.10.

For the set of 1730 molecules (Figure 4.10), there was no easy or automatic way to obtain a structure; therefore, chemists had to check and correct each name and associated structure manually. To obtain a structure for these molecules, searches were conducted in PubChem (http://pubchem.ncbi.nlm.nih.gov/), ChemSpider (www.chemspider.com), and Google. For some molecules, an associated CAS number was available, which allowed the use of SciFinder® (www.cas.org/products/scifinder), a tool for exploring the CAS REGISTRY database, to obtain their structure. Clearly, during the construction of the UCSD, structure conversion took a considerable length of time (in terms of months taken by one chemoinformatician) and the effort and time required should not be underestimated.

Once the structures were obtained and confirmed, a Pipeline Pilot protocol was developed to automate their importation (Figure 4.11). This protocol required an
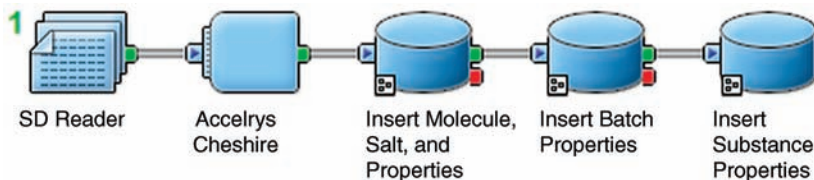
Figure 4.11 Accelrys Pipeline Pilot protocol to automatically insert an entire compound library.

input file in structure–data file (SDF) format in which the structures were described by a Connection Table in V3000 format. The SDF format was chosen because of its ability for attaching associated data to structures (e.g., project name, internal identifier, and scientist who works on this compound).

4.6
Linking Chemical Information to Analytical Data

There are many R&D companies that study the properties of complex compound mixtures such as plant extracts, which are often referred to as complex matrices. These activities primarily take place in companies operating in the nutraceutical, food flavoring, or cosmeceutical industries. At PMI, the major source of complex matrices for analysis originates from the aerosols (smoke) of smoking articles; another source of compounds for analysis is from tobacco plant extracts in the frame of metabolomics studies. The majority of compounds in such matrices are determined using chromatographic and mass spectrometric techniques. Cigarette smoke, for example, contains many thousands of compounds. Complex analytical methodologies are used to separate compounds, pick the peaks, deconvolute the spectra, and determine individual compounds.

In our concept, the UCSD comprises two separate but interlinked Oracle databases developed in-house: a compound database and a spectral database (Figure 4.12). The "C" in UCSD stands for compound database, which is referred to here as ChemDB, whereas the "S" represents its spectral part, called SpecDB. Distinct separation allows scientists to control the source and the occurrence of the compound, to check if the compound was determined using analytical methods, identify which methodology was used, and understand the level of confidence associated with the data.

Once a compound has been determined by spectrometric methods, its analytical spectrum is registered in the SpecDB and its structure is registered in the ChemDB. Whether the compound was extracted from a plant, synthesized in a laboratory, or retrieved from a text source, information is recorded at the *batch* level in the section called "Source of data." The system of batches, as presented earlier in this chapter, allows the registration of multiple pieces of laboratory-based information for a single molecule at the *batch* level, associated with the compound at the *molecule* level. It also implies that spectral information for a single molecule originating from different techniques should be considered as different batches and registered at the *batch* level.

At PMI, many analytical chemistry devices from suppliers such as Waters, ABSciex, Shimadzu, Leco®, and Bruker (*www.suppliername.com*) are used to determine the presence of aerosol constituents. The ACD/Labs software [10] is used as the principal software package for analytical chemistry. It allows data processing and interpretation for NMR, LC/GC/UV/MS, IR, and other types of spectra from different instrument vendors in a single environment, regardless of the original data format, which enables the comparison of spectra from different sources. For this reason, ACD/Labs software was selected and customized to provide the database and the interface for the SpecDB. If the spectrum of a compound is
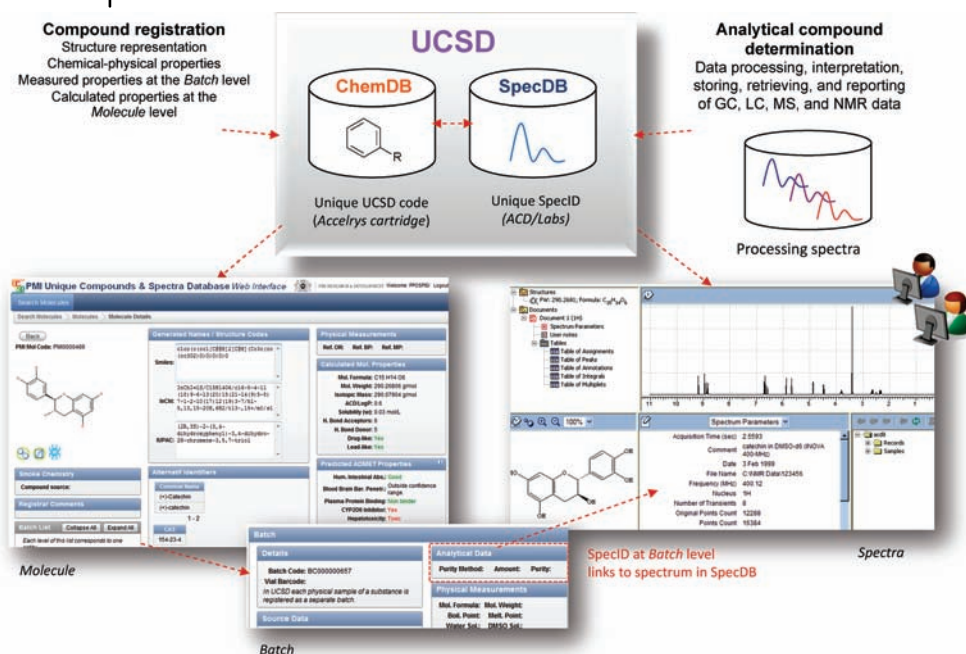
Figure 4.12 Linking chemical data to spectrometric data. The UCSD comprises two separate oracle databases, ChemDB and SpecDB, each having a unique compound and spectral identification code. Structural representation and associated compound properties are stored at the *molecule* and *batch* levels in ChemDB. If the compound is determined using analytical methods, relevant spectra are associated with the compound at the *batch* level, linked via a unique SpecID code.

known, a "SpecID" link for the associated spectrum is entered into the ChemDB, which directs the user to the SpecDB interface of the ACD/Labs software (Figure 4.12). If no SpecID link is present, it normally signifies that the compound has yet to be analyzed.

Analytical chemists often determine compounds with uncertain stereochemistry or isomeric composition, which they wish to register. Linking analytical spectra to unknown or partially known structural representations is very challenging. Usually, the chemist notes the putative name of the compound determined from the peak and writes a comment regarding the potential enantiomers. With the UCSD, this challenge is overcome by the use of two interlinked databases. The ChemDB part allows users to register "unknown" structures (using the label "no structure") or structures with unknown stereochemistries (using the enhanced stereochemistry labeling). Moreover, ChemDB allows the registration of known unknowns, as described in Section 4.4. In the SpecDB part, analytical spectra data are recorded as required by the scientist and linked to the batch record within the ChemDB. We believe that this separation of chemical and spectral parts within one bicameral

system further minimizes any ambiguity of structural representation, which enables precise registration and strengthens the database accuracy.

4.7
Linking Chemicals to Bioactivity Data

Once a system with a concept such as UCSD is implemented, it is of interest to link its content with chemical and biological data published in the literature and available in public databases. In the pregenomic era, public chemical databases were collections of compounds and their chemical names and structures. Basic chemical–physical properties, either measured or calculated, were generally published in the scientific literature one by one and not curated. In the first databases, the information associated with molecules was usually added manually. Later, these chemical databases were capable of depicting 2D chemical structures with their names, and they allowed file exports in SDF format. In addition to names, later databases introduced ID codes (CAS numbers, PubChem ID) and basic physicochemical properties, either measured or predicted by software. The 2D chemical structure itself became the common *unique denominator* of these chemical databases. Errors in the structure representation were not uncommon and the problem was accentuated with the requirement for representation of stereogenic centers, tautomers, or enantiomers and their mixtures.

After 2000, the industrial sector and public initiatives quickly moved the development of databases forward in the direction of creating robust data record systems in order to mine, analyze, and interpret the data. In industry, the arrival of high-throughput methodologies, such as combinatorial chemistry, high-throughput screening (HTS), and structure–activity relationship (SAR) studies, has resulted in databases being populated with millions of measured endpoints, which required the construction of robust databases. Management technologies for such record numbers have consequently improved; however, access to these data has been hard to obtain because companies kept data in nonpublic corporate registration systems. In general, the growth of chemoinformatics and the development of the Internet have provided new opportunities to publicize data. Namely, commercial databases such as SciFinder of the CAS REGISTRY^{SM} (www.cas.org/products/scifinder) provided such services and the CAS number became one of the most common identifiers. In 2004, the PubChem project (http://pubchem.ncbi.nlm.nih.gov) released to the public the PubChem database and the PubChem BioAssay database (www.ncbi.nlm.nih.gov/pcassay) within the NCBIs Entrez information retrieval system with its own unique code CID or AID. These advancements have changed the landscape of chemoinformatics for the end user chemists as well as biologists; chemical data has become more accessible, putting chemistry one step closer to biology.

Many databases subsequently became public and provided user-friendly Web interfaces (Table 4.2). Nowadays, there are several public initiatives and databases that store and allow searching high-throughput data, which take the form of complex

Table 4.2 Compound activity databases.

Name; Web source[a]	Public or commercial	Short description (October 2012); number of entries
Large databases		
PubChem; pubchem.ncbi.nlm.nih.gov	Public	Chemical information on the biological activities of small molecules; >35 million unique structures, >100 million substance records
PubChem BioAssay; ncbi.nlm.nih.gov/pcassay	Public	Contains bioactivity screens of PubChem chemical substances; >5 627 000 substances, >621 600 assays
ChemSpider; chemspider.com	Public	Chemical database with links to data sources, literature, patents, and chemical properties; >26 million compounds
ZINC; zinc.docking.org	Public	Commercially available compounds for virtual screening with vendor information, links to other large databases, and physical properties; >21 million compounds
ChEMBL; www.ebi.ac.uk/ChEMBLdb	Public	Database of bioactive compounds; >1 213 200 distinct compounds, >581 000 assays, >9000 targets
ChemBank; chembank.broadinstitute.org	Public	Includes data derived from small-molecule screens; >500 000 unique compounds
CAS SciFinder; www.cas.org/products/scifinder	Commercial	Tool with access to the world's largest collection of substances, reactions, and references (instant access to the CAS REGISTRY[SM]); >68 million substances, ~15 000 new substances added daily
Aureus Sciences; aureus-sciences.com	Commercial	Various compound, target, and pharmacological databases; >1.8 million biological activity points, >500 000 compounds
Thomson Reuters MetaDrug[TM]; genego.com/metadrug.php	Commercial	Compound-based pathway analysis system; >600 000 compounds, 65 000 proteins, >500 000 protein–compound interactions
Thomson Reuters Integrity; integrity.thomson-pharma.com/integrity	Commercial	Knowledge base of compounds with demonstrated biological activity; >330 000 compounds, >900 000 measured pharmacology values, >2600 targets, >140 000 patents
WOMBAT, WOMBAT-PK; sunsetmolecular.com	Commercial	Small molecule chemogenomics database; >331 800 ligand entries, 1996 targets; PK database >9000 of measurements
Accelrys MDDR and CMC; accelrys.com/products/databases/bioactivity	Commercial	Database with biological activities and therapeutic actions; >150 000 compounds, joint content with TR Integrity

Drug activity databases

BindingDB; bindingdb.org	Public	Contains measured binding affinities of mainly drug-like molecules; >910 000 binding data, >379 000 compounds, >6200 targets
DrugBank; drugbank.wishartlab.com	Public	Drug–target database with 150 fields on chemical/protein data; 6711 drug entries, 4227 proteins
SuperTarget; insilico.charite.de/supertarget	Public	Contains information from BindingDB, DrugBank, and SuperCyp; >195 000 compounds, >6200 targets, >332 000 drug–target interactions
Guide to PHARMACOLOGY – GRAC and IUPHAR database; guidetopharmacology.org	Public	Quantitative information on drug targets and the prescription medicines/experimental drugs. 1600 targets, >1800 small molecules
DTP from NCI; dtp.nci.nih.gov/webdata.html	Public	Developmental Therapeutics Program; >200 000 compounds, *in vitro* human tumor cell line assays for >43 000 compounds
Binding MOAD; bindingMOAD.org	Public	Contains ligands derived from PDB with measured binding data; >16 900 protein–ligand structures, 5630 structures with binding data
PDBbind; pdbbind.org.cn	Public	Experimentally measured binding affinity data for the PDB protein–ligand complexes; >15 000 ligands
PDSP Ki; pdsp.med.unc.edu	Public	Psychoactive drug pharmacological activity database; >55 400 K_i values, >7500 drugs and other compounds, 740 receptors
Therapeutic Targets DB (TTD); bidd.nus.edu.sg/group/ttd	Public	Provides information about the known and explored therapeutic targets, the targeted disease, the pathway information, and the corresponding drugs; > 17 816 drugs, 2025 targets
HMDB; hmdb.ca	Public	Human metabolome database of metabolites found in the human body, contains chemical, clinical, and molecular biology data; >37 300 metabolites, links to >5700 proteins and DNAs

Compound toxicity databases

ACToR; actor.epa.gov	Public	Aggregates data from >1000 public sources on >726 000 environmental chemicals
ChemIDplus/ToxNet; chem.sis.nlm.nih.gov/chemidplus	Public	National Library of Medicine database on toxicity data; >370 000 chemicals
DSSTox; epa.gov/ncct/dsstox	Public	Toxicity database of environmental relevance; >8000 chemicals
CTD; ctdbase.org	Public	Comparative Toxicogenomics Database; >12 100 chemicals and >28 900 genes with curated data

a) URLs are indicated shortened, but they are still active when copied directly into the browser address bar.

knowledge bases: PubChem from NCBI, ChEMBL from EMBL (www.ebi.ac.uk/ChEMBLdb), ChemSpider from RCS, or DSSTox (www.epa.gov/ncct/dsstox) from the EPA ToxCast[TM] project (www.epa.gov/ncct/toxcast). Some databases function as metadata centers; e.g., ChemSpider contains data from over 400 publicly available chemical databases and link their chemical structures to analytical and pharmacological properties. Similarly, in the field of toxicology, the project ACToR is compiling data (both quantitative and qualitative) from a large number of data collections; it aggregates data from over 1000 public sources (e.g., EPA databases, PubChem, other NIH and FDA databases, and databases from academic groups) on over 700 000 environmental chemicals. ACToR assembles all publicly available chemical toxicity data and it can be used to find data regarding potential chemical risks to human health and the environment (actor.epa.gov). Major compound activity databases are listed in Table 4.2, which includes a brief description, the content, and a link to their Internet home page. Some of the databases and their usage, in particular BindingDB, PubChem, and ChEMBL, are described in more detail in Chapter 2 [15].

Nowadays, more than 20 000 new compounds are published in medicinal and biological chemistry journals every year. It is crucial that these data do not rest in conventional texts, but are made available to be extracted and stored in databases for computational analysis. It is also important that, in addition to the development of more accurate text mining technologies, data in journals should become routinely available in machine-readable format to facilitate data mining and all regular updates of the existing medicinal chemistry databases with these data. Databases, on the other hand, should be able to regularly and automatically extract data from the literature and annotate their records with biological data. This would make both literature and databases less disparate and the systematic analysis of the way in which small molecules impact biological systems would be easier.

Retrieval of chemical structures from text sources remains, however, a very complex matter. In most cases, structures are published in journals as pictures and their names are either conventional (e.g., aspirin) or according to IUPAC naming conventions. Hence, ensuring a unique record for compounds is a challenge. In the UCSD concept, every molecule of interest is converted and imported into UCSD as described in Section 4.5.3. Records can point to external databases via external codes (PubChem, ChemSpider); however, the uniqueness of structures remains ensured solely by UCSD.

Although searching and data mining for chemical structures became routinely studied in the field of chemoinformatics, it is the bioactivity of compounds, the *molecular phenotype*, that is becoming of particular interest to data mining. In order to develop new therapeutics, it is important to understand the biological effect that the small molecule has. The following questions should be answered: What is the compound's target and mode of action (MOA)? What are the compound off-targets (selectivity)? What pathways and networks are modulated by the compound? What are the network changes that are related to diseases? What is the cause of any compound toxicity? And what is the therapeutic relevance to disease onset and progression? To answer these questions, chemical and biological databases are

getting more interlinked and customized to enable automatic data mining. Some advanced data mining technologies applied in chemoinformatics and computational biology are also presented in this book. The compound-related databases available today provide a plethora of useful links to other chemical, biological, and omics analysis databases that help with the discovery of new applications for the chemicals.

4.8
Conclusions

The Unique Compound and Spectra Database manages the registration of molecules in an efficient and nonredundant manner. It has the flexibility to register molecules with unknown structures or mixtures of compounds and at the same time can be used to register known structures with a precisely defined stereochemical configuration. This level of detail ensures the uniqueness of chemical records given primarily by its absolute 2D structure representation. The reliability of the database and the accuracy of the registration process are enhanced by the use of two staging areas and the system of batch assignment capability. We believe this process provides a higher level of molecule description and easier traceability of different entries. Also, linking the chemical database with its spectrometric information in relationship to the different entities stored in the system is of the high importance, especially for companies or institutes working with complex matrices.

The UCSD system, as implemented in-house at PMI, provides enhanced control for the accuracy of the chemical data. With the structure representation handled separately in the UCSD, data entries enriched by properties determined in-house and results of experimental assays can be then linked to external public knowledge bases via their names or public codes, and data mined for biological activities, modes of action, and therapeutic outcomes. Closer integration of chemical and biological systems, together with the pathway analysis and network modeling, represents a step further toward the systems biology approach and its associated drug discovery efforts.

Acknowledgment

The authors express their gratitude to Lynda Conroy for editing this chapter.

References

1 Warr, W.A. (ed) (1989) *Chemical Structure Information Systems: Interfaces, Communication, and Standards, ACS Symposium Series 400,* American Chemical Society, Washington, DC.

2 Buntrock, R.E. (2001) Chemical registries– in the fourth decade of service. *Journal of*

Chemical Information and Computer Sciences, **41**, 259–263.

3 Gobbi, A., Funeriu, S., Ioannou, J., Wang, J., Lee, M.-L., Palmer, C., Bamford, B., and Hewitt, R. (2004) Process-driven information management system at a biotech company: concept and implementation. *Journal of*

Chemical Information and Computer Sciences, **44**, 964–975.

4 O'Donnell, T.J. (2009) *Design and Use of Relational Databases in Chemistry*, CRC Press, New York.

5 Weisgerber, D.W. (1997) Chemical abstracts service chemical registry system: history, scope, and impacts. *Journal of the American Society for Information Science*, **48**, 349–360.

6 Martin, E., Monge, A., Duret, J.A., Gualandi, F., Peitsch, M.C., and Pospisil, P. (2012) Building an R&D chemical registration system. *Journal of Cheminformatics*, **4**, 11.

7 Rodgman, A. and Perfetti, T.A. (2008) *The Chemical Components of Tobacco and Tobacco Smoke*, CRC Press, Boca Raton.

8 Accelrys, Inc., San Diego, USA. Accelrys Pipeline Pilot (2013) accelrys.com/products/pipeline-pilot

9 Chemistry Development Kit, The CDK Development Team. (2013) cdk.sourceforge.net

10 Advanced Chemistry Development, Inc., Toronto, ON, Canada (2013) acdlabs.com/.

11 Cho, Y.S., No, K.T., and Cho, K.H. (2012) yaInChI: Modified InChI string scheme for line notation of chemical structures. *SAR and QSAR in Environmental Research*, **23**, 237–255.

12 Gobbi, A. and Lee, M.-L. (2011) Handling of tautomerism and stereochemistry in compound registration. *Journal of Chemical Information and Modeling*, **52**, 285–292.

13 McMurry, J. (1989) *Essentials of General, Organic, and Biological Chemistry*, Prentice Hall, Englewood Cliffs, NJ.

14 Accelrys, Inc., San Diego, USA, Accelrys Cheshire (2013) accelrys.com/products/informatics/cheminformatics/accelrys-cheshire.html

15 Nicola, G., Liu, T., and Gilson, M.K. (2012) Public domain databases for medicinal chemistry. *Journal of Medicinal Chemistry*, **55**, 6987–7002.

Part Two
Analysis and Enrichment

5
Data Mining of Plant Metabolic Pathways

James N.D. Battey and Nikolai V. Ivanov

5.1
Introduction

5.1.1
The Importance of Understanding Plant Metabolic Pathways

Plants are of economic and scientific importance and as a consequence, understanding and being able to manipulate them is of great interest to a broad audience. For diverse evolutionary reasons, such as attracting pollinators and defense against herbivores and parasites, plants have evolved a complex array of secondary metabolites. With plants such as *Arabidopsis thaliana* containing an estimated 5000 compounds in their metabolism [1], the potential for discovering novel compounds with interesting properties is immense. The biological and chemical properties of plants are conferred by the biochemical machinery that is encoded in the genome and organized into biochemical pathways. Understanding their composition is the prerequisite for manipulating these pathways, thus creating plants with new chemical properties.

The importance of creating novel properties in plants is perhaps most apparent in the agricultural sector. Here, the primary focus is on optimizing crop yield, improving plant robustness toward stress factors such as salinity or heavy metals, resistance to pests, and obtaining plants with desirable pharmaceutical or nutritional properties, such as rice variants that can produce beta-carotene (provitamin A) [2]. Besides their use in the agricultural sector, plants have also been "rediscovered" by the pharmaceutical industry as a valuable source of novel, diverse compounds that could be mined for potential use as pharmaceutical drugs [3]. It is estimated that hundreds of thousands of compounds could yet be discovered in plants [4], which could complement and extend current compound libraries used for screening in search of new drug leads.

Furthermore, plants not only provide compound diversity, but they also supply the necessary biochemical machinery, that is, enzymes organized into synthetic pathways that lead to the production of these compounds. Although biochemical synthesis is efficient, limited production of valuable compounds by naturally

occurring plants gives rise to the requirement for reengineering some pathways. A prime example of the benefits of pathway engineering is artemisinin, a drug that was discovered in *Artemisia annua*. It is one of the most potent antimalarial drugs in combating malaria, but unfortunately, the yield from the plant is low, making the drug scarce to the point of making it prohibitively expensive in developing countries. The creation of transgenic yeast, which expresses the plant pathway genes that lead to the production of artemisinic acid, a precursor for the drug [5], opens up a new possibility of increasing production by pathway engineering. It is hoped that this will allow large-scale manufacturing of artemisinin, making it affordable to a far larger number of malaria sufferers.

Finally, the discovery of completely novel compounds, produced by reengineering or reassorting the existing machinery of the cell is an active field of research. For this "combinatorial biosynthesis," enzymes are combined in new ways to allow the synthesis of novel compounds that can be discovered by screening for predefined phenotypes. Approaches to the problem include not only random sampling of preexisting pathway components, but also the rational engineering of pathways *in silico* [6,7]. By modeling the possible landscape of biochemical networks, it may be possible to find alternative, more energy-efficient paths for producing natural compounds or it may be possible to even devise synthesis routes that lead to completely novel ones.

Pathway modeling can help understand the alterations that are brought about by changes to the enzymatic complement of a cell or organism. It is therefore a powerful tool for rationally engineering or reengineering pathways in a target organism. The principles applied are not limited to plants, but can be applied to any organism.

5.1.2
Pathway Modeling and Its Prerequisites

In silico simulations require detailed information about the structure of the metabolic network of the target organism, that is, all the molecules and the reactions by which they are interconverted. Several different types of metabolic modeling can be performed, each with different goals and outcomes. Each type of modeling has prerequisites in terms of the information needed for such a study.

> *Flux balance analysis* (FBA) is used to determine the flux of carbon through a metabolic network at steady state. It uses the structure of the metabolic network at the reaction level as a basis for computing these fluxes. This method can help identify bottlenecks in systems and determine the most efficient pathways through the metabolic network. Therefore, it can be a valuable aid in understanding a system and simulating how changes affect it. For plants, this approach has been used to analyze which pathways are predominantly used under stress conditions, such as varying aerobic conditions [8], and it can also help understand how mutations (gene/pathway deletions) may affect the metabolic network (Figure 5.1). FBA can be used to predict changes that will create plants with desired phenotypes. The requirement for FBA is a model of the target organism's metabolism, with a stoichiometrically complete representation of all the participating reactions.
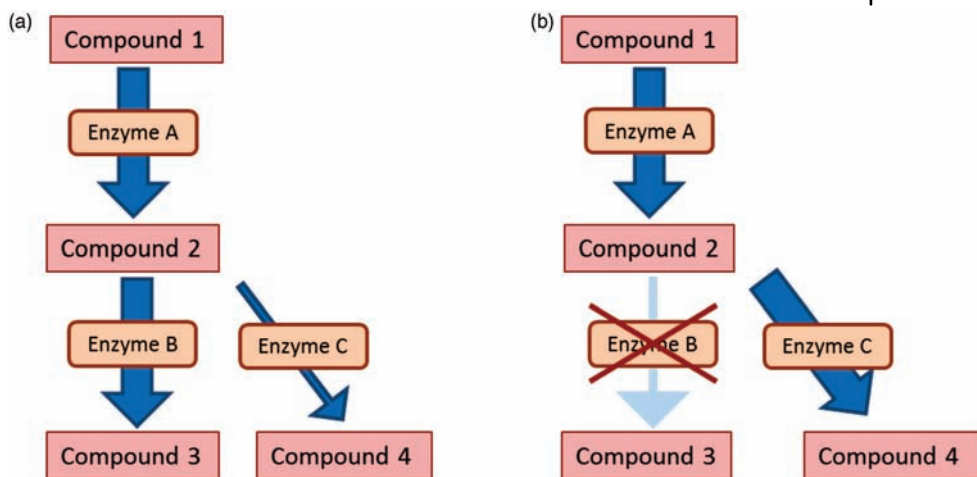
Figure 5.1 (a) A schematic representation of the analyses performed during FBA. (b) By modeling the elimination of an enzyme from a pathway, the carbon flux (represented by the thickness of the arrows) is diverted.

Metabolic flux analysis (MFA). MFA traces the flow through a metabolic network or molecular pathways of the cell. This method relies on isotope labeling data to determine the fluxes: when a cell is grown on a metabolite labeled with a heavy isotope at a given position, its metabolites will be differentially labeled depending on the pathways it passed through. To account for these isotopomers, the pathways in an organism have to be known, along with the atom transitions for each reaction (i.e., which atoms in the reactants are found at which positions in the products). This knowledge is necessary for modeling the isotopic variants of the metabolic intermediates that are produced depending on the pathway through which a substrate flows [9].

Metabolic design and prediction. Another growing field of interest is discovering potential pathways leading to the production of novel compounds. Rather than using a database containing known pathways, a list of known reactions is used to create a hypothetical network of compounds that can be interlinked by reactions; potential synthetic routes (paths) through this network can thus be identified and analyzed [10,11]. While current applications of this method are largely theoretical, it holds great potential for future developments. This method requires a repository of reactions from which potential metabolic pathways can be constructed by linking compounds to hypothetical reaction products.

5.2
Pathway Representation

Metabolism can be characterized in many different ways and can be represented in almost arbitrary detail, depending on the data available and the intended use. The

conventional method of representation is a graph-based structure, whereby the nodes represent the chemical compounds that are interconnected by the reaction edges. This metabolite-centric view is incomplete, as it neglects the fact that multiple reactants can be present on both sides of a reaction equation. A reaction-centric view has also been proposed, using reactions, or rather the enzymes that catalyze these reactions, as the nodes [12]. This however suffers from the difficulty that metabolites can participate in multiple reactions. Since a one-to-one relationship is either never complete or requires redundancy in the representation, the ontology generally chosen by the main databases uses a bipartite graph structure. In this case two classes of nodes, representing either reactions or compounds, are joined by edges linking all compounds to all reaction they participate in. These edges may be directional to indicate the physiological direction of the reaction (Figure 5.2).

The main challenges encountered in computationally capturing pathways using the bipartite graph structure are eliminating redundancy and ambiguity of the compounds and reactions. The following section treats these issues in more detail and discusses the approaches taken to resolve them.
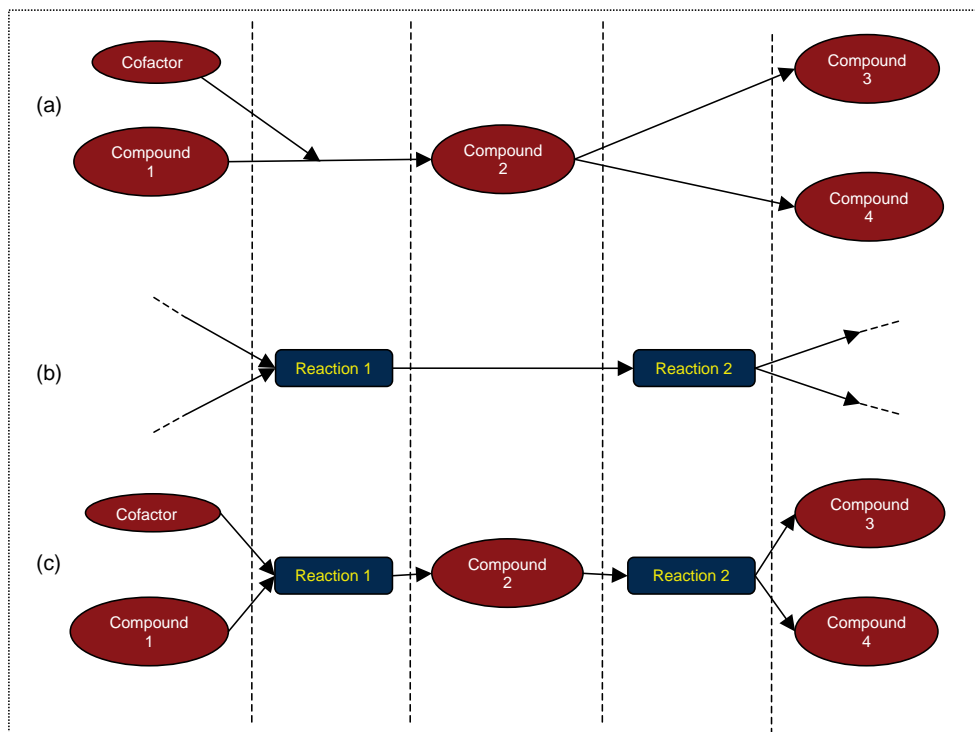
Figure 5.2 The different graph formats for pathways used for representing pathways. (a) The compound-centric method cannot represent multiple reactants well. (b) The reaction-centric view does not allow compounds participating in multiple reactions. (c) A bipartite graph allows any compound to participate in any number of reactions, and any reaction to have any number of reactants/ products.

5.2.1
Compounds

In the bipartite graph representation, there are two types of nodes, the metabolites and the metabolic reactions that link them. With this representation, it is necessary to unambiguously define molecules in order to guarantee that they are unique. This is important for numerous reasons, the main one being the need to avoid redundancy when merging data sources. Unfortunately, different data sources often contain different levels of detail in defining molecules, and this discrepancy in the information can lead to ambiguity and consequently duplication. The representation of the necessary information in computational form will be discussed in the following section.

5.2.1.1 The Importance of Having Uniquely Defined Molecules

Several aspects have to be taken into consideration when treating metabolites. There are numerous levels of representing the differences between compounds, all of which impact their biological effect and metabolic fate. At the most basic level, it is necessary to consider a metabolite's chemical formula. This is of importance for guaranteeing mass balance for reaction equations.

However, the existence of isomers gives rise to many more levels of complexity, which necessitate more detailed information. While constitutional isomers, that is, those where the connectivity of the atoms differs despite having the same chemical formula, have different chemical properties and can be treated as distinct compounds, stereoisomers are far more difficult to treat (Figure 5.3). Stereoisomers are often chemically indistinguishable, but biologically they can be very different. Many enzymes and receptor proteins are stereospecific, that is, they act only on, or in response to, one form of the molecule. For example, different enantiomers of carvone have different perceived flavors ascribed to them ((+)-carvone from caraway seed, (−)-carvone from spearmint) [13], suggesting the presence of multiple, stereoselective odorant receptors and associated signaling pathways [14]. For an extensive listing of enantiomers and their flavors, we refer the reader to the excellent website of Leffingwell [15]. Stereoisomerism also plays a role in the pathways a molecule enters into, since different enantiomer-specific degradation routes are possible, depending on the conformation.

Another physiologically important difference between molecules is the charge and protonation state. This is biologically relevant, as different cell compartments often operate at different pH values, and it is therefore necessary to take this difference into account in a computational representation. As the various databases may represent compounds in different protonation states (some may represent the compounds at the relevant pH value, whereas other may standardize all compounds by giving their protonation state at physiological pH), manual editing may be necessary to guarantee that the correct representation is used for subsequent calculations.

5.2.1.2 Representation Formats

Ideally a computer file format for storing chemical information needs to unambiguously define all of these levels of detail. The most informative representation of a

(-)-(R)-carvone

(+)-(S)-carvone

(R)-(-)-pentanol

(S)-(+)-pentanol

(R)-tetrahydrozerumbone

(S)-tetrahydrozerumbone
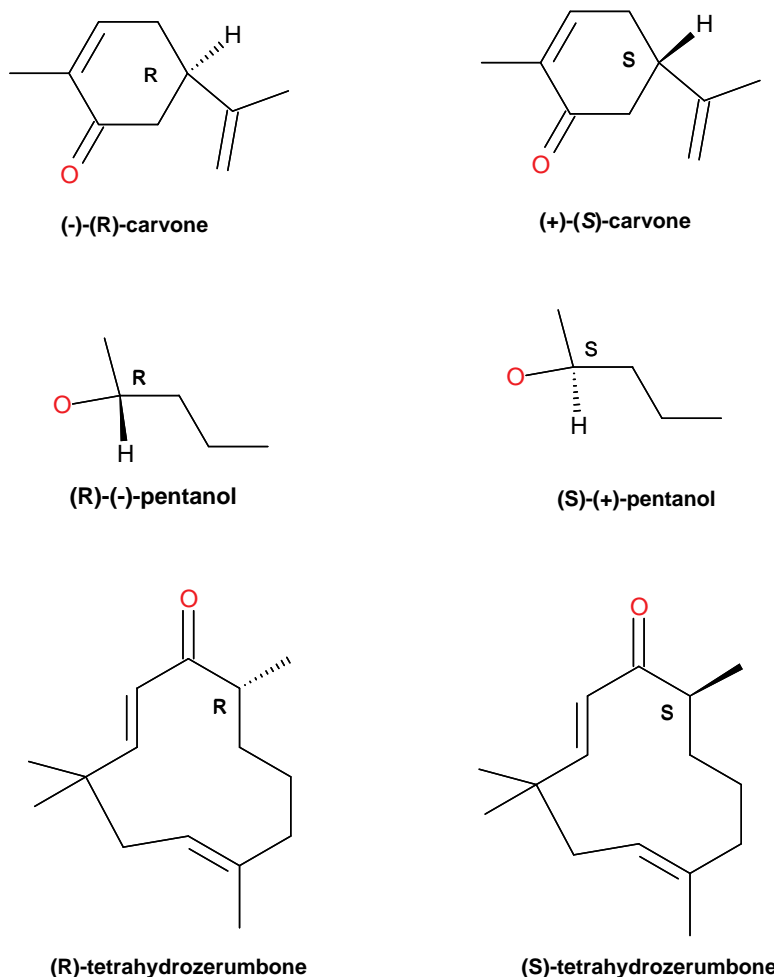
Figure 5.3 The biological importance of capturing stereochemistry. Different enantiomers can have very different biological properties that need to be taken into account. This may be important in areas such as fragrance development where strength and quality of a compound is determined by its racemic composition. The two racemers of carvone, for example, have very different flavors, as do the different enantiomers of 2-pentanol [16]. For tetrahydrozerumbone, the intensity of the fragrance varies strongly between the enantiomers [15].

molecule is its three-dimensional structure; this defines its atomic configuration and stereochemistry. Using the structure as an identifier is not useful as it is not easily human-readable; its size also makes it badly suited for storage and searching in a database. Instead of 3D structure information, a unique identifier is needed, which unambiguously allows one to identify a molecule and which can be used for efficient database searching. In the following section, three key ways to identify

molecules are discussed; this is not an exhaustive list, but rather a treatment of the advantages and drawbacks of the key formats.

Simplified Molecular Input Line Entry Specification (*SMILES*). SMILES codes [17,18] represent the structure of a molecule in string form, making it amenable to easy storage in a database. SMILES codes are generated from the molecular structure (both 2D and 3D structures can serve as input) of a molecule and contain the necessary information to reproduce the structure. Isomeric SMILES that define the configuration at stereocenters can also be generated. The SMILES code derived from a structure depends on the implementation of the algorithm. As a consequence, there can be a one-to-many relationship between molecular structures and SMILES codes if different algorithms are used. Another drawback is the fact that the format is dependent on commercial software, although open source alternatives of the algorithm do exist.

International Chemical Identifier (*InChI*). The International Union of Pure and Applied Chemistry (IUPAC) devised this[1] string-based representation for molecular structures, which circumvents some of the problems with the SMILES code. Again, the code can be generated from the structure and vice versa. Unlike SMILES, however, only one InChI code can be generated for each compound, making a string comparison sufficient to determine the identity of two molecules. It also presupposes idealized geometry, making it suitable for small molecules. It is human-readable to some extent, but not as easily as SMILES codes. One big advantage of the InChI is the layering of information, permitting increasingly complex levels of information about the structure to be incorporated into the string. For example, it can inform about missing details, such as lacking stereochemical information, in the structures used to generate the InChI. It also allows the charge and (de-)protonation state to be represented. The identity of two molecules can therefore be assessed using only one string comparison. As the InChI can become very long and unwieldy for efficient database searching, the so-called InChIKey can be generated: by means of lossy compression, an InChI code can be reduced to 27 characters, making it amenable for use as an efficient search index on database tables. One further benefit is that the nonproprietary nature of the format/algorithm allows it to be used freely and without restrictions.

CAS registry numbers. One of the most widespread identifiers for chemical substances is the registry number [19] assigned by the Chemical Abstracts Service (CAS), a division of the American Chemical Society. This number identifies compounds or mixtures. It is widely used to the point of being the de facto standard in many fields (e.g., labeling of substances by chemicals suppliers) and covers a very large compound library. It is an arbitrary number and thus does not contain, by itself, any information on the structure or even the general nature of the substance – it is merely a reference to an entry in the CAS repository, which itself contains the details of the molecular composition. The drawbacks of CAS numbers are due to their proprietary nature: they are assigned by CAS to new chemicals and therefore cannot be generated for new molecules that have not yet

---

1) The IUPAC International Chemical Identifier (InChI).

been processed by CAS. Also, their use in databases is limited and subjected to licensing provisions. Their ubiquity, however, have made it them preferred choice for many, including chemical suppliers.

5.2.1.3 Key Chemical Compound Databases

The foundation of any metabolic database is formed by the reference compound list that contains all metabolites mapped in the repository. When extending a database, it is necessary to have access to compound information beyond the existing ones. When new pathways are discovered and need to be integrated into the existing database, it is desirable to be able to obtain the molecular structures and unique identifiers for these compounds. While in some cases it is necessary to implement a system for entering new compounds "manually" and generating a unique code for them, a number of databases already exist that can be used for this purpose, and thus can save development time if they are sufficient for one's needs.

PubChem [20] is a public repository of chemical information that (in part) relies upon contributor data. It consists of three databases (BioAssay, Compound, and Substance), but for metabolic modeling, only Compound database is relevant. It has a convenient Web service relying on the Entrez interface for ease of access, but more importantly it can be downloaded for local storage and processing. Its considerable size and thus coverage makes it an attractive reference database for in-house use. PubChem supplies a unique identifier (an integer number without any relationship to the compound) for each chemical structure (CID), which takes into account differences in stereochemistry between otherwise chemically equivalent compounds. Essentially, the mapping is one-to-one between CIDs and InChI codes, meaning that when matching compounds based on their InChI code, the PubChem Compound database obviates the need for generating one's own unique identifier, provided the level of information included is sufficient. In addition, the database also contains an extensive list of synonyms for each compound; this includes common names and systematic chemical names. These make the information in the database amenable to human users, as well as supplying a powerful resource on which text mining operations can be based.

The CAS Registry (CAS Registry is a service mark of the American Chemical Society) is a reference database of chemical substances that is accessible via the SciFinder service.[2] It can contain complex chemical mixtures as well as pure compounds. This curated repository of chemical information supplies a unique, proprietary ID (known as the CAS registry number) for each compound or substance. The added value of the CAS registry, besides the ubiquity of the CAS identifier in the field of chemistry, is the additional information collected by curators about the substance. Its drawback compared to PubChem is the proprietary nature of the information and the restrictions this entails.

ChemSpider [21] is an extensively hand-curated chemical compound database that places strong emphasis on the correctness of structure-name assignments. Chem-Spider also contains properties of chemicals and links to many external databases. For

2) SciFinder, Chemical Abstracts Service, Columbus, OH.

each entry, multiple identifiers are available, distinguishing it from the CAS repository. While it started out as a community effort, it is now supported by the Royal Society of Chemistry [22]. Unfortunately, the database as a whole is not downloadable, and its access is limited to the Web portal. This can be a limiting feature as each compound has to be retrieved manually, and for larger data sets this process becomes prohibitive and a fully downloadable database may be preferred.

5.2.2
Reactions

In the bipartite graph representation of metabolic networks, the reaction forms the second class of nodes.

5.2.2.1 Definitions of Reactions

Essentially, a reaction is defined by the input and output compounds (reactants and products). The exact definition of a reaction may be different, depending on the level of detail required. In some cases, it may be that a multistep reaction is condensed into a single reaction, for example, in cases where only the reactants and end products are of interest or where the intermediate steps are unknown. A direction of the reaction may also be included in the definition, according to the predominant flow of reactants in a biological system.

5.2.2.2 Importance of Stoichiometry and Mass Balance

The principle of mass balance dictates that the number and identity of atoms entering into a reaction must be the same as those leaving, and thus a formal restraint is placed on the formulation of reaction equations. Some problems are observed when encountering reaction equations in the literature or in databases. Often, authors are only interested in tracing certain key compounds in a set of reactions; the equations may thus be incomplete in terms of mass balance. Furthermore, differences in the protonation state of reactants or products of equivalent reactions in different databases would be treated as separate reactions by computational methods that have not been devised to deal with this ambiguity. This source of ambiguity may lead to redundancy in databases. A good representation of molecules (i.e., a method that is resilient to ambiguities in the protonation state, such as use of InChI) can help avert this ambiguity. Otherwise manual curation may be needed to rectify ambiguity and duplication.

5.2.2.3 Atom Tracing

For some applications, such as MFA, it is necessary not only to track the quantitative flow of atoms through the network but also to know specifically which atoms in the initial metabolite will be at which position in the resulting compounds (Figure 5.4). To achieve this, tracing the atoms through each reaction must be possible. This is not necessarily apparent just from the molecular formulas of the reactants and products themselves, but has to be explicitly defined separately for each reaction. Currently, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database is the only
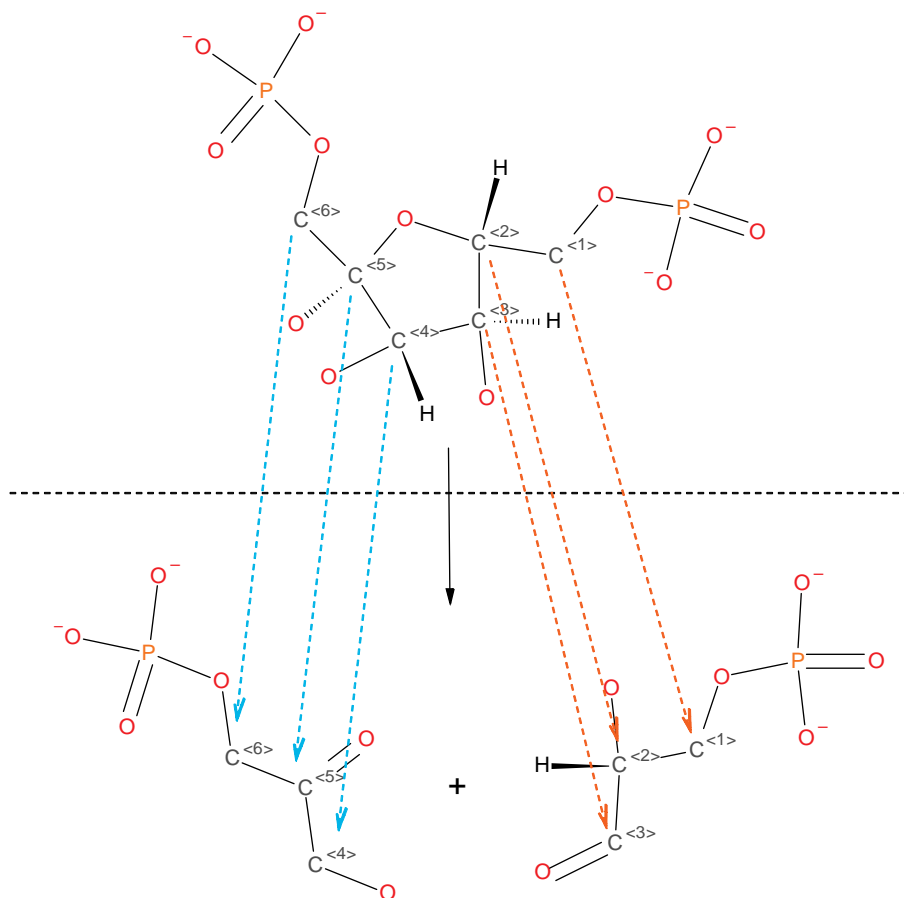
Figure 5.4 Tracking atom transitions for a reaction. One of the first steps of glycolysis is splitting fructose-1,6-bisphosphate into dihydroxyacetone-3-phosphate and D- glyceraldehyde-3-phosphate during glycolysis. Tracking the atom transitions allows the metabolic fate of the individual atoms to be recorded.

standard pathway repository providing such information, namely in the database RPAIR [23]. This resource is however not complete and in cases where the information is unavailable, manual assignment of the atom transitions may be necessary. The group of Peter Karp has developed a tool to automatically assign atom transitions to reactions [24], and applied this technique to the MetaCyc database. This tool is to be released as part of the Pathway Tools (PWT) software suite.

5.2.2.4 Storing Enzyme Information: EC Numbers and Their Limitations
In biological systems, enzymes catalyze reactions by stabilizing the transition state and thus lowering the activation energy. The standard definition/identifier for an

enzymatic reaction is the EC number – a four-digit number assigned by the International Union of Biochemistry and Molecular Biology (IUBMB). This number is supposed to unambiguously define a class of enzyme; there is no real way for the EC number to define the specificity of an enzyme for a particular substrate, aside from the last digit – this information may be only very general e.g. the reaction number E.C. (1.1.1.1), and is not a guarantee for unambiguity. When EC numbers are not available, enzyme names can provide an alternative; however, these are far more ambiguous in their specificity.

5.2.3
Pathways

Metabolic pathways are essentially a collection of subsequent reactions, whereby the products of the preceding reaction form the reactant for the next reaction.

5.2.3.1   How Are Pathways Defined?

Pathways are usually defined as the path taken by one key compound to reach another, such as the conversion of glucose to pyruvate during glycolysis. However, not all pathways are linear; many complex, branched pathways have also been defined in the literature. Particularly cofactors such as NAD/NADP and ATP are pervasive throughout the metabolism and so greatly increase the connectivity of metabolic networks.

5.2.3.2   Typical Size and Distinction between Pathways and Superpathways

The bipartite graph of all reactions and compounds forms the metabolic network of the cell. This can be subdivided into individual pathways. While this subdivision can be essentially arbitrary, there are certain rules that various pathway databases employ. Typically, up to half a dozen reactions form a pathway. Often, a pathway is defined as the set of reactions between key metabolic intermediates such as pyruvate, or key amino acids, which act as branching points to other pathways. Pathways may be joined to form superpathways in Pathway Tools. KEGG defines pathways by means of pathway maps, which are sub-maps of the entire metabolic network and may be of considerable size.

5.3
Pathway Management Platforms

Over the past few years, various software platforms have been developed to compile, process, and store a wealth of pathway information (Table 5.1 provides a brief overview of some of the most relevant ones). The spectrum of all the different software tools available is too broad for the scope of this chapter. We will focus on the two main pathway software packages that are most appropriate for plant pathway modeling: KEGG and Pathway Tools.

Table 5.1 Pathway management software and their features.

Name	Downloadable	Package type	Supplied databases	Entering new data	Organism-specific database creation	Pathway searching facility	Web site
Pathway Tools	Yes	Executable, includes integrated Web server	MetaCyc	Yes	Yes	Via application and Web interface	http://bioinformatics.ai.sri.com/ptools/
KEGG	Yes	Web site, local installation possible	KEGG	No	Yes, to a certain extent	Via Web interface	http://www.genome.jp/kegg
MetaCrop	No	Web site, remote only	MetaCrop	No	No	Via Web interface	http://metacrop.ipk-gatersleben.de/
WikiPathways	Yes	Executable	None	Yes	No	No	http://www.wikipathways.org/index.php/WikiPathways

5.3.1
Kyoto Encyclopedia of Genes and Genomes (KEGG)

The KEGG platform [25] is a large collection of databases covering a wide variety of biochemical information. The main database of interest here is the LIGAND database [26], which contains the chemical compounds reference library as well as the reactions connecting them. The Web-based interface provides facilities for searching compounds and reactions of interest and navigating and visualizing the pathway data. The KEGG platform has no facility for entering data, and as such its use is limited to read-only capability.

5.3.1.1   Database Structure in KEGG
At its core, the LIGAND database uses the bipartite graph structure reflecting the many-to-many relationships between compounds and reactions. Reactions are cross-referenced to a database containing enzyme information, thus linking metabolism with genomic data. By following the interconnected compound and reaction nodes of the graph, pathways can be reconstructed to arbitrary size by the user. The database does however provide its own pathways that consist of list of interconnected compounds and reactions. Many of these are also available in the form of graphical maps.

5.3.1.2   Navigation through KEGG
A string-based search will allow the user to identify possible compounds of interest. Each compound entry will contain references to the reaction it participates in, and each of these reactions will contain references to each participating compound. This way, one can navigate the metabolic Web, following the links between the entries. This nonvisual method of navigation is supplemented by the pathway networks provided by KEGG. Manually created maps of pathways, some of which are considerable in size, display the information in the KEGG database visually (Figure 5.5). The ATLAS feature of KEGG goes even further. This feature contains a large-scale map of a large number of interconnected pathways. These represent whole sections of the biochemical network of the cell and serve as an invaluable tool for exploring the metabolism.

5.3.2
The Pathway Tools Platform

The Pathway Tools package, developed by the research group of Peter Karp at Stanford Research Institute, provides a full framework for constructing, viewing and curating organism specific pathway databases. This framework allows the user to link metabolic and genomic data in what is referred to as a Pathway/Genome Database (PGDB) [27]. It comes with a comprehensive knowledge base (MetaCyc, described in Section 5.4.1.2), providing a solid foundation for building in-house, organism-specific databases.
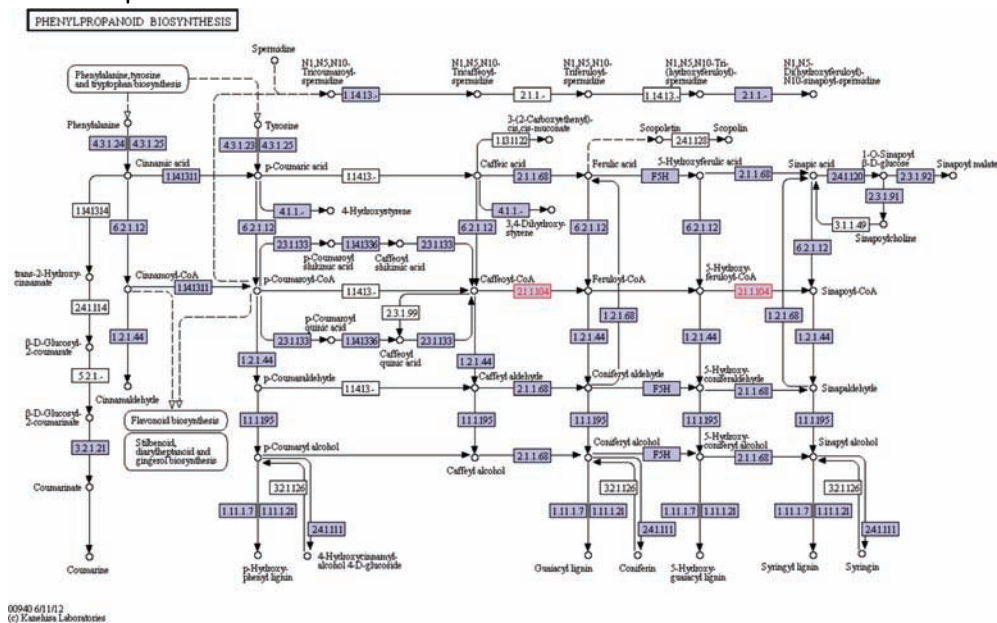
Figure 5.5 An example pathway from KEGG. The phenylpropanoid biosynthesis pathway (ID: ko00940), as it is displayed by the KEGG Web server software. The layout is static, though links to some enzymes and metabolites can be followed. Image from the KEGG database (http://www.genome.jp/dbget-bin/www_bget?ko00940), reproduced with permission. (c) Kanehisa laboratories.

5.3.2.1 Database Management in Pathway Tools

The Pathway Tools schema is build upon the bipartite graph structure outlined above. Compounds are defined by means of unique internal identifiers that are associated with a number of other standard identifiers such as the InChI code. Reactions link combinations of compounds as reactants and products. Building on this foundation, pathways are defined as sets of coupled reaction, and superpathways as sets of linked pathways. This nested definition allows information to be extracted in an intuitive, recursive procedure.

5.3.2.2 Content Creation and Management with Pathway Tools

The software front end to the platform has a Graphical User Interface (GUI) facility that gives the user multiple capabilities in managing pathway information. Entities such as compounds, reactions, and pathways can be entered into the system, making it a convenient system for storing in-house data for future use. Tightly coupled to the GUI is the PathoLogic, a tool that can be used to create organism-specific database from genomic data (see Section 5.2.2) by mapping enzymes identified in the organism to specific pathways in the reference database. The resulting database can be curated and augmented manually using the GUI editing facilities. In addition to the graphical features, Pathway Tools provides a programming interface for the language Lisp,

which not only gives the user access to many of the functions available via the GUI, but also allows the user to devise custom scripts and programs to automate task, perform complex queries, and extract the data in user-defined formats.

5.3.2.3 Pathway Tools' Visualization Capability

Pathway Tools gives the user the ability to browse pathways at different levels of detail (an example of a high-level, low-detail pathway is shown in Figure 5.6). The level of detail can be increased or decreased interactively by the user. Pathway Tools also offers the possibility of viewing an entire PGDB. Using the Web server facility of Pathway Tools, an overview of the metabolic map can be created and navigated using a Web browser-based tool, and search function allow compounds, reactions, and
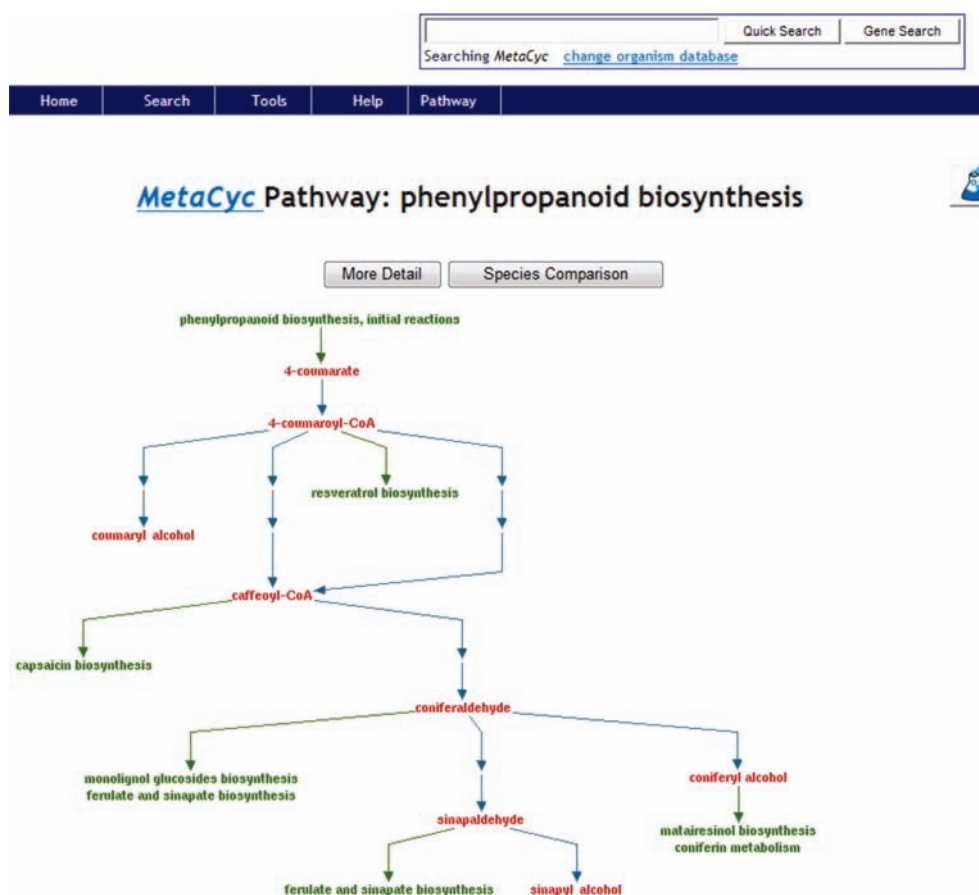
Figure 5.6 An example pathway from MetaCyc. The database contains numerous pathways of interest to plant biology (here the phenylpropanoid biosynthesis pathway is displayed), which can be visualized using the Pathway Tools server software. Note that "more detail" can be adjusted by the user. Image reproduced with permission.

pathways to be searched for and displayed. The online viewing facility of Pathway Tools also has the in-built capability of mapping gene expression data to metabolic pathways, allowing these two sources of information to be coupled visually.

5.4
Obtaining Pathway Information

All specific pathway information is essentially gleaned from the published scientific literature. Much metabolic information, particularly on the core metabolism, has been already compiled in a computer-accessible form. For more specialized areas of metabolism, such as the secondary metabolism of nonmodel organism plants, information may have to be compiled separately. Here, we give an overview of the standard resources and tools at a bioinformatician's disposal for collecting metabolic information.

5.4.1
"Ready-Made" Reference Pathway Databases and Their Contents

Compiling high-quality information from the scientific literature creates high-value resources that can be mined with ease using modern technologies. This process is however a labor-intensive and costly exercise. Fortunately, curators have already made significant efforts in compiling information in the field of life sciences. For example, the protein function information in Swiss-Prot has made it an indispensable resource for researchers. For metabolism, there are two key database resources, KEGG and MetaCyc, that contain manually curated information compiled from the scientific literature, on pathways from multiple organisms, including plants. In addition, several smaller initiatives exist, where information for selected organisms was collected and compiled into a resource. Table 5.2 gives an overview of the most relevant databases in the area of plant metabolism.

5.4.1.1 KEGG
As already mentioned, the KEGG LIGAND database is a manually curated reference database that covers reactions and ligands from a broad range of organisms. It provides a reference knowledge base for metabolic and biochemical pathways, to which gene information can be mapped [35].

5.4.1.2 MetaCyc and PlantCyc
MetaCyc is a universal reference database that is manually curated by experts. It contains pathway information from all kingdoms of life, drawn from the large body of evidence in the scientific literature. While it is available in various formats, including a machine-readable flat file format, it is best used with the Pathway Tools software. The PlantCyc database [32] is a specialized reference database for use in Pathway Tools, which is limited to organisms from the plant kingdom (a number of

Table 5.2 Metabolic pathway databases relevant to plant research.

Name	Type	Maintainer	Pathway Tools compatible	Manually curated	Key publications	Web site
MetaCyc	Universal	Pathway Tools group, Stanford Research Institute	Yes	Yes	[28,29]	http://metacyc.org/
KEGG	Universal	Kanehisa Laboratories, Kyoto University	No	Yes	[25]	http://www.genome.jp/kegg/
MetaCrop	Plant specific	Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)	No	Yes	[30]	http://metacrop.ipk-gatersleben.de/
AraCyc	Species specific	Plant Metabolic Network (PMN)	Yes	Automated build + annotation	[31]	http://www.arabidopsis.org/biocyc/
PlantCyc	Plant specific	PMN	Yes	Yes	[32]	http://plantcyc.org/
PMN database collection	Species specific (AraCyc, *Medicago truncatula*, and Poplar)	PMN	Yes	Yes	See Web link	http://plantcyc.org/
Gramene collection	Species specific (rice, sorghum, maize, and *Brachypodium distachyon*)	Gramene team	Yes	Yes	[33]	http://www.gramene.org/pathway/
SolCyc collection	Species specific (coffee, pepper, tomato, potato, and tobacco)	Sol Genomics Network	Yes	Automated builds	[34]	http://solgenomics.net/

reactions and compounds in these two databases overlap). The authors' analyses show that fewer false positives were identified during the creation of plant-specific pathway genome databases (PGDBs), when creating a plant-specific database using PlantCyc, compared to using the universal database MetaCyc alone [32].

5.4.1.3 MetaCrop

MetaCrop is an "information system" [30] that provides access to a detailed, manually curated database of metabolic information of a range of crop plants. Pathways of interest can be searched for using the Web interface provided on the resource Web site [36]. While it is, strictly speaking, not a database, in that it is accessible only through the Web site, it does allow for metabolic pathway information to be downloaded in a machine-readable file format (Systems Biology Markup Language, SBML).

5.4.2
Integrating Databases and Issues Involved

As the information contained within the various databases is complementary, it may be desirable to merge their content, so as to get a more complete picture of the metabolism. Significant efforts to integrate databases have been made, as for example for integrating the MetaCyc and KEGG family of databases [37], however these developments are not necessarily applicable to all available metabolic databases. It thus may be necessary to devise in-house protocols for merging various data sources. Thereby it is important to avoid duplicating entities in the resulting database. Some of the key issues in this merging process are discussed in the following.

5.4.2.1 Compound Ambiguity

The definition of reaction equations and consequently pathways depends on the correct identification of compounds. Their correct identification is therefore essential to integrating metabolic information from different sources, otherwise duplication and redundancy make mining these data difficult or even impossible. The various databases contain compound information at different levels of detail, depending on the format used for storage and the accuracy requirements for their intended purpose. Insufficient stereochemical information, for example can make it impossible to match the same compounds in different databases. As seen previously, a further cause of problems is the protonation state of the compounds. Different databases may pursue conflicting policies in this regard, as some will choose a physiological pH as the default while others may choose to represent models in their neutral state. A further cause may be differences in the file formats: different versions of the InChI code for example can cause problems, as they may contain different levels of detail on stereochemistry and charge of the molecule.

5.4.2.2 Reaction Redundancy

In the case of reactions, merging databases is also subject to a number of problems, due in no small part to the ambiguities involved in identifying the participating
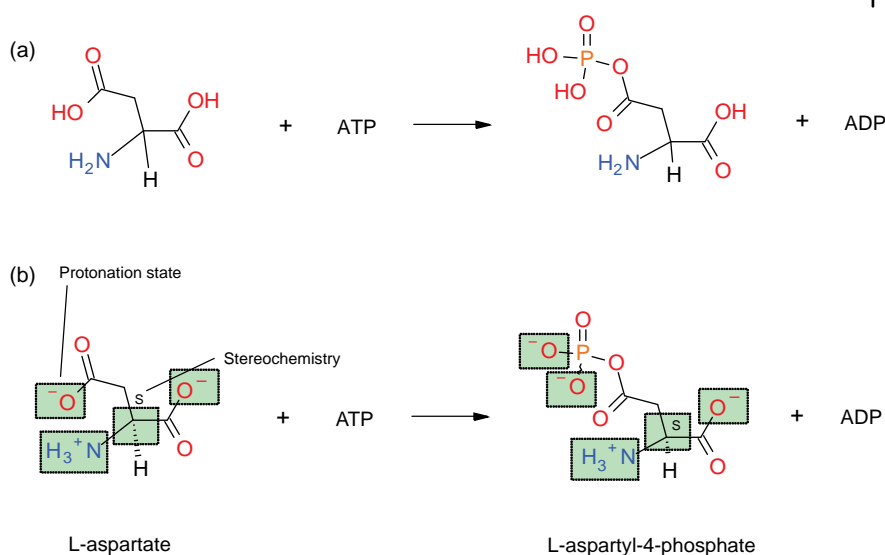
Figure 5.7 Two potential representations of the same reaction. Potential sources of ambiguity which could be encountered when merging databases are highlighted. In this example (aspartate kinase, EC 2.7.2.4), differences in protonation state and stereochemistry of the reactants and products compounds could lead to the reactions not being identified as equal. (a) This representation is ambiguous with regard to stereochemistry and the molecules are protonated. (b) This representation contains information on stereochemistry and deprotonation.

compounds. Reactions from different databases may be the same, but the compounds are defined at different protonation states (Figure 5.7). Ambiguities in compound stereochemistry are another source of problems. For many reactions, stereochemistry is not important, as only one enantiomer predominates in nature and the other is not prevalent. Nevertheless, there are cases where the metabolism of a compound is stereospecific. Sometime a database will leave the stereochemical identity ambiguous, making the exact identity of the reaction unclear.

5.4.2.3 Formats for Exchanging Pathway Data

To facilitate the transfer of knowledge between databases and programs that process these data, a file format for pathway information is required. Several formats have thus been devised, albeit for slightly different purposes. The Systems Biology Markup Language (SBML) [38] is a format suited mainly for metabolic pathway modeling applications. It contains information on compounds and their coupling via reactions. Kinetic information on the reaction rates can also be included, which is necessary for many applications that do not use the steady-state assumption of flux balance analysis. Another important exchange format is BioPAX [39], a community standard for transferring pathway information. BioPAX can encode not only metabolic pathway information, but also signaling pathways.

5.4.3
Adding Information to Pathway Databases

Ready-made pathway databases contain a vast wealth of information that is curated and well documented. However, due to the extent of information in the literature, it is not realistic to expect all the available information to be mined. Many pathways, particularly those that are not very well studied, for which information is sparse or very new, are not covered in these databases. If this information is required, it is necessary to collect this information by oneself. In the following section, standard procedures and methods of extracting this information are outlined and discussed.

5.4.3.1 **Manual Curation**
High-quality information is still best extracted by an expert in the respective field. A manual curator reading literature on biochemical pathways can recognize the terminology and the context of the statements. What is also very important is being able to assess the level of experimental evidence and make an informed decision about whether a reaction is likely to be valid. Many hypothetical reactions and pathways may also be described in an article, and only a curator can detect this from the sometimes complex context in which such a statement appears. In addition, much of the metabolic information in scientific articles is in the form of figures. These contain not only drawings of molecules but also schematic representations of reaction as flowcharts with compounds represented as molecular drawings rather than text. Chemical structure recognition has been a field of research for some time now [40,41], but it is not a process that is currently applied routinely using automatic protocols. It is still a field of interest and active research, and tools for chemical structure recognition from images boast steady progress [42].

In order to collect these data in a computationally useful manner, tools for input are required. There are many tools that facilitate this process, which are easy to use and guarantee the quality of the input data.

WikiPathways [43] is a tool enabling teams of biologists to annotate pathways and networks in a collaborative manner. It provides a fast and easy-to-use WYSIWYG (What You See Is What You Get) interface for visualizing and editing pathways in graph form, which can be stored on a central server accessible to all the members of a community [44]. The software uses a schema, which can be saved in the GenMAPP Pathway Markup Language (GPML), a custom XML file format, that allows knowledge to be exchanged between users, as well as permitting machine processing of the information. The model used for storing pathway information is a compound-centric view of metabolism, whereby the compounds are represented as nodes and the reactions are represented as edges. This form of representation has the advantage of being easily understood by scientists without training in computer sciences, but the drawback is that it is hard to represent the many-to-many relationship of chemical compounds and reactions already discussed.

5.4.3.2 Automated Methods for Literature Mining

Reading journal articles or other sources of information is time-consuming and is not cost-effective when the volume of data is large. The development of methods for automatically extracting information from text (Natural Language Processing or NLP) has been the focus of much research, and has in recent times received a lot of attention from the computational biology and chemistry community. For metabolic pathway reconstruction as such, some tools can be adapted to the task. To the best of our knowledge, the only tool devised specifically for the task of extracting chemical reaction pathways is the ChemicalTagger [45]. (The article in Ref. [45] contains a very informative overview of the entire text mining workflow.)

The task of NLP can be roughly divided into two parts: first, assigning a syntactic function (meaning) to the words, and second, understanding the relationships among the words. For the first part, one can match strings matching against a dictionary of chemical names. Alternatively, one can use machine learning (ML) methods aimed at recognizing relevant words. String matching has the advantage of being exact; its drawbacks are the partial incompleteness of the reference dictionaries, ambiguity of synonyms (many-to-many relationships between compounds and names), and the vast number of names and synonyms, which can make the task computationally very expensive. Machine learning methods, such as the OSCAR3 program [46], for recognizing relevant terminology can be a viable alternative; they are more flexible in that they only need a dictionary to be trained and can subsequently recognize previously unseen terms. This makes them more flexible and they require less memory; however, their downside is the relatively high error rates in correctly identifying terms. For the second part, extracting the structural relationship among the entities, it is necessary to parse the syntax of the sentence, leading to a tree representation of the sentence being analyzed (syntactic tree). This represents the relationships between the chemical entities in the text. Using one of the two NLP techniques, chunking or deep parsing, this grammatical structure can be extracted, and a relationship graph can be constructed based on this tree.

A number of limitations are associated with text mining:

- Correct chemical entity recognition is vital.
- Completeness is not always guaranteed; multistep pathways may be coerced into one step, so several intermediate reaction steps may be missing.
- The stoichiometry is not necessarily specified in the text portion of an article; therefore, either manual curation or mapping against a reference database of reactions is required.
- A reaction described in the text may be hypothetical, and this may only be apparent from the context of the sentence in the whole article; this information can be obtained only by a human.
- A reaction described may only be tentative, and the level of evidence cannot easily be judged by an automated method.
- The grammar of a sentence may not be complete, potentially leading to ambiguities in interpreting the relationships among the entities.

Text mining techniques are powerful methods for extracting large amounts of information from the literature, but the linguistic subtlety and ambiguity can prove intractable to current machine learning(ML) methods. The results of any ML process should therefore always be reviewed by a curator.

5.5
Constructing Organism-Specific Pathway Databases

In this section we outline possible ways of generating an organism-specific database, given the genome of an organism. Ideally, it would be desirable to infer all pathways based on experimental evidence obtainable for the organism of interest. Such comprehensive information is not generally available for any single organism, but rather needs to be obtained from related species. Based on such heterogeneous information, metabolic networks can be assembled. In the following section, the workflow necessary to achieve this goal is outlined and the limitations that need to be born in mind are discussed (an example of workflow is outlined in Figure 5.8).
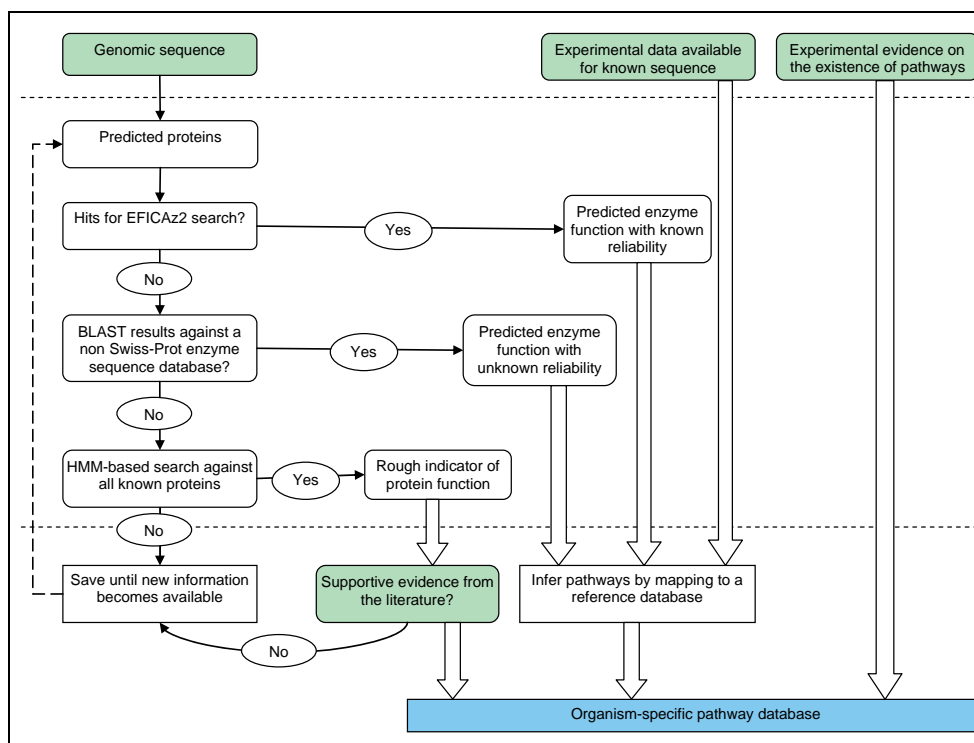
Figure 5.8 A possible workflow for creating a metabolic pathway database for a given organism.

5.5.1
Enzyme Identification

To infer the existence of a pathway in an organism, it is necessary to identify reactions that take place in this organism. This is determined based on the presence or absence of enzymes catalyzing these reactions. Such evidence for the presence of enzymes can be experimental (i.e., from the literature, see Section 5.4.3) or predicted from genomic data using bioinformatics methods.

5.5.1.1 Reference Enzyme Databases

ENZYME database and UniProtKB/Swiss-Prot. The key source of information on proteins is the Swiss-Prot knowledgebase [47], which is created and maintained by a team of experts. It contains curated information on proteins that is derived from the literature and includes enzyme function information together with references to the literature supplying these data. The information is based on experimental evidence, rather than remote homology-based inferences of proteins function in the way that, for example, the TrEMBL database [48] is annotated. As such it may be limited to well-characterized proteins and organisms, but the annotation can be relied upon to be of high quality. Swiss-Prot is thus an excellent source for collecting functional data and identifying candidate enzymes for a given organism.

RESD sequence database. The Reference Enzyme Sequence Database (RESD) [32] was built by the PlantCyc team to complement their metabolic database. It provides reference enzymes drawn from all kingdoms and acts as a reference database for assigning functions to genomic data by homology areas such that it is geared mainly towards the purpose of building PGDBs in an automated fashion.

BRENDA. Likewise, the BRENDA database [49,50] contains information on enzymes and has the advantage of including tentative EC numbers for proteins not yet approved by the IUBMB. Kinetic information is also collected and curated from the literature, making it a more comprehensive resource than those providing only qualitative function assignments. In order to circumvent the cost of "manual" curation, an automated extension of BRENDA has been developed that employs text mining procedures to compile functional information; these boast higher coverage at the cost of lower reliability [51]. BRENDA allows organism-based enzyme searches via its Web interface and, like Swiss-Prot, is a useful tool for identifying known enzyme information for an organism.

5.5.1.2 Enzyme Function Prediction Using Protein Sequence Information

The aforementioned databases contain a wealth of information on enzyme function, compiled by curators from the literature. The experiments needed to determine enzyme function are costly and time-consuming, and so it is not feasible to produce this information for every organism. In the era of genome sequencing, sequence information is becoming available for organisms that are currently not well experimentally characterized, and bioinformatics methods can be used to bridge

this knowledge gap by extrapolating from known information. The bioinformatics techniques used to predict function mostly rely on sequence homology to make functional inferences, but more complex methods, which take into account the protein structure and the shape and properties of the active site, have also been described [52]. Sequence homology-based methods use the assumption that similar sequences have similar function. Depending on the level of similarity, different levels of detail can be gleaned from homologous sequences.

Evolutionary profiles generated from the alignment of closely related sequences can be used to create Hidden Markov Models (HMM) for screening databases of novel sequences. Such a classification may give the user an idea about the protein fold family a protein belongs to, as well as the general function or substrates the enzyme is likely to act on. It will most likely not specify the level of detail required for assigning enzyme function, which is a prerequisite for constructing metabolic pathways.

Generally, a high level of similarity is required before being able to assign an enzyme function to a protein. A BLAST [53] search will identify all closely related sequences to the target sequence. Functional annotation is achieved using a reference sequence database that contains adequate annotations to assign the enzymatic activity. The Swiss-Prot database contains entries for manually annotated sequences, many of which also contain EC number, making it a very valuable generalized resource for the purpose of function prediction.

However, simple homology-based annotation with BLAST does have its limits. More than 60% sequence identity is required to achieve 90% accuracy in EC number assignment [54]. To address this problem, EFICAz [55] and its successor EFICAz2 [56] were developed. It combines multiple sources of information using a machine learning method. The information used includes HMM profiles for known enzyme families and specific functional information in the form of PROSITE patterns [57]. This increases the reported average accuracy to over 90% at a level of 40% identity or above of the target protein to its closest homologue in the set of homologues used to identify it.

When creating a metabolic database with a considerable proportion of predicted enzymatic steps, it is essential to bear in mind a number of caveats regarding these functional assignments. The assumption of sequence similarity implying function may not always be valid; there are numerous scenarios where this does not hold (Figure 5.9). Highly divergent orthologues (homologous proteins with the same function, which were formed during a speciation event), for example, may boast a very high number of neutral mutations and thus have low sequence identity, although the key residues responsible for activity may be conserved; assignment of function by homology in such cases is very difficult [58]. Convergent evolution, that is, where two enzymes of different origins have acquired the same function [59], is also a problematic case, in that functional assignments by homology are impossible. Conversely, paralogues (homologous proteins with a different function, which were formed during a gene duplication event) may differ in terms of the key residues involved in activity despite having accumulated very few mutations at other sites. In addition, the presence of pseudogenes with significant similarity to functional enzymes may lead one to erroneously believe that an enzyme function
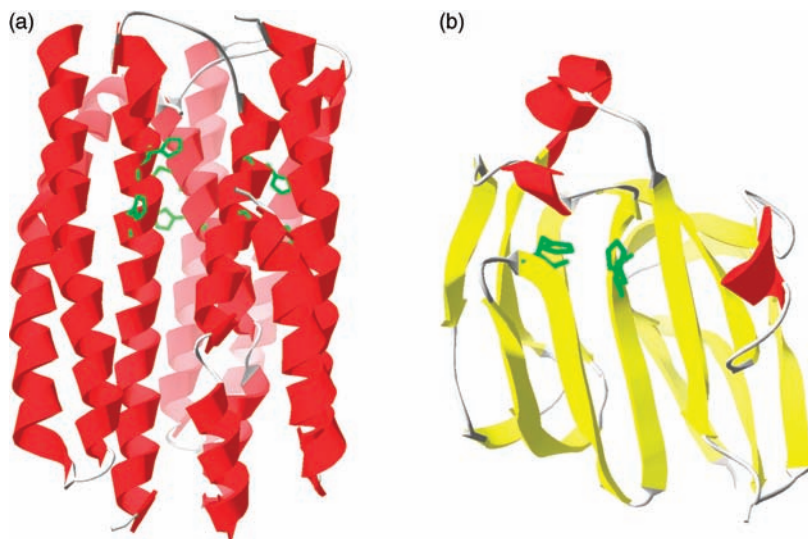
Figure 5.9 Vastly different protein folds provide a problem for annotating enzyme function by homology. These bacterial enzymes [structures taken from the Protein Data Bank, accession codes 2E2A (a) and 1GGR (b)] have different folds, but are annotated as having the same function (enzymatic reaction EC 2.7.1.69).

is present in an organism. Transcriptional analysis can establish whether or not a gene is expressed and therefore functional. Thus sequence identity by itself is not always a good indicator of function. This problem may be circumvented to a certain degree by a "mutual best hit BLAST search," that is, performing an all-against-all BLAST search between two organisms and accepting only BLAST hits for pairs of proteins, which are at the top of each other's search results. This procedure is limited by the requirement of having the entire genome sequence available: if the true top hit is missing from the database, one may falsely assign function to a lower quality hit. A further drawback of using only sequence information is that substrate specificity of an enzyme may not be clear; even if a mechanistic assignment can be made (first three digits of the EC number), the actual target metabolite of the enzyme (the fourth digit of the EC number) may be erroneous.

5.5.1.3 Enzyme Function Inference Using 3D Protein Structure Information

The function of an enzyme is determined by its three-dimensional structure. The arrangement of the catalytic and binding pocket residues defines its function and substrate specificity. While general function may be determined from the fold of the protein and its reaction mechanism may be inferred from conserved residues, its specificity can only be understood when one considers the enzyme's structure. Using structural information, one can perform computational screening methods to determine the predicted binding free energies (or affinities) of ligand for the enzyme. This computational search, generally referred to as "molecular docking,"

is used in drug discovery to identify molecules that are likely to bind to a given protein, and can be used as lead candidates for future investigation [60]. In function prediction, the situation is analogous; assuming that the enzyme is highly specific for its target ligand, the molecule with the highest affinity should be its natural ligand. Using the high-energy intermediate state of the ligand improves the prediction accuracy [61], as enzymes act by stabilizing this transitional form. This has been performed successfully from uncharacterized enzymes from *Thermotoga maritima* [52]. Although this technique is not widely used, it is a technology that holds promise in the future. Such data can supplement, or make more precise, the functional assignments made by means of the primary sequence and in turn use this knowledge to infer the presence of certain pathways in the organism.

5.5.2
Pathway Prediction from Available Enzyme Information

The information from the genome annotation provides the basis for establishing the metabolic database for the plant of interest. The basic idea of creating a model of the metabolic network of an organism is to map the available enzyme information to known reference pathway data. There are numerous considerations in doing this, and thus many different levels of complexity to this process.

5.5.2.1 Pathway "Painting" Using KEGG Reference Maps
The arguably simplest way of constructing an organism-specific database is using KEGG. The KEGG pathway maps can be created for an organism by simply mapping enzyme information onto the "wiring diagram" representation of the metabolic networks.

5.5.2.2 Pathway Reconstruction with Pathway Tools
Pathway Tools provides a tool PathoLogic specifically for assembling the metabolic network of a given organism based on genomic information and annotation. This process differs substantially from the KEGG method; while KEGG provides universal wiring diagrams containing all information from all organisms, while Pathway Tools uses a number of descriptions, including the presence or absence of enzymes, pathway connectivity, and taxonomy, in order to predict whether a pathway is present in an organism [62].

5.5.3
Examples of Pathway Reconstruction

There are a number of pathways built for a wide range of organisms, and which are available to the public. The BioCyc collection, for example, contains hundreds of organism-specific databases built using MetaCyc as a reference database. Among these are a number of plant databases, particularly for model organisms and crop plants. Some of these have extensive manual curation performed on them as, for example, in the case of AraCyc, the currently best annotated one among the plant databases. Other

high-quality automated builds exist such as the SolCyc collection that encompasses a number of automatically generated pathways derived from automated annotation of plants within the Solanaceae. Besides the MetaCyc-based collections, numerous specific models have been created for purposes such as flux balance analysis, and have been made publicly available, as, for example, the barley metabolic map [8].

A metabolic map can be readily generated for a plant whose genome is sequenced or in the process of being sequenced. Our in-house efforts have concentrated on generating pathways from the draft genome of tobacco. We have assembled a pipeline to assign enzyme function to predicted protein sequences, using EFICAz2 and BLAST. The identified enzymes were mapped to reference pathways using the Pathway Tools software component PathoLogic. This PGDB contains around 300 pathways using our current level of coverage. This is comparable in size to the PGDB generated for the fully sequenced (albeit far smaller) tomato genome. We anticipate that with increased quality of the genome draft, we will attain a higher number of enzymes and thus pathways within tobacco. The typical limitations apply to our method as well: while tobacco is well-characterized, coverage of its metabolism is not complete, particularly its secondary metabolite pathways.

5.6
Conclusions

Pathway databases are useful tools in modeling and understanding plant metabolism, and also act as useful repositories for knowledge gleaned from the literature. The best pathway software tools allow the user to store information on compounds, reactions, and pathways in a searchable, extendable manner. By using data from reference databases, it is possible to generate an organism-specific database from scratch for an organism for which the genomic sequence is available. Such databases can then be improved and extended by hand, complementing the data with additional information from new experiments or from the literature. The picture of the metabolism is usually incomplete: (i) only a small fraction of all pathways have been characterized and are present in the reference databases, and (ii) even if the pathway has been elucidated, not all enzymes can necessarily be identified from the genome. There is a distinct lack of coverage of the more specialized secondary pathways. With the ease of use of these tools and ease of producing genomic data, we anticipate that organism-specific metabolic pathway databases will rise sharply in number and will become an integral component of research.

References

1 Oksman-Caldentey, K.M. and Inze, D. (2004) Plant cell factories in the post-genomic era: new ways to produce designer secondary metabolites. *Trends in Plant Science*, **9** (9), 433–440.

2 Ye, X. *et al.* (2000) Engineering the provitamin A (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. *Science*, **287** (5451), 303–305.

3 Wilkinson, B. and Micklefield, J. (2007) Mining and engineering natural-product biosynthetic pathways. *Nature Chemical Biology*, **3** (7), 379–386.

4 Pichersky, E. and Gang, D.R. (2000) Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. *Trends in Plant Science*, **5** (10), 439–445.

5 Ro, D.K. *et al.* (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, **440** (7086), 940–943.

6 Finley, S.D., Broadbelt, L.J., and Hatzimanikatis, V. (2010) *In silico* feasibility of novel biodegradation pathways for 1,2,4-trichlorobenzene. *BMC Systems Biology*, **4**, 7.

7 Henry, C.S., Broadbelt, L.J., and Hatzimanikatis, V. (2010) Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. *Biotechnology and Bioengineering*, **106** (3), 462–473.

8 Grafahrend-Belau, E. *et al.* (2009) Flux balance analysis of barley seeds: a computational approach to study systemic properties of central metabolism. *Plant Physiology*, **149** (1), 585–598.

9 Zamboni, N. *et al.* (2009) (13)C-based metabolic flux analysis. *Nature Protocols*, **4** (6), 878–892.

10 Moriya, Y. *et al.* (2010) PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Research*, **38** (Web Server issue), W138–W143.

11 Hatzimanikatis, V. *et al.* (2005) Exploring the diversity of complex metabolic networks. *Bioinformatics*, **21** (8), 1603–1609.

12 Horne, A.B. *et al.* (2004) Constructing an enzyme-centric view of metabolism. *Bioinformatics*, **20** (13), 2050–2055.

13 Leitereg, T.J. *et al.* (1971) Evidence for the difference between the odours of the optical isomers (4)- and (-)-carvone. *Nature*, **230** (5294), 455–456.

14 Kirner, A. *et al.* (2003) Concanavalin A application to the olfactory epithelium reveals different sensory neuron populations for the odour pair D- and L-carvone. *Behavioural Brain Research*, **138** (2), 201–206.

15 Leffingwell, J.C. (last accessed: 2013) http://www.leffingwell.com.

16 Brenna, E., Fuganti, C., and Serra, S. (2003) Enantioselective perception of chiral odorants. *Tetrahedron: Asymmetry*, **14** (1), 1–42.

17 Weininger, D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, **28** (1), 31–36.

18 Weininger, D., Weininger, A., and Weininger, J.L. (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences*, **29** (2), 97–101.

19 CAS REGISTRY and CAS Registry Number FAQs (last accessed: 2013). http://www.cas.org/content/chemical-substances/faqs.

20 Bolton, E.E. *et al.* (2008) Chapter 12 PubChem: integrated platform of small molecules and biological activities, in *Annual Reports in Computational Chemistry*, Elsevier, pp. 217–241.

21 Pence, H.E. and Williams, A. (2010) ChemSpider: an online chemical information resource. *Journal of Chemical Education*, **87** (11), 1123–1124.

22 RSC acquires ChemSpider (2009) http://www.rsc.org/AboutUs/News/PressReleases/2009/ChemSpider.asp.

23 Shimizu, Y. *et al.* (2008) Generalized reaction patterns for prediction of unknown enzymatic reactions. *Genome Information*, **20**, 149–158.

24 Latendresse, M. *et al.* (2012) Accurate atom-mapping computation for biochemical reactions. *Journal of Chemical Information and Modeling*, **52** (11), 2970–2982.

25 Ogata, H. *et al.* (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **27** (1), 29–34.

26 Goto, S., Nishioka, T., and Kanehisa, M. (1999) LIGAND database for enzymes, compounds and reactions. *Nucleic Acids Research*, **27** (1), 377–379.

27 Karp, P.D., Paley, S., and Romero, P. (2002) The Pathway Tools software. *Bioinformatics*, **18** (Suppl. 1), S225–S232.

28 Karp, P.D. *et al.* (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Research*, **28** (1), 56–59.

29 Karp, P.D. *et al.* (2002) The MetaCyc database. *Nucleic Acids Research*, **30** (1), 59–61.

30 Grafahrend-Belau, E. *et al.* (2008) MetaCrop: a detailed database of crop plant metabolism. *Nucleic Acids Research*, **36** (Database issue), D954–D958.

31 Mueller, L.A., Zhang, P., and Rhee, S.Y. (2003) AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiology*, **132** (2), 453–460.

32 Zhang, P. *et al.* (2010) Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiology*, **153** (4), 1479–1491.

33 Youens-Clark, K. *et al.* (2011) Gramene database in 2010: updates and extensions. *Nucleic Acids Research*, **39** (Database issue), D1085–D1094.

34 Bombarely, A. *et al.* (2011) The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Research*, **39** (Database issue), D1149–D1155.

35 Bono, H. *et al.* (1998) Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. *Genome Research*, **8** (3), 203–210.

36 MetaCrop (last accessed: 2013). http://metacrop.ipk-gatersleben.de.

37 Lee, T.J. *et al.* (2006) BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics*, **7** (1), 170.

38 Hucka, M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19** (4), 524–531.

39 Demir, E. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, **28** (9), 935–942.

40 McDaniel, J.R. and Balmuth, J.R. (1992) Kekule: OCR-optical chemical (structure) recognition. *Journal of Chemical Information and Computer Sciences*, **32** (4), 373–378.

41 Ibison, P. *et al.* (1993) Chemical literature data extraction: the CLiDE Project. *Journal of Chemical Information and Computer Sciences*, **33** (3), 338–344.

42 Valko, A.T. and Johnson, A.P. (2009) CLiDE Pro: the latest generation of CLiDE, a tool for optical chemical structure recognition. *Journal of Chemical Information and Modeling*, **49** (4), 780–787.

43 Pico, A.R. *et al.* (2008) WikiPathways: pathway editing for the people. *PLoS Biology*, **6** (7), e184.

44 Kelder, T. *et al.* (2009) Mining biological pathways using WikiPathways web services. *PLoS One*, **4** (7), e6447.

45 Hawizy, L. *et al.* (2011) ChemicalTagger: a tool for semantic text-mining in chemistry. *Journal of Cheminformatics*, **3** (1), 17.

46 OSCAR3 (last accessed: 2013). http://sourceforge.net/projects/oscar3-chem/.

47 Boutet, E. *et al.* (2007) UniProtKB/Swiss-Prot. *Methods in Molecular Biology (Clifton, NJ)*, **406**, 89–112.

48 Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, **28** (1), 45–48.

49 Schomburg, I., Chang, A., and Schomburg, D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Research*, **30** (1), 47–49.

50 Scheer, M. *et al.* (2011) BRENDA, the enzyme information system in. *Nucleic Acids Research*, **39** (Database issue), D670–D676.

51 Chang, A. *et al.* (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Research*, **37** (Database issue), D588–D592.

52 Hermann, J.C. *et al.* (2007) Structure-based activity prediction for an enzyme of unknown function. *Nature*, **448** (7155), 775–779.

53 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215** (3), 403–410.

54 Tian, W. and Skolnick, J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of Molecular Biology*, **333** (4), 863–882.

55 Tian, W., Arakaki, A.K., and Skolnick, J. (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Research*, **32** (21), 6226–6239.

56 Arakaki, A.K., Huang, Y., and Skolnick, J. (2009) EFICAz2: enzyme function

inference by a combined approach enhanced by machine learning. *BMC Bioinformatics*, **10**, 107.

57 Sigrist, C.J. *et al.* (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics*, **3** (3), 265–274.

58 Bork, P. and Koonin, E.V. (1998) Predicting functions from protein sequences – where are the bottlenecks? *Nature Genetics*, **18** (4), 313–318.

59 Gherardini, P.F. *et al.* (2007) Convergent evolution of enzyme active sites is not a rare phenomenon. *Journal of Molecular Biology*, **372** (3), 817–845.

60 Höltje, H.D. *et al.* (2008) *Molecular Modeling: Basic Principles and Applications*, Wiley-VCH Verlag GmbH.

61 Hermann, J.C. *et al.* (2006) Predicting substrates by docking high-energy intermediates to enzyme structures. *Journal of the American Chemical Society*, **128** (49), 15882–15891.

62 Dale, J.M., Popescu, L., and Karp, P.D. (2010) Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics*, **11**, 15.

6
The Role of Data Mining in the Identification of Bioactive Compounds via High-Throughput Screening

Kamal Azzaoui, John P. Priestle, Thibault Varin, Ansgar Schuffenhauer,
Jeremy L. Jenkins, Florian Nigsch, Allen Cornett, Maxim Popov, and Edgar Jacoby

6.1
Introduction to the HTS Process: the Role of Data Mining

One of the main goals of high-throughput screening (HTS) is to identify, via a high-throughput process [1–4], small molecules that interact with a protein or a system of proteins. HTS uses cell-based assays or cell-free systems to test millions of compounds stepwise. When a target is selected for screening, an initial assay development is required before testing the full compounds collection (primary HTS). In primary HTS, the compounds are usually tested at a single concentration using an assay format where the detected signals are normalized to remove the signal of the buffer or other interacting agents (known agonist or antagonist). This step requires a large infrastructure – screening automation and information technology (IT) – and generates a large number of transformed and corrected data. During this step, statistical analyses are the key components to identify active compounds [5,6]. In fact, to derive a list of active hits, one has to set a threshold between active and nonactive compounds. This threshold is usually intentionally set very low in order to cover the entire active chemical space at the cost of dragging in more false positives. The confirmation step is usually run in triplicates or at multiple compound concentrations. In order to detect false positives, it may use a counter-screen (different HTS readout) or a secondary screen (different target). Hits are then promoted to the validation step and tested at different concentrations in order to produce dose–response curves (DRCs) for the target and the countertarget. The final hit list is generated and annotated using defined criteria (DRC parameters, analytical data, *in silico* contributions).

In the past 10 years, the Novartis Lead Finding Platform has run many hundreds of HTS for different target families [kinases, proteases, GPCRs (G protein-coupled receptors), PPIs (protein–protein interactions)] generating every year, more and more data (% activity, AC50, Ki, etc.). The average size of the primary hit list was multiplied by almost a factor of ten in the past decade while the capacity for validation has only doubled. For each HTS project, we put in place *in silico* support that initially helps the teams to manage information about hits (chemical classes,
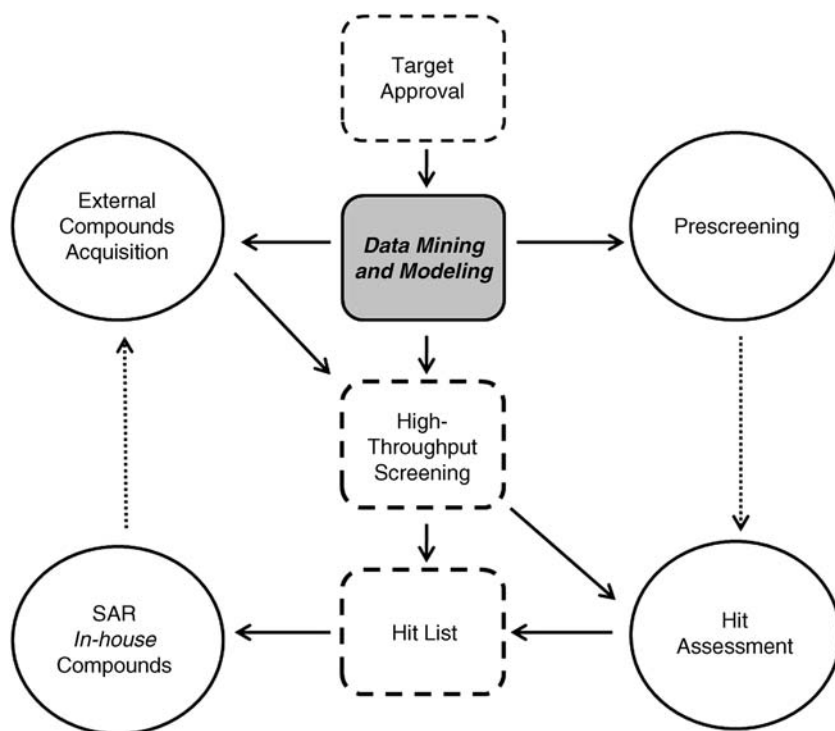
Figure 6.1 Data mining and modeling integrated in the hit finding process. When a target is approved for a HTS, data mining (internal or external database search for known modulators) and modeling techniques (docking, pharmacophore search, similarity search) are applied before running the HTS to support acquisition of external compounds and/or to design a focused set screening when it is necessary. It is also applied after HTS to annotate, assess, and cluster hits and derive structure–activity relationships for compound series.

known chemotypes, frequent hitters (FHs), etc.) [7–9]. Following up on hits using counterscreens and secondary assays can significantly reduce the number of primary or confirmed hits. In some cases the reduction of the number of hits can be achieved by sampling chemical classes or using historical data to deprioritize some hits. Indeed, one can learn from the data generated from each HTS in a historical fashion independent of the projects (collection of chemically unattractive compounds, fluorescents, quenchers, cytotoxics, etc.). As of today, data mining (DM) based on chemoinformatics approaches (as e.g., molecular modeling) are fully integrated in every step of the process (Figure 6.1). DM has become easier partially due to the development of a data architecture called "Hit-Hub" (see below) and different Pipeline Pilot[1] protocols, which allows us to extract information and build models from HTS data. Herein, we will also give examples of techniques and strategies for DM and modeling applied to help drive hit finding.

1) Accelrys, Inc., San Diego, USA.

6.2
Relevant Data Architectures for the Analysis of HTS Data

HT screens typically produce large amounts of data in a short amount of time. The type and amount of data obtained depend on the type of screening experiment performed, which leads to an inherent heterogeneity in the data produced over time. For example, a typical biochemical screen performed at one fixed concentration will produce a list of the activity readout for each molecule tested; however, a high-content imaging screen, where each well from hundreds of screening plates is photographed, produces much more data with higher information content. It is obvious that these two types of screens require vastly different approaches from the data processing and analysis viewpoints. Moreover, for data analysis purposes, for example, hit assessment, it is desirable that data from one screen or screening format can be put in the context of the results of previously performed screens in order to optimize the use of all available information. In addition to the data that an organization produces, analysis of HTS data can benefit greatly from incorporation of public domain information about the screening compounds. The provision and contextualization of HTS hits with external data, for example, licensed bioactivity databases or the PubChem BioAssay repository, are therefore of prime importance [10].

Any IT infrastructure to support HTS data analysis has to meet several requirements:

1) the ability to integrate results from different screening technologies;
2) the ability to provide information from previous HTS campaigns;
3) the ability to provide information obtained from sources outside the screening organization;
4) the ability to keep up to date with the production of internal and external data; and
5) the ability to retrieve information in a structured format quickly to allow efficient computational analyses.

This list of requirements is not particularly detailed nor exhaustive, but rather a generic assembly of typical attributes of any bioassay warehousing system.

6.2.1
Conditions (Parameters) for Analysis of HTS Screens

The efficient and meaningful analysis of HTS data requires a host of related pieces of information to be readily available. Ideally, this would cover all relevant aspects of the screen, ranging from the purity of screened samples to the general assay conditions, and performance of all samples in all previous screens. In the following, we explore briefly what these latter, seemingly simple, attributes – purity, assay conditions, and previous performance of samples – entail.

6.2.1.1 Purity
The goal of screening is to identify a molecular entity that perturbs a biological test system, the assay, in a specific way. The samples to be tested are typically

ordered by and delivered to the scientist performing the assay. This implies that the screener has to trust the sample source. For example, small molecule compounds are typically kept in a central solution repository as DMSO stock solutions at a concentration of several millimoles per liter. The solutions delivered for screening are typically at much lower concentrations, obtained by dilution of the stock solutions. This implies that the average lifetime of a stock solution should be long enough for it to be used for numerous screens, typically over the course of several years. Such long-term storage of compounds can result in their degradation, thereby introducing uncertainty in the contents of the samples tested. Reliable analysis of hit lists, as well as the entire rationale of HTS, is jeopardized if it cannot be ascertained what is perturbing the assay is what its label says. Thus, it is desirable to pass every compound through an analytical quality control procedure before the actual screen, or otherwise subject the stock solution to regular checks. This can be done in high-throughput mode with modern liquid chromatography systems coupled to a mass spectrometer (LC–MS), for example. Depending on the presence or absence and intensity of the signal, a decision can be made if the compound is *present*, and to what extent it is *pure*. Compounds that are no longer present in the sample can hardly be responsible for any observed assay signal.

6.2.1.2 Assay Conditions

To enable efficient data mining across screens that are performed over an extended period of time, it is essential to capture a minimum set of information about every assay. This typically includes the assay format, detection technology, purpose of the assay (e.g., primary, secondary, and counterscreen); however, other parameters such as the pH at which the assay was performed, eventual cofactors used and the concentration at which they were used (e.g., ATP in competition experiments involving kinases), the nature of gene constructs used in the protein production, presence and location of mutations in the target protein, cell lines employed for cellular screens, and so on can all be required attributes at one point in any data mining activity. For example, many compounds have the potential to undergo acid- or base-catalyzed chemical rearrangements. Even if prior to the screen the identity and purity of a compound has been established, the actual active entity could be the product of a pH-dependent reaction. Such events can only be inferred from the data if a large-enough sampling to analyze is available – in this case, all assays are with recorded pH values. Another example would be the detection of frequent hitters for specific detection technologies. Certain compounds will always interfere with certain detection methods, for example, fluorescent compounds might interfere with a fluorescence-based readout. Regardless of detection technology, some compounds will be active in practically any assay due to their inherent physicochemical properties. One such compound would be the flavonoid quercetin that, owing to its many polyphenolic groups, has the ability to have nonstoichiometric effects on a variety of proteins. Knowledge about assay conditions is therefore required to be able to determine which compounds are simply promiscuous and which ones interfere with particular assay technologies.

6.2.1.3 **Previous Performance of Samples**

Further to the examples provided in Section 6.2.1.2, any information about previous performance of samples in any screen is useful for hit assessment. We consider under *previous performance* characteristics known from internal or external sources that include, but are not limited to, the following: (ant)agonistic interactions with other proteins (off-target effects, polypharmacology) or mRNAs in the case of a siRNA screen, biological pathways in which known targets occur, known mode of action (MOA), selectivity indices with respect to related targets such as isoforms of receptors (e.g., selectivity toward one of the three opioid receptors), cell lines that the compound has been tested in, patents that the compound is mentioned in, clinical studies where the compound was used in and the corresponding medical indication, profiling panels that a compound has been submitted to (e.g., safety profiling panels), known toxic metabolites prone to stem from a compound, ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties, or source of the compound (e.g., a particular combinatorial library). Overall, any piece of information that can help to promote, or deprioritize, a single compound or entire compound series.

Overall, the data required for an efficient and proficient analysis of the results of a HTS screen come not only from the HTS campaign in question but rather they also need to comprise many more aspects of the molecular entities screened. Any data-driven approach to hit assessment is therefore preconditioned on the availability of comprehensive annotations for the screened molecular entities. Such annotations have to be aggregated from a number of different sources, and they have to be interlinked accordingly to allow meaningful questions to be answered in an equally meaningful way.

6.2.2
Data Aggregation System

We developed a custom extraction–transformation–load (ETL) system called Hit-Hub for the aggregation of internal Novartis and external data such as those mentioned in Section 6.2.1. The key requirements for this system were that it maximizes simplicity, automation, and robustness. These requirements are a natural result of the nature of the various upstream data sources feeding into the final repository:

1) Most, if not all, of the upstream data sources are changing on a constant basis, in terms of both structure and content.
2) The data to be aggregated/integrated are extremely heterogeneous, ranging from analytical compound data to protein sequences and protein–protein interactions in biological pathways.
3) There are few, if any, unifying concepts across the various data sources. A high degree of maintainability, therefore, implies a high degree of automation to focus on the more important task of data integrity and interoperability, as opposed to mere data availability.

The resulting ETL system is a tightly knit, version-controlled collection of UNIX shell and Python scripts that perform extensive checks and normalization operations on incoming data before loading into a PostgreSQL database. During the checking, cleaning, and normalization stages, the main focus is on establishing consistency and a common vocabulary across data sources. For example, all chemical structures are subjected to the same cleanup procedure before calculating IUPAC InChIKeys; moreover, after loading, every data source that provides chemical structures must provide an "InChIKey" field. This is an easy, yet extremely useful way to establish chemical interoperability. A similar procedure is employed to establish a common set of biological identifiers across all data sources. In practice, this entails the verification of every single identifier provided, including Entrez Gene IDs and Gene symbols, UniProt accession numbers, RefSeq protein and nucleotide accessions, enzyme commission (EC) numbers, as well as plain-text protein names such as "beta-lactamase." To the fullest extent possible, all of these identifiers are checked for discontinuation or deletion – Entrez Gene IDs as well as UniProt entries may disappear at any point in time – as well as internal consistency. If, for example, Gene ID and UniProt accession numbers are provided, it is ensured that they refer to the same biological entity in the same taxonomy. Analogically to the chemical structures, by this we ensure the provision of a common set of identifiers that is present for each data source containing biological entities such as drug targets. This extensive normalization establishes the basis for rapid database queries in chemical as well as biological space across all internal and external data.

6.3
Analysis of HTS Data

In this chapter, we will show two examples on how data analysis can be used in the context of HTS. The first example uses historical HTS data to learn how hits *behaved* in previous assays, so one can annotate them for future hit lists. The second example uses external or in-house data or sometimes predictions applied to a hit list in order to generate a hypothesis of their modes of action.

6.3.1
Analysis of Frequent Hitters and Undesirable Compounds in Hit Lists

Over the years, a large amount of HTS data has been collected. They have been used mainly for lead finding projects. Chemists who assess the results of HTS experiments quickly realized that some compounds are frequently found in different hit lists. These frequent hitters are compounds that can be artifacts interfering with the screening readout (e.g., fluorescent compounds) or the cell used for screening (detergent, membrane disruptor, or aggregate-like). They can interact noncovalently with proteins or form aggregates [11–13]. Such compounds are usually removed by a counterscreen, but some still survive especially when compounds are tested at a single concentration. FHs can also be compounds that are promiscuous [14] or

having privileged substructures toward certain targets families (e.g., kinase inhibitors and GPCR ligands) [15–18]. These nonselective hits are usually removed by a secondary screen.

To identify such frequent hitters, we used data mining of our historical HTS data through Hit-Hub. We retrieved all compounds (Set_FH) that hit 20% or more of the historical HTS assays. A compound is considered a hit in a primary assay if the absolute value of the calculated Z-score is greater than 4, we kept only hits that have been tested in more than 50 assays. We also retrieved compounds (Set_not_FH) that hit less than 1% of the assays panel using the same conditions.

Set_FH contains around 4000 compounds that have been historically tested in the HTS projects. Compounds with known kinase inhibitor scaffolds comprise 70% of the set. This is expected, since the historic HTS involved kinases. Some of the large classes in the remaining 30% are lipophilic compounds with high molecular weights (on average 496 MW) compared to Set_not_FH (on average 371 MW). In order to find out if there are any chemical fragments enriched in one set compared to the other, we retrieved the most frequent fragments in the Set_FH compounds and compared their frequency to the Set_not_FH compounds. Examples of such fragments are listed in Figures 6.2 and 6.3. The most enriched (enrichment factor is the ratio of the portion of a fragment in the considered set divided by the portion of the same fragment in the
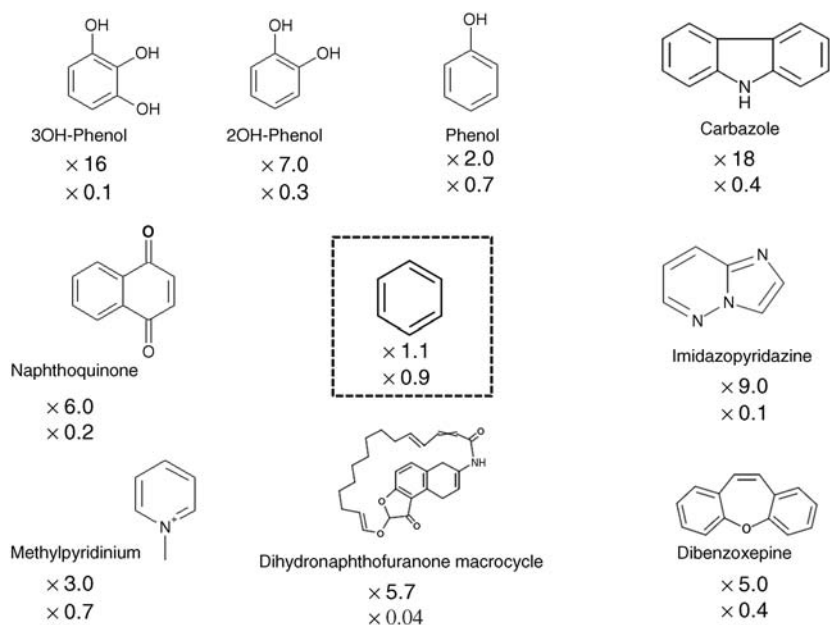
Figure 6.2 Enrichment factors of some fragments most frequently found in Set_FH without kinase inhibitors (phenyl ring is given as a reference). Upper values are the enrichment factors in the Set_FH set and lower values are enrichment factors in the Set_not_FH set. Enrichment factor is the ratio of the portion of a fragment in the considered set divided by the portion of the same fragment in the whole set.
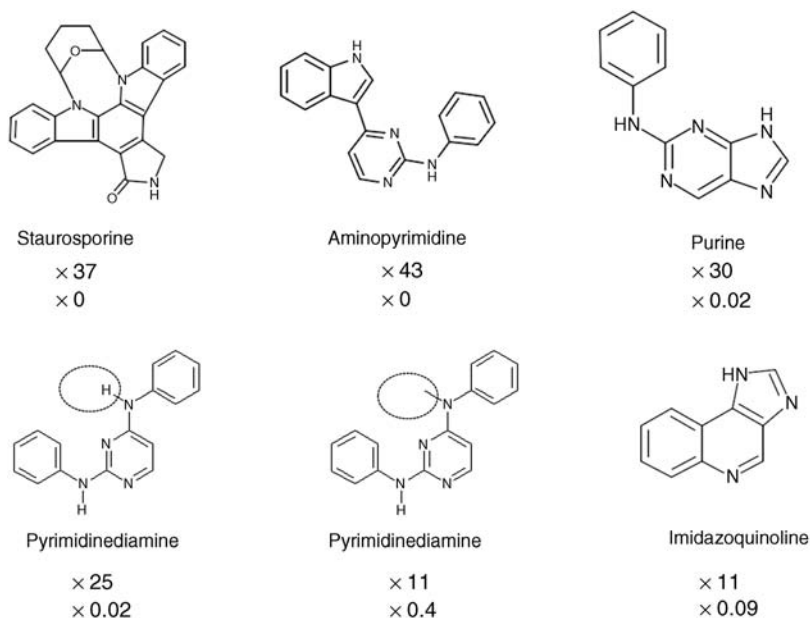
Figure 6.3 Enrichment factors of some fragments most frequently found in Set_FH with kinase inhibitors only. Upper values are the enrichment factors in the Set_FH set and lower values are enrichment factors in the Set_not_FH set. Enrichment factor is the ratio of the portion of a fragment in the considered set divided by the portion of the same fragment in the whole set.

whole set) fragment in the FH set without kinase inhibitors is the carbazole ring (enrichment factor ×18 in the Set_FH), followed by the trihydroxyphenol (×16) and imidazopyridazine ring (×9) (Figure 6.2). Among the phenol fragments, there is an increase of the enrichment factor with the number of hydroxyl groups. These and quinones can be subject of cysteine nucleophilic attacks. Another privileged substructure of GPCRs, dibenzoxepine, is enriched ×5 in the Set_FH.

The most enriched fragments found in the kinase-like FH are listed in Figure 6.3. Most of them are well-known kinase scaffolds such as aminopyrimidines (×43), staurosporine (×37), purines (×30), and imidazoquinoline (×11). Compounds containing such fragments are most likely to be hits in HTS primary data, but not only in kinase assays (Figure 6.4). They are usually flagged in the hit list and the chemistry team decides whether to follow them or not. One can notice that some fragments (pyrimidinediamine in Figure 6.3) shows greater selectivity when potential hinge region interactions are reduced (e.g., by replacing a hydrogen by methyl group).

Another way to make the already described approach more systematic is to mine the data by building a NB (naive Bayesian) classifier to differentiate between FH and non-FH. This approach and other methods were used to enrich or predict true active compounds in the hit lists [19–22]. We built two NB models using FCFP_6 fingerprints: one model (M1) was built using the set without known kinase inhibitors and a
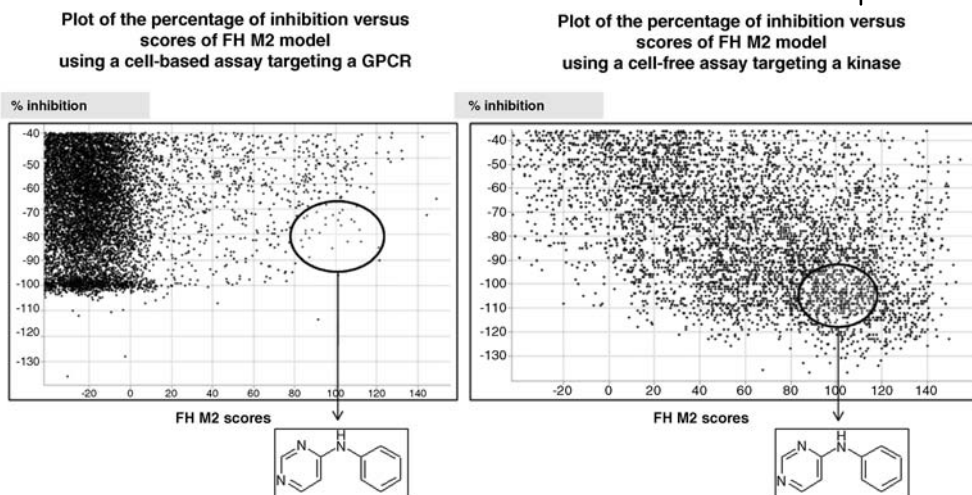
Figure 6.4 Example of two different hit lists scored with Model M2. *X-axis*: FH M2 model scores. *Y-axis*: primary percentage of inhibition. Circled points are compounds with pyrimidine substructure.

second model (M2) was built using only the known kinase inhibitors (the reason to make two sets is to show examples of target FH and assay FH). The models were trained using Set_not_FH as a control for inactive (not FH). Then, we applied the models to different hit lists. In Figure 6.4 we report the prediction of FH of primary hits from a cell-based and cell-free assay for two different targets. Obviously, some kinase-like frequent hitters are also found in a screening for a cell-based GPCR FLIPR assay. The explanation of such observations is that some of these kinase inhibitors can act in a cellular pathway and disturb calcium signaling as observed in the FLIPR signal. Such compounds usually drop out after dose–response validation, but not necessarily after a single concentration testing in a counterscreen or a secondary screen.

As mentioned before, compounds tested in HTS can have unexpected behaviors such as noncovalent competitive binding, interference with the assay via covalent modification or aggregation, or simply not be soluble enough. Another question arising is *What are the compounds that have been rejected from the hit lists and why?* To answer such a question, we have looked at 10 historical HTS campaigns from different assay techniques and assigned different chemists to the triaging of the hit lists. This task is not simple because the hit lists are different in terms of size (for example, if the list is short, few compounds are rejected), the project status (if there is a need for another screening campaign to back up a clinical candidate, there are high expectations for the lead nomination), the diversity of hits found, the nature of the target (e.g., a new target lacking tool compounds), and, last but not least, the chemist's background. For instance, a compound could be rejected because the chemist previously worked on the compound series and knows about potential toxicity, stability, or reactivity issues. In general, compounds are deprioritized because of
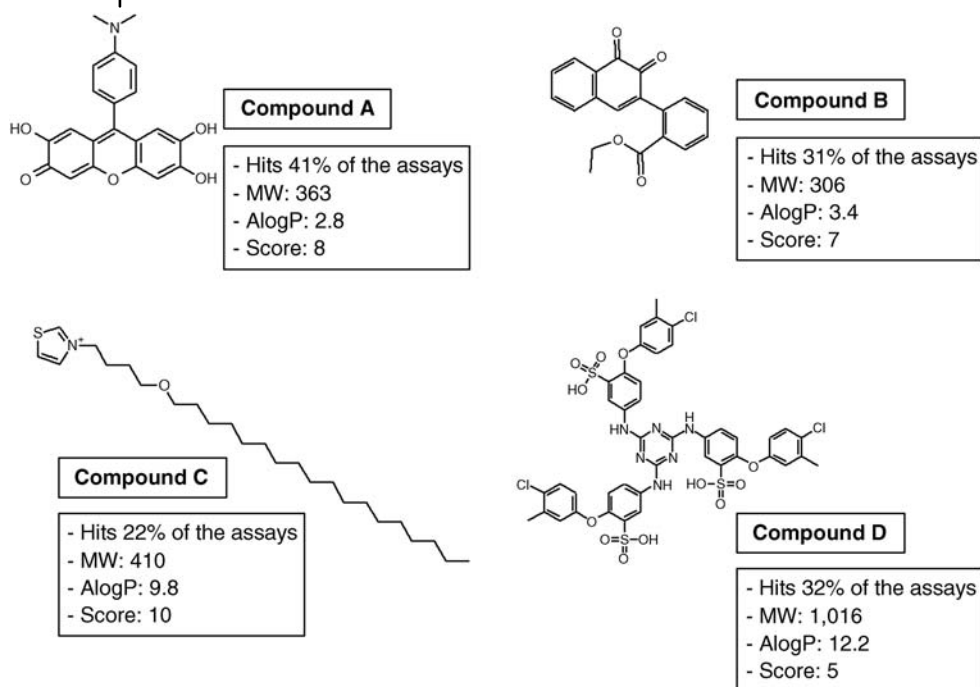
Figure 6.5 Examples of compounds found in hit lists. High-scoring compounds are most likely undesirable in the hit lists. Compound A: highly fluorescent compound. Compound B: chelating and redox benzoquinone substructure. Compound C: positively charged and highly lipophilic compound; such compounds are attracted by negatively charged cell membranes and are unspecific FH, subject to aggregation. Compound D: Large compound prone to aggregation.

the lack of potency, specificity, and, in some projects, drug-likeness (in the latter case, a simple Lipinski filter can be applied).

For each of the 10 HTS campaigns, we build a NB model for compounds that have been rejected from the hit lists and a voting score is then assigned to the hit. For example, if a compound has a high NB score (>0) in each of the 10 models, it will have a score of 10, it means that some chemical feature in the compound was found in 10 lists and rejected. The compound may exhibit frequent hitters behavior throughout the assays and have uninteresting "med chem" properties (Compound C, Figure 6.5). On the other hand, if a compound has a score of 1, it means it was rejected from one list, but it was not necessarily found in the other nine lists, so the nine others chemists did not have a chance to reject it. The voting score can be applied to future hit lists to help rank compounds that chemists can visually check. For practical reasons, instead of looking at thousands of hits, one can flag the top 100–500 high-scoring hits that can be checked by a team of chemists in order to decide to follow on them or not.

6.3.2
Analysis of Cell-Based Screening Data Leading to Mode of Mechanism Hypotheses

Phenotypic screens are typically high-throughput compound screens designed to identify compounds that modulate biological processes where the direct targets are unknown. Such screens may involve transformed cell lines, primary cells, tissues, or whole organisms (e.g., bacteria, yeast, fruit flies, and zebrafish). Phenotypic screens are special and distinct from biochemical HTS in a number of ways. First, the screening readout does not measure direct activity of a particular purified or recombinant protein, but rather a biological phenomenon. Second, phenotypic assays may have an intended target, but may also be influenced by other targets that modulate the assay. Third, the goal of a phenotypic screen may be to ultimately discover and validate a novel drug target that affects a desired biological response rather than to find lead compounds. Therefore, less tractable chemical matter may be acceptable in the final hit list. Finally, polypharmacology of compounds (i.e., activity against more than one target) may be relevant or even critical for efficacy of some compound hits.

In addition to these special considerations, the technologies and readouts used in phenotypic screens are quite diverse. Assay technologies and readouts range from single-point data to complex, multiparametric readouts. Typical cell-based assays measure changes in proteins from their gene expression levels, their mRNA translation and stability, as well as posttranslational states (e.g., phosphorylation, methylation, and ubiquitination), and translocation or localization. For example, HCS (high-content screening) can yield rich information about multiple cytological parameters, such as cell growth or protein states. Perhaps the most common phenotypic screen in industry is the RGA (reporter gene assay) in which a reporter gene such as luciferase is placed under the control of the promoters of a target of interest. Compounds can be assayed in RGAs based on their ability to stimulate or antagonize the reporter gene under pathway-stimulating conditions. In recent years, phenotypic screens have gone beyond single genes to gene signatures as a deeper readout of pathway regulation or as proxy readouts for disease modulation. For example, multiple mRNA (typically six–nine genes) can be detected with multi-analyte profiling beads to measure the impact of compounds on genes "moving" up or down in a pathway transcriptional signature with respect to "housekeeping" genes not in the signature.

Importantly, all of the aforementioned differences from biochemical HTS necessitates a very different workflow for analyzing the hit lists from phenotypic screens. For example, typical experimental follow-up to a biochemical HTS involves counter-screens, selectivity assays, or "orthogonal" secondary assays that use a different screening technology as well as biophysical or kinetic assays that raise confidence in a physical interaction with the biochemical target. Cellular activity is not guaranteed in this case. Therefore, biochemical hit list "triaging" focuses on prioritizing compounds with lead-like properties that are likely true binders and have potential to show efficacy under less simplistic conditions. In contrast, phenotypic screens implicitly weed out compounds without cellular activity. Cellular HTS hits already

possess many of the desirable physicochemical properties that would be scrutinized in a biochemical HTS hit list. Instead, the objective of assessment shifts toward understanding potential MOA of hits.

The first step to elucidating compound MOA is to annotate all HTS hits with existing compound bioactivity knowledge. For understanding or even predicting compound MOA, it is imperative to *know what we know* about every chemical structure. This is not possible without an extensive knowledge base of compound bioactivities, which enables hit annotation. Indeed, even annotating hits with *low-hanging fruit* information, prior bioactivities already known for compounds, is tricky. As described earlier, our compound knowledge base Hit-Hub, containing up-to-date and integrated assay bioactivities across internal, public, and commercial data sources, enables automation of compound annotation in large batches. Annotation that scales to the data set size using automation is particularly important for phenotypic screens, where hit rates from HTS can approach 1–2%, or 10–20 000 compound hits per 1 million compounds screened. Clearly, individual compound lookups are not feasible for hit lists of this size.

Thus, aside from the requisite knowledge base, a mechanism for automated annotation is needed. Using a canonical chemical structure representation that is identical for both the HTS hit list and the compound knowledge base, one can join compound annotations on the fly. Automating annotation by a simple join operation on chemical structure is trivial with data pipelining tools such as Pipeline Pilot.[1] For canonical chemical structure representation, the IUPAC InChI or InChIKey is often preferred (see above). As HTS screening decks are often populated by legacy project team compounds and reference compounds from literature and patents, as well as known drugs, there are typically a significant number of compounds in any given HTS hit list about which some prior knowledge exists that can aid the triaging effort.

Compound annotation should usefully answer questions such as follows:

1) In which prior assays, targets, or cell lines have this compound been active?
2) What were the activity values and what are the data sources (papers, patents)?
3) Are there known compound-induced phenotypes?
4) Has this compound reached preclinical, clinical, or marketed drug status?
5) Does this compound frequently hit a particular target class, assay technology, or assay format?
6) What is the origin of the compound and is it publicly known or available?
7) Is this compound sample pure and has it passed LC–MS characterization previously?

All of these questions, while important, pose a challenge for providing human-digestible annotation, especially for large hit lists. Thus, it is critical that the underlying bioactivity knowledge base employs semantic standards such as controlled vocabularies, taxonomies, or ontologies to represent facts. For example, if all assay targets are represented as Entrez Gene IDs in the bioactivity knowledge base, it becomes trivial once a phenotypic HTS hit list is annotated to sort compounds based on their prior known targets and ask questions like *Are any targets overrepresented among my hits?* Ideally, assay metadata such as targets, cell lines, assay technologies,

and result types can all be subjected to controlled terminology and supported by a BioAssay Ontology. As a practical matter, to properly capture assay metadata, infrastructure must be provided to enable scientists to "register" assays at the time they are created, run, or reported. At Novartis, we have created ARES (assay registration system), a Web-based system that allows HTS assay owners to record assay metadata in a controlled way using an internally developed BioAssay Ontology. ARES allows assays to be compared to one another, and importantly, for data to be compared globally across all assays. Unfortunately, extensive curation of prior assay data is necessary if a new BioAssay Ontology and assay registration system is imposed. However, the long-term payoff for HTS data querying and mining is immense: bioactivity data from every individual HTS becomes part of the global compound annotation data to inform all future HTS hit lists.

Occasionally, HTS hits are similar to known bioactive compounds with only subtle dissimilarities. For instance, a hit in a phenotypic screen may strongly resemble a known drug except for a change in a single atom; would a biologist want to know that a hit on their screen is one atom different from a drug? When no bioactivities are known for an exact structure, the annotation of bioactivities of close analogues is a reasonable probabilistic substitute, particularly for early MOA hypothesis generation. Fortunately, the chemoinformatics field has produced many chemical similarity algorithms and metrics, which can be exploited to identify similar bioactive compounds. In addition to similarity, statistical modeling on the chemical structures of ligands from known targets and target classes can be used to score hits from a phenotypic screen, with respect to their most likely targets [23]. In combination, known targets and Bayesian predicted targets for a hit list produce a classic "gene list" that resembles those produced by Affy-type microarray experiments. Therefore, methods developed in the bioinformatics field for assessing gene lists, such as pathway enrichment analyses [e.g., gene set enrichment analysis (GSEA)], can be repurposed for assessing a phenotypic screen hit list.

In this vein, we have developed a data pipeline capable of annotating phenotypic hit lists with both known quantitative activity on targets and Bayesian predicted targets (Figure 6.6). The input is a hit list of compounds from a phenotypic screen. The hits are annotated with a direct lookup to a bioactivity knowledge base using InChIKey. Known targets are annotated as an array of Entrez Gene IDs. Second, Bayesian predicted targets are optionally added (e.g., top five predictions where Bayes score >10). Next, a Gene Ontology (GO) term enrichment component returns Gene Ontology terms that are enriched for the given list of genes. Enrichment is calculated using Fisher's exact test with the Bonferroni correction for multiple testing [24]. The user can specify the GO term type, as well as set upper and lower bounds for number of genes belonging to the terms to be enriched. This is useful in eliminating uninteresting, very broad, or very specific terms. All terms with a *p*-value less than 0.05 are returned, along with the corresponding genes from your list, which belong to each enriched term. In this example, validated hits from a PubChem screen for histone demethylase JMJD2E were chemically clustered and annotated, and GO molecular function and GO process enrichment were computed. Voltage-gated potassium channel activity and carbonate dehydratase
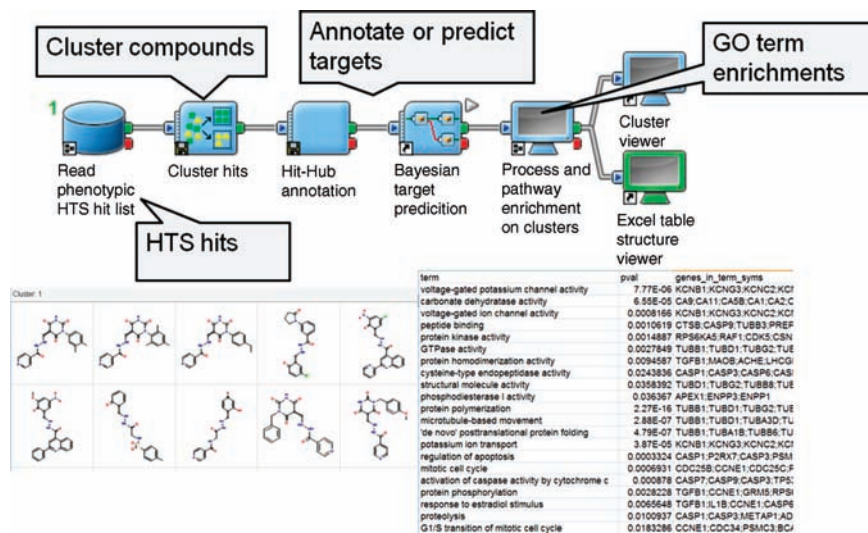
Figure 6.6 Pipeline Pilot protocol for annotating compounds from a phenotypic HTS hit list. Compound hits are clustered by scaffold and annotated with all prior known bioactivity data as well as Bayesian predicted targets. The significant targets per cluster are subjected to pathway and process enrichment to elucidate the likely MOA of the cluster.

activity were enriched for compounds in the largest chemical cluster (some representatives are described later). This example suggests that phenotypic data analysis does not have to be "black box" and can incorporate prior knowledge of chemotype bioactivities to help the biologist and the chemist further manually triage compound hits for novelty or obvious connections to phenotypes.

6.4
Identification of New Compounds via Compound Set Enrichment and Docking

6.4.1
Identification of Hit Series and SAR from Primary Screening Data by Compound Set Enrichment

The objective of HTS is not the identification of a final drug with optimal binding, selectivity, and pharmacokinetic properties, but rather to deliver a chemical starting point that will be further developed in a lead optimization process. In the conventional approach mentioned in the introduction, when primary data are available, an activity cutoff is fixed in order to identify the primary hits. Primary hit activities are then confirmed and quantified by dose–response measurement. Only afterwards are compounds classified (often by clustering methods) in order to identify series of related compounds. Series are usually preferred to singletons because they bring
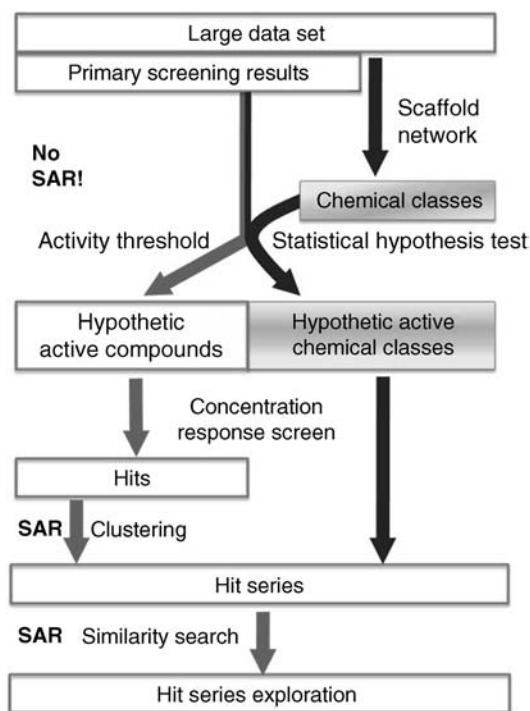
Figure 6.7 Integration of compound set enrichment in a screening campaign. All screening data are used in order to identify hit and latent hit series (right workflow). Singletons are not treated by compound set enrichment, but are still identified by a normal cutoff-based approach (left workflow). Reprinted with permission from Ref. [25].

SAR (structure–activity relationship) that can be used as a starting point for lead optimization. Whereas more than one million compounds have been tested for primary activity, very few of the data points generated are used for hit series identification. We recently proposed the CSE (compound set enrichment) [25] method that analyzes all primary screening data in order to identify hit series. Compounds are first clustered using a scaffold-based classification [25–27]. Then activity distribution of each scaffold is compared to the background distribution (Figure 6.7).

A major interest of CSE enrichment is that it does not use any compound activity cutoff [28]. Thus, even if highly active compounds are absent from the data set, the method can still detect a significant shift of activity. This is possible because within an active series, weakly active compounds are usually more frequent than the very active ones. Interest of these weakly active compounds was shown first by Mestres and Veenman [29]. Small modifications of such latent hits can transform them into hits ("latent hit promotion"). However, in their original paper, requirement of a known pharmacophore was required. With CSE, no prior knowledge about active

compounds, target structure, or even the target itself (phenotype screening, for example) is required to identify these latent hit series. Identification of hits requires testing new compounds with this scaffold. SAR exploration might help to select new compounds with optimal side chains (see below).

Describing and identifying active series of compounds by evaluating scaffold activity has many advantages. As scaffolds are usually the most rigid part of a molecule (rings), it is easy and highly relevant to superimpose compounds by active scaffolds. Active scaffolds identified by CSE are scaffolds for which a significant shift of activity is observed compared to the background distribution, but compounds do not necessarily have optimal side chains for maximum activity. Thus, not all compounds within a series are active. For this reason, compounds from a series usually display a range of activities. This particularity makes easy and relevant SAR extraction. One way to highlight informative SAR is by looking at highly similar compounds. To illustrate this approach, we picked two compounds from the same series with IC_{50} smaller than $10 \mu M$ (examples are from the PubChem assay AID 893; see Ref. [26] for a detailed analysis of this assay with CSE). The two nearest neighbors of these compounds are represented in Figure 6.8 (Series A1 and A2).

Exploration within these series shows importance of the methoxy group in *para*-position of the phenyl group (pairs 1–2 and 4–5 in orange). It suggests a hydrogen bond acceptor in this area and SAR transposition within a series is relevant. For series A1, we also observe that switching from tetrahydrofuran to furan (aromatization) induces a loss of activity (compounds 1–3 in green). For series A2, pyrrolidine seems better than piperidine (4–6 in blue). This might be due to a steric clash with the piperidine. These observations suggest that evaluating compound series activities without applying any activity cutoff can enable easy SAR identification.

CSE evaluates and annotates activity of all scaffolds described in a data set. This makes possible navigation in scaffold space described by a scaffold network. We observed that some active scaffolds directly connected together form active communities in the network. These areas, called active islands, are highly useful because they group related scaffolds together and make easier overview of results from CSE. This approach was used to develop a scaffold network visualization tool in which each active island can be explored separately. The representation retained is intuitive for chemists and helps identify strong scaffold SAR (Figure 6.9). Each active island is represented by a network. By clicking on nodes and/or edges, corresponding scaffolds are displayed as well as compounds with these scaffolds and their activity distribution.

Another interest of these islands is about SAR transposition between series. The family relationships of all scaffolds in the same island make easy the superimposition of different series of compounds. For example, series A and B have a common parent active scaffold, represented in red in Figures 6.8 and 6.9. The best compounds in series B have actually weak activity (25.12 μM). The optimal side chains according to the SAR of series A1 and A2 are not present in compounds of series B. Pharmacomodulation of compounds from series B are suggested by SAR transposition from series A1 and A2, as illustrated in Figure 6.8.
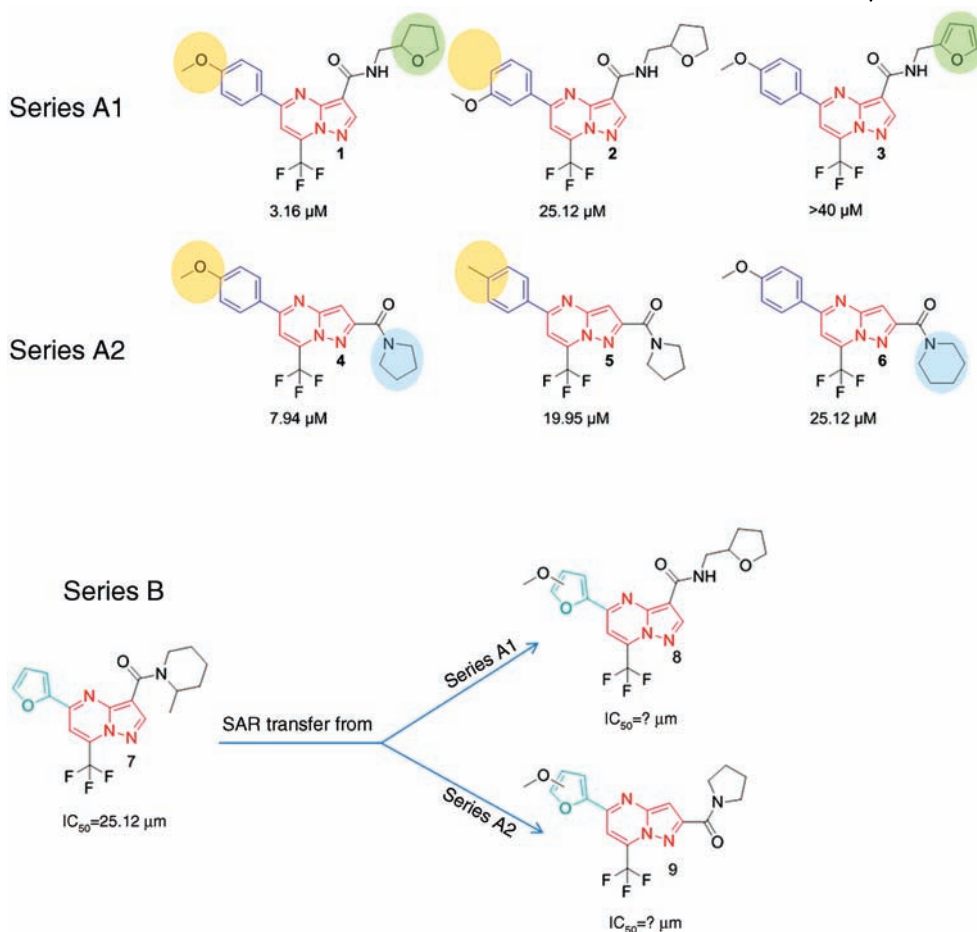
Figure 6.8 A/SAR within active series. Extraction and analysis of side chains SAR are facilitated by compound superimposition according to their common scaffold (represented in red and blue). B/SAR transposition from series A1 and A2 to series B. According to scaffold network, series A and B are related together by a common parent scaffold (in red). Series A and B are annotated as active by CSE. It suggests that compounds from these series can be overlapped according to this parent scaffold and, thus, optimal side chains from series A could be applied to series B in order to promote latent hits into hits (hypothesis not verified for these series).

6.4.2
Molecular Docking

Not only ligand-based molecular modeling techniques, but also chemoinformatics approaches are often used to enrich hit lists or to annotate them by chemical similarity comparison (similarity to a pharmacophore, similarity to known ligands of the target of interest, chemical clustering, etc.). In this last section, we illustrate
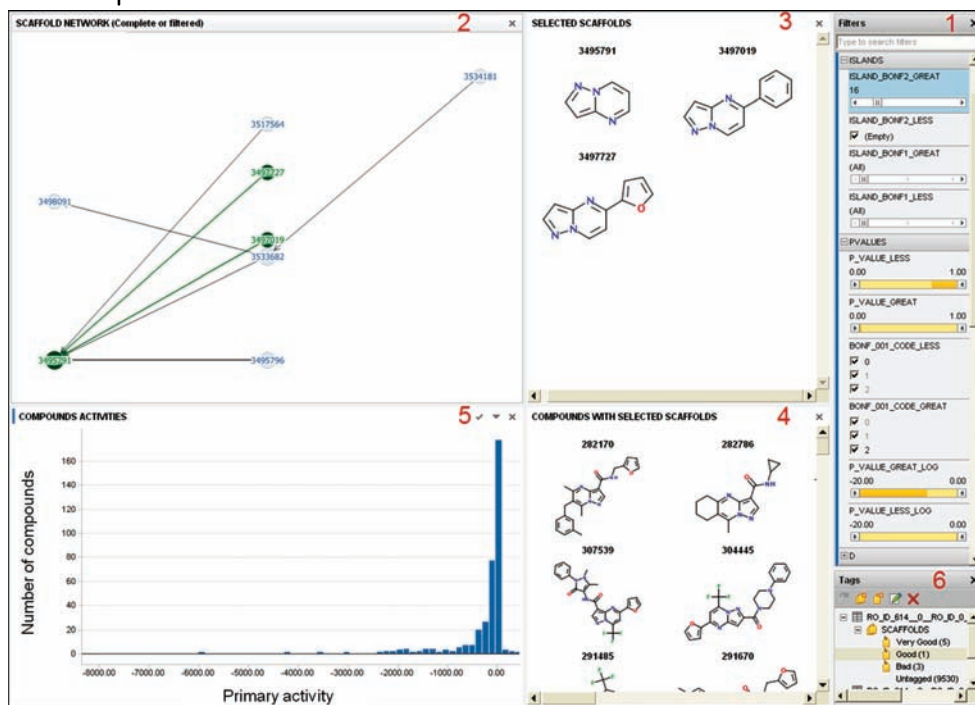
Figure 6.9 Scaffold network visualization tool. As the scaffold network contains usually a lot of scaffolds, it is better to use first filters (1) to only represent a small part of it. In this example, an active island has been selected and is displayed as a network (2). By clicking on nodes and/or edges, corresponding scaffolds are displayed (3) as well as compounds with these scaffolds (4) and their activity distribution (5). It's possible to annotate scaffolds and compounds using tag facilities proposed by Spotfire (6).

using tankyrase how molecular docking (Figure 6.10) can be used prior to a HTS experiment to provide a focused compound set. Tankyrase is a poly-ADP-ribosylating enzyme that acts on axin in the Wnt pathway leading to axin's degradation via the ubiquitin–proteasome pathway, making tankyrase a potential cancer target [30]. Tankyrase uses NAD as the ADP-ribose donating substrate. In 2009, Karlberg *et al.* [31] solved the cocrystal structure of tankyrase-2 with the inhibitor XAV939 and PDB accession code 3KR8.

We used *in silico* screening to select vendor compounds that might inhibit tankyrase by targeting the NAD binding site. Our collection of vendor compounds consisted of a little over 700 K SMILES that had already been filtered for lead-like properties, availability from reliable vendors, and absence from our own compound collection. Pipeline Pilot[1]) scripts were written to check for and remove structural duplicates, compounds with >8 freely rotatable bonds, compounds with unusual atoms, and compounds with excessive tautomerization (>1000). This removed just under 4% of the molecules, mostly due to too many freely rotatable bonds.

Raw Structure File

≤ 2 unknown chiral centers
≤ 6 freely rotatable bonds

↓

Filtered Structure File

generate SMILES
generate 3D coordinates

↓

Unbiased Structure File

expand chiral centers
generate all plausible tautomers
generate alternative charge states
remove exact duplicate structures
minimize energy

↓

Docking Set

docking

↓

Docked Poses

project-specific filters (MW, PSA, clogP, etc.)
sort by docking score
remove low-scoring isomers
apply docking score cutoff

↓

Unique Poses for Analysis

Figure 6.10 Scheme for generating an unbiased docking set of structures and analyzing/filtering the docked results. Docking is best carried out on compounds whose 3D structures are fully defined and that are not excessively flexible. For large structural databases, it is usually more convenient to prepare the docking set once and apply program-specific filters after docking. Each individual structure of the final selection should be visually examined and docked into the target protein for displaying a reasonable fit and ligand conformation.

Expansion of the remaining compounds for unknown chiral centers, tautomers, and alternative charge states with Pipeline Pilot roughly doubled the number of structures to 1.4 million. Since we were planning to perform the docking with Glide (Schrödinger, LLC, Portland, USA), which generates conformers on the fly during docking, we did not have to precalculate conformers. Generation of 3D structures was done with CORINA (Molecular Networks, GmbH, Erlangen, Germany). Energy minimization was carried out with MacroModel (Schrödinger) to clean up the structures and remove any structures that the Schrödinger force fields would not be able to handle. In addition, a final check to remove duplicate structures generated by tautomer expansion was done. CORINA failed to generate 3D structures for 0.008% of the input SMILES, 0.006% of the 3D structures failed energy minimization, and 0.04% of the structures were removed as duplicates.

An in-house crystal structure of the PARP catalytic domain of tankyrase in a complex with an in-house compound was used for docking. Although Glide allows the user to define pose constraints, we did not apply any in this case since we were interested in new chemical classes with potentially new binding modes. Because we

were planning to dock 1.4 million structures, we first ran Glide in HTVS (high-throughput virtual screening) mode, which examines fewer conformations, but is roughly 10-fold faster than its standard precision (SP) mode. Very few compounds failed docking in the second round (0.007%). Applying chemical filters developed by the tankyrase project team removed around 5500 compounds, leaving ∼744 000 acceptable poses. Since there were usually multiple copies of each molecule (isomers) due to the chirality, tautomer, or ["and" nor "or"] charge expansion, the compounds were sorted and only the single top scoring isomer was kept. Consensus scoring [32] with five different docking scores was performed on these unique molecules.

Besides being interested in finding active tankyrase inhibitors, we were also interested in developing SAR of the potential inhibitors we found. Because of this, and to avoid adding singletons to our compound collection, we purchased groups of related compounds (12–15 members) rather than individuals. The docked molecules were sorted by their normalized CScores (Tripos, Inc., St. Louis, USA) and the top 20 000 compounds were taken and clustered with Cluster3D [33], using a similarity threshold of 0.7. The average CScore of the top 15 members of each cluster was calculated and the top 47 clusters, 701 compounds in total, were selected for ordering. Of these, 55 (7.8%) were not ordered (small vendors), 150 (21.4%) were unavailable, and 496 (70.8%) were delivered and placed in the Novartis compound archive. IC_{50} determination showed 13 of these to be active ($IC_{50} < 50\,\mu M$), including 2 with submicromolar activity. This gives a validated hit rate of 2.6% compared to the 0.4% HTS hit rate against the entire Novartis collection. Results such as these indicate that *in silico* screening can be a useful tool for generating focused libraries, for either selecting commercial compounds to complement one's in-house collection or creating a limited set of in-house compounds for screening in cases where a full-blown HTS is either impossible or undesirable.

6.5
Conclusions

In this chapter we have given a broad view of how our in-house data from HTS can be accessed and used. The goal is to learn from historical data to help drug discovery project teams make better decisions. HTS generates a very large amount of data and currently most of it is dormant and unused. We demonstrate by using data mining techniques how to turn this data to knowledge. Such examples range from tracking frequent hitters or undesirable compounds to "mode of action knowledge" annotations, but also using the data to create models to find false positives and true negatives in hit lists.

The challenges of data mining are not only the growing amount of data, but also the way to communicate the results to a team from different scientific backgrounds. In-house Web applications such as "HTS Explorer" linked to visualization software such as Spotfire[2] are becoming useful tools for compound-data navigation. Another

2) TIBCO Software Inc., Palo Alto, USA.

big challenge is how to capture decision making on data and institutional knowledge. To overcome those challenges, in-house initiative such as NDFI (NIBR Data Federation Initiative) will provide technical foundation for data collection, integration, search, and reuse. The goal of this initiative is to create a framework for linking pathways, genes, proteins, and compounds, introducing to scientists – both chemists and biologists – the ability to interrogate data using a probabilistic or knowledge-based approach across many domains of science.

Other initiatives linked to public data mining are taking place. An example of such initiative is the Open PHACTS consortium (http://www.openphacts.org), funded by the Innovative Medicine Initiative (IMI), which reflects collaborations between academic groups and pharmaceutical companies by applying semantic technologies to available data resources, creating an Open Pharmacological Space that can be navigated to identify new drug targets and pharmacological interactions. It will deliver a single view across different databases, and will be freely available to users.

Acknowledgments

References

1 Macarron, R. (2006) Critical review of the role of HTS in drug discovery. *Drug Discovery Today*, **11**, 277–279.

2 Inglese, J., Johnson, R.L., Simeonov, A., Xia, M., Zheng, W., Austin, C.P., and Auld, D.S. (2007) High-throughput screening assays for the identification of chemical probes. *Nature Chemical Biology*, **3**, 466–479.

3 Mayr, L.M. and Fuerst, P. (2008) The future of high-throughput screening. *Journal of Biomolecular Screening*, **13**, 443–448.

4 Macarron, R., Banks, M.N., Bojanic, D., Burns, D.J., Cirovic, D.A., Garyantes, T., Green, D.V.S., Hertzberg, R.P., Janzen, W.P., Paslay, J.W., Schopfer, U., and Sittampalam, G.S. (2011) Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery*, **10**, 188–195.

5 Malo, N., Hanley, J.A., Cerquozzi, S., Pelletier, J., and Nadon, R. (2006) Statistical practice in high-throughput screening data analysis. *Nature Biotechnology*, **24**, 167–175.

6 Gubler, H. (2007) Methods for statistical analysis, quality assurance and management of primary high-throughput screening data, in *High-Throughput Screening in Drug Discovery*, Methods and Principles in Medicinal Chemistry Series, vol. 35, 1st edn (ed. J. Hüser), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 151–205.

7 Yan, F., King, F.J., He, Y., Caldwell, J.S., and Zhou, Y. (2006) Learning from the data: mining of large high-throughput screening databases. *Journal of Chemical Information and Modeling*, **46**, 2381–2395.

8 Ling, X.B. (2008) High throughput screening informatics. *Combinatorial Chemistry & High Throughput Screening*, **11**, 249–257.

9 Kümmel, A. and Parker, C.N. (2011) The interweaving of cheminformatics and

HTS, in *Chemoinformatics and Computational Chemical Biology*, vol. 672 (ed. J. Bajorath), Methods in Molecular Biology, Humana Press, pp. 435–457.

10 Ling, X.B. (2008) High throughput screening informatics. *Combinatorial Chemistry & High Throughput Screening*, **11**, 249–257.

11 Roche, O., Schneider, P., Zuegge, J., Guba, W., Kansy, M., Alanine, A., Bleicher, K., Danel, F., Gutknecht, E.-M., Rogers-Evans, M., Neidhart, W., Stalder, H., Dillon, M., Sjögren, E., Fotouhi, N., Gillespie, P., Goodnow, P., Harris, W., Jones, P., Taniguchi, M., Tsujii, S., von der Saal, W., Zimmermann, G., and Schneider, G. (2002) Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *Journal of Medicinal Chemistry*, **45**, 137–142.

12 Feng, B.Y. and Shoichet, B.K. (2006) A detergent-based assay for the detection of promiscuous inhibitors. *Nature Protocols*, **1**, 550–553.

13 Baell, J.B. and Holloway, G. (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, **53**, 2719–2740.

14 Azzaoui, K., Hamon, J., Faller, B., Whitebread, S., Jacoby, E., Bender, A., Jenkins, J.L., and Urban, L. (2007) Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem*, **2**, 874–880.

15 Sheridan, R.P. (2003) Finding multiactivity substructures by mining databases of drug-like compounds. *Journal of Chemical Information and Computer Sciences*, **43**, 1037–1050.

16 Bondensgaard, K., Ankersen, M., Thøgersen, H., Hansen, B.S., Wulff, B.S., and Bywater, R.P. (2004) Recognition of privileged structures by G-protein coupled receptors. *Journal of Medicinal Chemistry*, **47**, 888–899.

17 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., and Roth, H.-J.

(2005) Key aspects of the Novartis compound collection enhancement project for the compilation of a comprehensive chemogenomics drug discovery screening collection. *Current Topics in Medicinal Chemistry*, **5**, 397–411.

18 Aronov, A.M., McClain, B., Moody, C.S., and Murcko, M.A. (2008) Kinase-likeness and kinase-privileged fragments: toward virtual polypharmacology. *Journal of Medicinal Chemistry*, **51**, 1214–1222.

19 Padmanabha, R., Cook, L., and Gill, J. (2005) HTS quality control and data analysis: a process to maximize information from a high-throughput screen. *Combinatorial Chemistry & High Throughput Screening*, **8**, 521–527.

20 Glick, M., Jenkins, J.L., Nettles, J.H., Hitchings, H., and Davies, J.W. (2006) Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive Bayesian classifiers. *Journal of Chemical Information and Modeling*, **46**, 193–200.

21 Harper, G. and Picketta, S.D. (2006) Methods for mining HTS data. *Drug Discovery Today*, **11**, 694–699.

22 Mballo, C. and Makarenkov, V. (2010) Using machine learning methods to predict experimental high throughput screening data. *Combinatorial Chemistry & High Throughput Screening*, **13**, 430–441.

23 Nidhi, Glick, M., Davies, J.W., and Jenkins, J.L. (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *Journal of Chemical Information and Modeling*, **46**, 1124–1133.

24 Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2008) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acid Research*, **37**, 1–13.

25 Varin, T., Gubler, H., Parker, C.N., Zhang, J.H., Raman, P., Ertl, P., and Schuffenhauer, A. (2010) Compound set enrichment: a novel approach to analysis of primary HTS data. *Journal of Chemical Information and Modeling*, **50**, 2067–2078.

26 Schuffenhauer, A., Ertl, P., Wetzel, S., Koch, M.A., and Waldmann, H. (2007) The

scaffold tree: visualization of the scaffold universe by hierarchical scaffold classification. *Journal of Chemical Information and Modeling,* **47**, 47–58.

27 Varin, T., Schuffenhauer, A., Ertl, P., and Renner, S. (2011) Mining for bioactive scaffolds with scaffold networks: improved compound set enrichment from primary screening data. *Journal of Chemical Information and Modeling,* **51**, 1528–1538.

28 Schuffenhauer, A. and Varin, T. (2011) Rule-based classification of chemical structures by scaffold. *Molecular Informatics,* **30**, 2–20.

29 Mestres, J. and Veeneman, G.H. (2003) Identification of "latent hits" in compound screening collections. *Journal of Medicinal Chemistry,* **46**, 3441–3444.

30 Huang, S.-M.A., Mishina, Y.M., Liu, S., Cheung, A., Stegmeier, F., Michaud, G.A., Charlat, O., Wiellette, E., Zhang, Y., Wiessner, S., Hild, M., Shi, X., Wilson, C.J., Mickanin, C., Myer, V., Fazal, A., Tomlinson, R., Serluca, F., Shao, W., Cheng, H., Shultz, M., Rau, C., Schirle, M., Schlegl, J., Ghidelli, S., Fawell, S., Lu, C., Curtis, D., Kirschner, M.W., Lengauer, C., Finan, P.M., Tallarico, J.A., Bouwmeester, T., Porter, J.A., Bauer, A., and Cong, F. (2009) Tankyrase inhibition stabilizes axin and antagonizes Wnt signalling. *Nature,* **461**, 614–620.

31 Karlberg, T., Markova, N., Johansson, I., Hammarström, M., Schütz, P., Weigelt, J., and Schüler, H. (2009) Structural basis for the interaction between tankyrase-2 and a potent Wnt-signaling inhibitor. *Journal of Medicinal Chemistry,* **53**, 5352–5355.

32 Charifson, P.S., Corkery, J.J., Murcko, M.A., and Walters, W.P. (1999) Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of Medicinal Chemistry,* **42**, 5100–5109.

33 Priestle, J.P. (2009) 3-D clustering: a tool for high throughput docking. *Journal of Molecular Modeling,* **15**, 551–560.

7
The Value of Interactive Visual Analytics in Drug Discovery: An Overview

David Mosenkis and Christof Gaenzler

The world is awash with data. Nowhere is this deluge more obvious than in the drug discovery process. In areas as diverse as genomics, screening, image analysis, and many others, data repositories are growing at exponential rates. Even the most seasoned expert cannot hope to inspect or even review every potentially relevant test result [1–3]. Many valuable computational approaches have been developed to exploit and derive value from these growing data sets, yet these add their own results to the growing mound of information. How can a scientist best utilize the available data resources to guide and inspire their research? Interactive visual analytics brings a powerful approach to have insights into and derive value from large amounts of data.

What do we mean by the term "interactive visual analytics?" We refer to the graphical depiction of data in a way that promotes exploration and insights. The interactive component means that a scientist can interact with the visualizations by filtering, highlighting, slicing, or other operations that facilitate exploration of the data. Such *ad hoc* exploration empowers scientists to follow their curiosity, pursue hunches, form hypotheses, and in general ask and answer many types of questions as quickly as they arise. This type of exploratory graphical analysis complements more traditional approaches such as statistical hypothesis testing, in which a well-formulated hypothesis is subjected to rigorous analysis to prove or disprove it. In an interactive graphical environment, analysts can generate and explore large numbers of unsubstantiated ideas and serendipitously stumble upon insights that can then be confirmed by more rigorous methods and used as the basis for further action.

In addition to *ad hoc* exploration, visual analytics can be a powerful way for communicating information and insights. Exploiting the amazing capacity of the human visual system to discern patterns and outliers, graphical representations can quickly and easily convey messages that even many pages of text would convey only with great difficulty.

We can distinguish approaches to visual analytics along a spectrum of *ad hoc* exploratory power, from static visualizations to a full *ad hoc* exploratory environment. Intermediate approaches along this continuum include guided interactive analyses, where scientists view preconfigured visualizations of complex data sets and have the

opportunity to adjust certain aspects of the presentation such as filtering to subsets of data, color-coding to reveal potential patterns, and drilling down to more detailed levels on subsets of the data. The presentation may even walk users through a series of steps in a guided analysis, giving them the ability to make certain choices, but within a framework that guides them through a best-practice approach tailored to the task at hand.

These approaches are immensely valuable in analyzing any kind of data, and indeed such approaches are fast becoming the norm in fields as diverse as energy exploration, financial services, sales, marketing, manufacturing, and many others. However, the remainder of this chapter will focus on application of visual interactive graphical approaches to analyzing data in the drug discovery process. The visual approaches described in the following sections are useful both for open-ended exploration and to help address specific, practical questions:

- What compound(s) do I synthesize next, either to directly improve desired properties or to fill out my understanding of the structure–activity relationship (SAR) landscape?
- Which compounds have the best profile to advance to the next stage?
- Which protein(s) are my compounds targeting? Which protein(s) should they be targeting?
- How diverse is my collection of compounds? What attributes do I need to enhance in my library design or in acquiring or synthesizing additional compounds?
- Which cheaper measurements (computational or laboratory) are good proxies for more expensive ones? When should these proxies be trusted?
- What distinguishes diseased from healthy subjects?
- What is the effect of drug dose on biological response?
- How do genetic alterations affect biological behavior?

7.1
Creating Informative Visualizations

There are a number of excellent sources containing guidelines for producing informative visualizations [4–6]. Best practices include the following:

- Eliminate any extraneous elements in visualizations. Anything that does not add information or clarity may distract from the real information content. Potentially extraneous elements include excessive labels on axes or other graph elements, redundant or uninformative legends, use of color that does not add information, and gratuitous graphical additions such as 3D effects.
- Pay attention to the scale used. Scaling from 0 conveys absolute values, but can diminish differences between values.
- What kind of visualization is best? It depends on the nature of the data and the nature of the desired insight. Here are some general guidelines for common situations:
  - For summarizing categorical data, a bar chart is often a good choice. Each category can be represented by a separate bar or a different color. Heights of the

bars can represent absolute values or relative values. See examples in Figures 7.7 and 7.8.

- If there are too many categories to fit comfortably in a bar chart, or if there is a desire to represent hierarchies within the data, a Tree Map is a good alternative.
- Use pie charts sparingly. In most cases, bar charts are better choices, because they can convey more information and people can more accurately perceive relative heights of rectangles than relative sizes of pie wedges.
- For seeing relationships and trends in two or more measures, consider a scatter plot. Plot two principal variables on the *X*- and *Y*-axes. You can add additional variables through use of color, marker shape, marker size, trellising, or adding a third dimension to create a 3D scatter plot. See examples in Figures 7.4–7.6.
- For displaying data values over time, a line chart is an intuitive visualization. Categories can be distinguished by different color lines or by separate trellis panels.
- Similar in appearance to a line chart, a profile chart can be a good way of showing a profile of various related values over a range of entities or categories. See an example in Figure 7.9. An alternative representation is a radar or spider plot, exemplified in Figure 7.10.

7.2
Lead Discovery and Optimization

7.2.1
Common Visualizations

7.2.1.1 **SAR Tables**

A common way for communicating information about chemical compounds is a structure–activity relationship table, or SAR table. In this tabular representation, each row represents a chemical compound, one column depicts the chemical structure, and other columns contain information about each compound. Figure 7.1 shows an example of a simple SAR table.

This ubiquitous format displays relevant properties and experimental results alongside chemical structures, but it does little to highlight interesting compounds or to facilitate insights into the relationship between the quantities and the structures.

A simple improvement, illustrated in Figure 7.2, applies color coding to the values.

This depiction gives immediate, easily interpreted insights into which compounds have favorable values for each measure, with colors shading from green to red indicating to what extent a value is outside the acceptable range.

We can further enrich a SAR table with concise graphic summaries of key quantities. In Figure 7.3, a line graph depicts selectivity of each compound
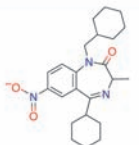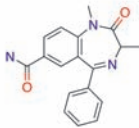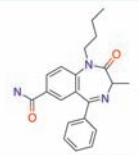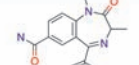
Corp_ID	Structure	Mol Wt	Solubility	In vivo Activity (%)	Bioavailability (%)
SF0180		397.5	7.6E-006	72.8	86
SF0217		307.4	5.2E-002	15.4	50
SF0218		349.4	1.2E-003	48.8	41
SF0219		363.5	5.0E-004	85.2	91

Figure 7.1 A simple SAR table.
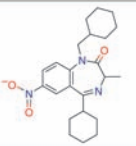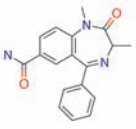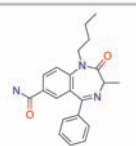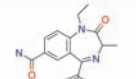
Corp_ID	Structure	Mol Wt	Solubility	In vivo Activity (%)	Bioavailability (%)
SF0180		397.5	7.6E-006	72.8	86
SF0217		307.4	5.2E-002	15.4	50
SF0218		349.4	1.2E-003	48.8	41
SF0219		363.5	5.0E-004	85.2	91

Figure 7.2 Color-coded SAR table.

Corp_ID	Structure	Mol Wt	Solubility	In vivo Activity (%)	Bioavaila bility (%)	Selectivity	Phys Props
SF0180		397.5	7.6E-006	72.8	86		
SF0217		307.4	5.2E-002	15.4	50		
SF0218		349.4	1.2E-003	48.8	41		
SF0219		363.5	5.0E-004	85.2	91		
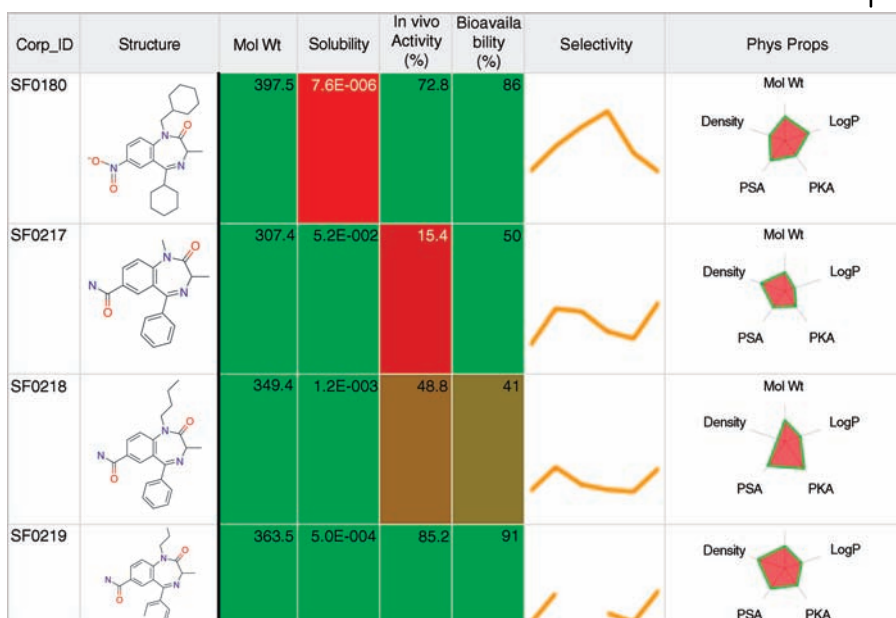
Figure 7.3   Color-coded SAR table with images.

against a series of receptors, and a radar plot shows the profile of each compound's physical properties. Now the SAR table starts to become a compact, information-rich representation of key information about compounds, as described in Ref. [1,7].

7.2.1.2 Scatter Plots

A scatter plot is a common and powerful way to see relationships between various properties or biological results. In its simplest form, a scatter plot shows two measures of interest on the *X*- and *Y*-axes and plots a point for each compound.

The scatter plot in Figure 7.4 depicts the effects of a library of compounds on both normal cells and cells treated with a growth inhibitor. The plot immediately reveals a strong correlation between cell functions in these two types of cells, with a minority of compounds that defy this correlation. In an interactive environment, a user might chose to highlight the outlier compounds, view their structures, and try to understand the structural basis for their different biological behavior.

The simplest use of scatter plots is to discover and highlight correlations and outliers between two variables. But they can also be highly effective for teasing out multivariate relationships, by exploiting various visual attributes of the graph, including color, shape, size, and trellis panels. Figure 7.5 explores the relationship between control and treated cell assay values, while also depicting four additional variables: the number of criteria met (by marker color), the year of
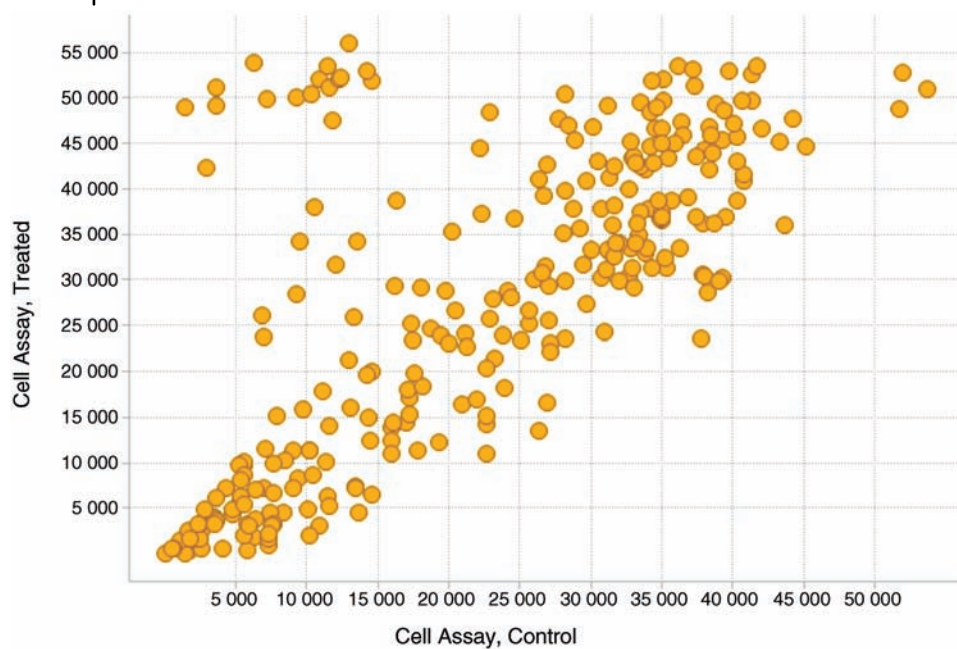
Figure 7.4   Simple scatter plot.
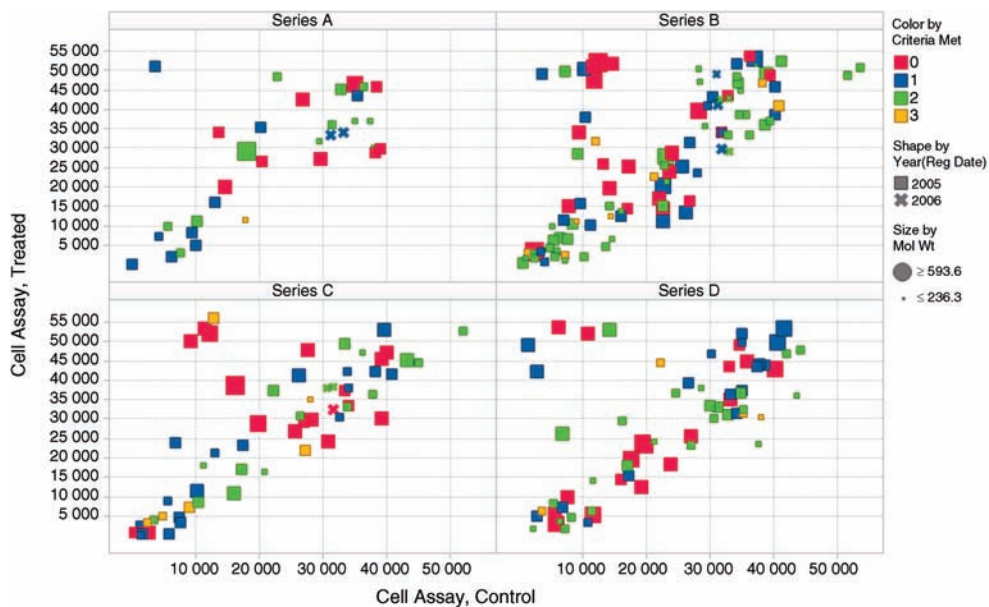
Figure 7.5   Multivariate analysis in a scatter plot.

registration (by marker shape), molecular weight (by marker size), and chemical series (by trellis panel).

From this richer representation, we can make a number of observations beyond the previous insight that the two measures are generally well correlated with some outlier compounds:

• Outlier compounds appear to have higher molecular weights (though not all high molecular weight compounds are outliers).
• Most compounds were registered in 2005, and there are not obvious distinctions between these and compounds registered in 2006.
• Few compounds meet three criteria. Some but not all of these compounds cause different behavior in treated versus untreated cells.
• Series A has just one outlier compound.

Additional variables can be depicted by interactively adjusting the range of values of one or more additional variables and by adding a third axis to the graph, as in Figure 7.6.

In this 3D plot, we can see that that compounds with lower cell assay values tend to have higher solubility, and that the outlier compounds (in the two cell assays) have
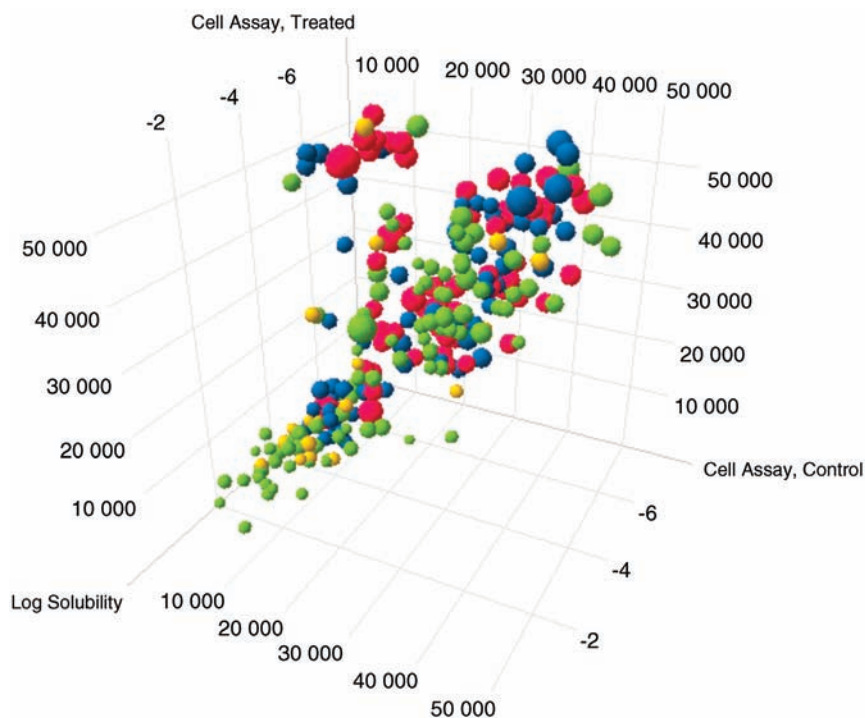
Figure 7.6   Three-dimensional scatter plot.
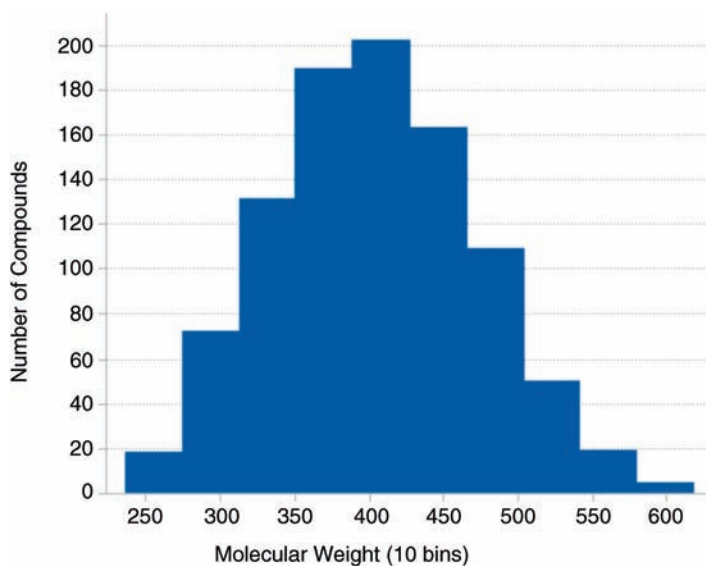
Figure 7.7 Simple histogram.

varying degrees of solubility. Such insights in 3D scatter plots are greatly aided by the ability to rotate the plot interactively.

7.2.1.3 Histograms

Another simple visualization that can quickly summarize large amounts of data is a histogram. Figure 7.7 shows the distribution of molecular weights in a library of 1000 compounds.

Adding additional visual dimensions to histograms can reveal more complex relationships or interesting subsets of data. Figure 7.8 summarizes the results of screening three series of compounds against four targets. It reveals that more compounds are active against 5-HT1b compared to the other receptors, and that compounds in series B are in general less active than the other two series.

7.2.2
Advanced Visualizations

7.2.2.1 Profile Charts

There are a number of graphical approaches to depicting multivariate data about individual compounds. We will expand on two such approaches that were briefly introduced in Figure 7.3 in our discussion of SAR tables. The first is a profile line chart that shows the relative activity of a compound against a series of targets. In Figure 7.9, we show an expansion of this idea using the same data set. Each line
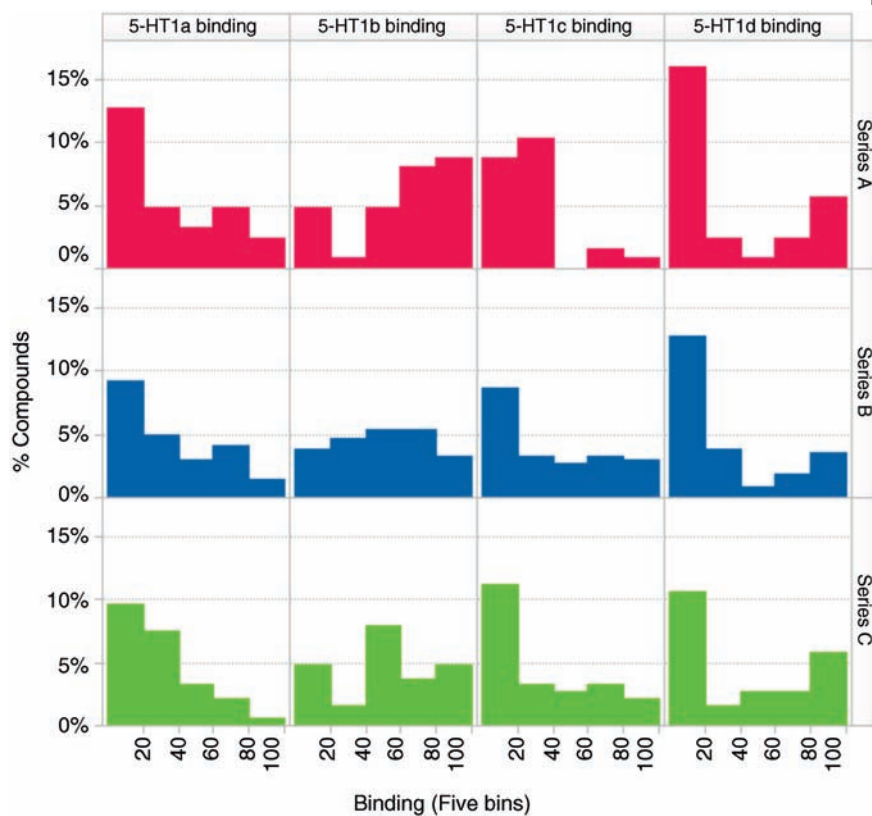
Figure 7.8 Histogram trellised by series and target.
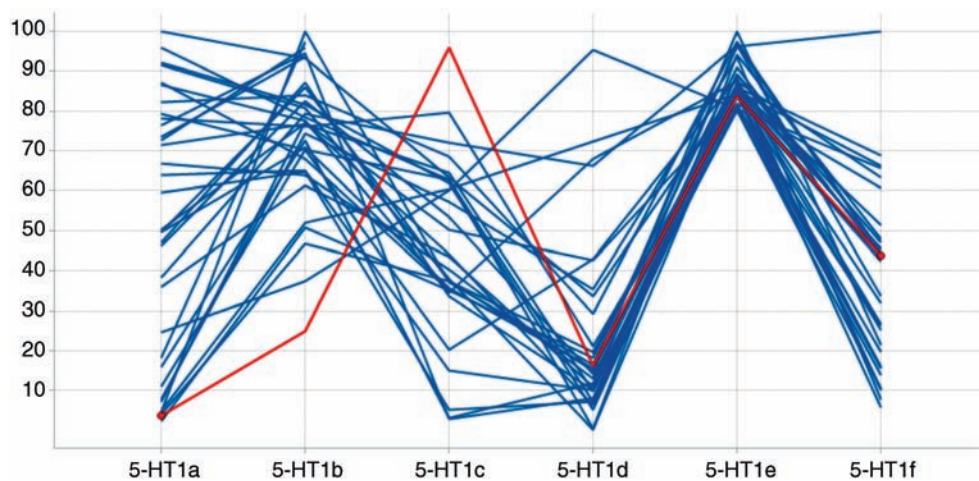
Figure 7.9 Selectivity profile.
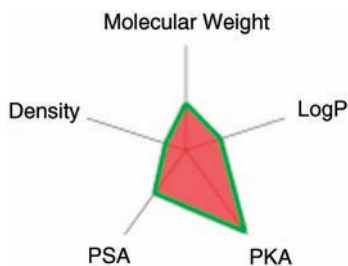
Figure 7.10 Radar plot.

represents a compound and shows the activity of that compound against six related receptors. All compounds in this graph show high activity against 5-HT1e, and many of them also show high activity against other receptors. For example, the red line represents a compound that is highly active against 5-HT1c as well.

Figure 7.10 depicts the values of five physical properties for a single compound in a radar plot (also called a spider plot). Each point in the polygon indicates, by its distance from the center, the relative value of that property relative to other compounds. This depiction makes it easy to pick out compounds whose radar plot shapes are either desirable or undesirable.

7.2.2.2 Dose–Response Curves

In drug discovery research, IC_{50} and EC_{50} values are key measures of the activity of compounds. These values are generally derived by measuring the response of compounds at several different concentrations and then fitting a logistic regression curve to these points to determine the concentration predicted to result in 50% of maximum activity. While often the derived IC_{50} values are used in further calculations or visualizations, it can be useful to view the curve fits in order to assess the quality of the data. Figure 7.11 shows an IC_{50} curve fit. Figure 7.11a shows the curve and derived IC_{50} using all the data. Note the highlighted point at concentration $= 1$, which is likely a bad measurement that adversely impacts the curve fit. From a visualization like this, it is easy to spot and remove bad data points and update the resulting curve fit, as illustrated in Figure 7.11b.

7.2.2.3 Heat Maps

Detecting and visualizing patterns across multidimensional large data sets remains a challenging problem. Although the color-coded SAR table already discussed can accommodate arbitrarily large data sets, the size of the table grows with the number of compounds, making it infeasible to visualize patterns across a large number of compounds. A heat map visualization can encapsulate information on an arbitrary number of compounds in a single compact view. It is ideal for visualizing patterns, particularly clusters, across large numbers of compounds.
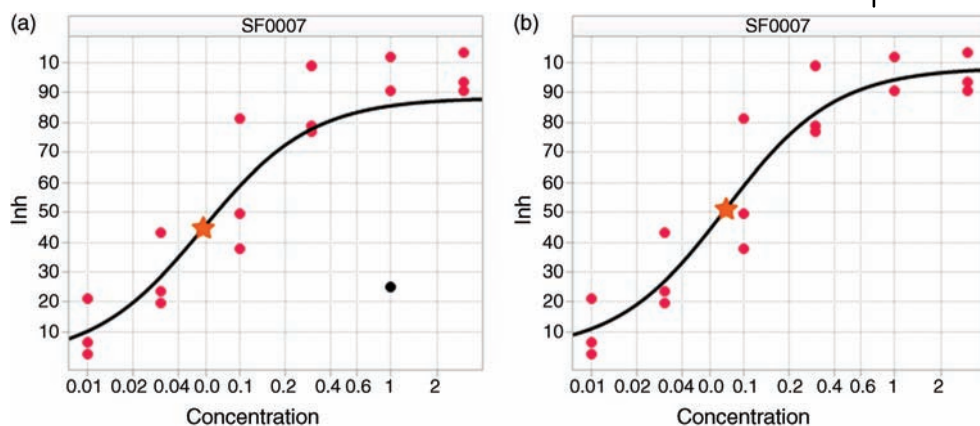
Figure 7.11 Dose–response curves. (a) All data. (b) Highlighted point removed.

The heat map in Figure 7.12 clusters together hundreds of compounds based on the similarity of their binding profiles to six different receptors. The heat map also shows the clustering of the receptors themselves based on the binding patterns of the compounds; this chemical-based clustering could provide insight into similarities among receptors.
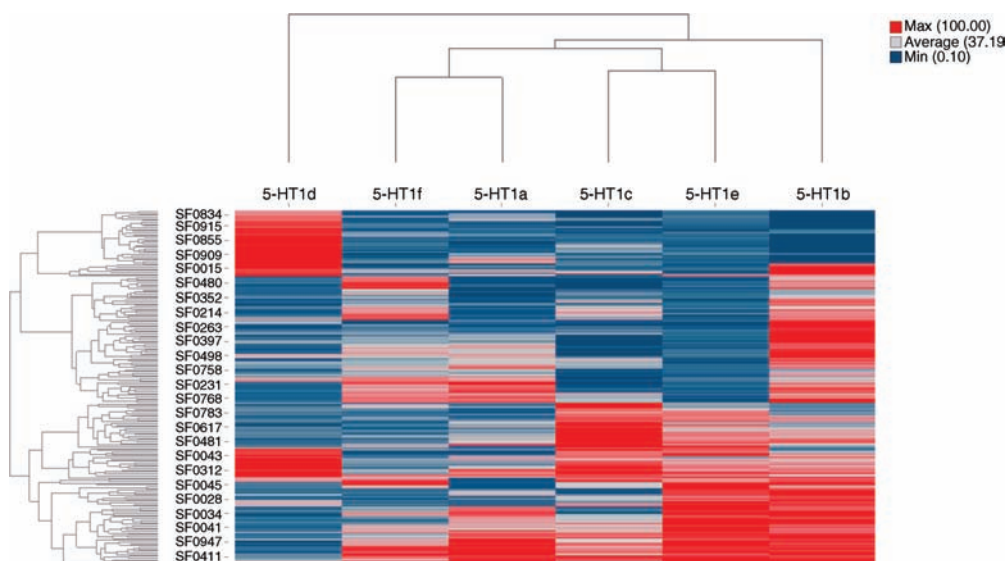
Figure 7.12 Heat map, showing clustering of compounds (rows) and receptors (columns).

7.2.3
Interactive Analysis

Each of the visualizations already described can be useful in itself as a static image to provide insights into important questions like structure–activity relationships. Visual analysis reaches its full potential, however, when performed in an environment that allows scientists to interact with the data and get real-time feedback. We have seen a few examples of this so far, as when we illustrated the ability to spot and omit bad data points from dose–response curves. Figure 7.13 illustrates an environment that gives scientists visibility into how well the compounds in a program are satisfying project criteria. Visualizations show the number of compounds that satisfy each criterion individually or in combination. The user has the ability to adjust any of the thresholds and to drill down to view a rich SAR table of the compounds that meet the adjusted criteria; in this illustration, the user chose to view all compounds that meet three out of the four criteria.

Another critical component of the interactive graphical experience is the ability to see cross-linking of highlighted data points across multiple graphs and to drill down to more detailed graphical views of selected data. Figure 7.14 illustrates a histogram of screening results alongside a scatter plot showing the relationship between assay results in normal versus treated cells. The user has highlighted in the histogram all

Figure 7.13 Compounds meeting multiple criteria, with adjustable thresholds.
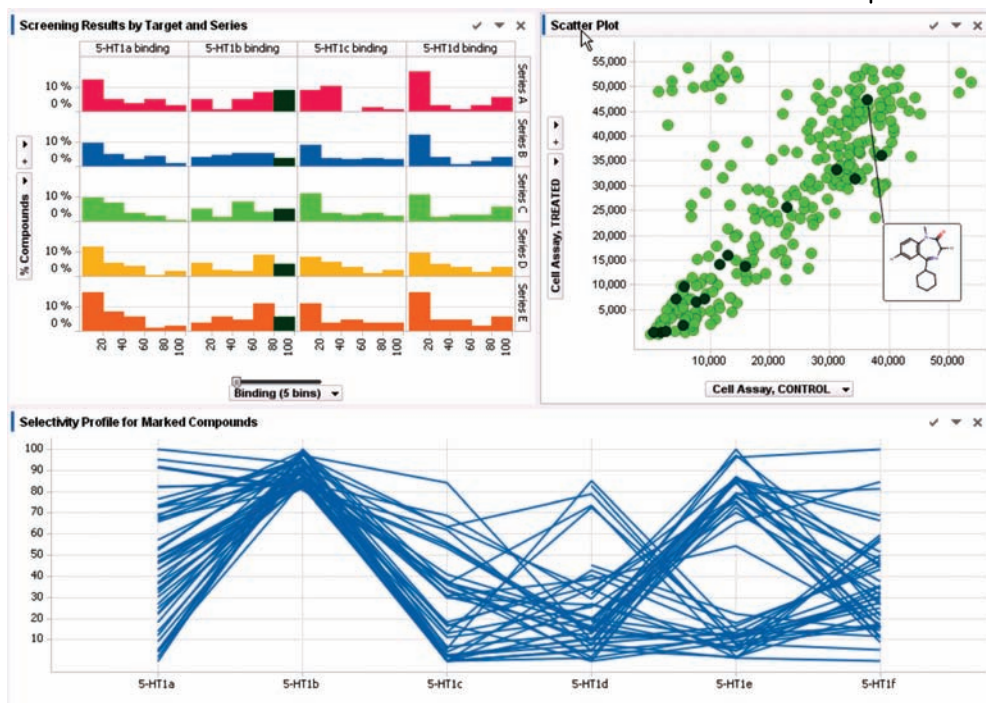
Figure 7.14 An interactive environment for understanding active compounds.

compounds that are highly active against the 5-HT1b target. The scatter plot automatically highlights the same compounds and the profile chart shows each compound's selectivity against a panel of six related receptors. This type of interactive environment is a great motivator for scientists to explore the next question that comes to mind. In this example, it might be to determine which compounds are not only *active* against the 5-HT1b target but also *selective* for that target. To answer this question, they could simply filter out any compounds that are active against the other five receptors, as shown in Figure 7.15.
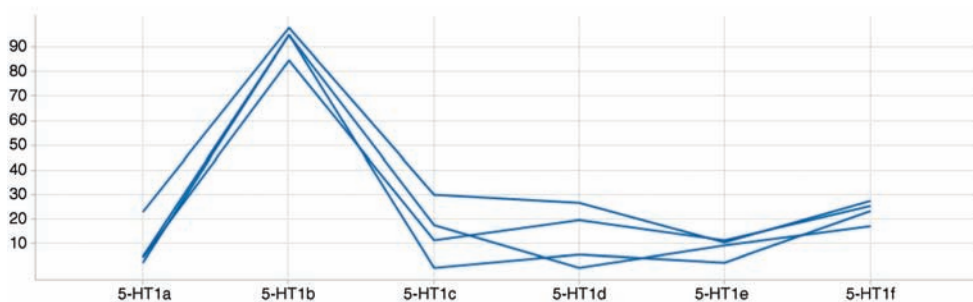
Figure 7.15 Filtering to selective compounds.

7.3
Genomics

7.3.1
Common Visualizations

Visualizations in genomics today are often directly based on the sequence of the molecule accompanied by functional annotation. In the past this was done by textual representation with minimal annotations using normal text formatting such as asterisks, superscript and subscript, fat, or italic.

The standard text for DNA is the GATC base representation. Also, proteins or polysugars are represented by a one-letter code. These sequences represent the most basic information for all other derived and aggregated visualizations of biological sequences.

Genomics sequence standards are stored in biological reference databases and function as gold standard for all other derived data sources or test results. These reference data sources also have the function of mapping the sequences to unique identifiers and versioning. Keeping these identifiers synchronized through all derived data sources is a major thread for all the information and test results we would like to visualize and analyze. Genomic test results are coupled to the sequence through common identifiers of biological databases.

Many modern genomic tests are based on binding of short, single-stranded D/RNA-molecules. There are numerous physical methods to measure sequence matches or mutations. The results of these tests are again interpreted in many different ways, from pure absence or presence of the molecules to functional properties of molecules.

High-throughput testing of molecular binding leads to computationally intensive result calculation and statistical tests. The compute step directly links the results to reference data sources and functional annotation. Statistical methods add confidence values and clustering for data mining. All steps add one or more layers to the test results. The result of such a test is in fact a result set that can be evaluated on its own, but most often is compared with other result sets.

The scientist has to apply sophisticated analytical methods to gain new insights into these complex data. Visualization and interactive analysis are key to allow fast, reliable, and flexible finding of results. This analysis is a variable multistep process, going through all combinations and layers of the result sets depending on the question the analyzing scientist has.

In this chapter we deal with interactive, graphical analysis of genomic result sets. The workflow to generate these result sets is relatively fixed for most tests used here. We will not discuss varying rules for data acquisition and assembly. Since the analysis process is very flexible, we only discuss the graphical tool set from basic to advanced and interactive, but that gives only limited guidance on when to apply which visualization and how simple or complex it is to create the visualization from the result set.

7.3.1.1 Hierarchical Clustered Heat Map
The idea of a heat map is to visualize a huge data table by color coding each cell based on cell value (Figure 7.16). An overview is generated by this graphical method, but
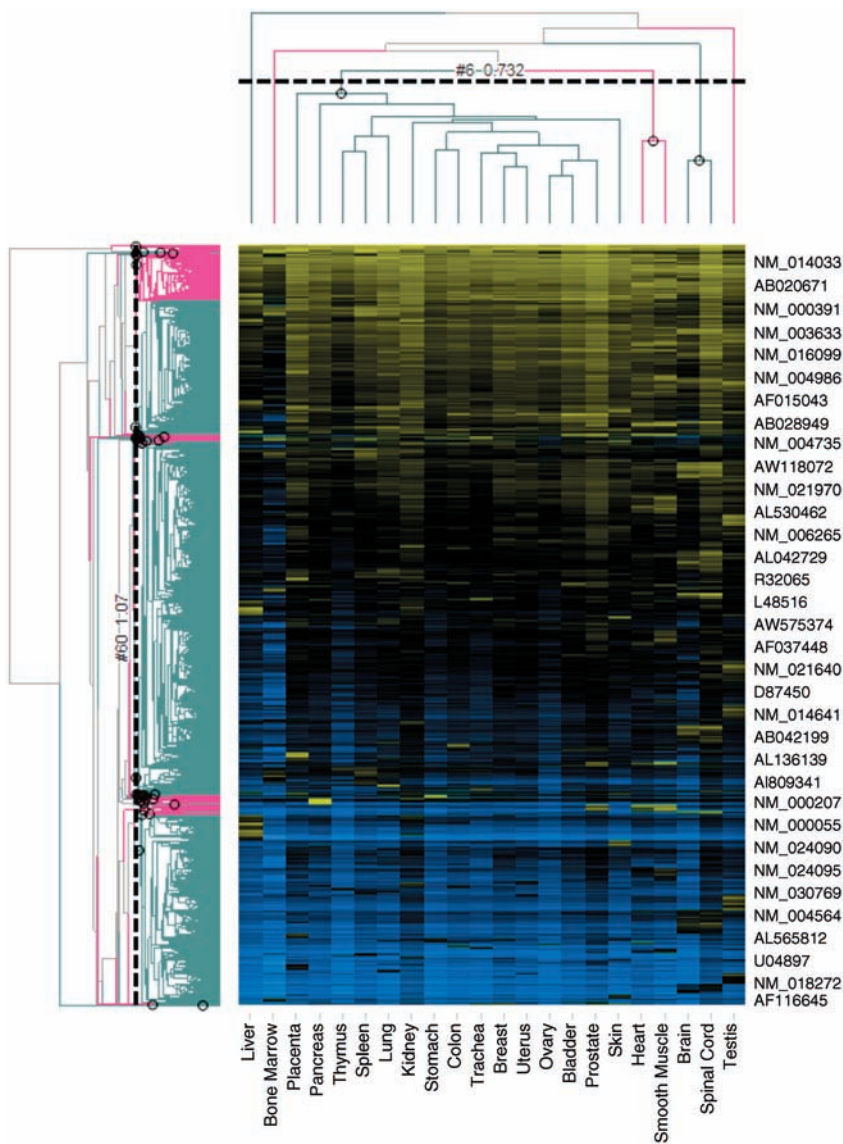
Figure 7.16 Heat map of gene expression in normal tissue. One gene per row and one tissue per column. Sorted by hierarchical clustering on rows and columns, grouping similarly expressed genes together and tissues with a similar expression pattern (expression profiles of human normal tissues http://www.ncbi.nlm .nih.gov/geo/query/acc.cgi?acc=GSE2361) [8].

since the results are unsorted we can apply statistics to sort or cluster the table according to the test results. Hierarchical clustering arranges items in a hierarchy with a tree-like structure based on the distance or similarity between them. The graphical representation of the resulting hierarchy is a tree-structured graph called a

dendrogram. Heat maps and dendrograms together are a widely used tool to analyze biological test results.

The interactivity comes into play when the analyst compares the clusters by looking at the trends in the heat map. Depending on the clustering method, similar results will be shown closer together.

7.3.1.2 Scatter Plot in Log Scale

Another way of visually grouping results together is a multidimensional scatter plot. The multiple dimensions come from the *x*- and *y*-axes, as well as from color coding and size and shape of the markers (Figure 7.17). Biological measurements are often best inspected in log scale for the axis, but also for the coloring or sizing. Although the differences in size of scatter plot markers can be seen very quickly by the human eye, it is nearly impossible to guess the right scale. Some software packages calculate the sizes by diameter, some by screen real estate. But, in general, these scatter plots are meant to quickly spot groups of outliers to the main point cloud. The human eye can immediately check the homogeneity of the data, although there might be
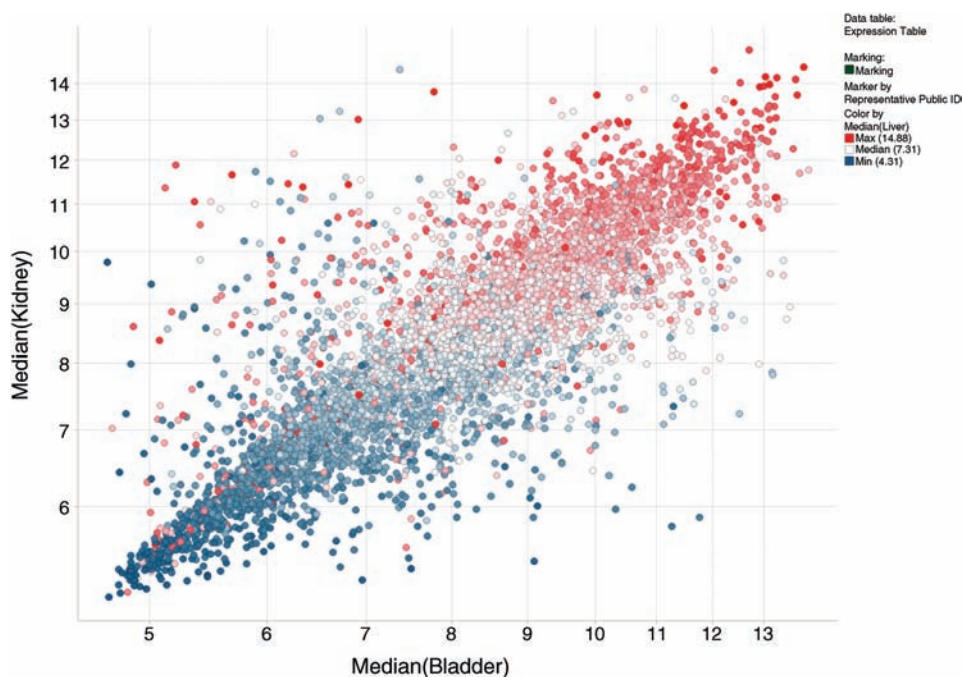
Figure 7.17 Log/log scatter plot of three expression profiles: bladder, kidney, and liver (color). The vast majority of the gene expressions are in the central data cloud. Highly expressed genes of kidney and liver that have a lower expression in bladder are in the upper left-hand corner in red. The lower right-hand corner indicates that the few genes high in kidney and low in bladder are all low in liver (expression profiles of human normal tissues http://www.ncbi.nlm.nih.gov/geo/query/acc. cgi?acc=GSE2361) [8].

millions of single data points shown in the scatter plot graph. More detailed visualizations or statistics need to be used to really identify more subtle differences between visually diverse groups. An additional possibility of subdividing scatter plot graphs by even another variable is the breakout or trellis plot. Here, we can draw the same graph multiple times for different test sets. Depending on the test layout, the trellis variables might be different set of cells or different treatment doses or times. The trellis possibility again enables the human eye to recognize and distinguish different patterns of multimillion data points very quickly.

If multiple tests were performed to come to a better quality of the result, a scatter plot can also be used to show only medians of, let us say, triplicate measurements. The variance of the measurements can again be used as a visual component in a chart, for example, using the size of the marker. By using a derived number for the somewhat incomparable size of a marker, the relative size is much more important and can again be spotted very quickly. Huge markers represent multimeasures which have a high variation and thus might not be reliable.

All aspects of the scatter plot are very well suited to quickly see trends and outliers in data clouds with less than a handful of parameters.

7.3.1.3 Histograms and Box Plots for Quality Control

In every workflow there is possibility to check for data quality while it is executed. Data coming from microarray experiments undergo a complex workflow of number crunching before the scientist can analyze the result. At some milestones, it makes sense to check for data quality and interfere if there are obvious outliers. The spread of a signal can be visualized in a histogram. In our example, we do an overlay of several histograms to verify the conformity of the signals (Figure 7.18). The line chart offers a good way of plotting multiple histograms in one single graph. In addition to the use of a line instead of a bar chart, we also binned the x-axis to smooth the line of the chart and to make it comparable. This makes the graph immediately understandable for the user and in our case it makes it easy to spot outliers, mark them, and remove them immediately from the downstream analysis.

The box plots used to compare lots of data points in one go are used here for comparing the individual experiments similar to the histograms already described. But the box plots have one more function, in this case a function that shows parts of the data workflow. Figure 7.18b shows all experimental data before a certain normalization step in the data workflow. Figure 7.18c shows the same data but after normalization. This can be seen in the distribution of the boxes of each experiment. They are scattered in Figure 7.18b and aligned in Figure 7.18. This means that after the normalization step, the data are more consistent and can be compared more easily than before.

7.3.1.4 Karyogram (Chromosomal Map)

Originally karyograms were used in cytogenetics and cancer research for showing if there are changes in the number or shape of chromosomes in metaphase (karyotype). Distinct karyotypes could be linked to phenotypes, for example, trisomy with different chromosomal aberrations. Since genomics became more and more fine grained and has now reached the base level, karyograms also became more detailed.
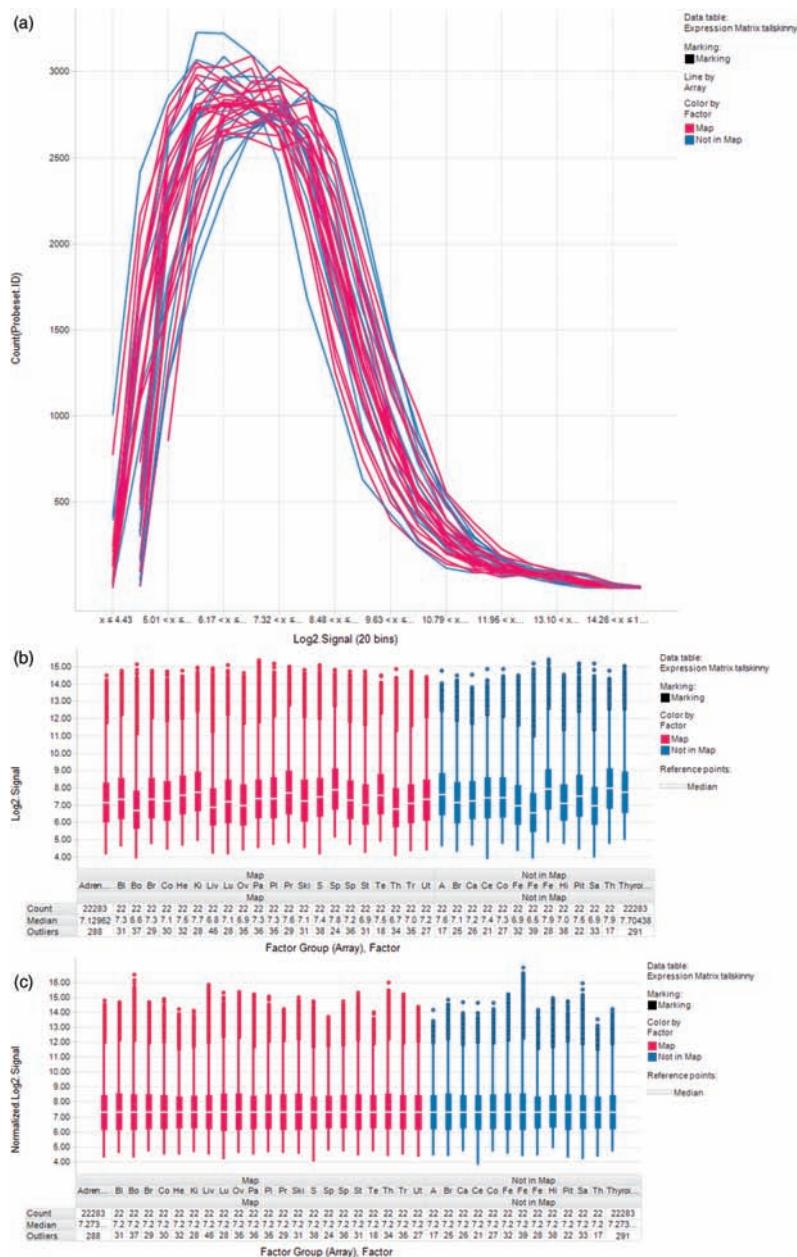
Figure 7.18 Quality control of the result sets. Expression values (signals) are mapped against the experiment. (a) The line chart shows the histogram of each experiment result. (b and c) The chart shows the signals before and after normalization – see the boxes (central 50% of the values) are aligned and thus better comparable in the analysis afterward (expression profiles of human normal tissues http://www.ncbi.nlm.nih.gov/geo/query/acc .cgi?acc=GSE2361) [8].
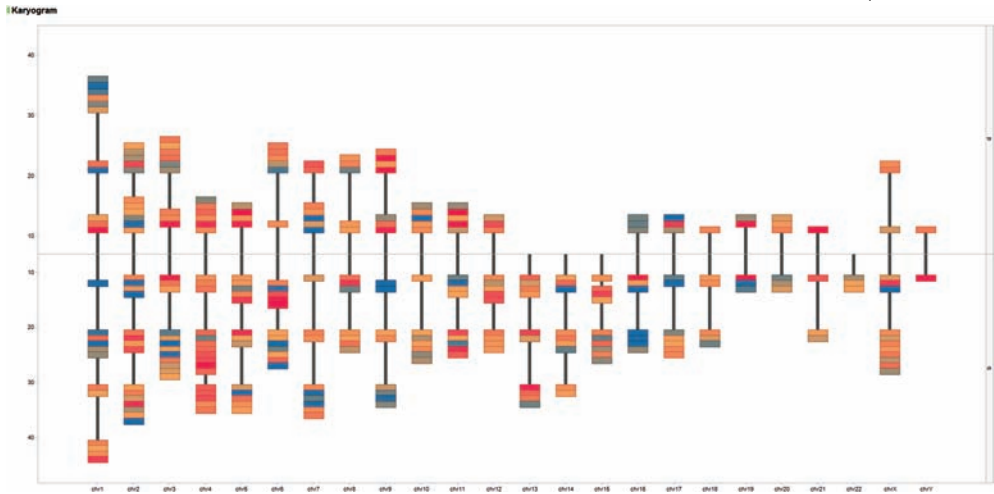
Figure 7.19 The karyogram shows the human chromosomes. The color-coded blocks are chromosomal bands, where genes of interest are located. The data set contains expression profiles of 22 organs. All genes above a certain expression level (signal) in the tested tissues are represented here. The color represents a calculated value normalizing the amount of genes per block/band versus the amount of tissues. Blue boxes represent stretches of highly expressed genes in many organs. Red boxes represent highly expressed genes in a small number of organs and thus organ-specific expression (expression profiles of human normal tissues http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2361) [8].

All information about alterations is still mapped to the genome to see the overall picture. An example could show the differences in single-nucleotide polymorphisms (SNPs) in two groups of cancer patients. The chromosomes are subdivided into multiple bins and the frequency of certain SNPs is represented by different colors (Figure 7.19). The usage of karyograms has not changed. Still researchers are mapping genetic information on the chromosomes and positions to identify changes related to diseases. In many cases, not only one but several types of biological information are combined and shown on the karyogram. An example of this would be the combination of SNP data and the information about gene promoter regions and a certain type or group of genes, for example, all genes belonging to a metabolic pathway and accordingly the expression levels per organ. Modern karyograms using this combination approach are often the starting point for a more detailed analysis using the type of interactive analysis described above in Section 7.2.3.

7.3.2
Advanced Visualizations

7.3.2.1 Metabolic Pathways
Metabolic pathways are used to describe the networks that occur in biological entities such as the cytoplasm of a cell. All pathways are linked and steered through
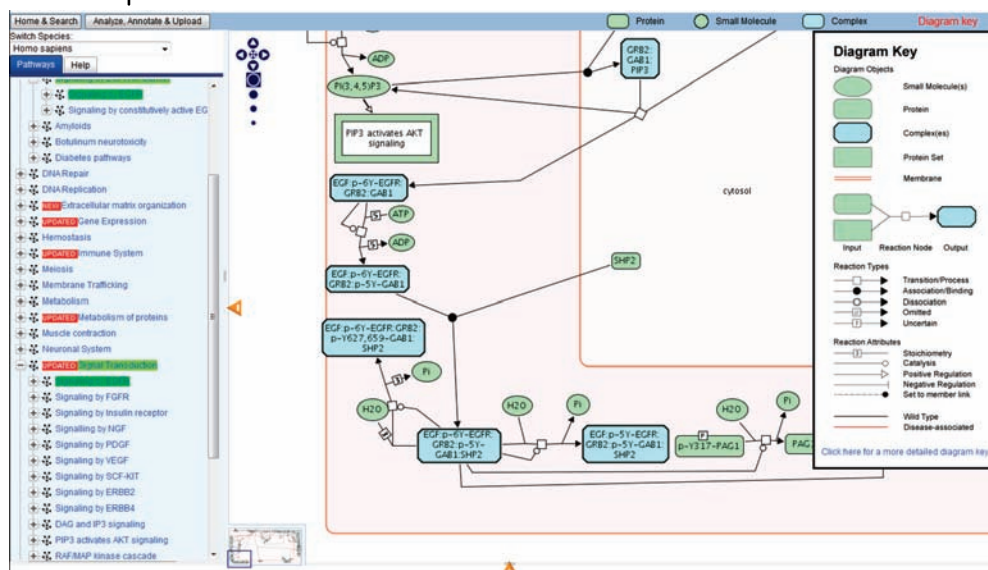
Figure 7.20 Signaling by EGFR. This is an interactive pathway map showing a part of a signaling cascade in a web browser (http://www.reactome.org/entitylevelview/ PathwayBrowser.html#DB=gk_current& FOCUS_SPECIES_ID=48887& FOCUS_PATHWAY_ID=177929&ID=177929).

the lack of abundance of the components. If one component is disturbed or not functioning, then normally other pathways are regulating this. Some key elements have no second option and are thus very susceptible to fail. Some of these are already known and described and linked to phenotypes. Only graphical pathways can really describe the reactions in living organisms. Many databases and approaches have been established around this (e.g., Reactome.org and pantherdb.org) (Figure 7.20). You can link the results of a gene regulation to the underlying pathways and can via this link see what the gene regulation is responsible for downstream in the pathway, or which events upstream can lead to this gene regulation.

This advanced visualization is not a direct result of one experiment, but a result of multiple experiments and their analysis and interpretation of the context. They could be seen as maps to put single genes or reactions into a broader context. The experimental results can be again put on top of the existing pathway to see how well it matches with the other results.

7.3.2.2 Gene Ontology Tree Maps

Gene Ontology (GO) [9] is a bioinformatics consortium providing a controlled vocabulary for genes and their products. The ontology comes in three parts: cellular components, molecular functions, and biological processes. Similar to metabolic pathways, the ontology provides a context to gene of interest. Again, you can build or extend ontologies yourself with new findings and so the tree grows and gets more
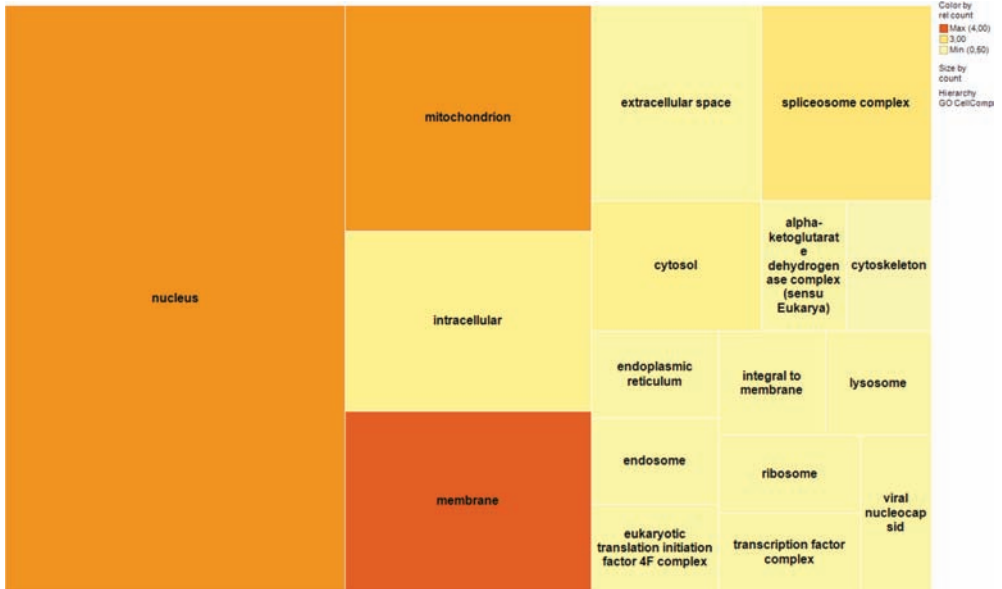
Figure 7.21 The Gene Ontology cellular component upper hierarchy level is represented as blocks of different size. The size is linked to the number of genes per GO category. The color of each block displays a calculated number of genes of interest per category. The darker the color, the higher the relative number of genes in the current selection, and the more important this use case. This represents already a drill down into a small subset of the data. The Tree Map in this figure is designed to handle a large amount of categories and give a very quick overview of and insight into the selected data set.

complex. The tree structure itself can also be regarded as a tool for visual analytics. As an alternative approach to the hierarchy-like structure in which the ontology is most often presented, one can represent the ontology in a Tree Map visualization (Figure 7.21). The visualization is using the ontology to draw the boxes. The size of the boxes is determined by the amount of entities, for example, genes belonging to the ontology group. If you link genes to ontologies, you can aggregate or cut the ontology at a certain level of detail and use this level to define the big boxes of the Tree Map visualization. If you are looking at a number of regulated genes or proteins by the results coming from an experiment, you can directly see where in the cell they are located/active, what functions they have, or in which processes they are involved. Normally you will see many groups, but only several prominent ones floating to the top left of the Tree Map. The bigger the box, the more it will land on the top left corner of the graph. Of course, one can do statistics on these ontologies as well; one very simple example would be a relative number achieved by dividing the actual number of genes counted in one box divided by the number of genes belonging to the entire group. This number could also be used in a Tree Map to color the boxes. One would expect many hits in big ontology groups, but if there are rather small groups with high counts, then this number would be higher indicating that this

function, compound, or process might be a relevant result to look after. So not only the sizes of the boxes in a Tree Map but also the associated calculations shown by color coding are of interest. The combination of size and color gives an immediate insight into the experimental results.

7.3.2.3 Clustered All to All "Heat Maps" (Triangular Heat Map)

We have discussed the use of heat maps earlier in this chapter. The kind of heat map we describe here has another layout and function (Figure 7.22). We have put patients onto the *X*-axis of the heat map as well as on the *Y*-axis of the heat map forming a grid. The number/color in the heat map results from a similarity computation across multiple factors. The grid will be useful when we apply a hierarchical clustering and similar patient combinations are grouped together suggesting that they have some things in common. With this approach, the data on both "sides" of the heat map are redundant and thus one half can be left out for presentation and the result would be a triangular heat map. In our example we have computed the similarity of patients among four groups and also found a distinct gene expression pattern in the associated heat map using sorting from the
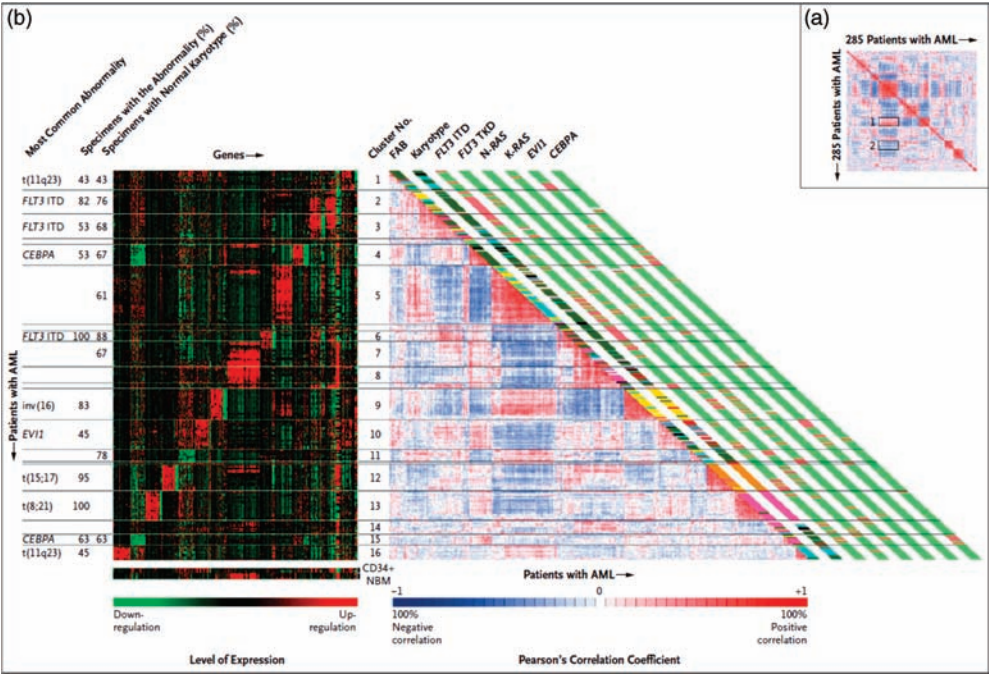
Figure 7.22 Correlation view of specimens from 285 patients with AML involving 2856 probe sets (part (a)) and an adapted correlation view (2856 probe sets) [right-hand side of part (b)], and the levels of expression of the top 40 genes that characterized each of the 16 individual clusters [left-hand side of part (a)].

clustering results of the triangular heat map (data and picture from Ref. [10]). Many different factors can be included into the triangular heat map and then afterward one can extend the annotation, for example, by gene expression or absence/presence of genomic markers or phenotypes. The triangular heat map can combine information from various sources and quality, not only focused on gene expression or other experimental results. Through clever use of statistics, we can again build an aggregated visualization we can start our data analysis from. We see the (hopefully) obvious clusters and can then drill down or across to other results or information that were not initially included in the first calculation or that were only included as an already aggregated value. The inclusion of medical history, vital signs, family history, and so on can be used to create the first all-to-all triangular heat map and via the patient or gene IDs, one can always link to other data.

7.3.3
Applications

7.3.3.1 Understanding Diseases by Comparing Healthy with Unhealthy Tissue or Patients

The biological networks in diseased tissue are disturbed. This might be visible in many different aspects. Taken together, gene expression patterns, genetic alterations such as SNPs, epigenetic factors, and environmental conditions can define a phenotype better than one of the factors alone. That is why every available piece of information needs to be combined and put into context. But the data could be very large and thus have to be stratified. One example of data stratification is to look only for those genes that are differentially expressed in the disease group versus the healthy group. Genetic alterations like SNPs or deletions can be filtered to those that are already described in literature and those that have an effect on the protein sequence or direct promoter region. Environmental factors such as working conditions or alcohol or smoking habits must be taken into account as well probably by adding additional blood test values of patients. Visualization is the key to create a big picture of such a comparison. Displaying different types of information together that otherwise would be too complex to integrate is one of the key applications for visual data analysis. Most of the visualization techniques described in this chapter are suitable in such an approach. The main point here is to show the differences between the groups instead of showing the data of both groups in parallel.

7.3.3.2 Measure Effects of Drug Treatment on a Cellular Level

In addition to comparing two groups, for example, patients and controls, the effect of different dosage of a treatment can be visualized (Figure 7.23). Different effects or different grades of one effect might be visible in a dose escalation experiment or in titrations. In many cases, this is a second step after the comparison of two groups. In addition, different time points can be measured after treatment with different doses. The measured effect can be the same in all doses, but with a different time course.

Median of samples over time
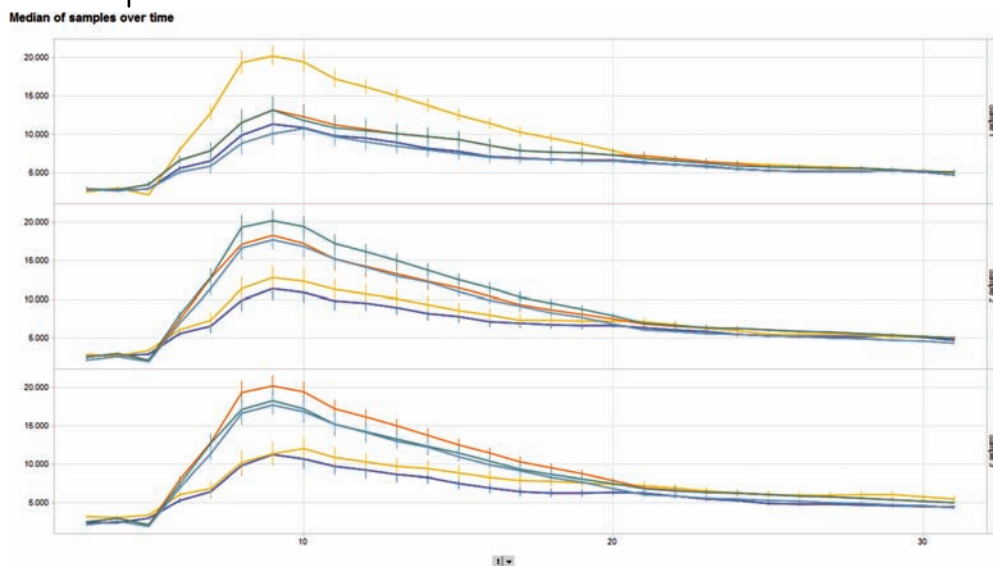
Figure 7.23 Comparison of many factors. A set of drug treatments is compared over time in three samples. Since all samples show the same curve shape, the tests are comparable. Now one can chose a drug by the desired effect based on several factors. The error bars represent the standard error of multiple measurements at the time point.

References

1 Allerheiligen, S.R.B. (2010) Next-generation model-based drug discovery and development: quantitative and systems pharmacology. *Clinical Pharmacology & Therapeutics*, **88**, 135–137.

2 Reid, A.J. (2011) *Nature Reviews Microbiology*, **9**, 401.

3 O'Donoghue, S.I. *et al.* (2010) *Nature Methods*, **7**, S2–S4.

4 Tufte, E.R. (2001) *The Visual Display of Quantitative Information*, Graphics Press.

5 Stele, J. and Iliinsky, N. (2010) *Beautiful Visualization*, O'Reilly Media.

6 Few, S. (2009) *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, Analytics Press.

7 Ritchie, T.J., Ertl, P., and Lewis, R. (2011) The graphical representation of ADME-related molecule properties for medicinal chemists. *Drug Discovery Today*, **16** (1–2), 65–72.

8 Ge, X. *et al.* (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics*, **86** (2), 127–141.

9 Ashburner, M. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**, 25–29.

10 Valk *et al.* (2004) *New England Journal of Medicine*, **350**, 1617–1628.

8
Using Chemoinformatics Tools from R

Gilles Marcou and Igor I. Baskin

While R is one of the most powerful statistical packages to date, it has been used in many areas from data mining, geography, economy, and social science to, most notably, biology and medical science. However, the basic objects of chemoinformatics being graphs, R was often disregarded to manipulate them and build meaningful statistical analysis and powerful models. The practice of R in chemoinformatics generally reduces to three main steps: (i) produce tables of descriptors using a chemoinformatics software (such as DRAGON [1] or MOE [2]), (ii) analyze them using R, and (iii) link the R output with chemical structures (such as CrossFire). This scheme is generally efficiently operated using flow-based programming packages dedicated to chemoinformatics, such as Pipeline Pilot or KNIME.

However, using chemoinformatics functions directly from R bring new degrees of freedom to solve problems: (i) it reduces the need to transfer chemical information through tables, thus saving time and memory; (ii) it allows retroaction of statistical manipulation to chemoinformatics tools. The most obvious example should be the interactive use of QSAR models from R: users should be able to design a molecule and get instantaneous predictions from R whenever they modify their drawings. An extension of the concept is *de novo* design.

This chapter will focus on the technical details on how to achieve such communications between chemoinformatics packages and R. The chapter will describe all of the three main communication protocols between R and external chemoinformatics tools. The first one is communication with shared libraries (.so files on UNIX systems and .DLL files on Microsoft systems). Since the rise of Java, communication is also possible with Java Archives (.jar files). The last and the least powerful communication method is the system call to software from R.

The chapter will illustrate these methods with simple examples demonstrating how R can be used to communicate with open and closed source solutions such as Rdkit [3], OpenEye [4,5], CDK (Chemistry Development Kit) [5], and ChemAxon [6]. It will focus on wrapping, compilation (R CMD SHLIB), managing external libraries (`dyn.load` and `dyn.unload`), and calling function (`.C`, `.Call`, `.External`). Concerning Java, the discussion will focus on `rJava`. Finally, the chapter will describe the function system.

8.1
Introduction

The objective of this chapter is to provide basic guidelines to extend R [7] with chemoinformatics [8–10] tools developed by third parties. The point of view chosen is to formulate the technical details in such a way so as to make them accessible to scientists who are not specialists in informatics. The present chapter assumes a Linux environment and GNU Compiler Collection. However, this is not a limitation since the concepts are identical for any other system (Windows, Mac) environment. Several examples are given to perform very basic tasks. Although the examples were designed to be simple for clarity, they are complicated enough to give an overview of the technical difficulties. Therefore, it is quite straightforward to generalize it as complex projects.

The chapter is structured in four sections (Figure 8.1). The first one describes the system call in order to run foreign software from R. The second one is dedicated to the shared libraries mechanism that allows access to third-party functions under some constraints. The third section explains how to override these constraints using wrapper functions. The last section is dedicated to the special case of Java.

8.2
System Call

System call consists in using from a host application (in this case R), second software.
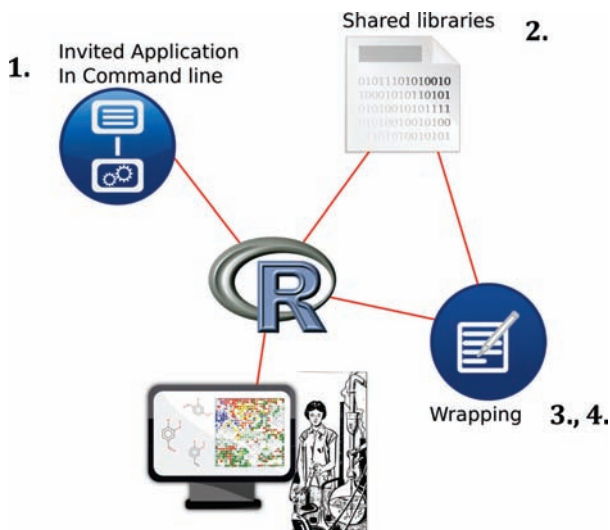
Figure 8.1 Principal concepts of using R in chemoinformatics. R can access third-party software components using three main concepts: 1. system call, 2. shared libraries, and 3.wrappers; 4. is dedicated to the case of Java.

8.2.1
Prerequisite

This functionality implies three constraints:

i) The invited application must have a command line.
ii) The host application should be able to send to the system the command line.
iii) The host application must be able to catch messages sent by the invited application. Other details are system specific.

8.2.2
The Command System()

Only some aspects of the commands are discussed here. For a complete description it is more convenient to use the internal help system of R [7]. The basic command line is the following:

```
system(command, intern = FALSE, ignore.stderr = FALSE, wait = TRUE,
input = NULL)
```

The `command` is a string referring to the invited application command line. R can wait for it to be terminated to continue if `wait` is set to `true`. In some sense, R allows input/output redirection. On one hand, the `input` keyword accepts a character vector, which is translated as a file that is submitted to the standard input of the invited application. On the other hand, it is possible to catch its output by setting the parameter `intern` as true in which case the result of the function is a character vector. In such a way it is possible to perform a sequence of application calls, catching the output of each one to be used as an input for the next one.

In case of error during execution of the invited application, several scenarios take place. First, if the `wait` and `intern` are `FALSE`, the result of the function is always "0." If `intern` is `FALSE` and wait is `TRUE`, the result is an error code "0" for no error. If `intern` is `TRUE`, the standard error is reported on the R console unless `ignore.stderr` is `TRUE`, in which case, it is discarded.

The behavior of `system()` is highly platform specific. In particular, care must be taken about environment variables. In general it is safe to assume that a new shell is invoked in which the command is executed. So, to take environment variables into account, they must be inherited.

8.2.3
Example, Command Edition, and Outputs

The following example will consist in the analysis of a small set of 37 compounds associated with vanilla fragrance displayed in Figure 8.3. In general, compounds containing the vanillin scaffold (**5** and Figure 8.2) possess a prefix *vanill*. The only exceptions are the veratraldehyde (**14**), which has its own specific name, and the vanillylidene acetone (**13**), a scaffold that differs a bit from vanillin. Thus, there are
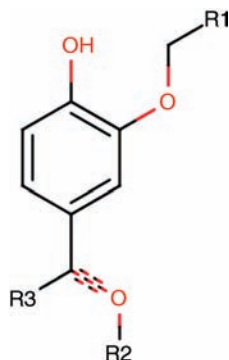
Figure 8.2 The vanillin scaffold. R1 is an alkyl chain. R2 can be an alkyl chain or nothing. R3 is an H or an alkyl chain.

12 compounds containing this particular scaffold. Also, the stereoisomerism is not considered in this example. Therefore the *para*-methoxycinnamaldehyde (**27**) and the (E)-*para*-methoxycinnamaldehyde (**31**) are duplicates. Compounds containing the vanillin scaffold are indicated using a red star (*).

Figure 8.3 Sample of 37 compounds possessing a vanilla fragrance. The vanillin itself is the compound **5**. Compounds containing vanillin scaffold are marked by red *.

The goal will be to compute a number of molecular descriptors that will be finally used to build a clustering model of vanilla fragrances. Meanwhile, some practical aspects of R are exemplified.

The command `system` will be illustrated using the applications `nam2mol` and `filter` from OpenEye [4,11,12]. The first application is used to translate a set of chemical names into a chemical sketch in a convenient file format. The second application is used to filter out compounds from an input list according to a set of rules. As a "side effect," filter can be used to compute a number of molecular descriptors.

First, the names are converted to SMILES [13] that are stored in a vector `smi` and the names themselves are stored in a vector `nme`.

`out<-system('nam2mol vanilla.txt', intern=TRUE, wait=TRUE, ignore.stderr=TRUE)`	Recover the output of `nam2mol` into a character vector named `out`
`aout<-strsplit(out,split=" ")`	The output is split into `aout`, a two-column array based on blank space
`smi<-c()`	An empty list `smi` is created
`nme<-c()`	An empty list `nme` is created
`for(i in c(1:length(aout)))` `smi<-c(smi,aout[[i]][1])`	The SMILES strings contained in the first column of `aout` are copied in the list `smi`
`for(i in c(1:length(aout)))` `nme<-c(nme,paste(aout[[i]][−1],collapse=" "))`	The names of the compounds contained in the last column of `aout` are copied in the list `nme`

Then the SMILES vectors are loaded as input for OpenEye `filter`.

```
system("filter -filter filter_nothing.txt -in - -prefix vanilla -out
vanilla_filter.smi -table vanilla.dat",input=smi)
```

Note that `filter` is waiting for standard input, which is fed by the content of the character vector `smi`. A number of descriptors are computed and stored in tab-separated format in the file `vanilla.dat`. This file can be read using standard input procedure. In the following example, the molecular descriptors provided by `filter` and those computed with ISIDA [14,15] (ISIDA is a chemoinformatics toolbox developed by the team of Prof. A. Varnek at Strasbourg University. It includes its own tool to compute molecular descriptors from a chemical structure: ISIDA descriptors) are used to perform a clustering of the data set using a bootstrapping *p*-value estimates based hierarchical algorithm [16].

`library(pvclust)`	Load the library containing the bootstrapping hierarchical clustering algorithm
`x<-read.table("vanilla.dat", head=TRUE,sep="\t")`	Load as a data frame, the file `vanilla.dat`

`xn<-c(x[3:180],x[182])`	Discard some columns such as the SMILES, the name, the qualitative assessment of solubility, and a binary filter flag
`xn<-as.data.frame(do.call(rbind, xn))`	Transpose the array `xn` and convert it as a data frame
`xn<-sapply(xn,as.numeric)`	Convert `xn` data to numeric type
`h<-pvclust(xn, method.dist="uncentered", method.hclust="mcquitty", nboot=100)`	Build the hierarchical clustering model
`x<-system("~/workdir/ Fragmentor2011/Fragmentor-i vanilla.sdf-o vanilla-t 3-l2-u 4-f STDO| ParseScript.sh", intern=TRUE)`	Compute ISIDA molecular fragment descriptors and store the standard output in `x`
`x<-read.table(textConnection(x), sep=";",header=TRUE)`	Interpret x as data frame in CSV format
`xn<-sapply(x[−1],as.numeric)`	Convert `x` data to numeric type
`h<-pvclust(xn, method. dist="uncentered", method.hclust="mcquitty", nboot=100)`	Build the hierarchical clustering model

The results are displayed in Figures 8.4 and 8.5. Using two sets of descriptors that are rather different, the clusters are also different. The ISIDA descriptors can be viewed as some hash function of the chemical graph. The hash function maps a chemical graph to a fixed length vector of integer. Each component of the vector is the count of one of the substructures enumerated in the whole chemical library. Besides, each atom of a compound is mapped to the set of these substructures it belongs to. Therefore, they are very efficient for substructure recognition. The positions of heliotropin (**25**) and 4-methoxysalicylaldehyde (**4**) in the cluster of vanillin (**5**) are therefore logical (Figure 8.5). On the other hand, `filter` descriptors are much more sensitive to volume as can be seen from the position of *para*-cresyl laurate (**19**) and not to multicomponents substances (**3** and **11**): the cluster correctly gathers ethyl vanillin (**2**), ethyl vanillin acetate (**3**), and ethyl vanillin isobutyrate (**11**).

Those examples are designed to illustrate the use of the command `system`. However, this strategy has a serious drawback. Each system call implies a lot of side effects (memory allocation, scheduling, etc.) that can use more memory and CPU than the process itself. As a consequence, the simpler is the invited application, the more often it is used in a particular process, the less interesting it is to use the command `system`. The only answer to this problem is to use compiled procedures or functions directly, from a `DLL` (Microsoft) or `o/so` (UNIX) file.
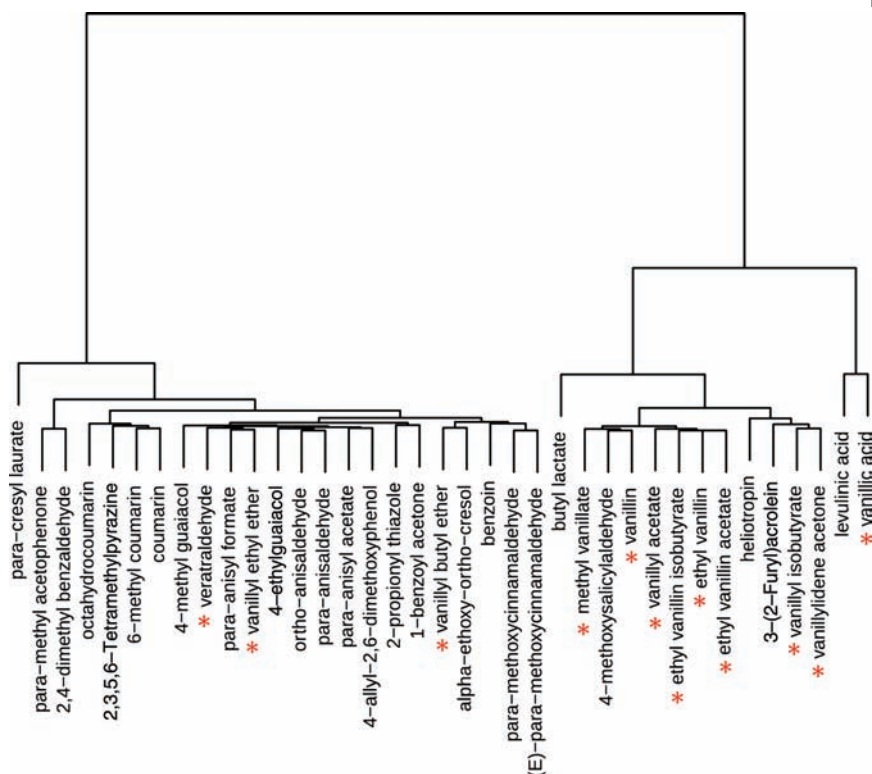
Figure 8.4 Dendrogram of organic compounds, all possessing a vanilla odor. Hierarchical clustering using bootstrapping *p*-value estimates. Dissimilarity measure: uncentered sample correlation; agglomerative method: McQuitty. Descriptors: OpenEye/filter. Vanillin scaffold containing compounds are marked.

8.3
Shared Library Call

This section presents the concept of shared library or dynamic library. For this concept to be understood, the architecture of such a library is described. This section describes what facilities R provides to access functions exposed by a library. This is a powerful but hard way to access those functions.

8.3.1
Shared Library

A shared library concept arises from the concern of splitting data representation and software implementation. The idea is to provide a system with a set of centralized functions that can be used by many different applications. A classical example is the window system of graphical operating systems: each graphical application should
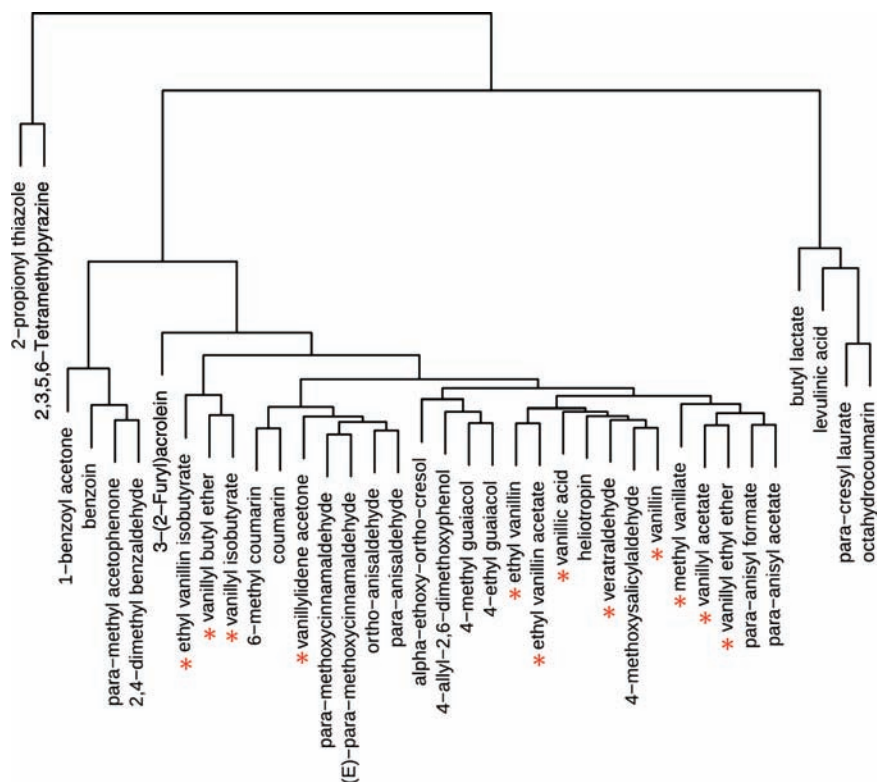
Figure 8.5 Dendrogram of organic compounds, all possessing a vanilla odor. Hierarchical clustering using bootstrapping *p*-value estimates. Dissimilarity measure: uncentered sample correlation; agglomerative method: McQuitty. Descriptors: ISIDA IAB(2-4). Vanillin substructures are marked.

rely on a basic set of functions drawing on screen the different widgets and setting the rules for interactions. It is particularly useful to decrease the size of software: those libraries that are useful to run the software can be loaded in memory only when needed and several processes can share the same chunk of memory containing the library to use its functionalities simultaneously.

In UNIX systems, files containing shared libraries are ended by *.so*, standing for Shared Object or Dynamic Shared Object. On Windows/OS2 systems, they are ended by *.dll*, standing for Dynamic Link Library.

Since the concept of shared libraries is very old and can be traced back to the 1950s, it has been implemented in many compilers and is supported by most operating systems. One of the earliest implementation of the idea can be found in FORTRAN II (1958), which allowed separating the compilation process into modules that were subsequently linked. However, shared libraries adopted their modern shape during the standardization process of the C language, leading to the C89 specification (1989) [17,18]. On this occasion, a number of essential functions were collected into a

document, the C standard library, describing interface standards and implemented in operating systems as a file such as libc.so (UNIX) and msvcm80.dll (Microsoft Visual Studio) [19]. For this reason and because of its design, the C language is frequently used as a template to access shared libraries.

Therefore, numerous useful functions are located in the enormous set of shared libraries that are available on a typical system. Once the correct file is located, in principle, it could be loaded and used into the R system. Some software editors provide well-documented functions in such format, allowing the end user to efficiently take advantage of them. This is the strength of such chemoinformatics libraries such as OpenEye.

8.3.2
Name Mangling and Calling Convention

As already discussed, software is an ensemble of functions that can be spread over a number of libraries. The linkage between the different parts of the software is done first at compilation time. Static libraries are linked this way once and for all. As opposed to shared libraries, their functions are to be included into the executable software. For a shared library, only the existence of the library is tested during compilation. The library itself is found and loaded during execution of the software at runtime.

The software in charge of finding, assembling, loading, and unloading libraries in these different steps is a *linker* – on UNIX system, this software is named ld. Therefore, a compiler needs to work tightly with the local linker of a platform in order to produce valid executables.

A compiler generates a lot of pieces of information in order to help the linkage of the different parts of the software. These pieces of information are mostly encoded into a new naming scheme stored in the static/shared library files. The names of the functions chosen by the programmer are transformed by the compiler to encode the additional information. This action of the compiler is called *name mangling* [20].

The primary information encoded in such a way is the *calling convention*, which is the manner in which parameters are passed to a given subroutine and how the results are returned. Other information concern data types, structures, and classes. For those languages that allow functions possessing the same name, some encoding is provided in order to distinguish between them when they are used.

In R, the most basic way to access functions requires them to use an interface (a name mangling) compatible with the C compiler or the Fortran compiler. Most modern languages possess controls in order to generate C interfaces. In the following section we will discuss only C interfaces. The name mangling of functions stored in a shared library can be checked using appropriate tools such as nm on UNIX systems. In general, if a function is declared in the original source code with a given name, this name should appear also in the library preceded with an underscore character (_). If it is not the case, it is likely that the library will not be loadable into R immediately. A small interface code is then necessary, usually written in C, called the *wrapper*. This task can be automated using tools such as SWIG (Simplified Wrapper and Interface Generator) [21]. The use of such tools is beyond the scope of this chapter.

To make things more concrete, let's take a look at OpenEye's OEChem library [4]. The toolkit can be provided as archive files (`.a`) that are static library or as shared library files (`.so`). An archive is simply a concatenation of static libraries. They can be converted into a shared library using the linker. For instance, on a Linux terminal, the command should be

```
    ld -shared -o liboechem.so -whole-archive
$OE_DIR/toolkits/lib/liboechem.a
```

However, whether the library is static or dynamic it is possible to look at the objects collected into the libraries using the following command on a Linux shell:

```
    nm $OE_LIB/liboechem*.so
```

In the huge amount of lines that are produced, one can find some references to classes (such as OEMolBase) and function names as, for instance, the SMILES parsing function, OEParseSmiles:

```
00000000001cac20 T _ZN6OEChem13OEParseSmilesERNS_9OEMolBaseEPKcbb
```

These are examples of name mangling. The first part of the line, including the single letter code, gives the location of the function.

8.3.3
dyn.load and dyn.unload

In order to use a function, the shared library containing it must be loaded in memory first and when it is no more needed, the memory should be released. This is the use of those two commands `dyn.load` and `dyn.unload`, respectively.

```
    dyn.load(x, local = TRUE, now = TRUE, . . . )
    dyn.unload(x)
```

The parameter `local` makes the symbols of the dynamic library local: other libraries do not have access to these symbols. This should stay to its default value TRUE. The parameter `now` forces the loaded library to check for resolving all symbols: the library checks for its own integrity. A value TRUE is used for debugging while for efficiency, it should be set to FALSE.

Another effect of these commands is to execute functions called `R_init_`*lib* and `R_unload_`*lib*, respectively, if these functions exist; *lib* is the name of the library file without the extension. This allows initialization of variables, allocation and releasing of memory, and so on.

These commands allow only loading shared libraries (`.so` and `.dll` files).

8.3.4
.C and .Fortran

Once a library is loaded, it is possible to call for an individual function using either the command `.C`, for functions compliant to C style, or `.Fortran`, for functions compiled using a Fortran compiler. Two other methods, `.Call` and `.External`, are provided, which are designed for interfacing with C functions unable to manipulate R internal objects, that is data structures used by R. These functions differ in the way the C code manages functions' arguments. The `.External` allows a special way to pass arguments to functions in the C code [22].

```
.C(name, ..., NAOK = FALSE, DUP = TRUE)
.Fortran(name, ..., NAOK = FALSE, DUP = TRUE)
.External(name,...)
.Call(name,...)
```

The first parameter, `name`, is just the name of the called foreign function or subroutine. Then values for the parameters of the function are passed in the same order as in the foreign function. The parameter `NAOK` indicates that nonnumbers (*NA*, *NaN*, and *Inf*) are allowed. `DUP` duplicates variables before passing them, in order to preserve them in case of modifications by the foreign function. This is specific to `.C` and `.Fortran`.

These calling procedures imply a number of limitations. First of all, it is not possible to use directly complex informatics structures such as objects. Besides, this implies correct interpretation of types. It is necessary to take care of type conversion when necessary, using the `mode` command to check the typing and the `as` command to force type interpretation. Table 8.1 resumes the correspondence between types in C, Fortran, and R [23].

For the C language, R provides also the `R.h`, `Rinternals.h`, and `Rdefines.h` header files to help conversion while support for complex numbers can be found in

Table 8.1 Type matching among R, C, and Fortran.

R	C	Fortran
integer	Int *	Integer
numeric	double *	double precision
	float *	real
complex	Rcomplex * typedef	double complex
	struct {double r; double i;}	
logical	Int * (FALSE=0, TRUE=1)	Integer (FALSE=0, TRUE=1)
character	char **	Character array, max size 255, size must be passed
raw	unsigned char *	Not allowed
list	SEXP *, void *	Not allowed
other	SEXP, generic pointer	Not allowed

the `Complex.h` header file. Using the header also allows using the command `.Call` (or `.External`), which simplifies the C code, the compilation step, and the R code. It is therefore recommended to use them.

It should be noted that these commands provide no means to access the result of a function. In other words, any result of a function should be in the parameter list, the type of the function, in C language, should be `void,` and only Fortran subroutines are accessible. The type `SEXP` designates a pointer in R and can be found in the R headers. It should be noted that it is possible to use generic pointers (noted by the keyword `void *`) in order to manage structures and classes when working, for instance, with object-oriented libraries.

8.3.5
Example

Let us suppose that a shared library `rchem.so` is available, providing some chemo-informatics function. A simple one could be to return an atom count parsing the SMILE of a molecule. The prototype of such function could be

```
void vAtomCount(char **smiles, int *nmol, int *count)
```

The first argument is an array of SMILES as C-style character strings, the second is the number of SMILES to process, and the last one is an array of integer containing the atom counts.

The following code exemplifies how to manipulate character strings, representing molecules in SMILES format with OpenEye from R:

```
>dyn.load('rchem.so')          Load the shared library rchem.so
>nmol<-2                        Set the number of molecules to 2
>smiles<-c("c1(cc(c(cc1)O)OC)  Store the SMILES character string
C=O","CCOc1cc(ccc1O)C=O")       for vanillin and ethyl vanillin
>counts<-.C("vAtomCount",smiles, Call the function vAtomCount
nmol,cnts= integer(nmol))$cnts
```

Note that the last parameter of the function call (`.C(...)$cnts`) is the return of the function. A variable `cnts` is created temporarily to receive the atom counts. The `$cnts` at the end of the command tells R that the value stored in this variable should be stored in the user variable `counts`.

8.3.6
Compilation

Because of the aforementioned limitations, it will often be desirable or even mandatory to write some libraries in C or Fortran in order to use them more easily in R. Such a piece of code is called a *wrapper*. It requires a compilation step that can be tedious especially when it is not very clear which libraries are to be included.

Some header adding functions and types facilitating communication with R will require options to the compiler that might depend on the local installation of R. Fortunately, R provides a tool for compilation that solves most of these difficulties.

The command, in a shell, is

```
R CMD SHLIB [options] [-o output] files
```

It can generate a shared library, given a set of object files. Additional options appended to the command line will be interpreted by the linker software (`ld` for Linux). It is also possible to use the same command to compile C, C++, Fortran, and Objective C/C++ code, and generate a shared library. It reads parameters from a file named `Makevars,` which must be created/edited in order to include nontrivial options to the preprocessor. However, in the case of complex projects, a more classical compilation protocol based on `Makefile` can be more convenient. In that case, the importance of the option `-fPIC` (or `-fpic`) should be stressed. This is a compiler option to generate position-independent code, which is mandatory for shared libraries. It allows the functions of the shared library to be stored at runtime in a definite memory address that will be accessible to multiple applications.

If compilation and no linking are required, the following command can be used instead:

```
R CMD COMPILE [options] files
```

Once compiled, the produced objects can be combined using the previous command.

8.4
Wrapping

As mentioned in the previous section, R facilities to access the functions of a library are restricted. Therefore, to satisfy such constraints or use data architectures from an object-oriented language, it is necessary to write a *wrapper*, a dedicated shared library that on one hand complies to R constraints and on the other hand communicates with the targeted libraries.

8.4.1
Why Wrapping

A number of interesting chemoinformatics toolkits (OpenEye, OpenBabel, RDKit, etc.) are written in C++, Fortran, or more unusual language. In general, it is not convenient to engineer the toolkit code, by modifying or adding functions, compliant with R constraints. If the toolkit is commercial, such operation is impossible, since the source code is typically closed. The reasonable solution is to write a

separate set of function prototypes that use C types compatible with R. At the linking stage, a shared library should be produced.

For instance, the OpenEye library provides a way to count atoms of a molecule object, which can be setup using a SMILES code. This is some kind of chemo-informatics "Hello world" example. However, such C++ construction cannot be used directly in R. So, a wrapping function is written in C++, which presents a prototype compatible with C and R in a file `atomcount.h`. The implementation of the function is written in pure C++ in a companion file `atomcount.cpp`:

`#include "atomcount.h"`	Header inclusion
`void cNbAtom(char **sml, int *nsml, int *cnt){`	The procedure `cNbAtom` requires an array of character string (`sml`) and the number of molecules in this array (`nsml`), and returns an array of integers containing the atom count for each molecule (`cnt`)
`  OEGraphMol mol;`	Variables declaration. The type
`  int i;`	OEGraphMol is specific to OpenEye
`  for(i=0;i < *nsml;i++){`	Start a loop parsing the array of SMILES sml
`    mol.Clear();`	Recycle the variable mol
`    OEParseSmiles(mol,sml[i]);`	
`    cnt[i]=mol.NumAtoms();`	
`  };`	
`};`	

The header uses `atomcount.h` that is

`#ifndef ATOMCOUNT_H`	Preprocessor instruction to avoid multiple
`#define ATOMCOUNT_H`	inclusion of the header
`#include "openeye.h"`	Loading OpenEye headers
`#include "oechem.h"`	
`#include "oesystem.h"`	
`using namespace OEChem;`	Setting the name spaces
`using namespace OESystem;`	
`using namespace std;`	
`#ifdef _LP64`	Preprocessor instructions defining the keyword
`#    define EXPORTCALL`	EXPORTCALL
`#else`	
`#    define EXPORTCALL __attribute__((stdcall))`	
`#endif`	
`extern "C"{`	Open the `extern C` context concerning linking conventions
`void cNbAtom(char **sml, int *nsml, int *cnt);`	Prototype of the function `cNbAtom`. The interface is fully compatible with C types and R
`}`	Close the `extern C` context

```
#endif                              Close preprocessor condition
```

The instruction `#ifdef _LP64` (`#ifdef _WIN64` for windows) is used to manage the `__attribute__((stdcall))`, controlling the way arguments are passed to a function. For 64-bit systems, there is one unified way to do it and for older systems, it must be specified that the C standard should be used. The second special instruction is `Extern "C"{}`. This is a C++ instruction specifying a linking convention: how to access memory, integrated type formats, and so on.

The compilation command line is

```
R CMD SHLIB atomcount.cpp -L${OEDIR}/toolkits/lib -loechem
-loesystem -loeplatform -lz -lpthread -lm
```

This command line will remain the same for all the following examples using OpenEye toolkit. Note the presence of −L and −l switches, indicating where OpenEye shared libraries are and to which ones should the software be linked. The OpenEye toolkit headers should be provided in a companion file used by the compilation command, the `Makevars` [22]. An example of such a file could be

`INCDIR = $(OEDIR)/toolkits/include` `INCS =-I$(INCDIR)`	Define the variable `INCDIR` and `INCS` indicating where OpenEye's header files are located
`PKG_CXXFLAGS =-m64-W-Wall -Wconversion-O3-fomit-frame-pointer-ffast-math $(INCS)`	Define the variable `PKG_CXXFLAGS`. The value of this variable is concatenated to any linking command. Here specifications about the 64-bit architecture of the machine, the levels of warnings and optimization, and the precise location of header files are given

The function can be used in R as described previously. An example of R session is

`> dyn.load("atomcount.so")`	Load the newly created shared library
`> smle<-c("c1(cc(c(cc1)O)OC)C=O","CCOc1cc(ccc1O)C=O")`	Store vanillin and ethyl vanillin in SMILES format as a character vector
`> l<-length(smle)`	
`> out<-.C('cNbAtom',as.character(smle),` `l, num=integer(l))$num`	Call the function `cNbAtom` and store in the variable out an integer vector containing the atom count of the input molecules

The notion of wrapper, therefore, defines two coding environments: the *R-side*, the code written in R, and the *C-side*, the code used to present a C interface to R at the level of the shared library.

8.4.2
Using R Internals

As mentioned before, R objects in use during a session can be accessed from the C-side. These objects are called the *R internals* [23]. The macro and the functions to communicate with R internals are defined in the `Rinternals.h` and `Rdefines.h` headers, defining the type `SEXP`. These headers should be included in the C-side. In that case, the `.Call` command should be used on the R-side. The previous example procedure counting atoms in a SMILES can be rewritten using this method.

`#ifndef ATOMCOUNT_H` `#define ATOMCOUNT_H`	Preprocessor instruction to avoid multiple header inclusion

`#include "openeye.h"`
`#include "oechem.h"`
`#include "oesystem.h"`
`#include "R.h"`
`#include "Rinternals.h"`
`#include "Rdefines.h"`
	Include headers. Note the presence of R-specific headers `R.h`, `Rinternals.h`, and `Rdefines.h`. The order of the headers should not matter but in practice R headers might cause compilation errors if not put at the end of the include lists

`using namespace OEChem;`
`using namespace OESystem;`
`using namespace std;`
	Define the name spaces

`#ifdef _LP64`
`#    define EXPORTCALL`
`#else`
`#    define EXPORTCALL`
`__attribute__((stdcall))`
`#endif`
	Preprocessor instruction to define the keyword `EXPORTCALL`

`extern "C"`	Open the `Extern C` context
`{`	
`  SEXP rNbAtom(SEXP sml, SEXP` `nsml, SEXP cnt);`	Prototype of a function counting the atoms of a set of SMILES encoded compounds. The function uses systematically `SEXP` type
`}`	Open the `Extern C` context
`#endif`	Close preprocessor conditional

In the include sections, it is recommended the R headers to be at the bottom of the list because some inclusions might overload some identifiers that might break on compilation. The order of inclusions usually does not matter except when working with large and complex toolkits: in that case, it might be crucial. Another concern is that R headers might include C headers, which might conflict with C++ ones. In that case, it may be necessary to define the macro `NO_C_HEADERS`, adding a line [22]

```
#define NO_C_HEADERS
```

The second point to note is that the function prototypes use `SEXP` types for all arguments and as return value.

The function implementation is

Code	Description
`#include "atomcount.h"`	Include the header atomcount.h
`SEXP rNbAtom(SEXP sml, SEXP nsml, SEXP cnt){`	Define the interface of the function rNbAtom

`  char *Psml;`
`  int *Pcnt;`
`  int Insml;`
`  int i;`
`  OEGraphMol mol;`
	Define variables: `Psml` is a character string, `Pcnt` is an array of atom count, `Insml` is the count of molecules in the character string, `mol` is an OpenEye structure for a molecule, and `i` is a counter

`  Insml=INTEGER_VALUE(nsml);` `  Pcnt=INTEGER_POINTER(cnt);`	Convert the input pointer parameter of the function of type `SEXP` to conventional C++ types
`  for(i=0;i<Insml;i++){`	Loop over all compounds in the array of SMILES
`    mol.Clear();`	Recycle the object `mol`
`    Psml=(char *)CHAR(STRING_ELT(sml,i));`	Convert one element of the array of SMILES to a conventional character string
`    OEParseSmiles(mol,Psml[i]);` `    Pcnt[i]=mol.NumAtoms();`	Parse the SMILES to a molecule object and store the atom count
`  };`	End of the loop
`  return(R_NilValue);`	Return the value nil to R
`};`	End of the function

The first example of R and C communication is the instruction `Insml=INTEGER_VALUE(nsml)`: a macro coerces the value of the parameter `nsml` to an integer, which is then copied into a standard integer variable. The instruction `Pcnt=INTEGER_POINTER(cnt)` coerces the parameter `cnt` to an array of integers – actually a pointer to the first element of such an array. The array is created on the R-side.

The management of the character vector `sml` is more delicate. An element of the vector is accessed using the macro `STRING_ELT`, then it is interpreted as a pointer to a C character string. Thus, the meaning of the instruction `Psml=(char *)CHAR (STRING_ELT(sml,i))`.

Finally, the instruction `return(R_NilValue)` returns a valid `SEXP` corresponding to the value `NULL` on the R-side. The return value as an `SEXP` is necessary.

Whatever command is used – .C, .Call, or others – C variables are not allowed to survive the conclusion of the procedure. Allocating memory space and keeping control on it during subsequent calls to library functions requires dedicated macros.

8.4.3
How to Keep an SEXP Alive

This problem occurs at least on two occasions: when creating an R object in the C-side or when allocating memory for a C++ object. Indeed, it is inefficient to allocate, initialize,

and deallocate objects too often. To escape this problem, R provides two tools: a protection mechanism from the garbage collector and R external pointer references.

The first problem is that, at the end of a subroutine, any unused memory space allocated in the C-side is collected by the R garbage collector and destroyed. Hopefully, the instruction PROTECT provided by R can be used. Any pointer in a PROTECT macro tells R that the object pointed to is in use, and shall not be destroyed. Before the end of the function, the command UNPROTECT(n) must be used in order to release the last n protected objects. The number of PROTECT instructions should match the overall number of released objects using UNPROTECT. Failing to do so will cause R to crash.

In order for R to manage a pointer to a memory space allocated in the C-side, it should be coerced to an *external pointer reference*. This kind of R object is a SEXP on the C-side and is produced using the function SEXP R_MakeExternalPtr(void *p, SEXP tag, SEXP prot). The parameter p is a general C pointer, which should be kept by R. Basically, the return value SEXP is the pointer p. However, a few more bits of information are added. First, a tag can be attached, which is an arbitrary information, such as a type information accessible from the R-side. Second, prot is an R structure, which is required to stay alive as long as the pointer. It can be useful if the pointer points to an R object created in the C-side, to ensure that the pointer will continue to point to a valid structure.

The pointer part of an external pointer reference can be accessed using void *R_ExternalPtrAddr(SEXP s). This function returns the address as a generic C pointer.

With these tools, the previous example can be rewritten in a set of functions that are flattening the C++ architecture of the library.

```
#ifndef ATOMCOUNT_H              Preprocessor instruction to avoid
#define ATOMCOUNT_H              multiple header inclusion
#include "openeye.h"             Include lists
#include "oechem.h"
#include "oesystem.h"
#include "R.h"
#include "Rinternals.h"
#include "Rdefines.h"
using namespace OEChem;          Define namespaces
using namespace OESystem;
using namespace std;
#ifdef _LP64                     Define the EXPORTCALL keyword
#    define EXPORTCALL
#else
#    define EXPORTCALL
__attribute__((stdcall))
#endif
extern "C"                       Open the extern C context
{
```

```SEXP MolCreate();```	Prototype of specialized functions to
```SEXP MolFree(SEXP extmol);```	allocate/deallocate (```MolCreate/```
```SEXP MolClear(SEXP extmol);```	```MolFree```) an OpenEye molecule
```SEXP MolSMILESSet(SEXP extmol,```	structure in memory, recycle the
```SEXP smle);```	structure (```MolClear```), load a molecule
```SEXP MolAtomCount(SEXP```	from its SMILES into the structure
```extmol);```	(```MolSMILESSet```), and count the number
	of atoms in it (```MolAtomCount```)
```}```	Close the ```extern C``` context
```#endif```	Close the preprocessor conditional

In the header, all atomic instructions (creation/deletion of molecule objects, cleaning, SMILES interpretation, and atom counting) are now separate functions. If a function uses internally a molecule object created previously, it requires a parameter SEXP, here called ```extmol```, which should contain a valid address.

The implementation is then quite straightforward.

```#include "atomcount.h"```	Include the header file ```atomcount.h```
```SEXP MolCreate(){```	Declaration of the function
	```MolCreate``` returning a SEXP
```OEGraphMol *mol;```	Declaration of variables. The variable
```SEXP out;```	```mol``` is a pointer to an OpenEye
	molecule object
```mol=new OEGraphMol;```	Create an OpenEye molecule object.
	The pointer ```mol``` points to the address
	of this object
```PROTECT(```	Copy the pointer ```mol``` to an R pointer
```out=R_MakeExternalPtr((void*)mol,```	and protect it
```R_NilValue, R_NilValue));```	
```UNPROTECT(1);```	Release the protected object
```return(out);```	Return the R pointer and end the
```};```	function

Following, this procedure creates a molecule object, here an ```OEGraphMol```. The pointer to this object is copied into an external pointer reference, ```out```. This pointer is ```PROTECTED```, so that it remains valid even after the function is ended.

```SEXP MolFree(SEXP extmol){```	Declaration of the function
	```MolFree``` returning a SEXP
```OEGraphMol *mol;```	The variable ```mol``` is a pointer to
	an OpenEye molecule object
```mol=```	Interpret the R pointer as
```(OEGraphMol*)R_ExternalPtrAddr(extmol);```	pointer to an OpenEye
	molecule object and then copy
	its value to mol

```delete mol;```	Deallocate the memory
```return(R_NilValue);```	Return a nil value and end the
```};```	function

This function is used to delete the molecule object. The pointer to the molecule is first read and then used to free the memory.

```SEXP MolClear(SEXP extmol){```	Declaration of the function ```MolClear``` returning a ```SEXP```
``` OEGraphMol *mol;```	The variable ```mol``` is a pointer to an OpenEye molecule object
``` mol= (OEGraphMol*)R_ExternalPtrAddr(extmol);```	Interpret the R pointer as pointer to an OpenEye molecule object and then copy its value to ```mol```
``` mol->Clear();```	Recycle the OpenEye molecule object
``` return(R_NilValue);```	Return a nil value and end the
```};```	function

The ```MolClear``` function reads the address to the molecule object and then uses a C++ semantic to use the ```OEGraphMol``` method to reset its content.

```SEXP MolSMILESSet(SEXP extmol, SEXP smle) {```	Declaration of the function ```MolSMILES```Set returning a ```SEXP```. Parameters are the pointer to the OpenEye molecule object (```extmol```) to which the SMILES string (```smle```) should be loaded
``` OEGraphMol *mol;```	The variable ```mol``` is a pointer to an OpenEye molecule object.
``` char* Psml;```	The variable ```Psml``` is a character string
``` Psml=(char *)CHAR(STRING_ELT(smle,0));```	Interpret the R pointer as a character string and copy its pointer to ```Psml```
``` mol= (OEGraphMol*)R_ExternalPtrAddr(extmol);```	Interpret the R pointer as pointer to an OpenEye molecule object and then copy its value to ```mol```
``` OEParseSmiles(*mol,Psml);```	Parse the character string and set the OpenEye molecule object to the chemical structure coded by the SMILES

`return(R_NilValue);` `};`	Return a nil value and end the function

The `MolSMILESSet` function uses the pointer to a molecule object and an R character vector provided as input. The object pointed to is used to store the molecule resulting from the parsing as a SMILES, the first string in the character vector.

`SEXP MolAtomCount(SEXP extmol){`	Declaration of the function `MolSMILESSet` returning a SEXP. The parameter is the pointer to the OpenEye molecule object (`extmol`) to which the SMILES string (`smle`) should be loaded
`OEGraphMol *mol;`	The variable `mol` is a pointer to an OpenEye molecule object.
`SEXP out;`	
`int *pout;`	The variable `pout` is a pointer to an integer and `out` is its R pointer copy
`PROTECT(out=NEW_INTEGER(1));`	Allocate memory for an integer and protect the pointer pointing to it
`mol=` `(OEGraphMol*)R_ExternalPtrAddr(extmol);`	Interpret the R pointer as pointer to an OpenEye molecule object and then copy its value to `mol`
`pout=INTEGER_POINTER(out);`	Copy the address of the new integer
`pout[0]=mol->NumAtoms();`	Compute the number of atoms and store the result at the allocated address
`UNPROTECT(1);`	Release the protected objects
`return(out);` `};`	Return the memory address containing the atom count

This implementation of `MolAtomCounts` reads the address of the molecule object. Then it creates an R integer list of 1 element. Since this R object shall be returned by the function, it is PROTECTED. A pointer to the first element of this list is set and can be used in standard C++ semantic to store the result of the `NumAtoms()` method of the `OEGraphMol` object.

The method proposed already gives access to methods of the objects, but it loses the object architecture of the object-oriented library: each method of an object needs a function having as argument a valid pointer to that object.

8.4.4
Binding to C/C + + Libraries

Some recent efforts have been made to facilitate the access to exposed functions and object-oriented interfaces in libraries. A major advance in this direction is the package rcpp [24], which facilitates the use and the access to C/C++ functions and classes in R.

8.5
Java Archives

This section describes the package rjava, which facilitates the integration of java modules into R. The integration of chemoinformatics module is already exemplified in Java with the rcdk package, which allows using some elements of the CDK inside R.

8.5.1
The Package rJava

rJava (see description in Ref. [25]) is a package that implements a low-level R interface to Java Virtual Machine (JVM) via Java Native Interface (JNI). rJava provides functions that are rather similar to the .C/.Call C interface and allows creating objects, calling methods, and accessing fields of Java objects from R. rJava also includes JRI (Java to R Interface), which provides means to call R functions from Java.

Using the package rJava, the JVM can be started from R by calling the function `.jinit()`.

```
library(rJava)
.jinit()
```

New Java objects can be created from R with the help of function `.jnew(class, ...)`.

```
s <- .jnew("java/lang/String", "Calling Java from R")
```

In this example, the Java String object with the value "Calling Java from R" has been created. This is completely equivalent to the following line of Java code:

```
s = new java.lang.String("Calling Java from R");
```

Note, however, two important differences. First, slashes are used inside the `.jnew` function instead of dots to specify the class name, as required by JNI. Second, the full class name should be specified in the `.jnew` function.

The value of the string variable s can be retrieved with the help of the `.jstrVal` function, which automatically calls `toString()` method and returns a string. The outputs of the commands are given starting with `[1]`

Table 8.2 Type matching among `.jcall` and Java.

Specification in `.jcall`	Java type
I	Integer
D	double (numeric in R)
J	long (numeric in R; to be marked by function `.jlong`)
F	float (numeric in R; to be marked by function `.jfloat`)
V	Void
Z	Boolean
C	char (integer in R)
B	byte (raw in R)
L<class>;	Java object of the class <class>, for example, Ljava/lang/Object;
S	Ljava/lang/String;
[type	Array of objects of type <type>

```
.jstrVal(s)
[1] "Calling Java from R"
```

Methods of the corresponding Java classes can be called using the `.jcall` function. As an example of a simple method, consider `length`, which returns the length of string.

```
.jcall(s, "I", "length")
[1] 19
```

This is equivalent to `s.length()` in Java but involves an additional parameter specifying the type of the returned value, as required by JNI. Table 8.2 specifies the required JNI types.

The same value can be obtained also via the R object connected to the java object, in this example, s. Note the use of the sign $ instead of the dot in Java.

```
s$length()
[1] 19
```

The next example illustrates passing parameters to methods of Java objects.

```
.jcall(s, "I", "indexOf", "Java")
[1] 8
```

This is equivalent to the instruction `s.indexOf("Java")` in Java. In this case the string parameter "Java" is automatically converted to java/lang/String object and passed to Java. The same result could also be obtained in R.

```
s$indexOf("Java")
[1] 8
```

The R command `.jcall` is needed at least once to create a Java object. An R object is created on this occasion connected to the Java object. In general the object can be manipulated as a standard R object mapping the Java object. In any case, the Java object can be accessed through the R command `.jcall`.

So, using this rather simple interface, R can efficiently be bridged to numerous libraries written in Java, including important libraries in the field of chemoinformatics, such as CDK.

8.5.2
The Package rcdk

rcdk (see the manual in Ref. [26] and tutorial) is an R package that allows the user to access functionality in the CDK [5], a Java library for chemoinformatics. This package allows the user, for example, to load molecules and calculate fingerprints and various molecular descriptors for them. In addition it allows the user to view chemical structures in 2D. This package keeps a part of its functionality in the auxiliary packages `rcdklibs` and `fingerprint`. It is based on the use of the rJava package in order to access the CDK Java library. This section describes some of the most basic facilities provided by rcdk.

The package is loaded into the R environment as usual:

```
> library(rcdk)
```

This loads automatically several additional packages: `rJava`, `rcdklibs`, `fingerprint`, `png`, and `iterators`.

The rcdk package provides two basic ways of loading chemical structures. First, they can be read from files in which they can be stored in various formats supported by CDK. Second, the structures can be generated from SMILES strings.

The following line loads a database of molecules from the file `vanilla.smi` into the array of molecules `mols`.

```
> mols <- load.molecules('vanilla.smi')
```

In this case all the loaded molecules are kept in computer memory. Individual molecules can be accessed as elements of this array.

```
> mol <- mols[[1]]
```

In order to work with huge databases, they can be read iteratively, one structure at a time, using the iterators mechanism.

```
> iter <-
iload.molecules('vanilla.smi',
type='smi')
```

Initialize the iterator `iter` to the first molecule of the file `vanilla.smi`

```
> while(hasNext(iter)) {
```
Start a loop. The command `hasNext` is false if the iterator points to the last molecule of the file

```
+   mol <- nextElem(iter)
+   ...
+ }
```
The command `nextElem` shifts the iterator to the next molecule in the file and sends it to the container `mol`

Alternatively, molecule objects can be obtained by parsing SMILES strings with the command `parse.smiles`.

```
> smiles <- c("c1(cc(c(cc1)O)OC)C=O","CCOc1cc(ccc1O)C=O")
> mols <- parse.smiles(smiles)
```

Chemical structures can be visualized in the form of tables.

```
> view.molecule.2d(mols)
```

In order to manipulate molecules, list of atoms and bonds can be extracted from them for further processing.

```
> mol <- parse.smiles(c('c1ccccc1'))[[1]]
> atoms <- get.atoms(mol)
> bonds <- get.bonds(mol)
```

CDK can be used for calculating a variety of molecular descriptors belonging to the following categories: topological, constitutional, geometric, electronic, protein, and hybrid. The list of available descriptor categories can be retrieved as follows.

```
> dc <- get.desc.categories()
> dc
[1] "electronic" "protein" "topological" "geometrical"
[5] "constitutional" "hybrid"
```

The names of descriptors (actually, the names of the corresponding Java classes that calculate them) belonging to the first category (e.g., the "electronic" descriptors) can be obtained using the following lines.

```
> dn <- get.desc.names(dc[1])
> dn
[1] "org.openscience.cdk.qsar.descriptors.molecular.
   APolDescriptor"
[2] "org.openscience.cdk.qsar.descriptors.molecular.
   BPolDescriptor"
[3] "org.openscience.cdk.qsar.descriptors.molecular.
   CPSADescriptor"
```

```
[4] "org.openscience.cdk.qsar.descriptors.molecular.
    HBondAcceptorCountDescriptor"
[5] "org.openscience.cdk.qsar.descriptors.molecular.
    HBondDonorCountDescriptor"
[6] "org.openscience.cdk.qsar.descriptors.molecular.
    TPSADescriptor"
```

This means that CDK can compute the following six "electronic" descriptors: APol, BPol, CPSA, HBondAcceptorCount, HBondDonorCount, and TPSA. So, each molecular descriptor is uniquely identified by its name. A set of descriptors can be evaluated for a molecule using the function `eval.desc`, which takes on as its input parameters the molecule and the list of full names of the required descriptors. Continuing the previous example, it is possible to compute the electronic descriptors listed in the variable `dn` for the molecule `mol` using the command

```
> allDescs <- eval.desc(mol, dn)
```

The function `eval.desc` returns a data frame with the descriptors as columns and the molecules as rows. The values of the descriptors can be used to develop QSAR/QSPR models in R.

Another very powerful feature of CDK is the ability to calculate several standard molecular fingerprints: Standard, Extended, EState, and MACCS (Table 8.3).

Molecular fingerprints can be computed by calling the function `get.fingerprint` with two input parameters specifying the molecule and the type of fingerprint.

The following example continues the hierarchical clustering example of compounds possessing a vanilla odor, demonstrating some functionality of R and CDK.

`> library(rcdk)`	Load the `rcdk` library
`> mols <-` `load.molecules('vanilla.smi')`	Load SMILES encoded compounds from the file `vanilla.smi`
`> nmes <-` `readLines('vanilla.txt')`	The names of the compounds are loaded into character vector `nmes` from the file `vanilla.txt`
`> fps<-lapply(mols,get.fingerprint,` `type="extended")`	Compute and store in the variable `fps`, extended fingerprints of depth 6 bits and length 1024 bits
`> fpm<-fp.to.matrix(fps)`	Convert fingerprint to matrix representation

`>h<-pvclus(t(fpm), method.dist=`
`"uncentered", method.hclus=`
`"mcquitty", nboot=100)`
	Compute the hierarchical clustering using bootstrapping *p*-value estimates

`>plot(h,label=nmes,print.num=FALSE,` `cex=0.6,cex.pv=0.3)`	Plot the dendrogram of the clustering

In this case, the resulting cluster dendrogram looks as.

Table 8.3 Fingerprints available in the rcdk package.

Fingerprint	Description	Length in bits
Standard	Path-based, hashed fingerprint	1024
Extended	Like the Standard one, but takes additionally into account rings in molecules	1024
EState	Structural key type fingerprint that checks for the presence or absence of 79 EState substructures	79
MACCS	MDL standard set of structural keys	166
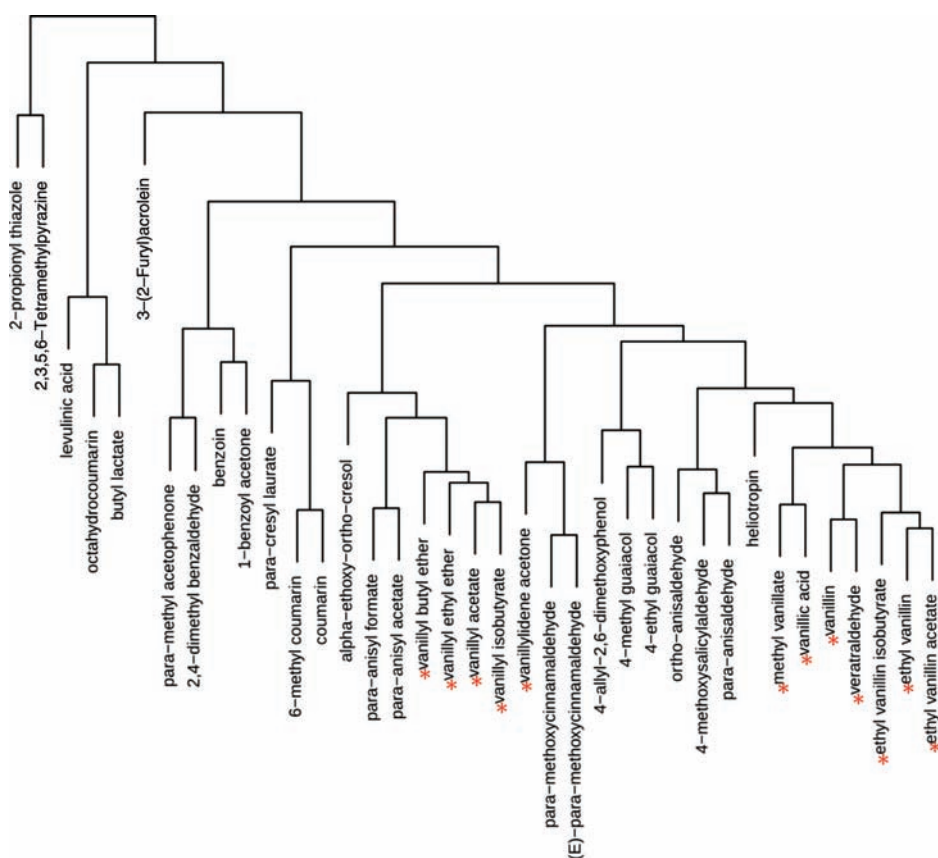
Figure 8.6 Dendrogram of organic compounds, all possessing a vanilla odor. The dendrogram is based on the chemical structures represented by CDK extended fingerprint. Hierarchical clustering using bootstrapping *p*-value estimates. Dissimilarity measure: uncentered sample correlation; agglomerative method: McQuitty. Fingerprint: CDK extended fingerprint (depth 6 bits and length 1024 bits). Vanillin substructures are marked.

The resulting dendrogram share some characteristics with the one obtained with ISIDA descriptors, which is conceptually close. For instance, both dendrograms agree to move away the 2-propyl thiazole (**28**), the butyl lactate (**34**), and the benzoin (**1**) clusters. However, it is less sensitive to multicomponent structures, since it recognized ethyl vanillin isobutyrate (**11**) as similar to ethyl vanillin. The clusters are also much more homogeneous and actually are missing some interesting similarities as illustrated by the position of heliotropin (**25**) and 4-methoxysalicylaldehyde (**4**). In general those two pictures (Figure 8.6) of the chemical space tend to disagree about objects of moderate similarity.

8.6
Conclusions

In this chapter we have described the ways of using popular chemoinformatics tools from R: through system calls, shared library calls, and calls to Java VM via JNI. Being a free and very powerful system for data processing and statistical calculation, R contains numerous packages implementing both the classical and the most modern machine learning methods and data visualization tools. Of course, this chapter is not comprehensive: there exists a variety of powerful chemoinformatics software and libraries implementing different ways to handle information in chemistry, including 2D and 3D information concerning molecular structures, calculation of various original descriptors, fingerprints, similarity measures, and so on. Integration of R with chemoinformatics tools using the methods described in this chapter can make R a very powerful and flexible platform for processing chemical information. It shall be noted also that R is vastly adopted by bioinformaticians and, of course, statisticians. Using R, therefore, simplifies scientific exchanges between these communities.

References

1 Todeschini, R. and Consonni, V. (2009) *Molecular Descriptors for Chemoinformatics*, Wiley-VCH Verlag GmbH, Weinheim.

2 Chemical Computing Group (2010) Molecular Operating Environment (MOE).

3 Landrum, G. (2011) RDKit: Open-Source Cheminformatics. http://www.rdkit.org.

4 OpenEye Scientific Software (2011) OEChem, Santa Fe, NM.

5 Steinbeck, C. *et al.* (2003) The chemistry development kit (CDK): an open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, **43** (2), 493–500.

6 ChemAxon (2011) JChem v5.9. http://www.chemaxon.com.

7 Team, R.D.C. (2011) *R: A Language and Environment for Statistical Computing*, Foundation for Statistical Computing, Vienna, Austria.

8 Varnek, A. and Baskin, I.I. (2011) Chemoinformatics as a theoretical chemistry discipline. *Molecular Informatics*, **30** (1), 20–32.

9 Engel, T. (2006) Basic overview of chemoinformatics. *Journal of Chemical Information and Modeling*, **46** (6), 2267–2277.

10 Gasteiger, J. and Engel, T. (2003) *Chemoinformatics: A Textbook*, Wiley-VCH Verlag GmbH, Weinheim.

11 OpenEye Scientific Software (2011) nam2mol, Santa Fe, NM.

12 OpenEye Scientific Software (2011) Filter, Santa Fe, NM.

13 Weininger, D. (1988) SMILES, a chemical language and information system: 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, **28** (1), 31–36.

14 Baskin, I. and Varnek, A. (2008) Fragment descriptors in SAR/QSAR/QSPR studies, molecular similarity analysis and in virtual screening, in *Chemoinformatics Approaches to Virtual Screening* (eds A. Varnek and A. Tropsha), RSC Publishing.

15 Varnek, A. *et al.* (2005) Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *Journal of Computer-Aided Molecular Design*, **19** (9–10), 693–703.

16 Suzuki, R. and Shimodaira, H. (2004) An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: how accurate are these clusters? in *The Fifteenth International Conference on Genome Informatics*, Pacifico Convention Plaza Yokohama, Japan, p. P034.

17 Ryder, B.G., Soffa, M.L., and Burnett, M. (2005) The impact of software engineering research of modern programming languages. *ACM Transactions on Software Engineering and Methodology*, **14** (4), 431–477.

18 Bergin, T.J. and Gibson, R.G. (1996) *History of Programming Language II*, ACM Press, New York (US).

19 Plauger, P.J. (1991) *The Standard C Library*, Prentice Hall.

20 Argner, F. (2011) Calling conventions for different C++ compilers and operating systems. http://www.argner.org.

21 Beazley, D.M. (2003) Automated scientific software scripting with SWIG. *Future Generation Computer Systems*, **19** (5), 599–609.

22 R.D.C. Team (2011) Writing R Extensions.

23 R.D.C. Team (2011) R Internals.

24 Eddelbuettel, D. and François, R. (2011) Rcpp: seamless R and C++ integration. *Journal of Statistical Software*, **40** (8), 1–18.

25 Urbanek, S. (2011) *rJava: Low-Level R to Java Interface*, R Foundation for Statistical Computing, Vienna, Austria.

26 Guha, R. (2007) Chemical informatics functionality in R. *Journal of Statistical Software*, **18** (6), 1–18.

Part Three
Applications to Polypharmacology

9
Content Development Strategies for the Successful Implementation of Data Mining Technologies

Jordi Quintana, Antoni Valencia, and Josep Prous Jr.

9.1
Introduction

Historically, biomedical research organizations have been able to face the challenge of how to discover and deliver new drugs efficiently. However, in recent years, while the cost of research and development has increased steadily, the number of new molecular entities entering clinical practice has remained stable or even decreased [1,2].

This has occurred in spite of the advent of new technologies such as QSAR, high-throughput screening (HTS), combinatorial chemistry, and functional genomics [3]. Combinatorial chemistry and HTS deliver the ability to synthesize and screen an ever-growing number of compounds simultaneously against an array of pharmacological targets. However, the difficulty in building effective and focused libraries and in filtering or designing libraries with the desired properties has limited the success ratio of these technologies so far [4]. QSAR models provide a way of obtaining an accurate description of useful properties that could be used for better drug design and as a means of filtering and designing combinatorial libraries. However, the lack of available large-scale numerical data sets has led to models with a reduced number of parameters and limited applicability domain. These in turn can only successfully describe a smaller set of compounds belonging to the same family, thus limiting the potential of QSAR [5].

These considerations explain the growing interest in the application of data mining technologies in the elucidation of the preclinical profile of small molecules. Data mining is defined as the automatic extraction of useful, often previously unknown information from large databases or data sets using advanced search techniques and algorithms to discover patterns and correlations in large preexisting databases. Therefore, the accuracy of a specific data mining application is closely related with the robustness of the mathematical model and the quality of the data set used to train or create that model [6]. The development and implementation of a robust content development strategy, including precise database design, accurate selection of information sources, realistic data curation, and update planning, will be key for the success of these techniques in solving challenges in biomedical research. This chapter will review in detail these considerations and explore two successful

case studies of databases created for the biomedical community with the potential application of data mining techniques in mind.

9.2
Knowledge Challenges in Drug Discovery

Drug discovery is currently one of the most knowledge-intensive disciplines in the world. The concept of a knowledge pyramid is a fitting example of this paradigm (Figure 9.1). The pyramid begins with a base of raw data, which are then transformed into structured information and turned into knowledge before the appropriate action is taken. It is relatively easy to create large amounts of data if one considers high-throughput chemistry, pharmacology, and genomics technologies, but the real value in the drug discovery chain comes from the interpretation of this knowledge [7,8].

One could state that the success of information products in the biomedical field is directly related to the quality and uniqueness of their content. Furthermore, access to content in the pharmaceutical field is shifting from a position where the competitive advantage no longer comes from the integration of information, but from the stickiness of information, since all professionals have to see content as part of their responsibilities. Information is to be considered a leading force in this domain as can be seen by the recent alliances between technology/bioinformatics organizations and drug research institutions.

To achieve this essential starting point one would need to combine in-house capabilities with the input of leading specialists in the scientific and medical fields worldwide to ensure the completeness and quality of any data training set.

When analyzing the requirements for the development of a quality training set, a number of considerations commonly apply. In the biomedical field, discrete compounds represent the essence of a specific R&D project leading to better and safer drugs.

Figure 9.1 The knowledge pyramid concept.

Therefore, the availability of chemical structure information, development phase data indicating where the compound stands in the race to the clinical, physicochemical properties or licensing availability, is one of the pillars of the strategy. However, this information would not be of any value without the knowledge of the interaction of these compounds with a variety of molecular targets, at least from a qualitative viewpoint and, ideally, from a quantitative perspective. The application of network theory to the analysis of biomedical data reveals an unexpectedly complex picture of drug–target interactions and that the topology of drug–target networks depends implicitly on data completeness, drug properties, and target families [9]. Due to the recent development of multiple therapeutic targets and pathway modulation opportunities, it is important to update the compound information with new biological data on a regular basis [10]. In addition, the access to specific ADMET (absorption, distribution, metabolism, elimination, and toxicology) information enriches any collection of compounds, if we keep in mind that a large number of drug R&D projects still fail to progress due to their limitations when they are applied to living organisms [11,12]. Finally, the intellectual property context plays an essential role when moving a project forward as does the scientific literature, which should be closely monitored for any novelty potentially affecting the collection of data on the compound [13].

In the case studies given in Section 9.3, we analyze successful developments in the biomedical knowledge field and the rationale behind their creation.

9.3
Case Studies

9.3.1
Thomson Reuters Integrity

Thomson Reuters Integrity[SM] (http://thomsonreuters.com/products_services/science/science_products/a-z/integrity/), originally launched as Prous Science Integrity in 2001, is a database system widely used by researchers in the pharmaceutical industry and government institutions to initiate and validate new drug discovery hypotheses (Figure 9.2). The system is organized into a number of interconnected knowledge areas that provide key up-to-date information on crucial facts for end users. The system can be used through Internet access and the provider also offers data sets for installation in corporate intranets. Other chemical databases are described elsewhere in this book; therefore, this section will focus on the fundamentals and a description of the Thomson Reuters Integrity drug discovery and development portal.

In 1958, Dr J.R. Prous, President and founder of the scientific publishing company Prous Science, launched its first publication, *Drugs of Today*. The aim of the company was to offer innovative information, communication, and educational products and services to the scientific and medical community by combining value-added contents with the most advanced information technology platforms. Prous Science pioneered in different areas of information technologies that made the company a leader in different market segments. For example, Prous Science was

Figure 9.2 Thomson Reuters Integrity home page.

one of the first companies to use CD-ROM technology to disseminate drug and medical information (early 1990s) and to incorporate Internet as a keystone of the company strategy in order to offer educational and communications products and services. Over the years, the company established different collaborations with a series of key players in the field of information distribution, including MDL Information Systems (now Accelrys), Dialog, Data Star, and CAS.

Based on this accumulated knowledge, Prous Science developed the Integrity portal to provide a unique knowledge solution designed to empower discovery and development activities. The portal integrates biological, chemical, and pharmacological data, which as of March 2013, covered the following:

- More than 395 000 drugs and biologics (93% with chemical structure) with demonstrated biological activity
- Target-based approaches to disease diagnosis and therapeutic intervention for more than 2200 targets
- More than 22 000 genes with documented disease associations that are potential targets for drug discovery

- More than 19 000 biomarker records with more than 750 000 known uses
- More than 109 000 organic synthesis intermediates from more than 24 000 synthesis schemes
- More than 1 290 000 numerical values from experimental pharmacology studies delineating drug–receptor and enzyme–target cell interactions
- Information on more than 8700 validated *in vivo* preclinical models of human diseases, toxicity, or target efficacy studies
- More than 539 000 numerical values on pharmacokinetics/metabolism with data on patent compounds and active metabolites
- Comprehensive information on more than 193 000 references to clinical studies of compounds currently under study for use in humans
- A background reference to more than 138 disease entities with full color multimedia illustrations
- Information for at least 9100 organizations active in the drug discovery and development fields
- More than 1 555 000 references to current literature, abstracts and proceedings from congresses and symposia, as well as company communications
- More than 215 000 patent families from 11 leading sources (including EP, JP, US, and WO)

The system is updated daily, and its refined drug information is integrated in a single, flexible resource. Integrity's key strength is the fact that it is developed, populated, and supported by a cross section of scientists with expertise in a variety of disciplines, including medicinal chemistry, pharmacology, organic synthesis, molecular biology, pharmacokinetics, and metabolism. Importantly, the information is complete and consistent, including comprehensive pipeline, patent, and reference data dating back to 1988. The data curation process relies on complete analysis of the above sources, where only the facts relevant to the advancement of biomedical research are selected. Very importantly, all the information published in the system is standardized using internal indexing systems and strict quality control procedures, including the input from key opinion leaders in different areas of research.

In 2007, Thomson Scientific acquired Prous Science and the Integrity platform was added to the Thomson portfolio, offering unparalleled drug discovery content and unique analytic functionality for chemists, biologists, and other professionals in the life sciences (http://science.thomsonreuters.com/press/2007/8411150/).

9.3.1.1 Knowledge Areas

The depth and breadth of content puts a wealth of refined, structured information at the user's fingertips, fostering innovative decision making. In order to reflect all the conceptual relationships that are evoked when designing new bioactive compounds or analyzing therapeutically or structurally related molecules, the Integrity content has been organized in a series of 13 subsets or Knowledge Areas as shown in Figure 9.3.

To illustrate the functionality of the Integrity system, a case study on the retrieval of information for dapagliflozin, a drug acting as an inhibitor of the sodium–glucose

Figure 9.3 Thomson Reuters Integrity Knowledge Areas relationship scheme.

transporter type 2 enzyme (SGLT-2) with therapeutic application in the field of diabetes, is presented below.

The Drugs & Biologics Knowledge Area provides essential chemical and pharmacological information, along with the development status of bioactive compounds in the drug pipeline (Figure 9.4). Searches can include criteria for targets, literature, or patent references associated with the compounds. The search fields available in this Knowledge Area include entry number (six-digit unique identifier in the system); drug name [research code, USAN (United States Adopted Name) or INN (WHO international nonproprietary name]; brand names; chemical name following the IUPAC rules; standard InChI and InChIKey representation; CAS registry number, molecular formula and molecular weight; highest phase attained by the compound in its pharmaceutical development and flag indicating if the compound is under active development; year of launch for compounds reaching the marketplace, including prescription and formulation data; organization responsible for the discovery and/or clinical development of the compound; therapeutic group; mechanism of action focusing on the interaction of the compound with an enzyme, receptor, or ionic channel, as well as modulation of specific biological pathways.

Very importantly, every record in the database includes its availability date in the system or when the record was last updated, with available detail on the update (Update button). Once a record is obtained, related Knowledge Areas can be accessed by clicking the corresponding button in the Related Information section at the bottom of the Web page.

Figure 9.4 Example of Drugs & Biologics search result for dapagliflozin, including a Product Summary that describes the pharmacological/therapeutic novelty of the compound as well as its clinical development milestones. Related Information boxes at the bottom of the screen permit linking to other Knowledge Areas in the system.

The Experimental Pharmacology Knowledge Area includes data from experimental studies that delineate drug–receptor and enzyme–target cell interactions (Figure 9.5). Information in the Experimental Pharmacology Knowledge Area dates back to as far as 1998. Searches can include criteria for bioactive compounds tested, associated literature, and patent references and include the following fields: pharmacological activity measured in the experiment, parameter (endpoint) and unit used, experimental protocol (materials and methods), and numerical results.

Figure 9.5 Example of Experimental Pharmacology SAR table, where SGLT-2 inhibition for different compounds can be observed, including specific experimental protocol, compound potency, and associated references.

The accompanying Experimental Models Knowledge Area allows users to access further details on the experimental techniques used in the studies. Each record includes a description of the experiment and the characteristics of the model. Individual summaries are provided for each *in vivo* experimental model and are linked to multiple experimental pharmacology results that were generated using that model. The records also contain summaries of all the drugs that were tested in the model.

The Pharmacokinetics/Metabolism Knowledge Area includes data from experimental and clinical studies that delineate the absorption, distribution, metabolism,

Figure 9.6 Example of Pharmacokinetics data (bioavailability, clearance, area under the curve, etc.) and metabolic scheme for dapagliflozin. Results are linked to the original references for further study.

and excretion (ADME) profile of a drug (Figure 9.6). Searches can include criteria for bioactive compounds tested or associated literature references. Information in the Pharmacokinetics/Metabolism Knowledge Area dates back to 2000. The following choice of fields is found in this Knowledge Area: condition where the compound has been tested, administered product and administration route (measured product), interacting agent in the case of combination therapies, parameter (endpoint) measured in the experiment, and numerical results. The availability of a large collection of numerical values associated with ADME properties of biologically active compounds is a key factor when designing new compounds where the pharmacokinetics and metabolism need to be optimized.

The Organic Synthesis Knowledge Area describes routes of synthesis for drugs currently on the market or in development. Searches can include criteria for end products, literature, or patent references associated with the syntheses (Figure 9.7). Information in the Organic Synthesis Knowledge Area dates as far back as the 1970s.

Figure 9.7 The Organic Synthesis scheme for dapagliflozin, including intermediates, reactants, and experimental conditions can be accessed, along with the corresponding references.
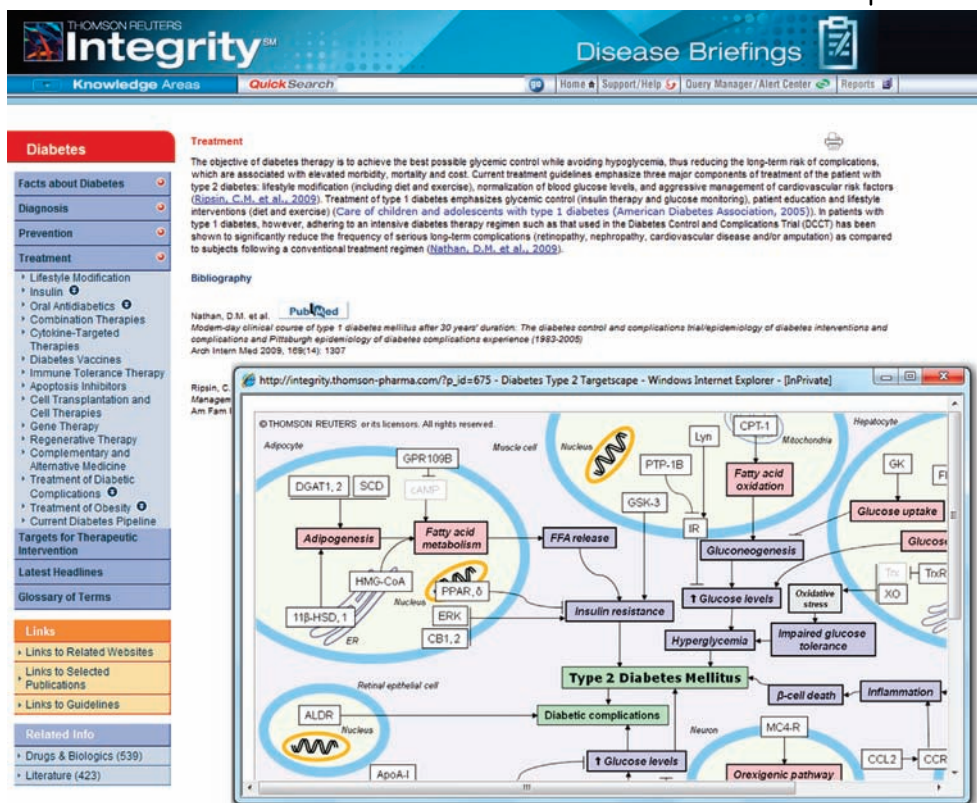
Figure 9.8 The Disease Briefings Knowledge Area provides key information for the compound in the context of its therapeutic activity. In this case, the diabetes disease briefing is shown, including a scheme (targetscape) delineating the pathophysiological basis of the disease. By clicking on a specific target the user can access active compounds for a specific mechanism of action.

The Organic Synthesis Knowledge Area includes the following choice of fields: synthesis description (text), intermediates used in the preparation of the compound, references, and supplier.

The Disease Briefings Knowledge Area consists of dynamic executive summaries on the current status and future trends in drug therapy for specific diseases (Figure 9.8). This area provides an executive view of a variety of therapeutic areas and is constructed in real time, based on the information available in the different database fields. Additional graphics and multimedia animations complement the text information available.

Targets & Pathways cover genes and their related proteins as targets for drug discovery (Figure 9.9). Information in the Targets & Pathways Knowledge Area dates back to 2004. Searches can include criteria for the product for the following choice of fields: target name, description, GenBank, Entrez, PDB and Swiss-Prot IDs, EC classification, and condition.

Figure 9.9 Example of the Targets & Pathways Knowledge Area, where information on the dapagliflozin target (SGLT-2) is shown. Users can link to the corresponding records in external databases and to the Integrity targetscape, where information on all the relevant targets for a certain disease is displayed in context. In addition, information whether the target has been validated (V) is being studied as a candidate (C) in clinical trials or is under exploratory (E) biological testing for a specific disease is specified.

The Biomarkers Knowledge Area links drugs, targets, and genes to biomarkers information from a diverse range of sources (Figure 9.10). Biomarkers information ranges from proof-of-mechanism to safety/toxicity studies or patient stratification applications. This area covers a wide range of biomarker types, from genomic, proteomic, and biochemical to cellular, physiological, and imaging biomarkers. Information on biomarkers is used at every stage of drug R&D, including disease risk detection, diagnosis, target identification, proof-of-mechanism, proof-of-concept, treatment/safety monitoring, and outcome measurement. Information in the

Figure 9.10 Example of SGLT-2 Biomarkers information in different diseases, including its role (diagnosis, disease progression measurement, etc.), the experimental technique used for the measure, the stage of research (from emerging status to late studies in humans), and the quantification of the described biomarker (+: increased, −: decreased, or +/−: presence). Links to the corresponding references or patents where the studies are described in detail are also provided.

Biomarkers Knowledge Area dates back to 2007, with additional selected back records. Searches can include criteria for a product, literature, or patent references associated with the biomarkers, such as biomarker name, biological process, product modifier, highest validity, population, role, and regulatory authority.

The Genomics Knowledge Area contains relationships between genes and diseases to gain insight into underlying biological mechanisms and to identify potential new drug targets (Figure 9.11). In fact, the identification of gene expression variations observed in human diseases permits the development of new compounds targeted to normalize the functioning of specific biological pathways. This Knowledge Area covers key facts on genes, diseases, and underlying biological mechanisms as targets for therapeutic intervention. The search fields include gene name, PDB and Swiss-Prot, GenBank or Entrez Identifiers, condition where the gene has a key participation, sequence, and polymorphism.

The Clinical Studies Knowledge Area includes information on clinical trials of drugs currently under study or in use in humans (Figure 9.12). Searches can include criteria for bioactive compounds tested or associated literature references.

Figure 9.11 Example of Genomics information related to the SGLT-2 target where gene polymorphisms associated with diseases are specified. Users can link to the corresponding records in external databases such as GenBank and KEGG, or systems biology solutions such as MetaCore™.

Information in the Clinical Studies Knowledge Area dates back to 2000 and includes the following fields: drug tested, study design, intervention type, and population number.

The Companies & Research Institutions Knowledge Area delineates essential information on public and private companies, academic centers, and research institutions that are active in the field of pharmaceuticals and biotechnology. Information on overall company sales/revenues, sales of launched products, products in development, and new patents is available. Searches can include criteria for products or patent references associated with the companies: company name, headquarters country, main activity, product annual sales, company economic data, number of employees, and key products and patents.
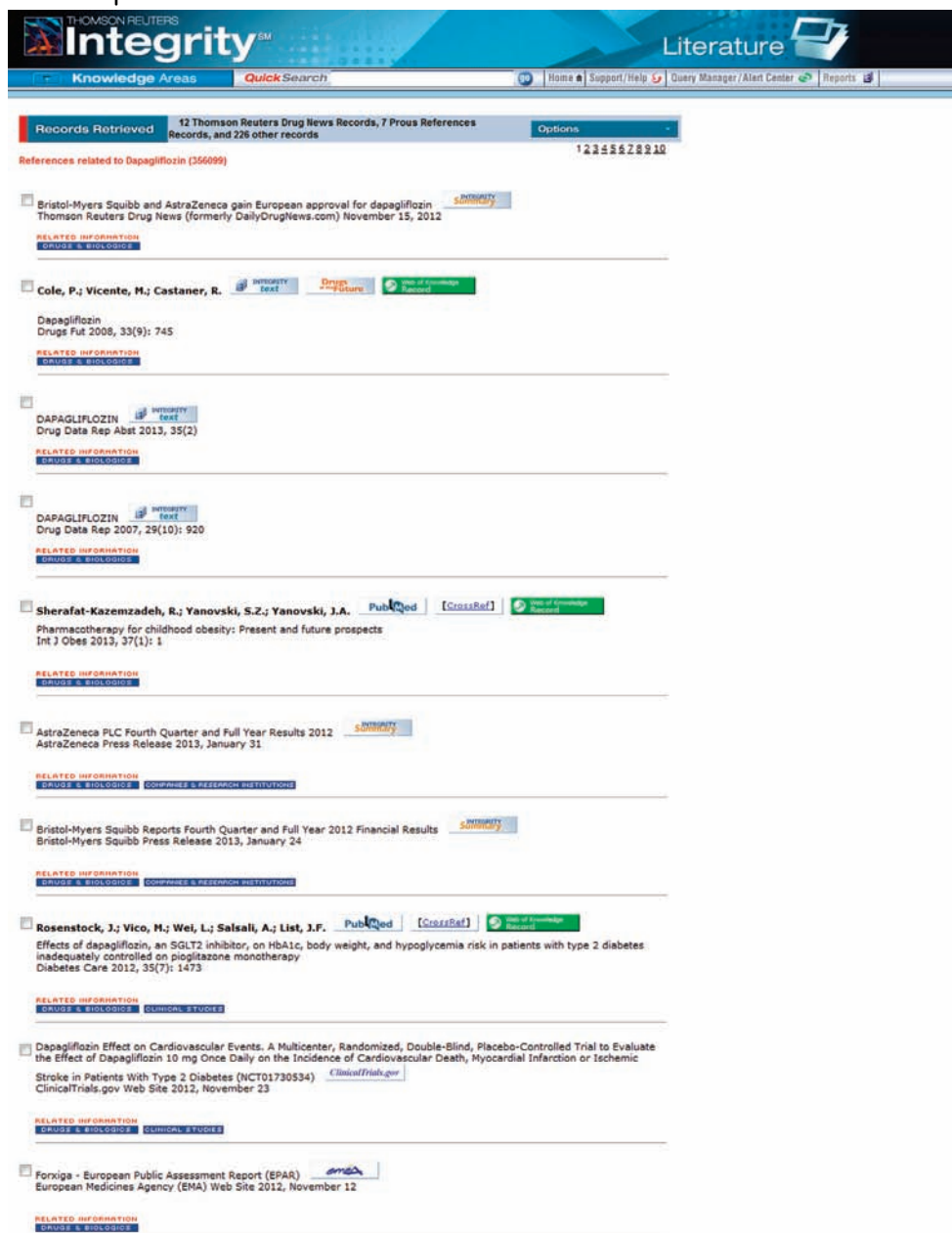
Figure 9.12 List of dapagliflozin clinical trials, with detailed information on number of patients, trial design, objectives, and conclusion. External links to references or public data are also provided.

In the Literature Knowledge Area, there are references to current biomedical literature, abstracts and proceedings from congresses and symposia, and company communications, as well as information on biomedical literature, congresses, and company communications (Figure 9.13). Searches can include bioactive compound criteria. Information in the Literature Knowledge Area dates back to 1988. The following fields can be searched: title, text, author, source, year, volume and issue/number, and congress edition.

The Patents Knowledge Area provides references to the most recent patent literature reflecting drug research activity throughout the world (Figure 9.14). Searches can include bioactive compound criteria. Information in the Patents Knowledge Area dates back to 1988. Selected fields include patent title, applicant name and data, country, inventor, patent number, publication date, expiration date, and priority data.

9.3.1.2 Search Fields

Each section in Integrity has three search field lines that can be set to any of the available field values (the menus have the default headings Select Value or Optional Value). All Knowledge Areas have at least two sections of search fields, one of which is the specific set of search fields pertaining to the Knowledge Area (e.g., fields such

Figure 9.13 References list for dapagliflozin selected from a variety of sources including biomedical literature, conferences, and company communications. External links are provided where available.

Figure 9.14 Example of dapagliflozin patent with access to complete patent document.

as "applicant" in Patents and "author" in Literature) and in most cases another will be the Product Section of the search field.

Thomson Reuters Integrity provides researchers with reliable, detailed information, from the perspective of a scientist, across multiple disciplines to support successful drug research and development. After selecting a Knowledge Area, the user defines the search strategy by combining the selected fields. Each field has an associated browse index containing available terms. Search fields can be combined using the appropriate Boolean operators. Integrity also has a chemical structure search feature that is compatible with four structure editors: Accelrys Draw, ISIS/ Draw (both from www.accelrys.com), ChemAxon Marvin Applet (www.chemaxon .com), and Cambridge Soft ChemDraw Plug-in (www.cambridgesoft.com).

9.3.1.3 Data Management Features

Integrity provides a quick search feature that can be used to retrieve terms across the full system, a statistics function that enables the filtering of retrieved data by further criteria, a query manager that makes it possible to save the search strategies so that they can be re-run automatically in the future, and a variety of listing and printing formats for easier organization of the information retrieved.

9.3.1.4 Use of Integrity in the Industry and Academia

Thanks to its specific Knowledge Area database concept and the possibility for the end user to navigate seamlessly across different types of data, the Thomson Reuters

Integrity system is used in many ways in the pharmaceutical industry or public research organizations. Examples include identification of newly described leads for emerging targets; rapid analysis of reported chemical diversity in a therapeutic area; finding of discrete experimental pharmacology or ADMET data; measuring the influence of biomarkers in drug discovery and development, or assessment of a specific clinical area in terms of intellectual property, published literature, or business development status. Furthermore, the availability of an extensive array of numerical values linked to chemical structures, which can be exported to spreadsheet format, has allowed new hypothesis in drug discovery to be initiated through the further use of SAR packages or data mining algorithms.

9.3.2
ChemBioBank

There have been several initiatives worldwide to develop additional content (i.e., new data and information derived from virtual screening and/or experimental high-throughput screening) in the area of chemical biology, which aim to understand in a comprehensive manner the interactions between chemical compounds (also called chemical probes) and biological entities (e.g., proteins, cells, pathways, and organisms), such as the Molecular Libraries Program in the United States (http://mli.nih.gov/mli/) or the EU-OPENSCREEN project in Europe (www.eu-openscreen.eu). Chemical biology may be considered as a starting point for further development of bioactive compounds, not only in pharmaceutical discovery but also in other areas such as veterinary medicine, agrofood, and so on.

We are hereby explaining details of the ChemBioBank (CBB) initiative for chemical biology in Spain (http://www.pcb.ub.edu/chembiobank/), which stemmed from this need to understand the chemical biology space in the search for bioactive compounds. While having similar goals to other chemical biology initiatives, CBB is unique in its focus on academic compounds as a source of innovative chemistry, and also unique in the inclusion of a virtual screening technique, polypharmacology screening, which addresses the issue of selectivity of chemicals versus diverse biological targets. The CBB proposal was initially promoted by three Spanish entities with complementary services: Parc Científic de Barcelona (PCB), offering synthetic and analytical chemistry services; Universidad de Santiago de Compostela (USC), offering experimental assay development and screening services; and Institut Municipal d'Investigació Mèdica (IMIM), offering virtual screening services. The proposal was divided into five areas summarized in Figure 9.15.

The following are the three main goals of the ChemBioBank initiative:

1) To build, organize, and maintain in a laboratory built at PCB a diverse, high-quality molecular library with compounds mainly from academic groups in Spain; the library will be accessible to the scientific community
2) To annotate the ChemBioBank library compounds with data derived from *in vitro* and *in silico* [experimental (HTS) and virtual screening] procedures

CHEMBIOBANK COMPOUNDS AND DATA WORKFLOW
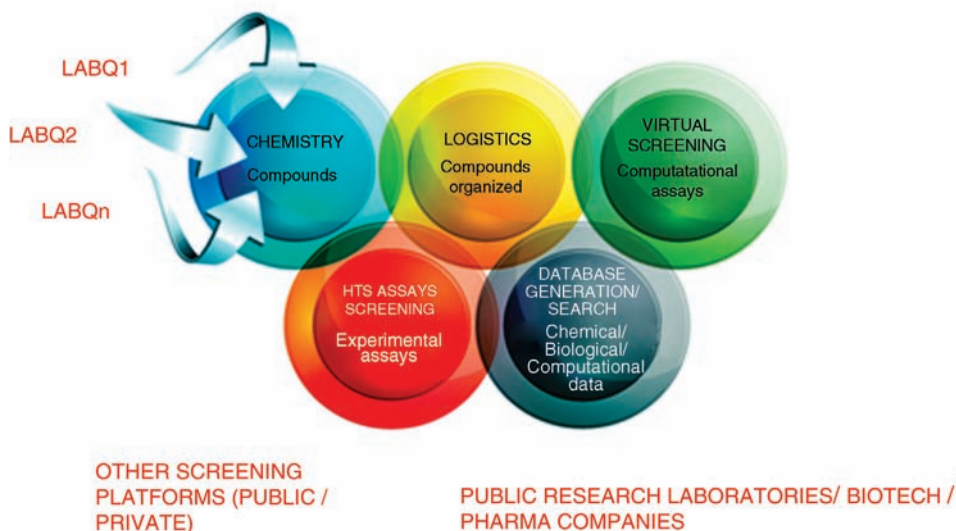
Figure 9.15 A summary of the five components of the ChemBioBank initiative.

3) To generate a remotely accessible ChemBioBank database containing the chemical structures of the library compounds and their virtual and experimental screening data

In order to achieve these goals, a ChemBioBank compound and data workflow were developed containing the following:

1) A new compound management laboratory set up at PCB, with capabilities to collect, analyze, organize, store, and distribute chemical compounds received individually in bar coded vials, or sets of compounds received in plate format, from external laboratories (either published compounds from academic laboratories or compounds selected from commercial library providers).

2) Once the received compounds are accepted into the ChemBioBank library, and after analytical chemistry quality control by LC/MS spectrometry, their chemical structures are profiled through a virtual screening procedure developed at IMIM, against around 4500 proteins to obtain a virtually annotated chemical biology hit-map.

3) The predicted chemical biology interactions are then validated experimentally through testing of virtual hits in biological assays developed at USC and other screening centers throughout Spain and Europe.

4) A new, remotely accessible ChemBioBank annotated database has been developed, containing the chemical structures of the ChemBioBank library compounds, their location in the compound management facility (vials, plates), and their virtual screening and experimental screening information. The data available for each compound will be publicly accessible through the Web, after

a 1 year moratorium to allow the original academic chemistry laboratories supplying the ChemBioBank library compounds to analyze the data obtained.

At the end of 2011, some of the results of this ChemBioBank initiative include the following:

1) The generation of a new 15 000-compound ChemBioBank library that contains around 4000 compounds of academic origin, 1100 compounds from the Prestwick Chemical Library (http://www.prestwickchemical.fr), and 10 000 compounds selected from three chemical library providers, on the basis of the maximization of both chemical and biological diversity. Since the compounds come from chemical catalogs and from academic laboratories, only a minor subset of the ChemBioBank library compounds are represented in the Thomson Reuters Integrity database

2) The characterization through virtual screening procedures of the interactions of the ChemBioBank library compounds toward 4500 proteins from the main families [G protein-coupled receptors (GPCRs), kinases, ion channels, nuclear receptors, transporters, and enzymes]

3) The generation of an annotated ChemBioBank chemical–biological logistic database, integrating database software tools developed by the companies IDBS (ActivityBase), for chemical and biological data registration; Titian (Mosaic), for compound management logistics registration; Chemotargets (iPhace), for virtual screening hit-map data registration and searching; and IDBS-InforSense, for remote access and chemical–biological data searches

Some applications of the ChemBioBank initiative include the following:

a) The ChemBioBank Virtual Screening profiling workflow process has been applied to the compounds from the Prestwick Chemical Library, which have reached the market or advanced clinical phases, in order to identify new mechanisms of action/new therapeutic applications for known compounds (reprofiling), as described in Figure 9.16

b) Around 4000 ChemBioBank compounds originating from academic laboratories have been virtually profiled toward targets of therapeutic interest. In one example [14], the hit-map obtained produced a signal for a set of compounds toward the family of G-protein-coupled adenosine receptors that led to the discovery of a set of new chemical scaffolds with selectivity for each of the adenosine receptor subtypes. This process that is led by the virtual screening polypharmacology application retrieves the association of chemical compounds to newly proposed protein targets, which, if validated, establishes new chemical scaffolds for protein targets of therapeutic interest

The ChemBioBank initiative has been undertaken in coordination with other chemical biology initiatives currently being developed in several European countries as part of the European Strategy Forum on Research Infrastructures (ESFRI) – funded project, the European Research Infrastructure on Open Screening Platforms, EU-OPENSCREEN (http://www.eu-openscreen.eu/). This project is
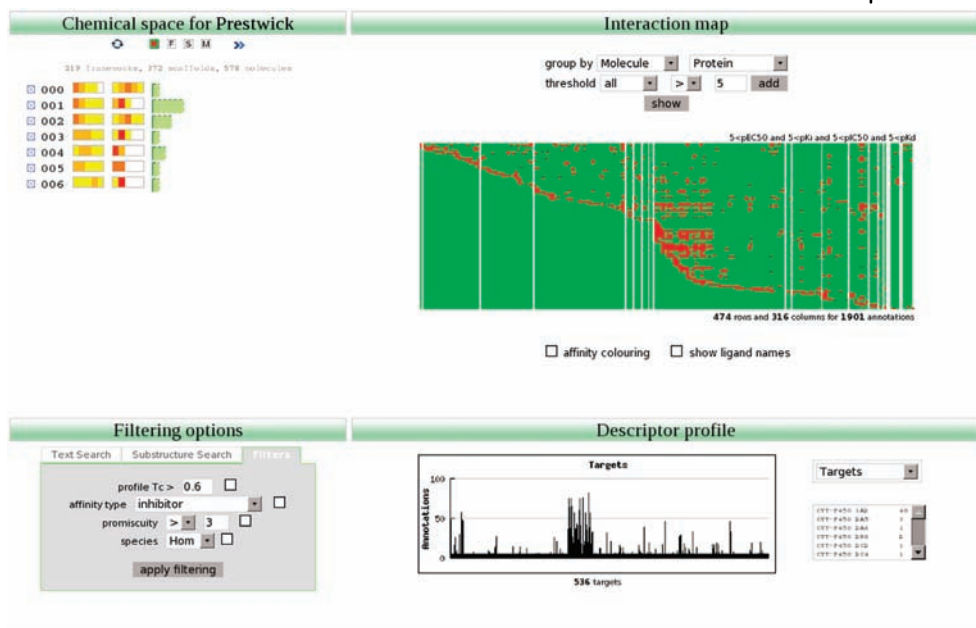
Figure 9.16 Virtual screening polypharmacology profiling of known compounds from the Prestwick Chemical Library (www.prestwickchemical.fr).

currently in its preparatory phase (November 2010–November 2013), with the goal of generating infrastructure (chemical library, distributed assay centers, and European chemical biology database) that may be able to complement the Molecular Libraries Program (MLP, a National Institutes of Health (NIH) Roadmap Initiative) in the United States, (http://mli.nih.gov/mli/), by merging the different chemical biology initiatives being developed in 12 European countries. The outcome of these initiatives will allow the scientific community around the world to discover new chemical probes, to start new drug discovery projects for unmet diseases, and have a repository of new chemical biology data for the analysis of structure–activity and structure–property relationships, and high-throughput data mining for systems biology.

9.3.3
Molecular Libraries Program

The Molecular Libraries Program (MLP) started by the National Institutes of Health (NIH) is another major initiative in the area of chemical biology and is summarized in Figure 9.17.

This MLP initiative started from an overall proposal from the NIH to improve the rates of success of new compounds reaching the clinic and eventually becoming effective therapeutics for diseases with unmet needs (http://www.mli.nih.gov). This
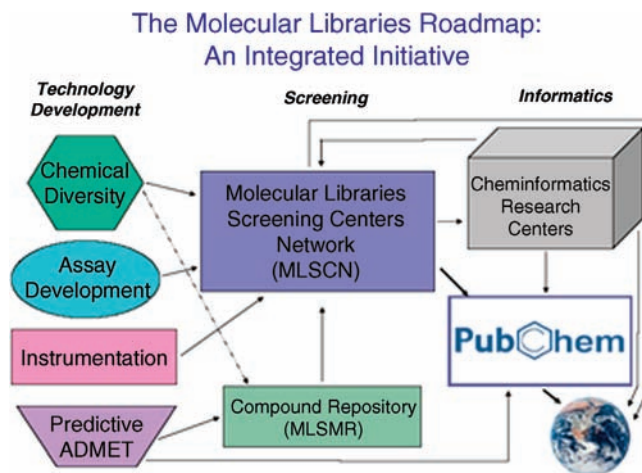
Figure 9.17 A summary of the three components of the NIH Molecular Libraries Program (MLP) (*source*: www.mli.nih.gov).

was based on the initial need for an overall understanding of chemical biology interactions. It led to the creation of a network of Molecular Libraries Screening Centers (MLSCN) (https://www.mli.nih.gov/mlscn/index.php) that developed assays to test selected compound repositories, with published chemical structures and assay data on a Web-accessible database, PubChem (http://pubchem.ncbi.nlm.nih.gov/).

In summary, all these chemical biology initiatives that are currently in development stage offer access to new chemical library compounds, innovative assays (both virtual and experimental screening), and chemical biology annotated databases, which are already contributing to the discovery of new bioactive substances. We may envision the application of innovative data mining and knowledge management technologies toward these large sets of data, which are expected to be openly accessible, such that the coverage of chemical biology interactions area is enlarged. Data mining technologies applied to these and other chemical biology annotated databases are described in Section 9.4.

9.4
Knowledge-Based Data Mining Technologies

As mentioned in the previous sections, the robustness of a predictive system for drug discovery and toxicity screening is closely related to the availability of large and diverse data sets of chemical compounds well characterized by their pharmacological properties. The knowledge associated with a training set of chemical structures can be expanded to include chemical descriptors, molecular fingerprints, or fragments, which can then serve to train machine learning-based predictive models.

Historically, a variety of machine learning algorithms were used with varying success in determining the pharmacological profiles of molecules. Key points in this area are the fact that the prediction needs to be made for a collection of hundreds of potential mechanisms of action in a very large and diverse molecular library, and that the small molecular entities of interest in pharmacology are frequently associated with several mechanisms of action. The mathematical problem of predicting the mechanisms of action of a queried molecular structure is formally known as a *multilabel classification* problem, where "labels" refer to the different mechanisms of action that can be predicted for a chemical compound. Multilabel classification has to be distinguished from the *single-label classification* problem. The latter refers to the problem of learning a model from examples (chemical structures), each one of which is associated with only one label of a set L of possible different labels. A query chemical structure will also be predicted with only one label. If $|L|$ is the number of different labels from which we choose a single label for each query chemical structure, $|L| = 2$ corresponds to the binary classification problem. Note that $|L| = 2$ and $|L| > 2$ are cases of single-label classification, since we are choosing only one label for each chemical structure.

In the multilabel classification problem, on the other hand, each chemical structure x_i of the training set is associated with a subset Y_i of the whole set L of possible labels, where $|Y_i|$ is the actual number of labels associated with x_i.

Two major tasks can be distinguished in multilabel learning. The first is classification itself, which for each query chemical structure produces a partition of the set of labels L into relevant and irrelevant labels; the second major task is ranking the labels according to their relevance to each chemical. Both tasks are related: if we have a ranking of labels based on a score, we can always divide the labels into relevant and irrelevant by applying a threshold to the score.

Given a training set for multilabel learning, one can ask to what degree the data set is multilabel. Two quantities measure this property of a data set, the label cardinality (LC) and label density (LD). Label cardinality is the average number of labels per example of the data set. Label density is the average percentage of the maximum number of labels $|L|$ that the examples in the data set actually have. Thus, label cardinality and label density are related by $LD = LC/|L|$.

The methods for multilabel learning can be classified as *problem transformation* methods and *algorithm adaptation* methods [15]. Problem transformation methods replace a multilabel learning task by one or more single-label classification tasks, for which there is a wide range of predictive algorithms. On the other hand, algorithm adaptation methods extend specific single-label learning algorithms to handle multilabel tasks.

9.4.1
Problem Transformation Methods

These methods are independent of the single-label predictive algorithm used after applying them. A number of simple problem transformation methods have been described in the literature [16,17]. Two crude transformations, not recommended in

general as they discard data, are the *select* and *ignore* transformations. The *select* transformation replaces each one of the multilabel examples by one of its associated labels, which can be chosen either randomly or following a criteria such as selecting the most or least frequent of the associated labels. The *ignore* transformation only keeps examples labeled with one label and discards every multilabel example. Another simple transformation method, the *copy* transformation, replaces each multilabel example (x_i, Y_i) by $|Y_i|$ single-label examples (x_i, λ_i), one for each label λ_i included in Y_i.

The *label powerset* transformation replaces the set of labels associated with each training example with a combined label, which is the union of the original labels. If the classifiers yield the probabilities of the predicted classes, in this case the probabilities of the combined labels, we can obtain a ranking of the original labels for a query instance from a score that is the sum of the probabilities of the combined labels [18]. One difficulty after applying the *label powerset* transformation is that if the number of combined labels is large compared to the number of original classes, even if we started with well-represented labels, we might end up with few training examples associated with the combined labels, which might hamper the quality of the predictions.

The *binary relevance* transformation replaces the original multilabel problem with $|L|$ binary classification problems, one for each label included in L. The binary classification for each label is trained with a data set created by considering the instances associated with the label as positives and all the other instances as negatives. Given this way of building the training sets for the binary classifiers, this method is sometimes said to follow a *one-versus-all* approach, meaning *one* label versus *all* other labels.

The *ranking by pairwise comparison* (RPC) transformation [19] replaces the multilabel problem with $|L|(|L| - 1)/2$ binary classification problems corresponding to the pairs of different labels in L. Each binary classifier for discriminating between a pair of labels is trained with the subset of examples in which either the first label or the second label appears, but not the two of them together. Finally, the labels are ranked by counting the votes of the binary classifiers for the different labels. Given that this method builds binary classifiers by pairing different labels, it is said to follow a one-versus-one approach. Fürnkranz and coworkers have proposed a natural method for applying a threshold to the RPC ranking and transforming it into a multilabel classification [20].

9.4.2
Algorithm Adaptation Methods

These are methods that extend single-label classifiers to deal with multilabel data. Some of them also apply the problem transformation methods already explained in the previous section.

A tree classifier or decision tree is a workflow in which leaves stand for class labels and forks between branches represent disjunctions of the decision process that takes to the classification labels. An adaptation of the C4.5 tree classifier has been introduced that handles multilabel data [21]. This adapted C4.5 algorithm allows multiple labels in the leaves of the tree and uses a modified formula for the entropy.

Adaboost is a machine learning ensemble method [22] that can be used with different learning algorithms. Adaboost builds a cascade of classifiers, tweaking subsequent classifiers in favor of those instances misclassified by previous classifiers. Although Adaboost was developed for boosting binary classifiers, Schapire and Singer have contributed two adaptations of Adaboost that can handle multilabel data: Adaboost.MR and Adaboost.MH [23]. The former, Adaboost.MR, uses the output of the cascade of weak classifiers to give a ranking of labels for each new example and aims to find a hypothesis that ranks the correct labels at the top. Adaboost.MH is a multilabel classification method that learns by minimizing the Hamming loss (i.e., the number of differences).

Zhang and Zhou adapted the feature subset selection method of backpropagation to solve the problem of multilabel learning [24]. This method starts by including all possible labels and leaves only the relevant ones at the end. To achieve this, they proposed an error function that takes the multiple labels into account.

A *support vector machine* (SVM) adaptation to the problem of multilabel ranking has been proposed that minimizes the ranking loss [25]. Later on, Godbole and Sarawagi introduced a number of improvements in the application of the SVM algorithm to multilabel learning [26]. On one hand, they propose using the *binary relevance* problem transformation with SVM classifiers, then adding the results of this first round of classification as new predictive variables to the original data set. This approach, which is a case of *stacking* (a method for combining classifiers), accounts for the potential dependencies among labels. On the other hand, Godbole and Sarawagi also propose removing all negative training instances of any class that is very similar to the positive class. These similarities are assessed on the confusion matrix estimated using a fast classifier on a holdout validation set. Finally, they improve the margin of the SVMs by removing very similar negative instances within a threshold distance from the learned hyperplane [27].

A number of adaptations of the *k-nearest neighbors* algorithm have been proposed that start by finding the closest k neighbors to the query instance and differ in the way they rank the labels of the nearest neighbor examples in order to obtain a prediction [28–31].

Thabtah and coworkers used association rule mining to build a classifier that they called MMAC [32]. MMAC learns an initial set of classification rules by association rule mining. It removes the examples associated with the learned rule set. It recursively learns a new rule set from the remaining examples until no frequent items are left. Rules with similar preconditions, but different predicted labels are merged into a multilabel rule. Finally, predicted labels are ranked by the support of the association rules that have predicted them.

9.4.3
Training a Mechanism of Action Model

Predicting the most probable mechanisms of action for a chemical structure query is a multilabel classification problem with the particularity that there is intrinsic

asymmetry in the data. This is because while there are many publications and patents that can be used as sources of information and that relate mechanisms of action to molecular structures, there are much less data about negatives, that is, mechanisms of action not being associated with molecular structures. This is evidently the consequence of negatives being much less interesting to readers than positives. Notice that in training sets that relate mechanisms of action to molecular structure, if a structure is not labeled with a given mechanism of action, it does not necessarily imply that the structure does not show this specific mechanism of action. This might cast doubt on using, for example, the *binary relevance* transformation method on a training set for mechanism of action because this method assumes that the training examples that are not labeled with a given label are necessarily negative for that label. Another more consistent approach for generating mechanism of action models would be to first use the *copy* transformation and then train a single-label coverage-based classifier for each label with only the training examples that contain that label. This approach also seems more natural for query chemical structures that will be predicted with no label, meaning that they have mechanisms of action not included in the model or that they are outside the applicability domain of the model.

Two software solutions that offer multilabel models for predicting the mechanisms of action of molecular structures are Poroikov's PASS [33] and Prous Institute's BioEpisteme (http://www.prousresearch.com). The latter is a flexible data and algorithmic integration platform with an intuitive Web interface that has been optimized to manage large data sets of chemical compounds through parallel computation. The system has also been evolved to cover important toxicity endpoints for bioactive compounds [34–37].

9.5
Future Trends and Outlook

In recent years, a variety of databases in the biomedical field, in both the private and the public domains, have been developed and enhancements to well-established systems have been made from both content and interface perspectives. In addition, collaborative developments, such as Open PHACTS that aligns and integrates proprietary and public data sources into a single system, have been presented (http://www.openphacts.org). Furthermore, the advances in genomics and personalized medicine are adding a new degree of complexity to the field. Therefore, a clear focus of the database, along with accurate selection and curation of the information and well-defined updating policy will be increasingly important in the successful deployment and use of drug discovery-oriented information systems.

From the technological perspective, distributed computing is a trend in computing that is meant to revolutionize the way data mining is performed. This is due to the fact that data warehousing technologies have provided a method of storing enormous volumes of data that have to be processed in order to extract useful information. There is also a technical reason that makes distributed computing a

requirement in the Internet age: answering user queries involving terabytes of data within a reasonable time frame requires distributed hardware, since terabyte processing is hampered not only by CPU power, but also by storage media speeds. Under the umbrella of the new software projects that are taking distributed computing to the next level [e.g., the Hadoop project (http://hadoop.apache.org)], a second generation of distributed databases commonly known as NoSQL, or nonrelational databases, are appearing. These databases are meant to be distributed across many nodes and are designed for performance and managing volumes of data that would collapse traditional relational databases (http://casandra.apache.org, http://hbase.apache.org). Such technology may enable even greater integration of data, while reducing the growing research costs and possibly potentiating the discovery of new molecular entities.

References

1 Paul, S.M., Mytelka, D.S., Dunwiddie, C.T., Persinger, C.C., Munos, B.H., Lindborg, S.R., and Schacht, A.L. (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, **9**, 203–214.

2 Garnier, J. (2008) Rebuilding the R&D engine in big pharma. *Harvard Business Review*, **86**, 68–76.

3 Williams, M. (2011) Productivity shortfalls in drug discovery: contributions from the preclinical sciences? *The Journal of Pharmacology and Experimental Therapeutics*, **336**, 3–8.

4 Geysen, H.M., Schoenen, F., Wagne, D., and Wagner, R. (2003) Combinatorial compound libraries for drug discovery: an ongoing challenge. *Nature Reviews Drug Discovery*, **2**, 222–230.

5 Kortagere, S. and Ekins, S. (2010) Troubleshooting computational methods in drug discovery. *Journal of Pharmacological and Toxicological Methods*, **61** (2), 67–75.

6 Ranjan, J. (2007) Application of data mining techniques in pharmaceutical industry. *Journal of Theoretical and Applied Information Technology*, **3** (4), 61–67.

7 Nonaka, I. (1991) The knowledge-creating company. *Harvard Business Review*, **69**, 96–104.

8 Loukides, M. (2011) What is data science? http://radar.oreilly.com/2010/06/what-is-data-science.html.

9 Mestres, J., Gregori-Puigjané, E., Valverde, S., and Solé, R.V. (2009) The topology of drug–target interaction networks: implicit dependence on drug properties and target families. Molecular BioSystems, **5**, 1051–1057.

10 Leeson, P.D. and Springthorpe, B. (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews Drug Discovery*, **6**, 881–890.

11 Nassar, A.F., Kamel, A.M., and Clarimont, C. (2004) Improving the decision-making process in the structural modification of drug candidates: enhancing metabolic stability. *Drug Discovery Today*, **9** (23), 1020–1028.

12 Hefti, F.F. (2008) Requirements for a lead compound to become a clinical candidate. *BMC Neuroscience*, **9** (Suppl. 3), S7.

13 Kesselheim, A.S. (2007) Intellectual property policy in the pharmaceutical sciences: the effect of inappropriate patents and market exclusivity extensions on the health care system. *AAPS Journal*, **9** (3), E306–E311.

14 Areias, F.M., Brea, J., Gregori-Puigjané, E., Zaki, M.E., Carvalho, M.A., Domínguez, E., Gutiérrez-de-Terán, H., Proença, M.F., Loza, M.I., and Mestres, J. (2010) *Bioorganic and Medicinal Chemistry*, **18** (9), 3043–3052.

15 Tsoumakas, G., Katakis, I., and Vlahavas, I. (2010) Mining multi-label data, in *Data*

Mining and Knowledge Discovery Handbook, 2nd edn (eds O. Maimon and L. Rokach), Springer, pp. 667–685.

16 Boutell, M., Luo, J., Shen, X., and Brown, C. (2004) Learning multi-label scene classification. *Pattern Recognition*, **37** (9), 1757–1771.

17 Chen, W., Yan, J., Zhang, B., Chen, Z., and Yang, Q. (2007) Document Transformation for Multi-Label Feature Selection in Text Categorization. Proceedings of the 7th IEEE International Conference on Data Mining, Omaha, NE, USA, October 28–31, 2007. pp. 451–456.

18 Read, J. (2008) A Pruned Problem Transformation Method for Multi-Label Classification. Proceedings of the 2008 New Zealand Computer Science Research Student Conference, Christchurch, New Zealand, April 14–17, 2008. pp. 143–150.

19 Hüllermeier, E., Fürnkranz, J., Cheng, W., and Brinker, K. (2008) Label ranking by learning pairwise preferences. *Artificial Intelligence*, **172** (16–17), 1897–1916.

20 Fürnkranz, J., Hüllermeier, E., Mencia, E.L., and Brinker, K. (2008) Multilabel classification via calibrated label ranking. *Machine Learning*, **73**, 133–153.

21 Clare, A. and King, R.D. (2001) Knowledge Discovery in Multi-Label Phenotype Data. Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery, Freiburg, Germany, September 3–5, 2001.

22 Freund, Y. and Schapire, R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, **55** (1), 119–139.

23 Schapire, R.E. and Singer, Y. (2000) BoosTexter: a boosting-based system for text categorization. *Machine Learning*, **39**, 135–168.

24 Zhang, M.L. and Zhou, Z.H. (2006) Multi-label neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, **18** (10), 1338–1351.

25 Elisseeff, A. and Weston, J. (2001) A kernel method for multi-labeled classification. *Advances in Neural Information Processing Systems, 14*, MIT Press.

26 Godbole, S. and Sarawagi, S. (2004) Discriminative Methods for Multi-Labeled Classification. Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Sydney, Australia, May 26–28, 2004.

27 Wolpert, D.H. (1992) Stacked generalization. *Neural Networks*, **5**, 241–259.

28 Spyromitros, E., Tsoumakas, G., and Vlahavas, I. (2008) An Empirical Study of Lazy Multilabel Classification Algorithms. Proceedings of the 5th Hellenic Conference on Artificial Intelligence Syros, Greece, October 2–4, 2008. pp. 401–406.

29 Brinker, K. and Hüllermeier, E. (2007) Case-Based Multilabel Ranking. Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6–15, 2007. pp. 702–707.

30 Zhang, M.-L. and Zhou, Z.-H. (2007) ML-kNN: a lazy learning approach to multi-label learning. *Pattern Recognition*, **40** (7), 2038–2048.

31 Luo, X. and Zincir-Heywood, A.N. (2005) Evaluation of Two Systems on Multi-Class Multi-Label Document Classification. Proceedings of the 15th International Symposium on Methodologies for Intelligent Systems, Saratoga Springs, New York, USA, May 25–28, 2005. pp. 161–169.

32 Thabtah, F.A., Cowling, P., and Peng, Y. (2004) A New Multiclass, Multilabel, Associative Classification Approach. Fourth IEEE International Conference on Data Mining (ICDM'04) Brighton, UK, November 1–4, 2004. pp. 217–224.

33 Poroikov, V. and Filimonov, D. (2005) PASS: prediction of biological activity spectra for substances, in *Predictive Toxicology* (ed. C. Helma), Taylor & Francis, pp. 459–478.

34 Matthews, E.J., Kruhlak, N.L., Benz, R.D., Contrera, J.F., Marchant, C.A., and Yang, C. (2008) Combined use of MC4PC, MDL-QSAR, BioEpisteme, Leadscope PDM, and Derek for Windows software to achieve high-performance, high-confidence, mode of action-based predictions of chemical carcinogenesis in rodents. *Toxicology Mechanisms and Methods*, **18** (2–3), 189–206.

35 Valencia, A. (2010) BioEpisteme: An *In Silico* Approach to Predict and Understand the Underlying Molecular Mechanisms Contributing to Toxicity Responses. XII International Congress of Toxicology, Barcelona, Spain, July 19–23. pp. 20–22.

36 Choi, S.S., Valerio, L.G., Jr., Kim, J.S., and Sadrieh, N. (2011) *In Silico* Predictive Model for Drug-Induced Phospholipidosis Using BioEpisteme Software. Society of Toxicology's 50th Annual Meeting &

ToxExpo, Washington DC, USA, March 6–10. Abst. 137.

37 Valerio, L.G., Jr., Moghaddam, S., Prous, J., Valencia, A., and Mensa, X. (2011) Computational Modeling for QT Prolongation: A Drug Cardiovascular Safety Endpoint of Paramount Importance. Society of Toxicology's 50th Annual Meeting & ToxExpo, Washington DC, USA, March 6–10. Abstract 146.

10
Applications of Rule-Based Methods to Data Mining of Polypharmacology Data Sets

Nathalie Jullian, Yannic Tognetti, and Mohammad Afshar

10.1
Introduction

The last decade has seen an unprecedented acceleration in the efforts for systematically identifying drug targets, benefiting from the sequencing of the human genome and a plethora of *omics* technologies, and the characterization of the interactions between known ligands and multiple targets. The seminal work by Paolini *et al.*, building upon their experience with Bioprint® and the endeavor to pull together large knowledge databases enabling comprehensive multiple target data mining, helped establish the word "polypharmacology" as a successor to chemogenomics [1,2]. This type of data is becoming more and more publicly available, as illustrated by PubChem, the public domain database that links a few million of small organic molecules with their activity in multiple bioassays [3,4].

Analyzing the similarity between the ligands that bind to multiple targets, rather than the sequence similarity between targets, has opened an additional dimension in searching for novel targets for known drugs [5–7]. The "unexpected" off-target activities may contribute to the desired activity or, on the contrary, be responsible for serious side effects, explaining in part the high attrition rate in drug development [8].

The occurrence of an adverse drug reaction (ADR) in humans is most likely related to the interaction of a compound and certain targets. Hence, a number of screening campaigns and lead optimization efforts have incorporated multiple targets in order to consider selectivity against potential undesired interactions. The required selectivity might be included within a family of targets (e.g., discriminating between related CDK kinases) or more generally it might consist in avoiding certain targets known to be linked to ADRs. One such selectivity target is the hERG potassium channel. It has been shown that hERG blockers are related to QT prolongation that induces cardiac toxicity [9].

Paolini *et al.* argue that, in some cases, it is a certain lack of specificity that constitutes the target profile; hence the question of whether one could design effective promiscuous drugs [1]. Biological activity against multiple targets might be an advantage when the related combination provides an interesting therapeutic outcome. This was shown to be the case for Nelfinavir, a compound active against a panel of kinase targets in oncology [10].

In both cases however, whether trying to improve the desired therapeutic effect or avoiding a negative side effect, the objective is to reach a desired "profile," that is, interacting with specific targets and "avoiding" others.

In a typical drug discovery setup, following initial screening and lead generation steps, molecules are designed based on the analysis of the structure–activity relationships (SAR). When starting a new project, the medicinal chemist extracts the SAR from the existing series of compounds with the purpose of generating new hypotheses as the basis for novel rational compound design. The SAR analysis consists in evaluating the impact of specific chemical fragments or combinations of fragments on a given bioactivity. A large number of computational methods have been developed to support this decision making process, mainly using predictive models based on molecular descriptors or on chemical fragments.

Key challenges have been linked to the difficulty of integrating multiple objectives (i.e., aiming for a profile that combines multiple desired and undesired activities across targets). When multiple targets are considered, this may lead to concurrent models. In addition, translating a computer-based prediction into a synthesizable molecule by the medicinal chemist continues to remain a challenge.

Rule-based data mining methods aiming to discover hidden patterns have long been applied to the market basket analysis in the retail business [11]. Similar methods such as Apriori, formal concept analysis (FCA), or substructure association, are attracting growing interest in drug discovery due to the ease of interpretation of the results [11–20]. In most of these applications, the chemical compounds are described by a set of substructure fragments. Indeed, it has been shown that substructure fragments are adequate descriptors for capturing the structural characteristics of chemical compounds [21,22].

Wolohan *et al.* combine a fragmentation method similar to RECAP and an adaptation of the Apriori algorithm to identify structural units that define highly active compounds in a chemical class [12,23]. Lounkine *et al.* report FragFCA, an adaptation of FCA to extract pairs or triplets of substructure fragments specific of a single or multiple activity profile [17]. They show that FragFCA is successful in identifying biologically relevant signature patterns for subsets of GPCR antagonists. Karwath and De Raedt illustrate an application of SMIREP to activity classification, a method based on generating simple rules from substructure fragments of active compounds [14]. Klopman and Tu describe MCASE, a system able to identify substructure descriptors named either biophores or biophobes, depending respectively on their positive or negative contribution to activity [24]. As illustrated in these papers, the rule-based methods can be adapted to consider chemical fragments as well as multiple activity endpoints.

In this chapter we illustrate the use of KEM®, a rule-based method using FCA for the systematic analysis of the relations between multiple activities and the contribution of chemical fragments to a desired pharmacology profile. The effort toward the development of antipsychotic drugs with less serious side effects is used as an example [25]. We report the analysis of 99 chemical compounds with activities measured against the σ-1 receptor, the dopamine D2 receptor, and the serotonin 5HT2 receptor. We characterize the relationships existing between the activities of

these targets. The rule-based system is used to derive relations that identify specific changes for the design of selective high-affinity σ-1 receptor binders with selectivity over the dopamine D2 receptor. This approach is then extended by incorporating an extra selectivity endpoint toward the serotonin 5HT2 receptor. We show that the rule-based framework is ideally suited for analyzing in detail a data set with multiple targets, leading to specific suggestion of 18 novel molecules predicted to achieve the desired profile.

## 10.2
Materials and Methods

10.2.1
Data Set Preparation

The data set is described as a table of categorical variables. The continuous numeric values of the activity variables are converted into discrete intervals. Each endpoint is described by at least two intervals for the learning process. Very often, a pharmacological activity can be binned into two intervals – activity = "yes" for active compounds and activity = "no" for inactive ones– or three intervals – activity = "high," activity = "medium," and activity = "low" – based on user-defined threshold values. The variables listing the constitutive chemical fragments are *per se* categorical variables. Commonly, a medicinal chemistry series may be represented as a Markush structure: a central scaffold and a set of functional groups named R-groups [26]. Multiple scaffolds can be used as far as the R-groups can be described from analogous substitution points. No enumeration of the library is required for the analysis, and the chemical fragments (scaffold and R-groups) can be coded as text strings or SMILES strings. The latter code is compatible with the display of the two-dimensional structures with KEM.

10.2.2
Preparation of the σ-1 Binders Data Set

The data set consists of 99 σ-1 receptor ligands extracted from Table III of Ref. [25]. They have the basic structure depicted in Table 10.1. The paper describes the SAR leading to the discovery of σ-1 ligands with selectivity over the D2 dopamine receptor. For each molecule, the data set includes the identities of the functional groups X, R1, and R2; the lengths of the carbon chains M and N; and the activity values for the σ, the D2 dopamine, and the 5HT2 serotonin receptors (Table 10.1). The chemical fragments are listed as SMILES strings. The "Sigma_KI" continuous variable is converted into a categorical variable by using the following threshold values: Sigma_KI is "low" when Sigma_KI is greater than 100 nM, Sigma_KI is "medium when Sigma_KI values are greater than 30 nM and lower than or equal to 100 nM, and Sigma_KI is "high" when Sigma_KI is lower than or equal to 30 nM.

Table 10.1 Description of the 99 molecules (scheme and R-groups).
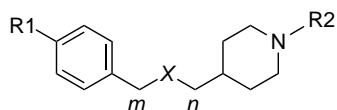
ID	M	N	X	R1	R2	Sigma_KI	D2_IC50	5HT2_KI
22	0	1	[*]S(=O)[*]	[*]F	[*]CC(C1CC1)	L	L	
23	0	1	[*]S(=O)(=O)[*]	[*]F	[*]CC(C1CC1)	H	L	L
10a	1	0	[*]C(O)[*]	[*]F	[*]C(c1ccccc1)	H	M	M
10b	1	0	[*]C(O)[*]	[*]H	[*]C(c1ccccc1)	H	L	L
10c	1	0	[*]C(O)[*]	[*]F	[*]CCC(c1ccccc1)	H	L	H
10d	1	0	[*]C(O)[*]	[*]F	[*]CCCC(c1ccccc1)	H	M	H
11b	1	0	[*]C(=O)[*]	[*]F	[*]C(c1ccncc1)	H	L	L
11c	1	1	[*]C(=O)[*]	[*]F	[*]C(c1ccccc1)	H	L	M
11d	1	0	[*]C(=O)[*]	[*]F	[*]CCCC(=O)(c1ccc(F)cc1)	M	H	H
18a	0	1	[*]O[*]	[*]F	[*]C(C1CC1)	H	H	H
18aa	0	1	[*]O[*]	[*]F	[*]C(c1ccccc1)	H	M	H
18ab	0	1	[*]O[*]	[*]Cl	[*]C(c1ccccc1)	H	L	M
18ac	0	1	[*]O[*]	[*]N(=O)(=O)	[*]C(c1ccccc1)	H	L	M
18ad	0	1	[*]O[*]	[*]OC	[*]C(c1ccccc1)	H	L	H
18ae	0	1	[*]O[*]	[*]C(F)(F)(F)	[*]C(c1ccccc1)	H	L	M
18af	0	1	[*]O[*]	[*]F	[*]C(c1ccc(F)cc1)	H	L	H
18ag	0	1	[*]O[*]	[*]F	[*]C(c1ccc(OC)cc1)	H	L	
18ah	0	1	[*]O[*]	[*]F	[*]Cc1ccc2ccccc2c1	M	L	M
18ai	0	1	[*]O[*]	[*]F	[*]C(c1ccncc1)	H	L	H
18aj	0	1	[*]O[*]	[*]F	[*]CC(c1ccc(Cl)cc1)	H	H	H
18ak	0	1	[*]O[*]	[*]F	[*]CC(C1CC1)	H	M	
18al	1	1	[*]O[*]	[*]F	[*]C(c1ccccc1)	H	L	M
18 a.m.	1	1	[*]O[*]	[*]OC	[*]C(c1ccccc1)	H	L	M
18an	1	1	[*]O[*]	[*]c1ccccc1	[*]C(c1ccccc1)	H	L	H
18ao	1	1	[*]O[*]	[*]H	[*]C(c1ccccc1)	H	L	L
18ap	1	1	[*]O[*]	[*]C(=O)OC	[*]C(c1ccccc1)	H	M	L
18aq	1	1	[*]O[*]	[*]F	[*]CC(c1ccccc1)	H	M	H
18ar	1	1	[*]O[*]	[*]F	[*]CCC(c1ccccc1)	L	L	
18as	1	1	[*]O[*]	[*]F	[*]C(c1ccc(C(=O)OC)cc1)	H	L	M
18at	1	1	[*]O[*]	[*]F	[*]C(c1ccc(Cl)cc1)	H	M	M
18au	1	1	[*]O[*]	[*]F	[*]C(c1ccc(c2ccccc2)cc1)	H	M	L
18av	1	1	[*]O[*]	[*]F	[*]C(c1ccc(O)cc1)	H	L	L
18aw	1	1	[*]O[*]	[*]F	[*]C(c1ccc(OCc2ccccc2)cc1)	H	M	H
18ax	1	1	[*]O[*]	[*]F	[*]C(c1ccncc1)	H	L	
18ay	1	1	[*]O[*]	[*]F	[*]C(C1CCCCC1)	H	L	M

18az	1	1	[*]O[*]	[*]F	[*]Cc1ccc2ccccc2c1	H	H	M
18b	0	1	[*]O[*]	[*]Cl	[*]C(C1CC1)	H	L	H
18ba	1	1	[*]O[*]	[*]F	[*]Cc1cccc2ccccc12	H	H	H
18bb	1	1	[*]O[*]	[*]F	[*]CCCCC	H	L	
18bc	0	2	[*]O[*]	[*]F	[*]C(c1ccccc1)	H	H	H
18bd	3	0	[*]O[*]	[*]H	[*]C(c1ccccc1)	H	L	
18be	3	1	[*]O[*]	[*]H	[*]C(c1ccccc1)	H	M	M
18bf	4	1	[*]O[*]	[*]H	[*]C(c1ccccc1)	H	L	M
18bg	5	1	[*]O[*]	[*]H	[*]C(c1ccccc1)	H	L	M
18bh	1	2	[*]O[*]	[*]C(CCCC)(CCCC)CCCC	[*]C(c1ccccc1)	H	L	L
18bt	0	1	[*]O[*]	[*]F	[*]CC(C1CC1)	H	L	H
18c	0	1	[*]O[*]	[*]OC	[*]C(C1CC1)	H	L	H
18d	0	1	[*]O[*]	[*]c1ccccc1	[*]C(C1CC1)	H	L	L
18e	0	1	[*]O[*]	[*]CO	[*]C(C1CC1)	M	L	H
18f	0	1	[*]O[*]	[*]C(CCCC)(CCCC)CCCC	[*]C(C1CC1)	H	L	L
18g	0	1	[*]O[*]	[*]C(O)C	[*]C(C1CC1)	M	L	
18h	0	1	[*]O[*]	[*]F	[*]C(C1CC1)	H	L	H
18i	0	1	[*]O[*]	[*]H	[*]C(C1CC1)	H	L	L
18j	0	1	[*]O[*]	[*]OC	[*]C(C1CC1)	M	L	L
18k	0	1	[*]O[*]	[*]SC	[*]C(C1CC1)	H	L	M
18l	0	1	[*]O[*]	[*]S(=O)(=O)C	[*]C(C1CC1)	M	L	L
18m	0	1	[*]O[*]	[*]N(=O)(=O)	[*]C(C1CC1)	H	L	M
18n	0	1	[*]O[*]	[*]CN	[*]C(C1CC1)	H	L	L
18o	0	1	[*]O[*]	[*]OCC	[*]C(C1CC1)	H	M	L
18q	0	1	[*]O[*]	[*]Oc1ccccc1	[*]C(C1CC1)	H	L	L
18r	0	1	[*]O[*]	[*]c1ccc(F)cc1	[*]C(C1CC1)	M	L	M
18s	0	1	[*]O[*]	[*]c1ccc(OC)cc1	[*]C(C1CC1)	L	L	
18t	0	1	[*]O[*]	[*]H	[*]C(C1CC1)	H	L	H
18u	0	1	[*]O[*]	[*]H	[*]C(C1CC1)	H	L	L
18v	0	1	[*]O[*]	[*]Cl	[*]C(C1CC1)	H	L	H
18w	0	1	[*]O[*]	[*]Cl	[*]C(C1CC1)	H	H	H
18x	0	1	[*]O[*]	[*]CCN	[*]C(C1CC1)	M	L	L
18y	0	1	[*]O[*]	[*]F	[*]C(C1C(C)C1)	H	M	H
18z	0	1	[*]O[*]	[*]F	[*]C(C1(C)C(Cl)(Cl)C1)	H	H	H
6a	0	1	[*]C(=O)[*]	[*]C(F)(F)(F)	[*]C(c1ccccc1)	H	L	L
6b	0	1	[*]C(=O)[*]	[*]OC	[*]C(c1ccccc1)	H	L	M
6c	0	1	[*]C(=O)[*]	[*]SC	[*]C(c1ccccc1)	H	M	M
6d	0	1	[*]C(=O)[*]	[*]O	[*]C(c1ccccc1)	H	L	L
6e	0	1	[*]C(=O)[*]	[*]c1ccccc1	[*]C(c1ccccc1)	M	L	H
6f	0	1	[*]C(=O)[*]	[*]CO	[*]C(c1ccccc1)	M	L	L
6g	0	1	[*]C(=O)[*]	[*]S(=O)OC	[*]C(c1ccccc1)	H	L	M
6h	0	1	[*]C(=O)[*]	[*]S(=O)C	[*]C(c1ccccc1)	M	L	M
6i	0	1	[*]C(=O)[*]	[*]F	[*]C(C1CC1)	H	L	H
6j	0	1	[*]C(=O)[*]	[*]Cl	[*]C(C1CC1)	H	L	L
6k	0	1	[*]C(=O)[*]	[*]OC	[*]C(C1CC1)	M	L	L

(*continued*)

Table 10.1 (Continued)

ID	M	N	X	R1	R2	Sigma_KI	D2_IC50	5HT2_KI
6l	0	1	[*]C(=O)[*]	[*]C(CCCC)(CCCC)CCCC	[*]C(C1CC1)	H	L	L
6m	0	1	[*]C(=O)[*]	[*]c1ccccc1	[*]C(C1CC1)	M	L	L
6n	0	1	[*]C(=O)[*]	[*]C(F)(F)(F)	[*]C(C1CC1)	M	L	M
6o	0	1	[*]C(=O)[*]	[*]NC(C)	[*]C(C1CC1)	M	L	L
6p	0	1	[*]C(=O)[*]	[*]N	[*]C(C1CC1)	L	L	
6q	0	1	[*]C(=O)[*]	[*]C#N	[*]C(C1CC1)	H	L	L
6r	0	1	[*]C(=O)[*]	[*]F	[*]C(c1ccc(C(F)(F)(F))cc1)	H	H	H
6s	0	1	[*]C(=O)[*]	[*]F	[*]C(c1ccc(F)cc1)	H	M	H
6t	0	1	[*]C(=O)[*]	[*]F	[*]CCC1 = CNC2 = CC = CC = C12	H	H	H
6u	0	1	[*]C(=O)[*]	[*]F	[*]CC(c1ccc(F)cc1)	H	H	H
6v	0	1	[*]C(=O)[*]	[*]F	[*]CC(c1ccccc1)	H	H	H
6w	0	1	[*]C(=O)[*]	[*]F	[*]CC(c1ccc(Cl)cc1)	H	H	H
6x	0	1	[*]C(=O)[*]	[*]F	[*]CC(c1ccc(C(F)(F)(F))cc1)	H	H	H
6y	0	1	[*]C(=O)[*]	[*]F	[*]CC(C1CC1)	H	H	H
7a	0	1	[*]C(O)[*]	[*]F	[*]CC(C1CC1)	L	L	
7c	0	1	[*]C(O)[*]	[*]SC	[*]C(c1ccccc1)	H	L	H
7d	0	1	[*]C(O)[*]	[*]OC	[*]C(c1ccccc1)	H	L	M
7e	0	1	[*]C(O)[*]	[*]C(F)(F)(F)	[*]C(c1ccccc1)	H	L	H
7f	0	1	[*]C(O)[*]	[*]F	[*]CC(c1ccccc1)	H	M	M

IC$_{50}$ and K_I values are listed as intervals (L, M, H).

For the dopamine D2 receptor, the intervals are defined as D2 = "high" when D2_IC50 ≤ 400 nM, D2 = "medium" when 400 nm < D2_IC50 ≤ 1000 nM, and D2 = "low" when D2_IC50 > 1000 nM. For the serotonin 5HT2 receptor, the intervals are as follows: 5HT2 = "high" when 5HT2_IC50 ≤ 100 nM, 5HT2 = "medium" when 100 nm < 5HT2_IC50 ≤ 400 nM, and 5HT2 = "low" for 5HT2_IC50 > 400 nM. In what follows, we have used _L, _M, and _H to designate low, medium, and high intervals respectively.

10.2.3
Association Rules

KEM uses a rule-based machine learning method that extracts the knowledge contained in a data set [27]. The technology is derived from the Galois lattice, also known as formal concept analysis theory [13,28]. The algorithm is able to detect association rules (or relationships) between specific values of categorical variables in large data sets.

The first set of rules characterizes the polypharmacology profile of the data set. These rules are of the following form:

```
act_target1="High" → act_target2="High"
```

This rule characterizes the compounds that have a high activity on target2 (right term or *consequent*) considering that they also have a high activity on target1 (left term or *antecedent*). The *support* and the *confidence* of the rules characterize the coverage of the dual profile in the data set. The *support* of a rule is the number of cases in the data set that contain both the *antecedent* and the *consequent*. The confidence of the rule is the probability that an item contains the *consequent* given that it contains the *antecedent*. When more than two targets are considered, these rules may contain multiple targets on the left and/or combination of targets on the right.

The second set of association rules identify molecular fragments (and combination of fragments) that appear frequently in the active molecules and are only detected rarely in the inactive molecules.

```
{RU1}: (R1="methyl") AND (R2="Cl") → activity="High"
```

This rule labeled {RU1} can be easily read as follows: IF a molecule has a methyl group at the R1 position AND a chlorine atom at the R2 position THEN the activity is "high". The (R1 = "methyl") and (R2 = "Cl") are the left terms of the rule and the (activity = "high") is the right term. The size (or length) of the rule is the number of left terms, that is 2 in the {RU1} example. In the {RU1} example, the support corresponds to the number of molecules containing a methyl group at the R1 position and a chlorine atom at the R2 position altogether with a high activity. The confidence of {RU1} explains how frequently the desired "high" activity occurs among molecules containing the combination of fragments such as a methyl group at R1 and a chlorine atom at R2 (for RU1). An additional parameter "lift" is also calculated that is defined as the relative confidence of the rule versus the probability of the occurrence of the consequent in the data set. Assuming *#right* is the number of examples supporting the right term, the *lift* formula is expressed as *left = confidence/#right*.

The output of the algorithm can be a very large set of association rules, on the order of thousands. Further filtering steps comprise a combination of methods to reduce the large number of rules to a focused subset of interest.

10.2.4
Novel Hybrid Structures by Fragment Swapping

The rules or combinations of fragments selected are matched on a set of molecules, and position-specific fragment changes are suggested leading to novel chemical entities. Typically, molecules with "low" or "medium" activity will be selected for optimization. The details of the KEM-Optimize process will be published elsewhere.

10.3
Results

The example illustrated here aims at the design of antipsychotic drugs with less serious side effects, starting with a set of 99 disubstituted piperidine σ-1 receptor ligands [25].

10.3.1
Rules Generation and Extraction

10.3.1.1   Rules Describing the Polypharmacology Space
The rules describing the polypharmacology space illustrate the relationships existing between the three endpoints for this chemical series (Table 10.2). They summarize the entire multiobjective context covered by the molecules of the project. The ideal polypharmacology profile is represented by a subset of 17 compounds supporting the rules PO1, PO3, PO5, PO7, PO11, PO16, and PO17 (Table 10.2). The low confidence (23%) rule PO11 illustrates the difficulty in reaching the desired selectivity: only 23% of the 79 compounds with a high sigma affinity have the desired selectivity profile against D2 and 5HT2.

The analysis of the relations between the Sigma and D2 shows that the data set does contain molecules with the desired profile (50 molecules) and that Sigma_H is associated with D2_L (Sigma_H → D2_L) with a confidence of 63% and D2_L is associated with Sigma_H (D2_L → Sigma_H) with a confidence of 72% (Table 10.2, rules PO3, PO7). However, the lift (relative probability) is lower than 1. This demonstrates that a low D2 does not increase the probability of identifying a molecule with high Sigma.

Furthermore, molecules combining a low D2 and a low 5HT2 are linked to a high Sigma with a confidence of 71%, but again the lift is below 1, indicating that these two conditions do not enrich the data set with additional molecules with high Sigma (Table 10.2, rule PO17). We note that a high affinity for D2 is likely to be associated with a high K_I against the sigma receptor: rule PO2 has a confidence of 93% and a lift of 1.17.

Table 10.2   List of the three rules compatible with the target profile (high σ-1 and low D2).

Rule ID	Combination	Sigma KI_nM	Sigma Support	Sigma Confidence	D2 IC50_nM	D2 Support	D2 Confidence
{A}	M_0 *AND* N_1 *AND* R2_[*]C(c1ccccc1)	H	13	81%	L	14	87%
{B}	M_0 *AND* R2_[*]C (c1ccccc1)	H	14	82%	L	14	82%
{C}	N_1 *AND* R2_[*]C (c1ccccc1)	H	22	88%	L	21	84%

We therefore need to extract additional features from compound structures that would help us identify subsets of molecules where the relative probability of D2_L and Sigma_H is increased. These subsets may be described by combination of fragments (also named "motifs") that have a favorable impact on the desired profile.

10.3.1.2 Optimization of σ-1 with Selectivity Over D2

The SAR consists of relationships between chemical fragments and activity endpoints. The SAR is converted into a set of 30 association rules characterized by a minimum confidence of 80% and a minimum support value of 10, compatible with a high σ-1 *or* a low D2 affinity. Of these 30 rules, only 3 rules are consistent with the desired profile of a high σ-1 *and* a low D2 activity (Table 10.3). The rule with the highest support, rule {C}, will be considered for further optimization. The 14 molecules with a high affinity for the σ-1 receptor but that do not exhibit the required selectivity over the D2 dopamine receptor are selected for the optimization process.

Table 10.3 Selection of rules representing the polypharmacology space.

Rule ID	Left		Right	Support	#Left	#Right	Confidence (%)	Lift
PO1	5HT2_KI_nM_L	→	D2_IC50_nM_L Sigma_KI_nM_H	17	27	50	63	1.20
PO2	D2_IC50_nM_H	→	Sigma_KI_nM_H	14	15	79	93	1.17
PO3	D2_IC50_nM_L	→	Sigma_KI_nM_H	50	69	79	72	0.91
PO4	D2_IC50_nM_L	→	Sigma_KI_nM_M	14	69	15	20	1.34
PO5	D2_IC50_nM_L	→	Sigma_KI_nM_H 5HT2_KI_nM_L	17	69	20	29	1.27
PO6	D2_IC50_nM_M	→	Sigma_KI_nM_H	15	15	79	100	1.25
PO7	Sigma_KI_nM_H	→	D2_IC50_nM_L	50	79	69	63	0.91
PO8	Sigma_KI_nM_H	→	D2_IC50_nM_M	15	79	15	19	1.25
PO9	Sigma_KI_nM_H	→	D2_IC50_nM_H	14	79	15	18	1.17
PO10	Sigma_KI_nM_H	→	D2_IC50_nM_H 5HT2_KI_nM_H	13	79	14	18	1.10
PO11	Sigma_KI_nM_H	→	D2_IC50_nM_L 5HT2_KI_nM_L	17	79	24	23	0.84
PO12	Sigma_KI_nM_M	→	D2_IC50_nM_L	14	15	69	93	1.34
PO13	D2_IC50_nM_H Sigma_KI_nM_H	→	5HT2_KI_nM_H	13	14	36	93	2.27
PO14	D2_IC50_nM_H 5HT2_KI_nM_H	→	Sigma_KI_nM_H	13	14	79	93	1.10
PO15	D2_IC50_nM_L 5HT2_KI_nM_H	→	Sigma_KI_nM_H	14	16	79	88	1.04
PO16	Sigma_KI_nM_H 5HT2_KI_nM_L	→	D2_IC50_nM_L	17	20	69	85	1.27
PO17	D2_IC50_nM_L 5HT2_KI_nM_L	→	Sigma_KI_nM_H	17	24	79	71	0.84

Left and *Right* are the *antecedent* and the *consequent* of the rule and *#Left* and *#Right* are the number of molecules in agreement with the profile defined by the *Left* term and the *Right* term, respectively.

Figure 10.1 Suggestion to synthesize molecule **1** (*right*) from molecule **6r** (*left*).

These are the molecules with the desired σ-1 profile and the undesired D2 profile: 18a, 18aj, 18az, 18ba, 18bc, 18w, 18z, 6r, 6t, 6u, 6v, 6w, 6x, and 6y.

Considering the two left terms of rule {C}, N = 1 and R2 = benzyl, the suggestions are based on a match and swap algorithm. For the molecules matching R2 = benzyl, it will suggest to replace N = 0 or N = 2 by N = 1. In a similar manner, it will suggest to synthesize a compound with a benzyl group at the R2 substitution point for molecules matching N = 1. This is the case for molecule 6r: KEM suggests one novel compound by substituting the R2 4-(trifluoromethyl)benzyl group with a benzyl (Figure 10.1). This novel compound corresponds to molecule **1** in the original paper [25]. It is reported with a K_I value of 1 nM against the σ-1 receptor and an IC_{50} of 1550 nM for the D2 receptor in agreement with our prediction of a high sigma and a low D2 [24]. The origin of the suggestion can be tracked down: it is supported by the 21 examples highlighted in gray in Table 10.1.

10.3.1.3 Optimization of σ-1 with Selectivity over D2 and 5HT2

The second step of our multiobjective experience is aimed at optimizing high activity toward the σ-1 receptor with dual selectivity over the D2 and the 5HT2 receptors. When the second selectivity endpoint (5HT2) is incorporated in the optimization process, of the 30 rules (confidence ≥ 80% and support ≥ 10) none are consistent with the full desired profile. A compromise consists in extracting rules matching a partial profile, that is, what are the rules that satisfy two of the three objectives? At this stage, the rules that satisfy a partial profile might be explored in more detail. If we consider the three rules listed in Table 10.3, they are consistent with a high σ-1 and a low D2 profile, but miss the selectivity over the 5HT2 receptor. Figure 10.2 shows the repartition of the three 5HT2 intervals (H, M, and L) for the molecules supporting rules {A}, {B}, and {C}. The number of compounds belonging to each interval is also reported above each bar. Of the 25 compounds supporting rule {C}, the majority (56%) has a medium range activity for 5HT2 and another 20% are in the low range activity (Figure 10.2). In other words, 76% of these molecules do not show a high IC_{50} toward the serotonine receptor. One compromise could be to consider that an activity in the medium range is acceptable for 5HT2, so rules {A}, {B}, and {C} become valid hypothesis for the profile [high σ, low D2, low/medium 5HT2]. An alternative option consists in redefining two intervals for 5HT2: "H" as before and "not_H" by combining L and M. A set of 19 molecules with the following selectivity profiles are selected for optimization: [high σ, high D2, high 5HT2] or [high σ, high D2, medium 5HT2].
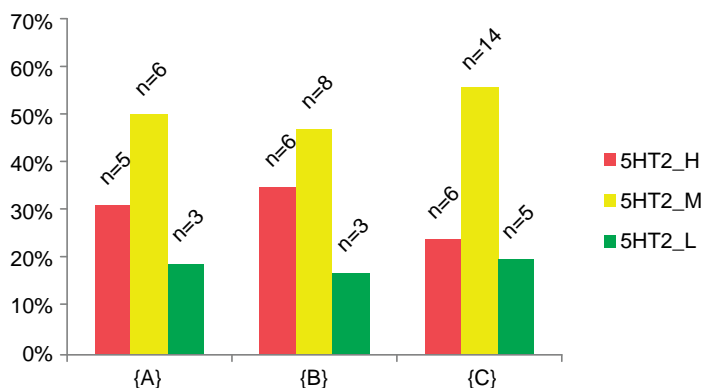
Figure 10.2 Representation of the confidence values (*Y*-axis) for the three rules {A}, {B}, and {C}, as defined in Table 10.2.

In this context, 18 novel molecules are suggested by KEM (Table 10.4). Two suggestions **mol3** and **mol18** correspond to molecules, described in the original paper, **7b** and **1**, respectively. The prediction is validated for compound **7b** as it is reported with a high K_I toward the σ-1 receptor (8 nM), a low IC_{50} toward the D2 receptor (5658 nM), and a medium IC_{50} for the 5HT2 receptor (323 nM) [25]. Molecule **1** matches the profile of a high K_I toward σ-1 and a low IC_{50} for D2 as already described, but it does not match the profile of a low or medium IC_{50} for the

Table 10.4 List of the 18 suggestions (novel structures).

ID	M	N	X	R1	R2
mol1	0	0	[*]C(O)[*]	[*]F	[*]CCCC(c1ccccc1)
mol2	0	1	[*]C(O)[*]	[*]F	[*]C(C1CC1)
mol3	0	1	[*]C(O)[*]	[*]F	[*]C(c1ccccc1)
mol4	0	1	[*]C(O)[*]	[*]F	[*]CC(c1ccc(Cl)cc1)
mol5	0	2	[*]C(O)[*]	[*]F	[*]C(c1ccccc1)
mol6	0	1	[*]C(O)[*]	[*]Cl	[*]C(C1CC1)
mol7	0	1	[*]C(O)[*]	[*]F	[*]C(C1C(C)C1)
mol8	0	1	[*]C(O)[*]	[*]F	[*]C(C1(C)C(Cl)(Cl)C1)
mol9	0	1	[*]C(O)[*]	[*]F	[*]C(c1ccc(C(F)(F)(F))cc1)
mol10	0	1	[*]C(O)[*]	[*]F	[*]C(c1ccc(F)cc1)
mol11	0	1	[*]C(O)[*]	[*]F	[*]CCC1=CNC2=CC=CC=C12
mol12	0	1	[*]C(O)[*]	[*]F	[*]CC(c1ccc(F)cc1)
mol13	0	1	[*]C(O)[*]	[*]F	[*]CC(c1ccc(C(F)(F)(F))cc1)
mol14	1	1	[*]C(O)[*]	[*]F	[*]CCCC(c1ccccc1)
mol15	1	1	[*]C(O)[*]	[*]F	[*]CC(c1ccccc1)
mol16	1	1	[*]C(O)[*]	[*]F	[*]C(c1ccc(OCc2ccccc2)cc1)
mol17	1	1	[*]C(O)[*]	[*]F	[*]Cc1cccc2ccccc12
mol18	0	1	[*]C(=O)[*]	[*]F	[*]C(c1ccccc1)

serotonin receptor (IC_{50} of 20 nM). The validation of the prediction for the other 16 novel structures will require experimental investigation.

10.4
Discussion

The data mining of a SAR table focuses on extracting the knowledge present in the data set in order to design novel active compounds. This is the principle of the rule-based method implemented in KEM: it is capable of rapidly extracting specific combinations of chemical fragments with known positive impact on a desired bioactivity profile. In general, this technique is well suited for an exhaustive analysis of combinations present in the data sets containing three or more diversity points and one or several endpoints. For a smaller data set, it ensures that the knowledge is extracted systematically with no user predefined biased query. The system is able to suggest direct chemical modifications for synthesis. The suggestions are based on unexplored novel combinations of known fragments.

In the first example, the substitution leads to a novel compound predicted to be compatible with the desired profile (a strong sigma binder selective over the dopamine receptor). This reflects an ideal situation when hypothesis consistent with all the objectives are extracted. The second example illustrates the analysis of the same data set with the incorporation of selectivity over the 5HT2 serotonin receptor. As an additional constraint is added, there is no consensus rule that matches the full desired profile, highlighting the difficulty of multiple objective optimization when the number of endpoints increases. In both applications, we show that suggestions derived from a selected subset of meaningful association rules are relevant hypothesis for optimization.

KEM is able to capture the SAR of a chemical series and to describe it as a set of rules, that is, the set of 30 rules in the first example. Only 6 of the 30 rules are consistent with the desired objective. The remaining 24 rules highlight R-groups or combinations of R-groups that are specific for a single endpoint. For example, one such rule associates the presence of a methylcyclopropyl at R2 and a low IC_{50} for the D2 receptor with a support of 29 molecules and a confidence of 90%. But this nitrogen substitution alone is not sufficient for the simultaneous optimization of the sigma activity (only 62% of the examples are reported with a high sigma KI). Similarly, the combination of a fluorine atom at the R1 position and a chain length of 1 carbon as the N-linker are associated with a high sigma KI for 35 molecules with a confidence of 90%. Only 46% of these examples reach the desired low IC_{50} for the dopamine receptor. These rules characterize the presence of concurrent local SAR models and they illustrate the complexity of the multiobjective problem.

When no consensus rules are compatible with a multiobjective profile, an alternative option consists in considering rules with lower confidence and then using the novel molecules to experimentally confirm or not the related hypothesis. The choice of the minimum threshold values for confidence and support will impact both the number of novel suggestions and their chance of achieving the objectives.

The minimum confidence threshold value will be directly linked to the predictive power of the rule. The selection of high confidence rules will strongly support the objectives. Nevertheless, at the early stage of a project, rules with lower confidence, around 50%, for example, might be used for exploring a hypothesis with the aim of validating it or not. When targeting simultaneous multiple objectives, the user may have to trade off with partial fitting as shown in the second example here. This will be the case in most projects as the molecules move forward the development path and more extensive profiling is required with incorporation of ADME/Tox targets.

Recently, a potential role in cancer cell apoptosis has been suggested for sigma ligands after evidencing that sigma receptors are overexpressed in rapidly proliferating cells [29]. This will certainly raise interest for the potential development of sigma ligands in oncology that will require the definition of an alternative polypharmacology profile compared to the profile previously identified for the development of antipsychotic drugs.

Various strategies can be followed for the optimization of a medicinal chemistry series, depending on the amount of data available. The rules extracted from KEM are based on chemical fragments that can be linked to the chemical building blocks used for synthesis. The interpretability of rules in general is complemented here by the use of descriptors (chemical fragments) that have real meaning to the medicinal chemists. A similar easy interpretation was reported with SMIREP that uses rules based on structural fragments to predict activity classification [14]. SMIREP does not consider R-groups listed in the SAR table, but rather it automatically produces a list of fragments present in the molecules. In the latter case, fragments may be more difficult to translate into synthesizable molecules.

Wassermann and Bajorath introduced a method called directed R-group combination graph (DRGC) to extract SAR patterns in a series of analogues [30]. The patterns defined as combinations of R-groups and characterized by their impact on the activity might then be used to design novel analogues. The examples presented in the paper are limited to small data sets (maximum size of 54 molecules) and a single activity endpoint, but the approach should be applicable to larger data sets.

Tamura *et al.* proposed an original approach where the molecules are first clustered based on their chemical similarity and then detailed information is extracted from each cluster based on expert rules and statistical methods [31]. One major advantage of the method is its ability to handle multiple targets and multiple scaffolds generated from a maximum common substructure search for each cluster. As compared to other methods, it is able to work with a large variety of complex fragments and it is thus applicable to large diverse databases such as the NCI anti-HIV data set.

The examples discussed above differ in both the way the structural fragments are generated and the way the rules are extracted. These methods use the knowledge rules to predict the activity class for new chemicals (MCASE and SMIREP), to extract the SAR (DrugPharmer), or to suggest novel molecules (KEM and DRGC graph). The substructure fragments are based on input table scaffold and R-groups for KEM and DRGC graph, while more general fragments are extracted by Tamura *et al.* The "matched molecular pairs" (MMP) method considers small structural differences

associated with activity changes to suggest novel compounds for synthesis [32]. This latter method does not require the identification of scaffold and it generates chemical fragments and hypothesis that are easily interpretable by the chemist.

With the hybrid structures generated by KEM, no new R-group chemistry is suggested in the case of a single core as all the fragments come from the existing knowledge database. However, it is possible to describe the R-groups by calculating properties of these R-groups and to use KEM to generate rules that suggest modifying properties. Although this enables to go beyond existing R-groups in the database, it also puts the burden of translating desired "property" into a new R-group on the chemist.

10.5
Conclusions

The examples presented here illustrate successful application of KEM rule-based methods for the development of novel σ-1 ligands selective over the D2 and 5HT2 receptors. With the aim of developing new cancer drugs, the kinome space has been intensively explored and it has been showed that many approved compounds are therapeutically relevant by targeting multiple protein kinases [33–35]. We believe that detailed rule-based analysis of polypharmacology data sets, either at the lead optimization level like here or in the broader profiling experiments, will help decipher the complex relations existing between multiple compounds and multiple targets.

The discovery of a new target for an existing drug may be derived from its polypharmacology profile and may help generate novel therapeutic applications in a field known as drug repositioning [5,36]. The discovery of a side effect might provide evidence of an undesirable "off-target" activity, thus leading to a new project with an adapted strategy where the initial "antitarget" becomes a new target of interest for another therapeutic indication. It is thus of great importance to identify the appropriate combination of targets by integrating the information derived from extensive analysis of systems pharmacology [37,38].

References

1 Paolini, G.V., Shapland, R.H.B., van Hoorn, W.P., Mason, J.S., and Hopkins, A.L. (2006) Global mapping of pharmacological space. *Nature Biotechnology*, **24** (7), 805–815.

2 Kresja, C.M., Horvath, D., Rogalski, S.L., Penzotti, J.E., Barbosa, F., and Migeon, J. (2003) Predicting ADME properties and side-effects: the BioPrint approach. *Current Opinion in Drug Discovery & Development*, **6**, 470–480.

3 Bolton, E., Wang, Y., Thiessen, P.A., and Bryant, S.H. (2008) Integrated platform of small molecules and biological activities. *Annual Reports in Computational Chemistry*, **4**, 217–241.

4 Chen, B., Wild, D., and Guha, R. (2009) PubChem as a source of polypharmacology. *Journal of Chemical Information and Modeling*, **49**, 2044–2055.

5 Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J., and Shoichet, B.K.

(2007) *Nature Biotechnology* **25** (2), 197–206.

6 Hu, Y. and Bajorath, J. (2010) Molecular scaffolds with high propensity to form multi-target activity cliffs. *Journal of Chemical Information and Modeling*, **50**, 500–510.

7 Hopkins, AL. (2008) Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, **4** (11), 682–690.

8 Kola, I. and Landis, J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nature Reviews. Drug Discovery*, **3**, 711–715.

9 Crumb, W. and Cavero, I.I. (1999) QT interval prolongation by non-cardiovascular drugs: issues and solutions for novel drug development. *Pharmaceutical Science & Technology Today*, **2** (27), 270–280.

10 Xie, L., Evangelidis, T., and Bourne, PE. (2011) Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. *PLOS Computational Biology*, **7** (4), e1002037.

11 Agrawal, R., Imielienski, T., and Swami, A. (1993) Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ACM Press, New York, pp. 207–216.

12 Wolohan, P.R.N., Akella, L.B., Dorfman, R.J., Nell, P.G., Mundt, S.M., and Clark, R.D. (2006) Structural unit analysis identifies lead series and facilitates scaffold hopping in combinatorial chemistry. *Journal of Chemical Information and Modeling*, **46**, 1188–1193.

13 Priss, U. (2007) Formal concept analysis in information science. *Annual Review of Information Science and Technology*, **40** (1), 521–543.

14 Karwath, A. and DeRaedt, L. (2006) SMIREP: predicting chemical activity from SMILES. *Journal of Chemical Information and Modeling*, **46**, 2432–2444.

15 Ghemtio, L., Devignes, M.D., Smail-Tabbone, M., Souchet, M., Leroux, V., and Maigret, B. (2010) Comparison of three preprocessing filters efficiency in virtual screening: identification of new putative LXRB regulators as a test case. *Journal of Chemical Information and Modeling*, **50**, 701–715.

16 Jullian, N. and Afshar, M. (2008) Novel rule-based method for multi-parametric multi-objective decision support in lead optimization using KEM. *Current Computer-Aided Drug Design*, **4**, 35–45.

17 Lounkine, E., Auer, J., and Bajorath, J. (2008) Formal concept analysis for the identification of molecular fragment combinations specific for active and highly potent compounds. *Journal of Medicinal Chemistry*, **51**, 5342–5348.

18 Auer, J. and Bajorath, J. (2008) Distinguishing between bioactive and modeled compound conformations through mining of emerging chemical patterns. *Journal of Chemical Information and Modeling*, **48**, 1747–1753.

19 Lounkine, E., Stumpfe, D., and Bajorath, J. (2009) Molecular formal concept analysis for compound selectivity profiling in biologically annotated databases. *Journal of Chemical Information and Modeling*, **49**, 1359–1368.

20 Tsunoyama, K. *et al.* (2008) Scaffold hopping in drug discovery using inductive logic programming. *Journal of Chemical Information and Modeling*, **48**, 949–957.

21 Raymond, J.W., Watson, I.A., and Mahoui, A. (2009) Rationalizing lead optimization by associating quantitative relevance with molecular structure modification. *Journal of Chemical Information and Modeling*, **49**, 1952–1962.

22 Baskin, I. and Varek, A. (2008) Building a chemical space based on fragment descriptors. *Combinatorial Chemistry & High Throughput Screening*, **11** (8), 661–668.

23 Lewell, X.Q., Judd, D.B., Watson, S.P., and Hann, M.M. (1998) RECAP – restrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of Chemical Information and Computer Sciences*, **38**, 511–522.

24 Klopman, G. and Tu, M. (1999) Diversity analysis of 14156 molecules tested by the National Cancer Institute for anti-HIV activity using the quantitative structure–activity relational expert

system MCASE. *Journal of Medicinal Chemistry*, **42**, 992–998.

25 Gilligan, P.J., Cain, G.A., Chistos, T.E., Cook, L., Drummond, S., Johnson, A., Kergaye, A.A., McElroy, J.F., Rohrbach, K.W., Schmidt, W.K., and Tam, S.W. (1992) Novel piperidine σ ligands as potential antipsychotic drugs. *Journal of Medicinal Chemistry*, **35**, 4344–4361.

26 Brown, J.L. (1991) The Markush challenge. *Journal of Chemical Information and Modeling*, **31** (1), 2–4.

27 Afshar, M., Lanoue, A., and Sallantin, J. (2006) Multiobjective/multicriteria optimization and decision support in drug discovery. *Comprehensive Medicinal Chemistry*, **4**, 767–774.

28 Liquiere, M. and Sallantin, J. (1998) Structural machine learning with Galois lattices and graphs. *Proceedings of the 1998 International Conference on Machine Learning*, Morgan Kaufmann, pp. 305–313.

29 Pal, K., Pore, S.K., Sinha, S., Janardhanan, R., Mukhopadhyay, D., and Banerjee, R. (2011) Structure–activity study to develop cationic lipid-conjugated haloperidol derivatives as a new class of anticancer therapeutics. *Comprehensive Medicinal Chemistry*, **54**, 2378–2390.

30 Wassermann, A.M. and Bajorath, J. (2012) Directed R-group combination graph: a methodology to uncover structure–activity relationship patterns in a series of analogues. *Journal of Medicinal Chemistry*, **55**, 1215–1226.

31 Tamura, S.Y., Bacha, P.A., Gruver, H.S., and Nutt, R.F. (2002) Data analysis of high-throughput screening results: application of multidomain clustering to the NCI anti-HIV data set. *Journal of Medicinal Chemistry*, **45**, 3082–3093.

32 Griffen, E., Leach, A.G., Robb, G.R., and Warner, D.J. (2011) Matched molecular pairs as a medicinal chemistry tool. *Journal of Medicinal Chemistry*, **54**, 7739–7750.

33 Knight, Z.A., Lin, H., and Shokat, K.M. (2010) Targeting the cancer kinome through polypharmacology. *Nature Reviews Cancer*, **10** (2), 130–137.

34 Bamborough, P., Drewry, D., Harper, G., Smith, G.K., and Schneider, K. (2008) Assessment of kinome space and its implications for kinase drug discovery. *Journal of Medicinal Chemistry*, **51**, 7898–7914.

35 Morphy, R. (2010) Selectively nonselective kinase inhibition: striking the right balance. *Journal of Medicinal Chemistry*, **53**, 1413–1437.

36 Jarvis, LM. (2006) Teaching an old drug new tricks. *Chemical & Engineering News*, **84**, 52–55.

37 Scheiber, J., Chen, B., Milik, M., Sukuru, S.C.K., Bender, A., Mikhailov, D., Whitebread, S., Hamon, J., Azzaoui, K., Urban, L., Glick, M., Davies, J.W., and Jenkins, J.L. (2009) Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *Journal of Chemical Information and Modeling*, **49**, 308–317.

38 Berger, S.I. and Iyengar, R. (2009) Network analyses in systems pharmacology. *Bioinformatics (Oxford, England)*, **25** (19), 2466–2472.

11
Data Mining Using Ligand Profiling and Target Fishing

Sharon D. Bryant and Thierry Langer

11.1
Introduction

The "magic bullet" concept of hitting a target responsible for a disease with a drug molecule tailored to act as a selective agent has been a therapeutic goal since Ehrlich [1] and one of the driving forces in modern drug discovery for several decades. With the rise of structural biology and molecular pharmacology, and the shift from *in vivo* to *in vitro* models in the initial evaluation of biological effects of molecules, the aim of obtaining absolute target specificity had become a goal that seemed within reach. However, there is evidence that drugs interact with many physiological targets and that polypharmacology bears essential importance on therapeutic efficacy [2–6]. In this light, discovering compounds exhibiting the "right" selectivity profile, that is, interaction with several targets or target hubs in a converging biological pathway, has become the holy grail in drug development. Recent examples in the kinase field illustrate this new paradigm. Although imatinib (Gleevec) and sunitinib (Sutent) were designed to be selective, later they were found to be more promiscuous than initially thought [7,8], which could explain why these molecules are successful therapeutically. As recently pointed out, searching for selectively nonselective kinase inhibitors when striking the right balance can deliver drug candidates with superior efficacy compared to inhibitors with high specificity for a single kinase [9].

As a result of the increasingly stringent regulatory environment, another trend in drug discovery has emerged within the last decades. Known as drug repurposing, it involves the use of old, already approved drugs for new indications (targets and diseases). A recent review reported success using this approach in the fields of pediatrics and pediatric hematology–oncology [10]. In fact, repurposing marketed drugs or compounds in the development of alternative indications is not a new concept in the pharmaceutical industry. Over the years, this strategy was realized serendipitously, whereas only recently systematic approaches based on computational analyses materialized. An interesting example of the latter involves the identification of a new target for sorafenib, known inhibitor of the tyrosine protein and Raf kinases. Using pharmacophore-based *in silico* screening, the multidrug resistance target ABDG2 was identified as a potential target and biological testing

confirmed that sorafenib indeed inhibited this target [11]. The number of publications detailing novel systematic approaches for computational drug repurposing discovery has grown significantly and they have been cited in several reviews [12–19]. In parallel to this growth, companies specializing in computational drug repurposing have also emerged.

Within this context, the prediction of drug polypharmacology has become an interesting, albeit highly challenging, task inspiring numerous efforts to characterize drug–target associations [20–25]. Although phenotypic and chemical similarities among molecules have been used by several groups to identify compounds with multiple targets [26,27], others have linked shared side effects to compounds for profile prediction [28]. In a seminal paper, Shoichet, Roth, and coworkers demonstrated that using a statistic-based chemoinformatics approach, it became possible to extend easily accessible associations in order to obtain a recalculated map able to predict new off-target effects [29]. For example, using this approach, they mapped 332 targets by 290 drugs interacting with at least two of the targets, thus resulting in a network with 972 connections [29].

Clearly, data mining using ligand profiling has become a useful tool in the hands of scientists involved in the search for new drugs or for optimization of lead compounds. In this chapter, we present an overview of the most useful *in silico* ligand profiling methods along with several application examples.

11.2
In Silico Ligand Profiling Methods

For a long time, computational chemists have been challenged by medicinal chemists and biologists to predict the affinity of a small organic ligand to a particular protein target, in order to provide decision support for hit to lead development and for lead optimization. Nowadays, the question no longer involves the prediction of a compound's affinity for a target, but rather it is to profile a ligand against a large collection of macromolecular targets and provide answers to queries, such as (i) to which proteins could the compound bind (target fishing) and (ii) what could the pharmacological profile of the compound look like (ligand profiling). Interestingly, virtual target profiles have been reported to outperform classical standard chemical similarity measurements in assessing whether two compounds are similar or not [30].

Principally, two different situations are taken into account when discussing *in silico* ligand profiling and target fishing methods. The more favorable situation occurs when the three-dimensional (3D) structure of the target protein is known. In this case, approaches commonly referred to as structure-based methods are applied. However, even in the absence of structural information about the target, so-called ligand-based methods can be employed for *in silico* profiling. In a recent review [31], the five most useful structure-based approaches for ligand profiling and target fishing are described, and listed by decreasing maturity level as follows: (i) protein–ligand docking, (ii) structure-based pharmacophore profiling, (iii) 3D binding site similarity-based profiling, (iv) profiling with protein–ligand fingerprints, and (v) ligand descriptor–based profiling. In the following sections, both structure- and

ligand-based methods for ligand profiling will be discussed, and the major advantages and issues of such approaches will be highlighted.

11.2.1
Structure-Based Ligand Profiling Using Molecular Docking

During the last two decades, molecular software docking programs have been used to support drug discovery extensively, and numerous examples of successful applications in different domains have been described. Docking programs aim to predict the three-dimensional binding orientation (the "pose") of a ligand in a protein binding site and compute a binding energy. However, after several years of application, it became clear that underlying scoring functions used by these docking programs did not accurately predict binding free energies and therefore did not precisely rank-order molecules by their predicted affinities [32]. Despite such disappointments in the computational chemistry community, structure-based docking tools and related scoring approaches are still a focus of development for guiding experimental design. Although protein–ligand docking as a virtual screening engine has been used to find novel ligands for pharmacologically interesting targets such as protein kinases or G protein-coupled receptors (GPCRs), the opposite paradigm, namely, finding novel targets for a pharmacologically interesting ligand (named also inverse screening) has been applied relatively only recently. For such an approach, a database of protein–ligand binding sites is required, together with a robust docking and scoring protocol, as well as a postprocessing script for ranking resulting targets by decreasing binding energy/scoring values.

An early example of the application of inverse docking involves natural product profiling [33]. In their study, Do *et al.* used the selnergy [34] profiling protocol based on the FlexX docking method [35] to identify cyclooxygenase type-1 (COX1), cyclooxygenase type-2 (COX-2), and peroxisome proliferator-activated receptor gamma (PPARγ), among about 400 manually selected proteins, as targets for meranzin, a major component isolated from *Limnocitrus littoralis* (Miq.) Swingle.

Another ligand profiling protocol named TarFisDock [36] based on the DOCK algorithm [37] offers a target database containing 698 protein structures covering 15 therapeutic areas and was available as a Web-based service for inverse docking studies. The authors of TarFisDock published an article indicating that the top 2 and 10% of candidate proteins predicted by their program to bind vitamin E respectively covered 30 and 50% of either already reported verified targets or those suggested by experiments. In addition, 30 and 50% of experimentally confirmed targets for 4*H*-tamoxifen appeared among the top 2 and 5% of the TarFisDock-predicted candidates, respectively [36]. Similarly, positive results were obtained when a combinatorial library of phthalimide derivatives was docked into a set of six guanine phosphoribosyl transferases (GPRT) [38]. Small molecular weight inhibitors of GPRT from the protozoan parasite *Giardia lamblia* were identified as potential starting points for the development of new antiparasitic agents [38].

Several success stories involving protein–ligand docking approaches were covered in a review by Rognan in 2010 [31]. However, one relevant study from 2001 not

mentioned in this review involved the prediction of potential side effect targets of small molecules [39]. Using a database of protein cavities developed from the Protein Data Bank (PDB) [40], docking was conducted by a procedure involving multiple conformer shape-matching alignments of a molecule to a cavity followed by molecular mechanics torsion optimization and energy minimization on both the molecule and the protein residues at the binding region. Potential protein targets were selected by molecular mechanics energy evaluation and, when applicable, a target's binding competitiveness against other ligands that bind to the same receptor site in at least one PDB entry was analyzed. The authors reported that 83% of the experimentally known toxicity and side effect targets for these drugs were predicted correctly, and only five experimentally confirmed protein targets were missed.

Since 2010, other studies involving target prediction using inverse docking methods have been published. For example, a large-scale *in silico* profiling experiment based on a 2D matrix of docking scores among all possible protein structures in yeast and humans and 35 important drugs from different therapeutic areas was published recently [41]. Another interesting study identified new potential direct targets of 2,2-bis(hydroxy-methyl)-3-quinuclidinone (PRIMA-1), a compound well known for its ability to restore mutant p53's tumor suppressor function that drives apoptosis in several types of cancer cells [42]. The authors identified oxidosqualene cyclase (OSC) as highest ranking human protein in their study, and used PRIMA-1 in combination with the known OSC inhibitor Ro 48–8071 to significantly reduce the viability of BT-474 and T47-D breast cancer cells. Thus, for the first time, Ro 48–8071 was shown to act as a potent agent in killing human breast cancer cells.

Protein–ligand docking for ligand profiling or target fishing can be considered an established method with many documented success stories. However, the major problem with docking-based *in silico* target screening remains in the preparation of a heterogeneous collection of binding cavities despite considerable progress in data curation and harmonization in the PDB [43] and other derived data collections, such as the sc-PDB [44]. Several steps, such as defining the position of polar hydrogen atoms, assigning a relevant tautomeric state, and atom typing of cofactors, are not straightforward for automatization. Moreover, the influence of the binding site on the ligand ionization state is difficult to anticipate. Modifying on the fly the protonation state of the ligand according to the binding site context would require a prior storage of all possible ionization states of both the ligand and the protein and is currently not available in most docking tools. In addition, the overall utility of such an approach, at this stage of history, is still somehow limited by the heavy computational effort needed. Authors recently estimated computation times ranging approximately between 25 h and 30 days for profiling one ligand, depending on the underlying docking technique [45].

11.2.2
Structure-Based Pharmacophore Profiling

In the ligand-based drug design, feature-based pharmacophore creation from a set of bioactive molecules is a frequently chosen and well-validated approach. In

contrast, structure-based pharmacophores lacked the reputation for a long time to be an alternative or at least a supplement to docking techniques. Nevertheless, screening using 3D pharmacophores as filters bears the advantage of being much faster than docking, which is of utmost importance especially in parallel and inverse screening campaigns. In addition, pharmacophores transparently provide the investigator with relevant information that is used by the screening algorithms to characterize the ligand–macromolecular interaction.

In fact, the concept of pharmacophores has been used in medicinal chemistry drug discovery research for many years [46]. It is based on the assumption that the molecular recognition of a biological target shared by a family of compounds can be described by a set of common features that interact with a set of complementary sites on the biological target. Such features are quite general and encompass hydrogen bond donors and acceptors, positively and negatively charged or polarizable groups, hydrophobic regions, and metal–ion interactions. Interestingly, they represent precisely the same elements that medicinal chemists imagine when designing compounds. However, the three-dimensional relationship between each feature in a pharmacophore model is another key component of the pharmacophore description, and sometimes is missing in the medicinal chemist's imagination, since most of them have been trained extensively to think about structures in two dimensions. Furthermore, as the feature-based pharmacophore concept is closely linked with the widely used principle of bioisosterism, it is quite understandable that medicinal chemists have largely adopted it when designing a bioactive compound series.

Although the first definition of the pharmacophore as a concept had been attributed to Paul Ehrlich, Van Drie [47] published that it was Kier who introduced the concept in the late 1960s and early 1970s [48,49] when describing common molecular features of ligands of important central nervous system receptors, followed by Höltje in 1974 [50]. In these early studies, the pharmacophore models were deduced manually and supported through the use of simple interactive molecular graphics visualization programs. Later, the diversity and steadily growing complexity of molecular structures that characterize drug discovery led to the development of sophisticated computer software programs for the determination, manipulation, and use of pharmacophore models. A considerable number of books, book chapters, and reviews [51–58] on this approach exist today. The most recent and comprehensive volume was published by Leach *et al.* [59] Still, the basic concept of pharmacophore models as simple geometric representations of key molecular interactions remains unchanged. Such feature-based pharmacophore models have found extensive use in medicinal chemistry for hit and lead identification as well as subsequent lead to candidate optimization. Although pharmacophore representations provide excellent design templates and are useful for rapidly screening compound libraries for new leads, like all other *in silico* approaches, they cannot explain everything about binding of ligand to the biological target.

The pharmacophore modeling software LigandScout [60] was developed initially as a rapid and efficient tool for automatic interpretation of ligand–protein interactions and subsequent transformation of this information into 3D chemical feature-based pharmacophore models. As an extension of this approach, parallel

pharmacophore-based ligand screening was introduced for the first time as an innovative *in silico* method to predict the potential biological activities of compounds [61]. Using LigandScout, the entire PDB was processed, and a pharmacophore database of validated structure-based pharmacophore models covering the most important targets and antitargets of interest for drug discovery was developed. In addition, validated ligand-based pharmacophore models for proteins that lack information about their three-dimensional structure were included.[1] Another pharmacophore-based approach has been described recently by Meslamani *et al.*, where a total of 68 056 structure-based pharmacophores were automatically derived from 8166 high-resolution protein–ligand complexes [62].

Screening ligands against a library of 3D pharmacophore models allows rapid profiling of compounds even before they are synthesized and drastically enhances the library design process. Several studies about pharmacophore-based ligand profiling [63–65] and target fishing [66–70] have been published so far. The results indicate that these methods can compete well with other approaches based on scalar descriptors or on molecular docking and scoring [71,72]. In addition, having the advantage that information can be traced back easily from virtual space toward molecular structure information, pharmacophore-based modeling and *in silico* profiling provide the solid basis for successful medicinal chemistry decision support.

11.2.3
Three-Dimensional Binding Site Similarity-Based Profiling

A common assumption in chemogenomics is that similar receptors bind similar ligands [73]. Analyzing binding site similarities in unrelated proteins can therefore be considered a possible route for finding new targets for existing ligands. Following the paradigm that similar ligands will bind to similar cavities, function and ligands for a novel protein may be deduced from structurally similar ligand cavities. Since binding site similarities are difficult to detect from amino acid sequences, efficient 3D computational methods for quantifying global or local similarities between protein cavities are a prerequisite. Such methods have been developed in the last decade and are the basis for ligand profiling by binding site similarity comparison [74].

Basically, all methods described for binding site similarity analysis follow the same three-step flowchart. First, the structures of the proteins to be compared are parsed into meaningful 3D coordinates in order to reduce the complexity of a pairwise comparison. Typically, only key residues/atoms are considered and described by a limited number of points, which are labeled according to pharmacophoric, geometric, and/or chemical properties of their neighborhood. Second, the resulting patterns are structurally aligned using, for example, clique detection [75,76] and geometric hashing methods [77,78], to identify the maximum number of equivalent points. Finally, a scoring function is applied to quantify the number of

---

1) PharmacophoreDB. The entire collection of structure and ligand based 3D pharmacophore models is available from Inte:Ligand GmbH, Austria. http://www.inteligand.com.

aligned features in the form of root-mean-square deviation (rmsd), residue conservation, or physicochemical property conservation.

One of the earliest binding cavity similarity-guided explanations of unexpected ligand cross-reactivity was made by Weber *et al.* [79]. Starting from the observation that many (COX-2) inhibitors share a common arylsulfonamide moiety with carbonic anhydrase (CA) inhibitors, already known COX-2 inhibitors were tested for binding to various CA isoforms and nanomolar binding affinities were revealed. A rational explanation of this cross-reactivity was obtained by comparing COX-2 inhibitor subpockets with a set of 9433 cavities using the CavBase descriptors [76]. For two out of three subcavities, CA subpockets were retrieved among the top-scoring entries. However, no global similarity could be detected between entire ligand binding pockets of both enzymes. Likewise, a systematic pairwise comparison of the staurosporine binding site of the proto-oncogene Pim-1 kinase with 6412 druggable protein–ligand binding sites [44] using the SiteAlign algorithm [80] suggested that the ATP-binding site of synapsin I (an ATP-binding protein regulating neurotransmitter release in the synapse) may recognize the pan-kinase inhibitor staurosporine [81].

In silico profiling using a comparison of protein–ligand binding sites is a rapid method that presents the noticeable advantage of taking into account protein space only. It avoids sampling the ligand conformational space and thus a potentially incorrect definition of a ligand's bioactive conformation. On the other hand, this approach can only be applied if the binding site comparison method is not too sensitive to variations in the atomic coordinates and in fact it has been found to be quite sensitive to the quality of the protein–protein alignment utilized for scoring binding site similarities. In cases where only *local* and *not global* similarities can be detected between two unrelated protein cavities, the approach is likely to fail. Although it is not mandatory, the binding site reference with which all active sites are compared should be cocrystallized with a drug-like ligand to avoid induced fit phenomena. This however is a problem common with any structure-based approach. The inherent fuzziness embedded in some binding site comparison tools renders them less sensitive to moderate induced fits (up to 3.0 Å rmsd deviations) [80] when compared to docking or pharmacophore searches.

11.2.4
Profiling with Protein–Ligand Fingerprints

Complex information describing binding topology of a small molecule to a biomolecular target can be encoded in fingerprints that represent vectors in which both the ligand and the protein cavity are encoded. Several successful ligand profiling studies using protein–ligand fingerprints are summarized in a recent review [31].

It is interesting to note that such combined fingerprints usually outperform corresponding ligand fingerprints when mining the target–ligand space [82]. Since this descriptor can be applied to a much larger number of receptors (e.g., orphan targets) than the ligand-based fingerprints, they represent a novel and promising way to directly screen protein–ligand pairs in chemogenomic applications. Whether

predictions are qualitative (binary association) or quantitative (pK_i), no information is derived about the putative binding mode of the protein–ligand under consideration. This is a considerable difference from the three already described approaches, but is not necessarily a drawback. Hence, ligand profiling does not require outputting structural information about protein–ligand complexes. Simply a target list that is as short and specific as possible is available to guide experimental validation. Only prospective applications can provide an indication of whether usage of protein–ligand fingerprints really represents a breakthrough when compared to either pure structure-based or ligand-based methods.

11.2.5
Ligand Descriptor-Based *In Silico* Profiling

Impressive progress has been achieved in solving X-ray structures of a large variety of protein receptors, including more recently membrane proteins such as GPCRs. However, at this stage, the three-dimensional structures of the majority of existing pharmacologically relevant targets remain unsolved. Therefore, purely ligand-based profiling methods still are of considerable relevance. Although pharmacophore-based profiling approaches can be used in both the absence and presence of target structural information, pure ligand descriptor-based approaches can be applied if the target structure is not known, or even if the target itself remains to be discovered. So-called ligand-centric approaches are still actively developed and used to predict the polypharmacological profile of bioactive compounds [83–86]. For recent reviews on ligand-based inverse screening approaches, the reader is referred to Refs [20,87].

Poroikov's PASS was probably the first attempt to predict a large variety of bioactivity profiles on a large scale. The initial publication dates from 1995 [88], and at present the system allows prediction of more than 4000 categories of biological activity, including pharmacological effects, mechanisms of action, toxic and adverse effects, interaction with metabolic enzymes and transporters, influence on gene expression, to name a few. The basis of PASS predictions is knowledge about structure–activity relationships of more than 260 000 compounds with known biological activities. QSAR models for each activity type have been generated and evaluated with a 95% average accuracy of prediction, derived by a leave-one-out cross-validation procedure for the whole PASS training set. The system is available as a web service[2] and there are many citations related to this approach.

Another successful ligand-based approach to mining the chemogenomics space has been reported in a seminal article by Gregori-Puigjané *et al.*. The authors describe the practical implementation and validation of their ligand descriptor-based approach to investigate the chemogenomic space of drugs. *In silico* target profiling of 767 drugs against 684 targets of therapeutic relevance was presented. The results revealed that drugs targeting aminergic G protein-coupled receptors displayed the most promiscuous pharmacological profiles. The authors detected cross-pharmacologies between aminergic GPCRs and the sigma, NMDA, and 5-HT$_3$

receptors, thus finding an augmentation of the potential promiscuity of predominantly analgesic, antidepressant, and antipsychotic drugs.

11.3
Summary and Conclusions

Clearly, data mining using ligand profiling and target fishing is a hot topic in modern pharmaceutical research. Chemogenomics, that is, the identification of all possible drugs for all possible targets, has emerged as a new paradigm in drug discovery in which efficiency in the compound design and optimization processes is achieved through the gain and utilization of already targeted knowledge [89]. As targeted knowledge resides at the interface between chemistry and biology, computational tools aimed at integrating the chemical and biological spaces currently play and will continue to play a central role in chemogenomics. Library design will profit from such approaches [90] as well as hit to lead expansion and lead optimization processes through prioritization of compounds with desired predicted pharmacological profiles with low risks due to potential off-target-mediated toxicity. We are at the beginning of a new age, where chemogenomics information is rapidly available even in open access formats [91,92] to everybody involved in the field of drug discovery. It is up to us to use this wealth of information in the most intelligent way.

References

1 Ehrlich, P. (1911) The theory and practice of chemotherapy. *Folia Serologica*, **7**, 697–714.

2 Roth, B.L., Sheffler, D.J., and Kroeze, W.K. (2004) Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nature Reviews. Drug Discovery*, **3**, 353–359.

3 Nobeli, I., Favia, A.D., and Thornton, J.M. (2009) Protein promiscuity and its implications for biotechnology. *Nature Biotechnology*, **27**, 157–167.

4 Marona-Lewicka, D. and Nichols, D.E. (2007) Further evidence that the delayed temporal dopaminergic effects of LSD are mediated by a mechanism different than the first temporal phase of action. *Pharmacology Biochemistry and Behavior*, **87**, 453–461.

5 Marona-Lewicka, D. and Nichols, D.E. (2009) WAY 100635 produces discriminative stimulus effects in rats mediated by dopamine D4 receptor activation. *Behavioural Pharmacology*, **20**, 114–118.

6 Peterson, R.T. (2008) Chemical biology and the limits of reductionism. *Nature Chemical Biology*, **4**, 635–638.

7 Rix, U. *et al.* (2007) Chemical proteomic profiles of the BCR-ABL inhibitors imatinib, nilotinib, and dasatinib reveal novel kinase and nonkinase targets. *Blood*, **110**, 4055–4063.

8 Hopkins, A.L. (2007) Network pharmacology. *Nature Biotechnology*, **25**, 1110–1111.

9 Morphy, R. (2010) Selectively nonselective kinase inhibition: striking the right balance. *Journal of Medicinal Chemistry*, **53**, 1413–1437.

10 Blatt, J. and Corey, S.J. (2013) Drug repurposing in pediatrics and pediatric hematology concology. *Drug Discovery Today*, **18**, 4–10.

11 Wei, Y. *et al.* (2012) New use for an old drug: inhibiting ABCG2 with sorafenib. *Molecular Cancer Therapeutics*, **11**, 1693–1702.

12 Adronis, C. *et al.* (2011) Literature mining, ontologies and information visualization for drug repurposing. *Briefings in Bioinformatics*, **12**, 357–368.

13 Dudley, J.T., Deshpande, T., and Butte, A.J. (2011) Exploiting drug–disease relationships for computational drug repositioning. *Briefings in Bioinformatics*, **12**, 303–311.

14 Haupt, V.J. and Schroeder, M. (2011) Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Briefings in Bioinformatics*, **12**, 312–326.

15 Moriaud, F. *et al.* (2011) Identify drug repurposing candidates by mining the Protein Data Bank. *Briefings in Bioinformatics*, **12**, 336–340.

16 Sanseau, P. and Koehler, J. (2011) Editorial: computational methods for drug repurposing. *Briefings in Bioinformatics*, **12**, 301–302.

17 Sardana, D. *et al.* (2011) Drug repositioning for orphan diseases. *Briefings in Bioinformatics*, **12**, 346–356.

18 Swamidass, S.J. (2011) Mining small-molecule screens to repurpose drugs. *Briefings in Bioinformatics*, **12**, 327–335.

19 Xu, K. and Coté, T.R. (2011) Database identifies FDA-approved drugs with potential to be repurposed for treatment of orphan diseases. *Briefings in Bioinformatics*, **12**, 341–345.

20 Bajorath, J. (2008) Computational analysis of ligand relationships within target families. *Current Opinion in Chemical Biology*, **12**, 352–358.

21 Oprea, T.I., Tropsha, A., Faulon, J.L., and Rintoul, M.D. (2007) Systems chemical biology. *Nature Chemical Biology*, **3**, 447–450.

22 Newman, D.J. (2008) Natural products as leads to potential drugs: an old process or the new hope for drug discovery? *Journal of Medicinal Chemistry*, **51**, 2589–2599.

23 Siegel, M.G. and Vieth, M. (2007) Drugs in other drugs: a new look at drugs as fragments. *Drug Discovery Today*, **12**, 71–79.

24 Miller, J.R. *et al.* (2009) A class of selective antibacterials derived from a protein kinase inhibitor pharmacophore. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 1737–1742.

25 Walsh, C.T. and Fischbach, M.A. (2009) Repurposing libraries of eukaryotic protein kinase inhibitors for antibiotic discovery. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 1689–1690.

26 Young, D.W. *et al.* (2008) Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nature Chemical Biology*, **4**, 59–68.

27 Wagner, B.K. *et al.* (2008) Large-scale chemical dissection of mitochondrial function. *Nature Biotechnology*, **26**, 343–351.

28 Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J., and Bork, P. (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

29 Keiser, M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature*, **462**, 175–181.

30 Bender, A. *et al.* (2006) "Bayes affinity fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept? *Journal of Chemical Information and Modeling*, **46**, 2445–2456.

31 Rognan, D. (2010) Structure-based approaches to target fishing and ligand profiling. *Molecular Information*, **29**, 176–187.

32 Ferrara, P. *et al.* (2004) Assessing scoring functions for protein–ligand interactions. *Journal of Medicinal Chemistry*, **47**, 3032–3047.

33 Do, Q.T. *et al.* (2007) Reverse pharmacognosy: identifying biological properties for plants by means of their molecule constituents: application to meranzin. *Planta Medica*, **73**, 1235–1240.

34 Do, Q.T. *et al.* (2005) Reverse pharmacognosy: application of selnergy, a new tool for lead discovery. the example of epsilon-viniferin. *Current Drug Discovery Technologies*, **2**, 161–167.

35 Rarey, M. *et al.* (1996) A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, **261**, 470–489.

36 Li, H. *et al.* (2006) TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Research*, **34**, W219–W224.

37 Ewing, T.J. *et al.* (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design*, **15**, 411–428.

38 Aronov, A.M. *et al.* (2001) Virtual screening of combinatorial libraries across a gene family: in search of inhibitors of *Giardia lamblia* guanine phosphoribosyltransferase. *Antimicrobial Agents and Chemotherapy*, **45**, 2571–2576.

39 Chen, Y.Z. and Ung, C.Y. (2001) Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand–protein inverse docking approach. *Journal of Molecular Graphics & Modelling*, **20**, 199–218.

40 Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**, 235–242.

41 Lee, M. and Kim, D. (2012) Large-scale reverse docking profiles and their applications. *BMC Bioinformatics*, **13** (Suppl. 17), S6.

42 Grinter, S.Z. *et al.* (2011) An inverse docking approach for identifying new potential anti-cancer targets. *Journal of Molecular Graphics & Modelling*, **29**, 795–799.

43 Henrik, K. *et al.* (2008) Remediation of the Protein Data Bank archive. *Nucleic Acids Research*, **36**, D426–D433.

44 Kellenberger, E. *et al.* (2006) sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *Journal of Chemical Information and Modeling*, **46**, 717–727.

45 Zheng, R., Chen, T., and Lu, T. (2011) A comparative reverse docking strategy to identify potential antineoplastic targets of tea functional components and binding mode. *International Journal of Molecular Sciences*, **12**, 5200–5212.

46 Langer, T. (2011) Pharmacophores in drug research. *Molecular Informatics*, **29**, 470–475.

47 Van Drie, J.H. (2007) Monty Kier and the origin of the pharmacophore concept. *Internet Electronic Journal of Molecular Design*, **6**, 271–279.

48 Kier, L.B. (1967) Molecular orbital calculation of preferred conformations of acetylcholine, muscarine, and muscarone. *Molecular Pharmacology*, **3**, 487–494.

49 Kier, L.B. (1970) Receptor mapping using MO theory, in *Fundamental Concepts in Drug–Receptor Interactions* (eds J.F. Danielli, J.F. Moran, and D.J. Triggle), Academic Press, New York.

50 Höltje, H.D. (1974) Molekular orbital berechnungen zur struktur des muscarin-pharmakophors. *Archiv der Pharmazie (Weinheim)*, **307**, 969–972.

51 Güner, O.F. (ed.) (2000) *Pharmacophore Perception, Development and Use in Drug Design*, International University Line, La Jolla, CA.

52 Mason, J.S., Good, A.C., and Martin, E.J. (2001) 3-D pharmacophores in drug discovery. *Current Pharmaceutical Design*, **7**, 567–597.

53 Langer, T. and Hoffmann, R.D. (eds) (2006) *Pharmacophores and Pharmacophore Searches*, Wiley-VCH Verlag GmbH, Weinheim, Germany.

54 Van Drie, J.H. (2003) Pharmacophore discovery: lessons learned. *Current Pharmaceutical Design*, **9**, 1649–1664.

55 Langer, T. and Wolber, G. (2004) Pharmacophore definition and 3D searches. *Drug Discovery Today*, **1**, 203–207.

56 Langer, T. and Hoffmann, R.D. (2006) Pharmacophore modelling: applications in drug discovery. *Expert Opinion on Drug Discovery*, **1**, 261–267.

57 Martin, Y.C. (2007) Pharmacophore modeling: 2. Applications, in *Comprehensive Medicinal Chemistry II, Computer-Assisted Drug Design*, vol. **4** (eds J. B. Taylor, D.J. Triggle, and J.S. Mason), Elsevier, Amsterdam, pp 515–536.

58 Laggner, C. *et al.* (2008) Pharmacophore-based virtual screening in drug discovery, in *Chemoinformatics: An Approach to Virtual Screening* (eds A. Varnek and A. Tropsha), Royal Society of Chemistry, Cambridge, UK, pp 76–119.

59 Leach, A.R. *et al.* (2010) Three-dimensional pharmacophore methods in drug discovery. *Journal of Medicinal Chemistry*, **53**, 539–558.

60 Wolber, G. and Langer, T. (2005) LigandScout: 3D pharmacophores derived

from protein-bound ligands and their use as virtual screening filters. *Journal of Chemical Information and Modeling*, **45**, 160–169.

61 Steindl, T.M. *et al.* (2006) Parallel screening: a novel concept in pharmacophore modelling and virtual screening. *Journal of Chemical Information and Modeling*, **46**, 2146–2157.

62 Meslamani, J. *et al.* (2012) Protein–ligand-based pharmacophores: generation and utility assessment in computational ligand profiling. *Journal of Chemical Information and Modeling*, **52**, 943–955.

63 Steindl, T.M. *et al.* (2007) Parallel screening and activity profiling with HIV protease inhibitor pharmacophore models. *Journal of Chemical Information and Modeling*, **47**, 563–571.

64 Markt, P. *et al.* (2007) Pharmacophore modeling and parallel screening for PPAR ligands. *Journal of Computer-Aided Molecular Design*, **21**, 575–590.

65 Schuster, D. *et al.* (2010) Predicting cyclooxygenase inhibition by three-dimensional pharmacophoric profiling. Part I: model generation, validation and applicability in ethnopharmacology. *Molecular Informatics*, **29**, 75–86.

66 Sciabola, S. *et al.* (2010) High-throughput virtual screening of proteins using GRID molecular interaction fields. *Journal of Chemical Information and Modeling*, **50**, 155–169.

67 Rollinger, J.M. *et al.* (2009) *In silico* target fishing for rationalized ligand discovery exemplified on constituents of *Ruta graveolens*. *Planta Medica*, **75**, 195–204.

68 Duwensee, K. *et al.* (2011) Leoligin, the major lignan from Edelweiss, activates cholesteryl ester transfer protein. *Atherosclerosis*, **219**, 109–115.

69 Schuster, D. *et al.* (2011) Pharmacophore-based discovery of FXR-agonists. Part I: model development and experimental validation. *Bioorganic and Medicinal Chemistry*, **19**, 7168–7180.

70 Nashev, L.G. *et al.* (2012) Virtual screening as a strategy for the identification of xenobiotics disrupting corticosteroid action. *PLoS ONE*, **7**, e46958.

71 Gregori- Puigjané, E. and Mestres, J. (2008) A ligand-based approach to mining the

chemogenomic space of drugs. *Combinatorial Chemistry & High Throughput Screening*, **11**, 669–676.

72 Paul, N. *et al.* (2004) Recovering the true targets of specific ligands by virtual screening of the Protein Data Bank. *Proteins*, **54**, 671–680.

73 Klabunde, T. (2007) Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *British Journal of Pharmacology*, **152**, 5–7.

74 Kellenberger, E., Schalon, C., and Rognan, D. (2008) How to measure the similarity between protein ligand-binding sites? *Current Computer-Aided Drug Design*, **4**, 209–220.

75 Kinoshita, K., Furui, J., and Nakamura, H. (2002) Identification of protein functions from a molecular surface database, eF-site. *Journal of Structural and Functional Genomics*, **2**, 9–22.

76 Schmitt, S., Kuhn, D., and Klebe, G. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *Journal of Molecular Biology*, **323**, 387–406.

77 Gold, N.D. and Jackson, R.M. (2006) Fold independent structural comparisons of protein–ligand binding sites for exploring functional relationships. *Journal of Molecular Biology*, **355**, 1112–1124.

78 Shulman-Peleg, A., Nussinov, R., and Wolfson, H.J. (2004) Recognition of functional sites in protein structures. *Journal of Molecular Biology*, **339**, 607–633.

79 Weber, A. *et al.* (2004) Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *Journal of Medicinal Chemistry*, **47**, 550–557.

80 Schalon, C. *et al.* (2008) A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins*, **71**, 1755–1778.

81 Defranchi, E. *et al.* (2010) Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements. *PLoS ONE*, **5**, e12214.

82 Weill, N. and Rognan, D. (2009) Development and validation of a novel protein–ligand fingerprint to mine

chemogenomic space: application to G protein-coupled receptors and their ligands. *Journal of Chemical Information and Modeling*, **49**, 1049–1062.

83 Mestres, J. *et al.* (2006) Ligand-based approach to *in silico* pharmacology: nuclear receptor profiling. *Journal of Chemical Information and Modeling*, **46**, 2725–2736.

84 Nettles, J.H. *et al.* (2006) Bridging chemical and biological space: 'target fishing' using 2D and 3D molecular descriptors. *Journal of Medicinal Chemistry*, **49**, 6802–6810.

85 Keiser, M.J. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nature Biotechnology*, **25**, 197–206.

86 AbdulHameed, M.D.M. *et al.* (2012) Exploring polypharmacology using a ROCS-based target fishing approach. *Journal of Chemical Information and Modeling*, **52**, 492–505.

87 Ekins, S., Mestres, J., and Testa, B. (2007) *In silico* pharmacology for drug discovery: methods for virtual ligand screening and profiling. *British Journal of Pharmacology*, **152**, 9–20.

88 Filimonov, D.A. *et al.* (1995) Computer-aided prediction of biological activity spectra of chemical substances on the basis of their structural formulae: computerized system PASS. *Experimental and Clinical Pharmacology (Russian)*, **58**, 56–62.

89 Mestres, J. (2004) Computational chemogenomics approaches to systematic knowledge-based drug discovery. *Current Opinion in Drug Discovery & Development*, **7**, 304–313.

90 Cases, M. *et al.* (2005) Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family. *Current Topics in Medicinal Chemistry*, **5**, 763–772.

91 Carrascosa, M.C., Massaguer, O.L., and Mestres, J. (2012) PharmaTrek: a semantic web explorer for open innovation in multitarget drug discovery. *Molecular Informatics*, **31**, 537–541.

92 Lui, X. *et al.* (2010) PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Research*, **38**, W609–W614.

Part Four
System Biology Approaches

12
Data Mining of Large-Scale Molecular and Organismal Traits Using an Integrative and Modular Analysis Approach

Sven Bergmann

12.1
Rapid Technological Advances Revolutionize Quantitative Measurements in Biology and Medicine

Understanding how genotypes impact on phenotypes is one of the central goals of biology. In the last decade, immense technological advances have been made, allowing measuring the properties and the behavior of biological systems with great accuracy. Whole-genome sequencing not only provides an inventory of genes, including their regulatory regions, but has also paved the way for high-throughput technologies that elucidate their genetic variability across populations and their transcriptional response subject to different genetic and environmental conditions.

Until recently, microarrays have played a leading role in cost-efficient measurements of genome-wide profiles of gene expression as well as genetic variants, like single-nucleotide polymorphisms (SNPs) and copy number variants (CNVs). This technology is now being superseded by ultrahigh-throughput sequencing, which allows not only affordable whole-genome sequencing but also very accurate molecular phenotyping of the transcriptome (RNA-seq), methylome (Me-seq), and interactome (ChIP-seq).

While microarrays originally were used most extensively to quantify gene expression in many model organisms and clinical samples, in recent years SNP arrays have become the tool of choice for genotyping large sample collections that usually had already been phenotyped for various parameters. This includes in particular human clinical cohorts, whose individuals had been classified as cases or controls for a particular disease or had been measured for almost any imaginable trait.

12.2
Genome-Wide Association Studies Reveal Quantitative Trait Loci

The presence of phenotypic and genotypic data for large collection of individuals or samples made possible the so-called genome-wide association studies (GWAS) searching for correlations between genetic markers and phenotypic traits. Such GWAS can be viewed as unsupervised data mining, since the genetic markers are

generally capturing a large portion of all frequent genetic variants and can be used to impute most of those that have not been directly measured. The standard procedure is to perform a statistical test for each genetic marker (whether directly measured or imputed) for its association (i.e., correlation) with the trait of interest. Markers that are proximal on the DNA are often correlated (or in the so-called linkage disequilibrium), such that GWAS typically (but not always) identify entire regions of several markers known as "trait loci." The motivation for this unsupervised approach is that statistically significant associations could reveal new candidate genes for playing a role in the phenotype of interest and that this would eventually lead to a better understanding of the genetic components of diseases and their risk factors, and potentially lead to new therapeutic avenues.

From the many GWAS that were performed in the last years, it became apparent that even well-powered (meta-)studies with many thousands (and even ten thousands) of samples could at best identify a few (dozen) candidate trait loci with highly significant associations. While many of these associations have been replicated in independent studies, each locus explains but a tiny ($<1\%$) fraction of the total genetic variance of the phenotype. Remarkably, all significantly associated loci as features for a combined model still miss out by at least one order of magnitude in explained variance of most phenotypes. Thus, while GWAS already today provide new candidates for disease-associated genes and potential drug targets, very few of the currently identified (sets of) genotypic markers are of any practical use for assessing risk for predisposition to any of the complex diseases that have been studied.

Various solutions to this apparent enigma have been proposed:

- First, it is important to realize that the expected genetic component of many traits has been estimated from studies using data from twins, sometimes several decades ago. These estimates are population-specific, depend on the total phenotypic variance (which depends on the environmental variance), and for many traits a wide range of estimates have been observed across populations. Thus, it has been argued that these estimates may be problematic [1].
- Second, the genotypic information is still incomplete. Most GWAS used microarrays probing only around half a million of SNPs, which is almost one order of magnitude less than 4 million common variants that have been identified from the HapMap [2] CEU panel. While many of these SNPs can be imputed accurately using information on linkage disequilibrium, there still remains a significant fraction of SNPs that are poorly tagged by the measured SNPs. Furthermore, rare variants with a minor allele frequency (MAF) of less than 1% are not accessed at all with SNP chips, but may nevertheless be the causal agents for many phenotypes [3]. Moreover, other genetic variants like copy number variations (CNVs) may also play an important role [4,5].
- Third, it is important to realize that current analyses usually only employ multilinear models considering one SNP at a time with few, if any, covariates, like sex, age, and principal components reflecting population substructures. This obviously covers only a small set of all possible interactions between genetic variants and the environment. Even more challenging is taking into account

purely genetic interactions, since the number of all possible pair-wise interactions already scales like the number of genetic markers squared.

12.3
Integration of Molecular and Organismal Phenotypes Is Required for Understanding Causative Links

Another plausible explanation for the fact that for most organismal phenotypes we can only explain a small fraction of their estimated genetic variance in terms of measured SNPs is related to the complexity of these phenotypes. Indeed, there is a long path from a genetic variant to a phenotype that is observed at the level of the organism (Figure 12.1). A variant nucleotide can have many effects: Exonic variants may disrupt proper transcription by generating a premature stop-codon or can alter an amino acid that is crucial for protein function, while intronic variants may affect splicing. Also, variants outside the transcribed region can modify the level of expression by altering regulatory sites for chromatin state, as well as transcriptional and posttranscriptional regulation. However, whatever the direct effect of a genetic variant is, it is first acting at the cellular level. Thus, the cell and therefore the tissue type play an important role by providing the chemical environment under which the variant is exerting its effect. This environment may differ not only across different cell types but also as a function of the organismal environment (day or night, after or before meals, etc.) or age.

It is important to realize that regulatory networks have evolved to function robustly under external and internal perturbations. Any effect of a genetic variant on the organismal phenotype is propagated through these networks. This propagation, in particular if it involves crucial cellular functions, is likely to induce compensatory effects mediated by regulatory circuits like feedback loops. Moreover, robust functions are often achieved by "backup systems," alternative pathways that can at least partially compensate each other [6–8]. Thus, for the vast majority of variants segregating in a population, the resulting macroscopic phenotypic variation is expected to be small, since variants giving rise to dramatic effects reducing individual fitness would have been quickly purged from the population. Indeed, rare monogenic diseases arise from such variants that alter a gene product (or its expression level) in a way that cannot be compensated. For example, maturity onset diabetes of the young (MODY) [9] refers to any of several hereditary forms of diabetes caused by mutations in a single gene disrupting insulin production. MODY is often referred to as "monogenic diabetes" to distinguish it from other forms of diabetes involving more complex combinations of causes involving multiple genes (i.e., "polygenic") and environmental factors. Such common diseases are usually governed by a large number of variants, each of which has a small, if any, effect, and only a (unfortunate) combination of them can lead to a systemic breakdown of homeostasis.

Thus, in general, it is not surprising that the effects of genetic variability are more pronounced "upstream" at the molecular level than "downstream" at macroscopic
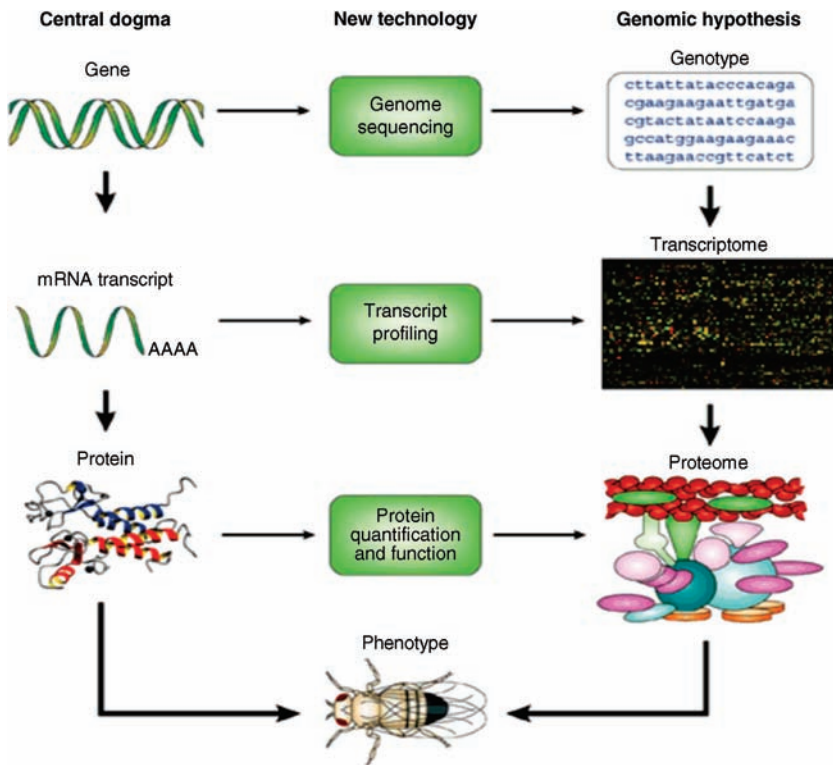
Figure 12.1 From genotype over molecular to organismal phenotype. According to the "Central Dogma" (left), the double-stranded DNA contains coding regions (or "genes," top) from which messenger RNA ("mRNA," middle) is transcribed, which in turn is translated into proteins that play a pivotal role in generating the organismal phenotype (bottom). New high-throughput technologies allow quantifying the information content at each level (middle). This quantification is done at the genomic level comprising all, or a large fraction, of the genome, the transcriptome, and the proteome, respectively (left). The latter two consist of "molecular phenotypes" that are highly dynamic and well regulated. Better quantification of this system as a whole, combined with in-depth knowledge about the individual elements, has been proposed as the prerequisite for understanding the emergence of complex organismal phenotypes. (This figure is taken from http://www.nature.com/nrg/journal/v3/n1/box/nrg703_BX1.html with permission.)

level of the organism. Indeed, recent GWAS for gene expression data have shown that the transcript levels of many genes (in cultured cell lines) can be explained in part by the genotypes of SNPs in the vicinity of these genes [10]. Interestingly, the fraction of the explained variation ($\sim$30%) of such SNPs is much higher than for any of the complex phenotypes that have been considered by the recent GWAS. Similar evidence comes from GWAS with metabolomic phenotypes [11,12] showing that single SNPs may explain up to 12% of the observed variance in some serum metabolite concentrations and up to 28% of certain concentration ratios as a proxy for enzymatic activity [11].

An alternative to the direct association of organismal phenotypes with genotypes is thus to construct and integrate molecular networks defining the molecular states of a system that underlie a particular phenotype or disease. By molecular state we mean a set of molecular phenotypes that take a particular configuration (e.g., a set of metabolites and genes that are all upregulated with respect to some base level). In order to construct these networks and identify those states indicative of particular organismal traits, large cohorts have to be phenotyped at both the molecular and the macroscopic levels. Indeed, several studies characterizing gene networks have demonstrated how genetic loci associated with expression traits can be combined with clinical data to infer causal associations between expression and disease traits [13]. For example, Chen *et al.* [14] reported an approach to uncover the components of coexpression networks that respond to variations in DNA associated with obesity-, diabetes-, and atherosclerosis-related traits. The genetics of gene expression in different tissues and its effect on obesity-related traits were studied by Emilsson *et al.* [15].

12.4
Reduction of Complexity of High-Dimensional Phenotypes in Terms of Modules

Molecular phenotypes, like the aforementioned mRNA and metabolite concentrations, provide much more direct information on the impact of genotypic variation than the resulting organismal phenotypes. However, in general, the number of molecular observables (e.g., the number of genes or metabolites) is much larger. Moreover, their measurements are often noisy. Thus, assigning genes or metabolites into groups and considering the group average have the following advantages:

1) It reduces the complexity of such data, since the number of groups is typically much smaller than the number of individual elements.
2) It may provide biological focus if the individual elements share common features (e.g., genes belonging to the same metabolic pathway).
3) It may provide insights into the structure of the underlying regulatory network (e.g., groups of gene being organized in a hierarchical manner).

The advantages have been well recognized for large-scale gene expression data and a multitude of methods has been developed to identify groups (or "modules") from such data [16]. A general advantage of studying properties of modules, rather than individual elements, relies on a basic principle of statistics: The variance of an average is proportional to $1/N$, where N is the number of (statistical) variables used to compute its value, because fluctuations in these variables tend to cancel each other out. Thus, mean values over the elements of a module or between the elements of different modules are more robust measures than the measurements of each single element alone. This is not only relevant for the noisy expression data produced by microarrays but also for mass-specific quantification of protein or metabolite concentrations.

A common challenge in the analysis of large and diverse collections of molecular profiles lies in the context-dependent nature of regulation. For example, certain genes are only expressed in specific cell types, and their expression levels may depend on

external conditions (e.g., stresses or drug response) or *internal* conditions (e.g., circadian rhythm or developmental stages). Usually genes are coordinately regulated only in specific experimental contexts, corresponding to a subset of the conditions in the data set. Most standard analysis methods classify genes based on their similarity in expression across *all* available conditions. The underlying assumption of uniform regulation is reasonable for the analysis of small data sets, but limits the utility of these tools for the analysis of heterogeneous large data sets for the following reasons: First, conditions irrelevant for the analysis of a particular regulatory context contribute noise, hampering the identification of correlated behavior over small subsets of conditions. Second, genes may participate in more than one function, resulting in one regulation pattern in one context and a different pattern in another. Thus, combinatorial regulation necessitates the assignment of genes to several context-specific and potentially overlapping modules. In contrast, most commonly used clustering techniques yield disjoint partitions, assigning each gene to a unique cluster [16].

12.5
Biclustering Algorithms

To take these considerations into account, molecular profiles should be analyzed with respect to specific subsets. Algorithms for the coclassification of molecular concentrations and conditions (e.g., samples in a study) have been pioneered in the context of expression data [17–23]. Such algorithms usually yield "transcription modules" (another common term is "bicluster") consisting of *sets of coexpressed genes together with the conditions over which this coexpression is observed*. The naïve approach of evaluating expression coherence of all possible subsets of genes over all possible subsets of conditions is computationally infeasible, and most analysis methods for large data sets seek to limit the search space in an appropriate way. For example, Getz *et al.* [19] introduced a variant of biclustering based on the idea to perform standard clustering iteratively on genes and conditions. Their coupled-two-way-clustering procedure is initialized by separately clustering the genes and conditions of the full matrix. Each combination of the resulting gene and condition clusters defines a submatrix of the expression data. Instead of considering all possible combinations, two-way clustering is then applied to all such submatrices in the following iteration. Other biclustering methods, like the Plaid Model [21] and Gene Shaving [22], aim to identify only the most dominant bicluster in the data set, which is then masked in a subsequent run to allow the identification of new clusters. The SAMBA (Statistical-Algorithmic Method for Bicluster Analysis) biclustering method [23] combines graph theory with statistical data modeling. While each method has its advantages and disadvantages [24], a common challenge is their scaling properties in terms of CPU time and memory usage when applied to large data. This has been our main motivation to develop our own clustering tools, which we describe now in more detail.

The *Signature algorithm* (Figure 12.2) [20] was designed to test whether a set of candidate genes exhibits a coherent expression over a subset of the microarray data, thus already taking context-specific regulation into account. These test sets are
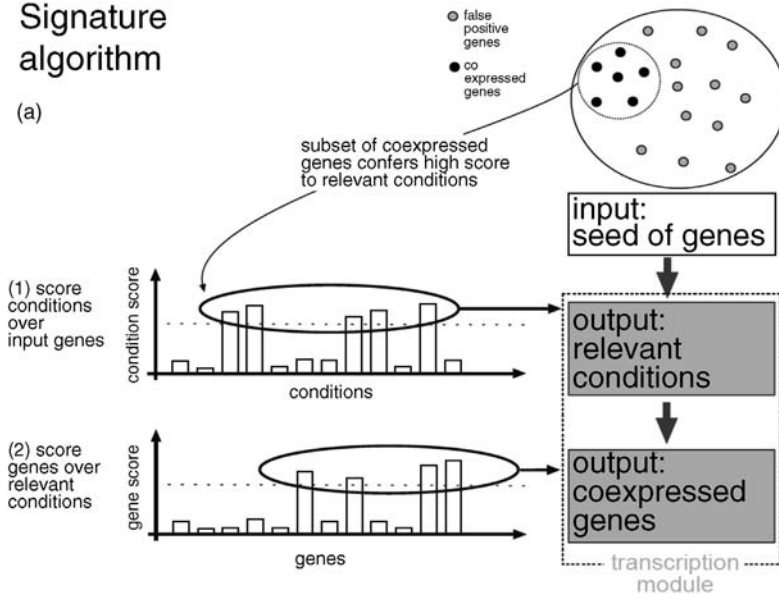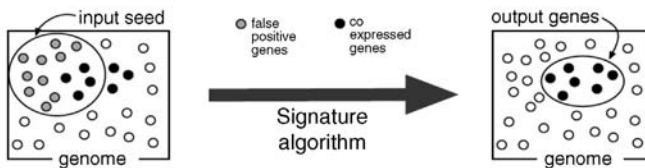
Figure 12.2 The Signature algorithm requires as input a set of genes, some of which are expected to be coregulated based on additional biological information such as a common promoter binding motifs or functional annotation. (a) The algorithm proceeds in two steps: In the first step, this input seed is used to identify the conditions that induce the highest average expression change in the input genes. Only conditions with a score above some threshold are selected. In the second stage of the algorithm, genes that are highly and consistently expressed over these conditions are identified. The result consists of a set of coregulated genes together with the regulating conditions and is termed "transcription module." (b) The output contains only the coregulated part of the input seed, as well as other genes that were not part of the original input but display a similar expression profile over the relevant conditions.

constructed by integration of additional biological data, including functional annotations and regulatory sequence information. In order to provide a more global modular picture of the transcription program, this algorithm was later extended into an iterative scheme (the *Iterative Signature algorithm* (ISA)) (Figure 12.3) that allows an efficient modular decomposition of large-scale expression data (typically tens of thousands of gene probes tested over hundreds of conditions) even in the absence of any *a priori* information [25]. The ISA is one of the state-of-the-art methods for these types of data according to various performance measurements [19,20] and has been

Iterative Signature Algorithm
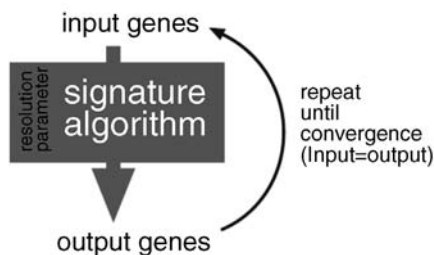
Figure 12.3 The Iterative Signature algorithm (ISA) is an extension of the Signature algorithm and is designed to reveal hierarchies of coregulatory units of varying expression coherence. This approach is applicable also in the absence of biologically motivated seeds, in which case the iterative scheme is initialized by many sets of randomly chosen input genes. The output genes determined by the Signature algorithm are reused as input. This procedure is iterated until input and output converge. Each resulting "transcription module" is self-consistent: Its genes are most coherently coexpressed over the module conditions, which, in turn, induce the most coherent expression of the module genes.

employed for numerous biological studies [19,21–23]. Briefly, the ISA uses a set of data to identify a compendium of modules, consisting of coherently behaving elements (transcripts, protein, and metabolites) as well as the experimental conditions (samples) for which this coherent behavior is the most pronounced. Specifically, by coherent we mean that for a given condition all elements of a module are either all induced or suppressed with respect to some baseline level. The ISA has the following advantages: (i) The elements and samples can be assigned to multiple modules (while standard clustering produces mutually exclusive units). (ii) Requiring only coherent behavior over a subset of samples allows picking up subtle signals of context-specific and combinatorial coregulation, which may be too weak to be extracted from the correlations over all samples that are used by many clustering algorithms. (iii) Since the ISA does not require the calculation of correlation matrices, it is highly efficient computationally and is thus applicable even to very large data sets. We recently made available a comprehensive Bioconductor [26] software package including a highly optimized implementation [27] of the ISA in R as well as a number of tools for module annotation and visualization [28].

12.6
Ping-Pong Algorithm

High-throughput technologies are now used to generate different types of data from the same biological samples. A central challenge lies in the proper integration of such data. To this end, we proposed the concept of *comodules* (CMs), describing coherent patterns across paired data sets and conceive several modular methods for their identification. We proposed the *Ping-Pong algorithm* (PPA) (Figure 12.4) and
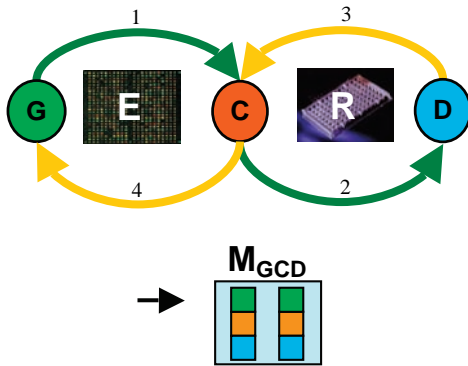
Figure 12.4 The Ping-Pong algorithm starts with a candidate set of genes (G, green) and uses the available expression data E to identify the cell lines (C, orange) for which these genes exhibit a coherent expression (arrow 1). In the next step, the response data R are employed to select drugs (D, blue) that elicit a similar response in these cell lines (arrow 2). This set of drugs is then utilized to refine the set of cell lines by eliminating those that have an incoherent response profile and adding others that behave similarly across these drugs (arrow 3). Finally, this refined set of cell lines is used to probe for genes that are coexpressed in these lines (arrow 4). This alternating procedure is reiterated until it converges to stable sets of genes, cell lines, and drugs. We refer to these sets as comodules M_{GCD} (green, orange, and blue boxes), which generalize the concept of a module from a single to multiple data sets.

other modular schemes for the identification of such comodule. The main idea of the PPA is to extend the iterative scheme of the ISA to two data sets that share one common dimension. For example, a collection of samples that was studied using two different assays would provide such data. In its original application to the drug response [30] and gene expression profiles [31,32] from the NCI60 cell lines, the PPA was used to integrate molecular with cellular phenotypes. However, this is just one of many possible applications. For example, the shared dimension could also be constituted by a set of genes whose expression (or any other feature) is quantified in two different sample collections (see the following sections for details).

12.7
Module Commonalities Provide Functional Insights

Both modules and comodules group elements together. Often biological insight can be gleaned by identifying features that are common to these elements. For example, genes may share similar function if coexpressed or similar binding sites if this coexpression is induced by a common transcription factor. Drugs attributed to the same module may have the same target or interfere with the same pathway. Similarly, metabolites of the same module may be involved in the same or related interactions. Thus, in order to annotate (co-)modules, one needs to identify such commonalities. This has been pioneered for genes. Using the growing body of information on the function of gene products [33–35], it is feasible to provide an initial annotation of transcription modules

based on automated functional enrichment analysis. Specifically, overrepresentation of genes belonging to the same functional category in one module suggests its association with this function. Overrepresentation can be quantified in terms of a *p*-value, based on the total numbers of elements in the category and the module, as well as their intersection. Usually these *p*-values are computed using Fisher's exact test. A number of tools (e.g., FUNSpec [36], MAPPFinder [37], or FatiGO [38]) for online enrichment analysis have been published. Functional categories for many human and mouse genes are provided, for example, by the Gene Ontology (GO) project [39] and associations with metabolic pathways are available at the KEGG [40] database. Although functional annotations are incomplete, and sometimes even wrong, very small *p*-values usually indicate a functional link for the module.

12.8
Module Visualization

Standard hierarchical clustering still remains the default analysis tool for large sets of biological data, despite the limitations of this analysis method for large-scale data [17–23]. One reason for this is that the widely used representation of
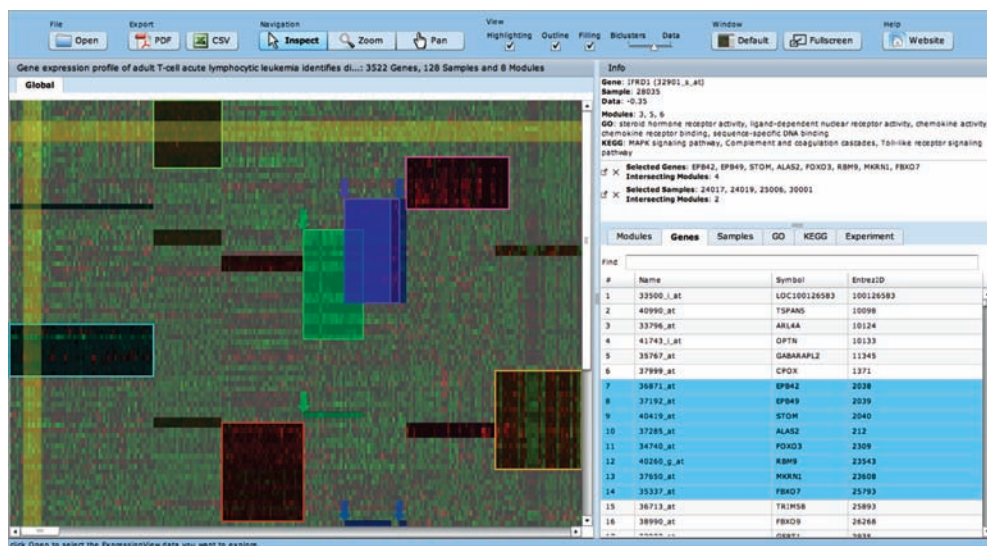
Figure 12.5 Screenshot from ExpressionView (available online at http://www2.unil.ch/cbg/index.php?title=ExpressionView). Similar to the standard biclustering, our tool visualizes expression levels of all the genes in the data set (columns) under many experimental conditions (rows) using a color code. However, the order of genes and conditions has been optimized in order to highlight the coherent expression patterns that are apparent only over a subset of the entire data set (i.e., transcription modules). Genes (as well as conditions or modules) can be selected in the window on the right. Modules are clickable, providing detailed information on their genes and conditions, as well as automated annotation in terms of enriched GO categories and KEGG pathways.

expression data in terms of a reordered color-coded matrix with dendrograms delineating the clusters and their hierarchy has the advantage of being exceedingly simple. In particular, many biologists apparently appreciate that the original expression values (or ratios) are shown in this presentation (somewhat akin to the fact that showing the image of a gel shift experiment is still a must, although quantification software for gels has existed for some time now). Accordingly, we have designed a new visualization tool *ExpressionView* [28] that presents modules as rectangles that denote its genes and arrays on the actual expression data (Figure 12.5). Since it is in general impossible to represent more than two mutually overlapping modules in this manner, we have developed an algorithm that minimizes the fraction of genes or arrays that appear as disconnected module fragments. Thus, our tool maintains the aforementioned simplicity of the common cluster representation, while allowing an intuitive presentation of overlapping groups of genes and arrays.

12.9
Application of Modular Analysis Tools for Data Mining of Mammalian Data Sets

The ISA has been applied for a number of analyses of gene expression data, by both ourselves [20,24,29,41–47] and others. Here, we provide two recent examples, highlighting the usefulness also in the context of limited sample size [43] and the new generation of RNAseq expression data [47]. Subsequently, we briefly review the original application of the PPA, integrating gene expression with drug–response data from the NCI60 panel [29]. Finally, we provide an alternative application, where this approach was used to identify cross-species transcript.

1) *Modular analysis of fibroblast expression profiles increase power to detect dysregulated units in patient samples*

In collaboration with the Reymond laboratory [43], we recently applied the ISA to a collection of 96 public expression profiles from human fibroblasts to establish sets of genes coexpressed in this cell type. We used these transcription modules to identify differential expression of fibroblast samples from individuals with Williams–Beuren syndrome (WBS) with respect to properly matched controls. As it turns out, it is much more powerful to query for differentially expressed modules (i.e., comparing the mean expression across module genes) than the individual genes. In particular, the modular approach has the advantage that much less tests have to be performed, which helps to overcome the burden of multiple hypotheses testing. Another advantage is that any associated module provides a context in terms of its enrichment with functional categories. For example, dysregulated modules in WBS highlighted the role of process related to the extracellular space and immune response in the pathophysiology of WBS. This case study illustrated the usefulness of a modular approach even in the context of a small set of samples by projecting them on modular components of related large-scale data, which are becoming available for many cell types.

2) *Using the ISA to study evolutionary dynamics of mammalian transcriptomes*

Genome analyses can uncover protein-coding changes that potentially underlie the differences between species, although many of the phenotypic differences between species are the result of regulatory mutations affecting gene expression. In this collaborative study [47] headed by Prof. Henrik Kaessmann group (CIG and UNIL), we used high-throughput RNA sequencing to study the evolutionary dynamics of mammalian transcriptomes in 6 major tissues (cortex, cerebellum, heart, kidney, liver, and testis) of 10 species from all major mammalian lineages. The study shed new light on the extent of transcriptome variation between organs and species, as well as potentially selectively driven expression switches that may have shaped specific organ biology. Notably, for the first time we applied the ISA to RNAseq data identifying transcriptional units (modules), including subsets of orthologous genes that have conserved (coherent?) expression patterns across different sets of organs in certain species or lineages.

Information on all the modules in the whole data set and in a primate-specific data set was made available in a searchable database at http://www.unil.ch/cbg/ISA/species. We found 33 organ-specific modules with conserved expression levels among species, 145 modules specific to an organ or organ pair with lineage-specific expression patterns (i.e., only observed in evolutionary closely related species), and 658 modules that show no clear relation to specific phylogenetic groups and/or affect multiple organs [47]. The organ-specific modules were enriched with genes typically involved in processes matching the function of the respective organs. The organ-specific modules with lineage-specific expression patterns provided clues to the organ biology of different mammals. For example, we found 25 nervous tissue modules that evolved distinct expression levels along the major terminal branches of the mammalian phylogeny. Notably, CNS-specific modules in the nonprimate mammals often showed altered expression in both nervous tissues, suggesting a tight functional and evolutionary link between them in mammals. Interestingly, the only lineage with brain-specific expression modules in the primate data set was that of humans. The genes in the largest of the four human-specific brain modules are involved in various neurologic processes, many of which are related to neuron insulation probably reflecting the larger proportion of myelinated axons (white matter) in the human prefrontal cortex compared to that of other primates.

3) *Identification of gene–drug links through integrative modular analysis of NCI60 data sets*

As the initial application of our PPA to real data, we performed a comodule analysis of gene expression and drug–response data from the NCI60 project. For this study, 60 tumor cell lines had been phenotyped using both microarrays [31,48,49] and assays monitoring their growth when subjected to a large number of chemical compounds [50,51]. Thus, each cell line was characterized by two profiles, one for the expression of each gene, and one for its resistance to each drug. A simple way to search for gene–drug links would thus be to connect genes whose expression profile correlates with the response profile of certain drugs (across cell lines). However, as we showed in our paper [29], this

approach has very little power to predict true positives (using DrugBank as reference). Studying different approaches that modularize the two data sets, we showed that the PPA predicts drug–gene associations significantly better than other methods (refer to Figure 1a and b of Ref. [29]). In this setting, comodules contain sets of genes that are coexpressed across some of the cell lines as well as drugs that affect the growth of exactly these cell lines. Candidates for gene–drug interactions are scored by how often and how strongly a given gene–drug pair was attributed to a comodule. For example, if some gene was overexpressed only in cell lines derived from one type of tissue and a certain drug would slow growth only in these cell lines, then this would establish some evidence for a gene–drug link, which however would only rely on the matched tissue specificity and not necessarily on any molecular interactions. However, although some comodules reflected the tissue origin, others grouped cell lines from different origins, in which case the associated gene sets were more likely to be markers of the common dysregulated pathways of the underlying cancers. Clearly, the study could only provide proof of principle for predicting gene–drug links from large-scale data. Nevertheless, already based on this small collection of samples, we could provide interesting new insights into possible mechanisms of action for a wide range of drugs, which in principle suggest new targets and could eventually lead to novel therapies.
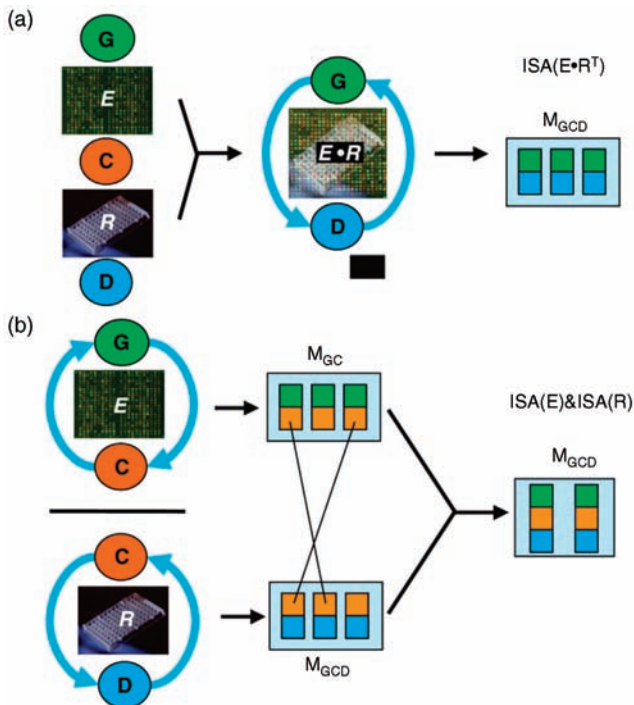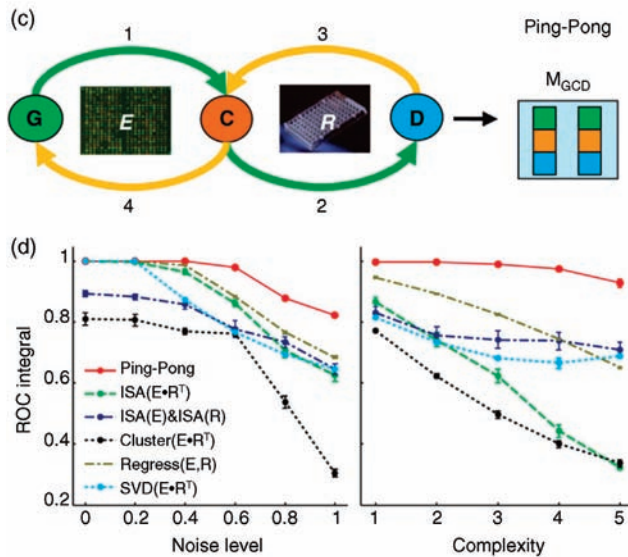
4) *Comparative modular analysis of gene expression in vertebrate organs*

The degree of conservation of gene expression between homologous organs largely remains an open question. Several recent studies reported some evidence in favor of such conservation based on similarity measures computed across all orthologous genes. However, expression levels of many genes are not informative about organ specificity. In order to overcome this limitation, in collaboration with the Robinson lab we used our PPA for the identification of interspecies comodules of organs and genes. In fact, there are two possibilities for the dimension shared by the human and mouse expression data: orthology relations between genes (Figure 12.6a) or organ homology (Figure 12.6b). First, we ran the PPA on the data sets matched through one-to-one orthologous gene pairs. Thus, the comodules consisted of orthologous genes and the mouse and human organs where these genes were expressed. We found that the nervous system (in particular, amygdala, cerebral cortex, hypothalamus, and spinal cord) formed a distinct module in both mouse and human. We also observed several other functionally related comodules, like a comodule containing kidney and liver, a comodule related to the immune system (including lymph node and thymus), a female reproductive system comodule (ovary and uterus), and a respiratory system comodule (lung and trachea). Second, we ran the PPA on the data sets matched through homologous organ groups (HOGs) [52]. The resulting comodules consisted of sets of homologous organs and (potentially different) sets of mouse and human genes with coherent overexpression in these organs, which were enriched in, but not limited to, orthologous genes. In addition to system-specific comodules, we also found organ-specific comodules. Each of these comodules contained genes whose function was clearly related to the respective organs, confirming that overexpressed genes tend to play important roles in a given organ or organ system.
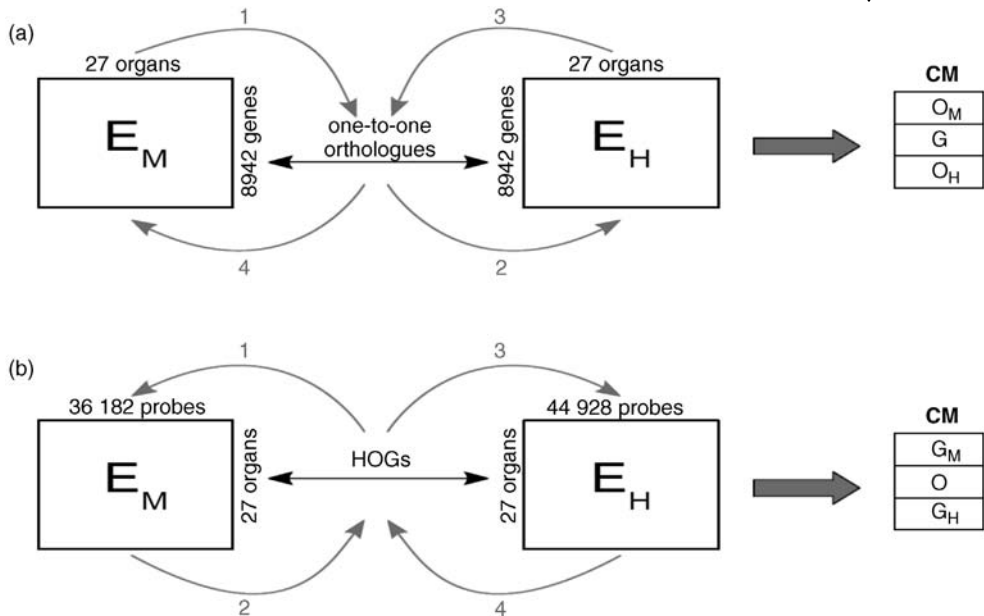
Figure 12.6 Scheme for the two possible ways for the integrative modular analysis of human (E_H) and mouse (E_M) expression data. The two data sets can be matched either by orthology relations between genes (a), giving rise to comodules (CM) consisting of mouse (O_M) and human (O_H) organs, as well as the one-to-one orthologous genes (G) or (b) by organ homology such that the comodules include all mouse (G_M) and human genes (G_H), as well as homologous organ groups (HOGs or O).

12.10
Outlook

High-throughput data acquisition technologies have created the potential for new insights into biological systems. However, the hope to better understand the regulation of these systems and eventually predict their response will only materialize with adequate computational tools to process and visualize the vast amount of data these technologies produce. Medical research is rapidly adopting high-throughput technologies to characterize clinical samples and, therefore, requires appropriate computational tools to interpret the results and use them for diagnostic purposes.

Modules provide the building blocks of the regulatory network. A systems biology approach aims at not only identifying (and annotating) these units but also describing the relationships between them in order to reveal the structure of the entire network. Module relationships can be defined in many ways: the extent of common elements or functional categories enriched for these elements describes static intermodule relations. However, since our (co-)modules typically also include the samples (conditions) over which coherent patterns are detected, they potentially

also allow to provide insights into dynamic relationships between modules. For example, the induction of certain transcription modules may be mutually exclusive. For example, in yeast, modules pertaining to growth are inversely regulated to modules related to stress. Similarly, certain metabolite groups may reflect opposing metabolic states (like hypoxia and hyperoxia).

We believe that a modular approach to large-scale data will be instrumental in reducing the complexity of large-scale biomedical data and also provide a means for integrating different types of omics data. For example, future applications of the Ping-Pong algorithm [29] can extend to the analysis of data sets covering different types of gene regulation (e.g., posttranscriptional modifications or protein expression). In this case, comodules would include sets of genes that are coregulated at multiple instances, as well as subsets of samples where this coregulation occurs.

References

1 Schonemann, P.H. (1997) On models and muddles of heritability. *Genetica*, **99**, 97–108.

2 Frazer, K.A. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

3 Goldstein, D.B. (2009) Common genetic variation and human traits. *The New England Journal of Medicine*, **360**, 1696–1698.

4 McCarroll, S.A. (2008) Extending genome-wide association studies to copy-number variation. *Human Molecular Genetics*, **17**, R135–R142.

5 Beckmann, J.S., Sharp, A.J., and Antonarakis, S.E. (2008) CNVs and genetic medicine (excitement and consequences of a rediscovery). *Cytogenetic and Genome Research*, **123**, 7–16.

6 Gu, Z. *et al.* (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature*, **421**, 63–66.

7 Conant, G.C. and Wagner, A. (2004) Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. *Proceedings of the Royal Society of London Series B*, **271**, 89–96.

8 Kafri, R., Bar-Even, A., and Pilpel, Y. (2005) Transcription control reprogramming in genetic backup circuits. *Nature Genetics*, **37**, 295–299.

9 Rehman, H.U. (2001) Diabetes mellitus in the young. *Journal of the Royal Society of Medicine*, **94**, 65–67.

10 Stranger, B.E. *et al.* (2007) Population genomics of human gene expression. *Nature Genetics*, **39**, 1217–1224.

11 Gieger, C. *et al.* (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genetics*, **4**, e1000282.

12 Suhre, K. *et al.* (2011) A genome-wide association study of metabolic traits in human urine. *Nature Genetics*, **43**, 565–569.

13 Dermitzakis, E.T. (2008) From gene expression to disease risk. *Nature Genetics*, **40**, 492–493.

14 Chen, Y. *et al.* (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature*, **452**, 429–435.

15 Emilsson, V. *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature*, **452**, 423–428.

16 Ihmels, J.H. and Bergmann, S. (2004) Challenges and prospects in the analysis of large-scale gene expression data. *Briefings in Bioinformatics*, **5**, 313–327.

17 Bittner, M., Meltzer, P., and Trent, J. (1999) Data analysis and integration: of steps and arrows. *Nature Genetics*, **22**, 213–215.

18 Cheng, Y. and Church, G.M. (2000) Biclustering of expression data. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, **8**, 93–103.

19 Getz, G., Levine, E., and Domany, E. (2000) Coupled two-way clustering analysis of gene microarray data. *Proceedings of the*

National Academy of Sciences of the United States of America, **97**, 12079–12084.

20 Ihmels, J. *et al.* (2002) Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, **31**, 370–377.

21 Lazzeroni, L. and Owen, A. (2002) Plaid models for gene expression data. *Statistica Sinica*, **12**, 61–86.

22 Hastie, T. *et al.* (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, **1**, Research0003.

23 Tanay, A., Sharan, R., and Shamir, R. (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics (Oxford, England)*, **18** (Suppl. 1), S136–S144.

24 Ihmels, J., Bergmann, S., and Barkai, N. (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics (Oxford, England)*, **20**, 1993–2003.

25 Bergmann, S., Ihmels, J., and Barkai, N. (2003) Iterative Signature algorithm for the analysis of large-scale gene expression data. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, **67**, 031902.

26 Reimers, M. and Carey, V.J. (2006) Bioconductor: an open source framework for bioinformatics and computational biology. *Methods in Enzymology*, **411**, 119–134.

27 Csardi, G., Kutalik, Z., and Bergmann, S. (2010) Modular analysis of gene expression data with R. *Bioinformatics (Oxford, England)*, **26**, 1376–1377.

28 Luscher, A. *et al.* (2010) ExpressionView: an interactive viewer for modules identified in gene expression data. *Bioinformatics (Oxford, England)*, **26**, 2062–2063.

29 Kutalik, Z., Beckmann, J.S., and Bergmann, S. (2008) A modular approach for integrative analysis of large-scale gene-expression and drug–response data. *Nature Biotechnology*, **26**, 531–539.

30 Weinstein, J.N. *et al.* (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science*, **275**, 343–349.

31 Staunton, J.E. *et al.* (2001) Chemosensitivity prediction by transcriptional profiling. *Proceedings of the*

National Academy of Sciences of the United States of America, **98**, 10787–10792.

32 Shankavaram, U.T. *et al.* (2007) Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Molecular Cancer Therapeutics*, **6**, 820–832.

33 Wheeler, D.L. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **33**, D39–D45.

34 Mewes, H.W. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research*, **32** (Database issue), D41–D44.

35 Bairoch, A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Research*, **33**, D154–D159.

36 Robinson, M.D., Grigull, J., Mohammad, N., and Hughes, T.R. (2002) FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 35.

37 Doniger, S.W. *et al.* (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biology*, **4**, R7.

38 Al-Shahrour, F., Diaz-Uriarte, R., and Dopazo, J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics (Oxford, England)*, **20**, 578–580.

39 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**, 25–29.

40 Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Research*, **32**, D277–D280.

41 Bergmann, S., Ihmels, J., and Barkai, N. (2003) How to use genome-wide expression data to learn from yeast about gene regulation in higher eukaryotes. *Yeast (Chichester, England)*, **20**, S276–S276.

42 Bergmann, S., Ihmels, J., and Barkai, N. (2004) Similarities and differences in genome-wide expression data of six organisms. *PLoS Biology*, **2**, E9.

43 Henrichsen, C.N. *et al.* (2011) Using transcription modules to identify

expression clusters perturbed in Williams–Beuren syndrome. *PLoS Computational Biology*, **7**, e1001054.

44 Ihmels, J., Bergmann, S., and Barkai, N. (2003) Yeast expression at your fingertips. *Yeast (Chichester, England)*, **20**, S285–S285.

45 Ihmels, J., Bergmann, S., Berman, J., and Barkai, N. (2005) Comparative gene expression analysis by differential clustering approach: application to the *Candida albicans* transcription program. *PLoS Genetics*, **1**, e39.

46 Ihmels, J. *et al.* (2005) Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science*, **309**, 938–940.

47 Brawand, D. *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.

48 Ross, D.T. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, **24**, 227–235.

49 Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R., and Kohane, I.S. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 12182–12186.

50 Scherf, U. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nature Genetics*, **24**, 236–244.

51 Shi, L.M. *et al.* (2000) Mining and visualizing large anticancer drug discovery databases. *Journal of Chemical Information and Computer Sciences*, **40**, 367–379.

52 Parmentier, G., Bastian, F.B., and Robinson-Rechavi, M. (2010) Homolonto: generating homology relationships by pairwise alignment of ontologies and application to vertebrate anatomy. *Bioinformatics (Oxford, England)*, **26**, 1766–1771.

13
Systems Biology Approaches for Compound Testing

Alain Sewer, Julia Hoeng, Renée Deehan, Jurjen W. Westra, Florian Martin,
Ty M. Thomson, David A. Drubin, and Manuel C. Peitsch

13.1
Introduction

Industry and academia are faced with the challenge of evaluating the health risks associated with long-term exposure to drugs and/or consumer products. This challenge results from the fact that the health risks of new drugs and/or consumer products are assessed over a short period of time, while their health effects may become apparent after long-term exposure only. The focus on long-term effects requires not only the ability to extrapolate the short-term observations to long-term outcomes but also the ability to translate the potential risks identified from a combination of *in vivo* and *in vitro* experimental systems to human populations. Over the last decade, it has become apparent, and increasingly accepted, that understanding the biological mechanisms underpinning the activity of compounds is key to explaining toxic effects and adverse events [1], and is also a necessary component of the knowledge required to predict risk. Rodents are frequently used for *in vivo* testing, and while these species often show major ADMET (absorption, distribution, metabolism, excretion, and toxicity) differences with humans [2], they remain an important source of preclinical data that together with human-derived cell and tissue culture data can facilitate an early assessment of toxicity risk. The challenge formulated above can therefore be subdivided into two distinct aspects: first, the discovery of the biological mechanisms (biological network perturbations) that link short-term observations with long-term outcome(s), and second, the translation of mechanistic observations derived from experimental systems to humans and their populations. Mechanistic understanding is the key to addressing both dimensions.

Typically, epidemiological studies are used to correlate biological impact with long-term outcome, but are not designed to elucidate the causal chain of events that link the two. Furthermore, the systems-based variations within this causal chain represent the basis for the translation of model systems data to human relevance. Systems-based variation includes interspecies and interindividual differences as well as differences in complexity between *in vitro* and *in vivo* systems. Fortunately, due to recent technological advances in molecular biology and computational

science, it is becoming much easier to generate, organize, and interpret high-throughput data and mechanistic information, making it increasingly feasible to account for the majority of the biological system when measuring the effects of a substance or product. The resulting perturbations can be observed in the context of the biological network, thereby allowing a systems-wide understanding of the mechanisms leading to disease.[1] Based on this framework, tools and biomarkers designed to measure the comparative risk of toxic substances can be developed.

Our objective is to establish a set of novel methods that quantitatively measure biological impact [designated as the biologic impact factor (BIF)] from systems-wide data such as transcriptomics and proteomics [3]. These methods use comprehensive, mechanistic causal biological network models as the substrate for quantitative data analysis to identify mechanism of action and assessment of biological impact at the pharmacological/toxicological level. The impact of a specific biological network perturbation caused by a single, or a mixture of, biologically active substance(s) is determined for every described node (molecular entity) of the network, thereby identifying causal mechanistic effects induced by the substance(s). Because our approach is based on the collection and analysis of systems-wide experimental data, this quantitative method is capable of measuring the activation of multiple biological networks that are perturbed by the active substance(s). This enables a quantitative and objective assessment of each molecular entity (or node) in the described biological network(s) that can serve – alone or as part of a signature – as a molecular biomarker closely expressing the overall state of perturbation (activation or inhibition compared to control). Every biological network in the system and its correlation with events such as disease onset or disease progression can be accounted for. Furthermore, our approach enables the quantitative comparison of biological impacts across individuals and species at the mechanistic level representing an advantage over gene-level comparisons that are confounded by genomic/genetic variations. This capability provides a means to facilitate the translation of *in vivo* and *in vitro* model system biology to human biology.

This approach provides both potential predictive capabilities and an explicit listing of all associated assumptions through deterministic scoring algorithms. The algorithms, metrics, and biological network models presented here will be further developed and published in peer-reviewed sources to enable public access. We believe the approach enables the application of network pharmacology and systems biology beyond toxicological assessment [1,4–7] and into areas such as drug development, consumer product testing, and environmental impact analysis. Here, we outline five steps of the strategy and the progress made to date (Figure 13.1) [3].

---

1) These processes embody precisely the "data mining" aspect of the approach described in this chapter: unstructured high-throughput experimental data are mined using biological network models to provide interpretable information in the form of perturbed biological mechanisms, which will be used for compound assessment.
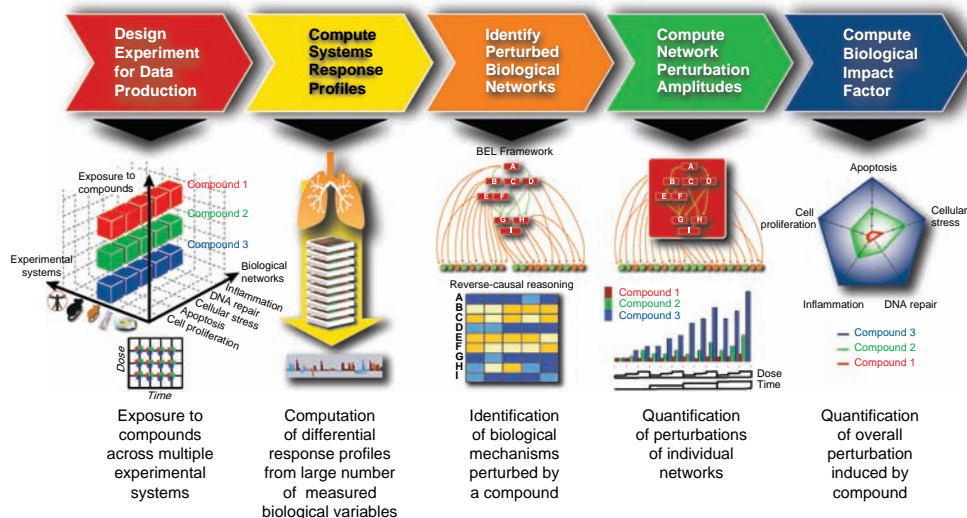
Figure 13.1 The five-step strategy for assessing the biological impact of exposure to compounds from systems-wide data using biological network models. Steps 1–4 provide a framework for the identification of biological networks that are perturbed by short-term exposure to compounds. In step 5, these results are summarized into the biological impact factor (BIF) that can potentially link the observations of early effects with long-term health impacts.

13.2
Step 1: Design Experiment for Data Production

The first of the five steps of the strategy for compound assessment is focused on the generation of high-quality systems-wide experimental data. It takes into consideration a number of components: (i) the experimental system, (ii) the conditions of exposure to the test substance, (iii) the appropriate assays necessary to monitor the effect of exposure on the system (perturbations), (iv) the technology platforms used to measure the perturbations, and finally, (v) the execution of the experiment itself. In this section, the choices made for the above-mentioned components are presented and the reasons underlying them are explained. While fundamental for the data generation (Figures 13.3 and 13.4), the technology platforms are only very briefly discussed.

To assess the potential human health risks associated with a compound, data collected from clinical studies are the most relevant. However, compound-related diseases may take decades to manifest and it is often difficult or even impossible to obtain longitudinal human data sets. Therefore, we have to rely heavily on animal (preclinical) models, as well as on models based on cellular and organotypical (3D) *in vitro* cultures, to generate data. These experimental systems allow us to gain insights into the biological perturbations caused by the compound, identify mechanism-specific biomarkers for use in human studies, and eventually link these mechanisms to the onset of disease for impact assessments. Although the *in vitro* and preclinical systems are known to have many shortcomings, we propose that by
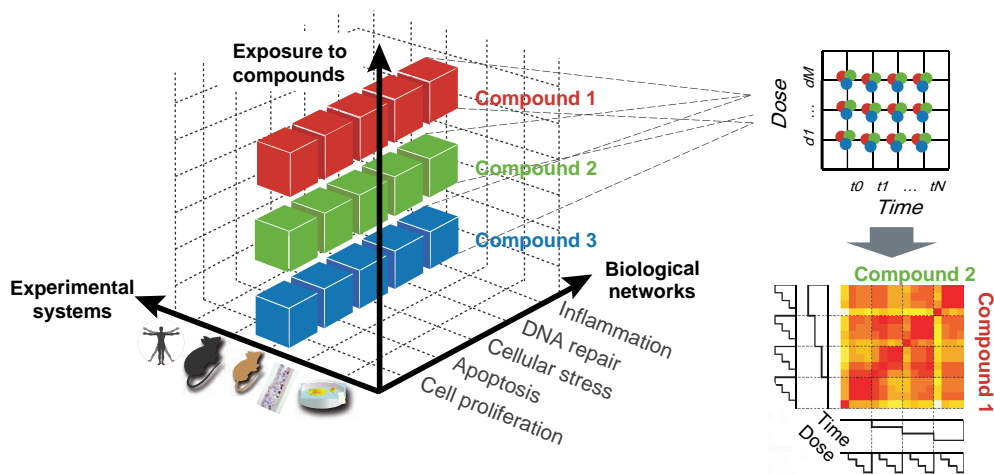
Figure 13.2 The "cubic" experimental design space, accompanied by the standardized time- and dose-dependent exposure regimen and a comparability heat map between two compounds.

taking a systematic approach to their use we can minimize these issues and begin to close the gaps in understanding of *in vivo* human biology.

Our strategy is based on a systematic "cubic" experimental design space, as represented in Figure 13.2. The three dimensions are (i) the experimental systems, (ii) the exposure to the compound(s), and (iii) the (perturbed) biological networks, which will be measured and are often related to particular cellular functions. These three dimensions are discussed here in more detail:

Experimental systems: For optimal utility, the biological experimental systems need to fulfill two complementary purposes. First, the animal models to be used should reproduce at least some features of the human disease and be adequate for the exposure regimens required. Second, the cellular and organotypical systems should reflect the cell types and tissues involved in the disease etiology. Priority should be given to primary cells or organ cultures that recapitulate the *in vivo* biology as much as possible. Furthermore, it is crucial to match each human *in vitro* culture with the equivalent culture derived from the animal models used. These constitute a "translational parallelogram" from animal model to human biology using the matched *in vitro* systems as intermediate hubs [8].

Exposure to the compound: Ideally, the exposure matrix must be well characterized chemically, even in the case of a complex mixture. The goal is to recreate an exposure regimen, that is, the dose and duration of exposure that most realistically mimic the human situation. It is therefore imperative to define a set of standard exposure regimens to be applied systematically to the well-defined experimental systems. Furthermore, appropriate biological assays should be used to obtain time- and dose-dependent data to capture both early and late events and ensure that a representative dose range is covered. If this standardization is fulfilled,

translatability questions involving a fixed treatment can be more easily addressed (Figure 13.2, comparability heat maps between two cubes).

Biological networks: Biological networks constitute the only dimension of the experimental design cube that is not a true experimental factor, but rather a knowledge-driven filter applied to the measurements made during the experiment. Using biological networks as a translatability factor significantly enhances the power of the resulting comparisons. It also opens an additional comparability direction that connects the short-term disease onset manifestations and the associated long-term disease risk. This aspect will be discussed in Section 13.6. The use of biological networks to provide context and improve data relevance is supported by recent advances in the so-called network view of diseases that describes a biological system in terms of a limited set of networks that are diversely perturbed in the case of diseases [9–12]. A more detailed discussion about the particular types of biological networks used in implementing this strategy and how they are assembled is given in Section 13.4.

Once the experimental system and the compound exposure regimen have been established, the appropriate technology for measuring the corresponding systems-wide effects must be selected. Since they simultaneously and rather exhaustively measure a high number of individual molecular entities (typically several tens of thousands), high-throughput technologies are clearly preferred. However, high-throughput approaches yield more noise in the data than the low-throughput approaches and, therefore, careful implementation of quality controls (QCs) must accompany all steps of the application, as shown in Figures 13.2 and 13.3. In principle, high-throughput systems-wide measurements of gene expression, protein expression, and posttranslational modifications such as phosphorylation can be generated to assess the effects of the applied treatment on the considered biological system. In practice, gene expression is currently the most widely used approach. To complement the high-throughput systems-wide assessment, additional assays are used to characterize the functional outcomes of the biological system. An example is presented in Section 13.5 where fluorescence-based techniques were used to quantify *in vitro* cellular stress and immune responses. Although animal models and cellular systems do not always directly translate to human disease, some of the key system elements can be reproduced and these observations represent a major asset in understanding how biological network perturbations can lead to disease. Considerations regarding the potential of model systems to translate to human health benefit are further developed in Section 13.6 constituting the last component of the five-step strategy outlined in Figure 13.1.

The final aspect to be considered concerns the actual execution of data-generating experiments. Figure 13.3 shows a typical workflow to measure *in vivo* gene expression in a nonhuman species. The figure illustrates the complexity of these experiments that involve the collection of a variety of samples at multiple time points. The test compound is subjected to a standardized time- and dose-dependent exposure regimen for the assessment of the two quantitative measures. Zero dose is used to represent control conditions. The experiment is performed in multiple replicates to optimize
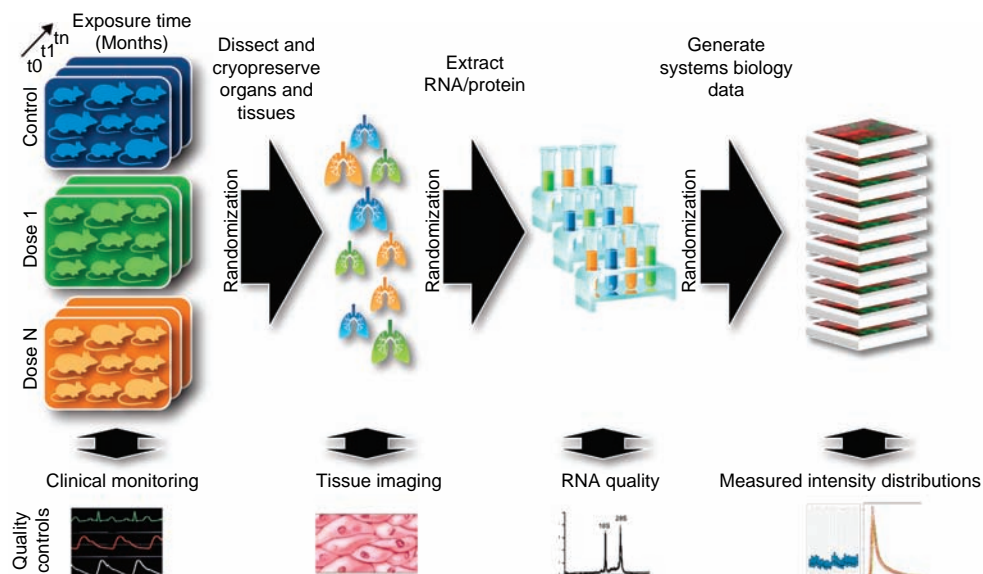
Figure 13.3 The workflow of the experiment for data production.

statistical power during analysis, and a randomization and blocking strategy is employed to minimize variability and bias in the data. As shown in Figure 13.3, it is particularly important to randomize the sample grouping during the intermediate steps of organ dissection and RNA extraction to avoid batch effects that can mask the actual treatment-induced signal [13]. When measuring gene expression with widespread DNA microarrays, the single-channel technology (such as Affymetrix GeneChip®) circumvents the additional complications incurred by having to account for both channels when using dual-channel arrays. Finally, as indicated on the bottom of Figure 13.3, each biological sample must be controlled for quality. Once the experiment has been successfully completed and all quality controls deemed satisfactory, the generated data are ready for input into the next stage of the process where the calculations of the systems response profiles (SRPs) are made.

13.3
Step 2: Compute Systems Response Profiles

The quality-controlled measurements generated in the experimental stage constitute a SRP for each exposure in a given experimental system. Systems-wide data generated by high-throughput technologies are used to elucidate the mechanistic impact of biological perturbations on the experimental system. The SRP expresses the degree to which each individual molecular entity is changed as a consequence of the exposure of the system to the tested compound and is the result of rigorous quality controls and statistical analysis. In this way different data types (transcriptomics for messenger
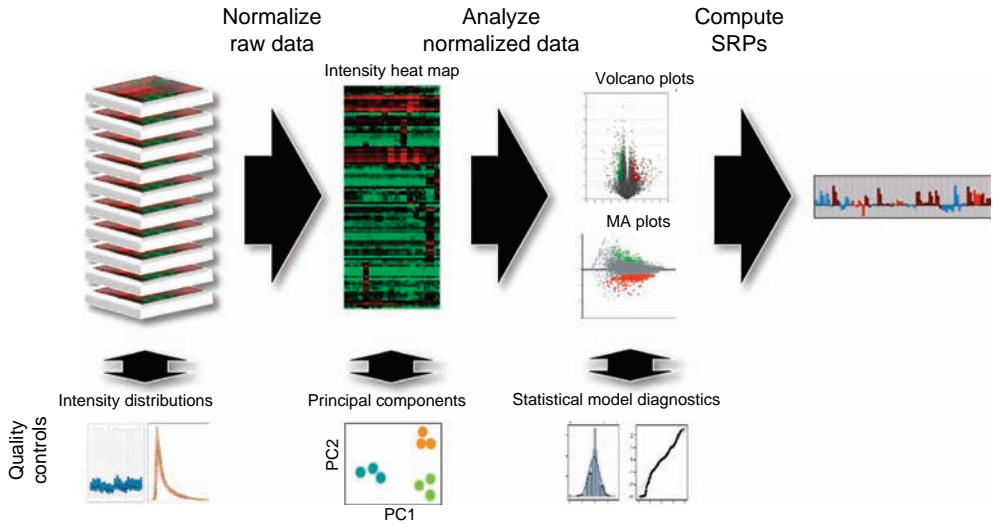
Figure 13.4 The pipeline for computing systems response profiles (SRPs).

RNAs and microRNAs, proteomics/phosphoproteomics, metabolomics, etc.) can be integrated and coanalyzed to provide the most accurate possible quantitative representation of the biology. However, several processing steps are necessary in order to satisfy the various requirements expected from such multidimensional data, such as the raw data normalization, the choice of appropriate statistical models, and the acquisition of rich experimental meta-information in a standardized format. The following section discusses various aspects of the SRP computational workflow.

The primary input of this workflow is a data set obtained from a high-throughput profiling technology. Gene expression data probing messenger RNAs (mRNAs) will be used as a concrete example, but these considerations are also valid for data obtained from other technologies such as exon or tiling microarrays, as well as RNA sequencing (RNAseq). The experimental studies used to generate the data are assumed to be optimized for quality and relevance, while all measuring devices are adjusted and validated according to manufacturers' directions. The goal of the processing workflow is twofold. First, it transforms the input data (raw data) into the appropriate SRPs, accompanied by measures of their statistical significance. Second, it computes quality metrics at the various processing stages, in order to confirm the appropriateness of both data and applied transformation methods. Figure 13.4 gives a schematic view of the workflow, and its components are explained in the following (more details are given in Refs [14,15]).

Raw data normalization: The input raw data contain intensity measurements performed directly on the images of the array ("probe level" for Affymetrix GeneChip). However, they do not yet provide the intensity values of actual genes ("probe set level" for Affymetrix GeneChip). In addition, between-array

comparisons might contain array-specific biases due to independent measurements of each array. The goal of raw data normalization is to generate mRNA-based intensities that can be compared across all arrays in the experiment. An efficient method to achieve this task is the RMA (robust multichip average) algorithm, based on the reasonable assumption that probe intensity distributions are identical across all arrays [16]. The first step consists of background subtraction where the effects of parasite hybridization on the microarray probes are reduced. An improvement of this step has led to the GCRMA (GeneChip robust multichip average) algorithm, which explicitly takes into account the nucleotide content of the probes in the evaluation of the background contributions [17]. The next step is the actual normalization based on the quantile normalization algorithm that exploits the above assumption of identical probe intensity distributions across all arrays. The last step is the summarization, which computes an estimate of the actual mRNA abundance based on the intensities of the multiple matching probes and using the median polish algorithm [18].

Systems response profile calculation: Normalized data constitute the input of the actual SRP calculation, together with the experimental design details that contain the relationships between all measured samples. For the sake of clarity, the simplest case of a pairwise comparison between one group of "treated" samples and one equally sized group of "control" samples is considered (see Section 13.1). More complex designs are common and they can be handled in a manner similar to the pairwise comparison, as long as linear models are used appropriately in the calculation [19]. In a pairwise comparison context, the SRP measures the effect of the applied treatment at the gene level by comparing against a group of control samples that did not undergo the treatment. Specifically for a given gene, the response consists of the difference between mean $\log_2$ intensities of the group of treated samples and mean $\log_2$ intensities of the group of control samples. This quantity is usually referred to as gene differential expression. The associated statistical significance is provided by the *t*-statistic taking into account the expression variance within each group. In the case of microarray experiments, the number of samples is often small and so the variances are difficult to be estimated accurately. A solution to this problem is provided by the moderated *t*-statistic, which improves the specificity of the SRP statistical significance by using empirical Bayes methods [19]. Another solution is the significance analysis of microarrays (SAM) approach based on bootstrapping computations and shrunk *t*-statistic [20]. The SRP specificity is further increased by applying multiple testing corrections, for example, the Benjamini–Hochberg correction [21], to account for the fact that thousands of genes are measured on the microarray. At the end of the process, the SRP is characterized by the differential expression values of all genes measured on the microarray, complemented by their statistical significance, usually in terms of *p*-values or false discovery rates (FDRs).

Quality controls: The data quality is controlled at three different levels and specific features are examined in each case. At the raw data level, several within-array metrics are computed, which allow detecting possible hybridization in homogeneities on the microarray as well as sensitivity issues in the intensity range. Normalized

data are used to perform between-array comparisons and thereby identify possible outlying samples. Multivariate approaches like principal component analysis (PCA) are used to verify the consistency of the data, typically by showing that in reduced dimensional space the samples belonging to the same treatment group are closer to one another than to samples from different treatment groups. Normalized data can also be corrected for possible (nonconfounding) experimental batch effects, if this information is available. This operation is important because it reduces the fraction of data variance that is not due to the test treatment, and therefore increases the statistical significance of the downstream-calculated SRPs. Several algorithms have been developed for performing this task such as the Combat method that uses an empirical Bayes approach [22]. At the SRP level, the assumptions underlying the statistical models used in calculations must be, at least partly, verified. This information is derived from so-called diagnostics plots. For the *t*-test described above, MA-plots, volcano plots, QQ-plots, histograms of the residuals, and so on are used to visualize this information.

Several studies have shown that the reproducibility of published gene studies remains low [23]. In order to ensure optimal acceptance by the both the scientific community and the regulatory authorities, particular attention must be given to aspects of the process of computing SRPs that can improve reproducibility. These are discussed in the following:

Scientific relevance: Recent studies have shown that there are some components of the pipeline for computing SRPs that could possibly be improved with regard to both normalization and statistical evaluation [24,25]. However, it has also been shown that the current version of the pipeline presented above continues to perform well compared to alternative approaches [26]. In addition, the choice of pipeline components represents a compromise between confirmed value and top performance, and the current approach has emerged as a consensus within the microarray analysis community. While there is improvement for scientific correctness, this consensus is a strong indicator of its potential to generate meaningful results that will continue to find acceptance in science and regulatory communities. An example of compromise is seen with the fact that the assumptions underlying the RMA normalization algorithm are not always satisfied [27]. However, this does not put into question the value of RMA itself since this has been successfully used in a large number of studies with over 3000 citations of the original article in subsequent papers.

Technical standardization: In order to facilitate the reimplementation of the processing pipeline, it is prudent to use the free and open-source Bioconductor software [28]. This software is based on the R programming language used in statistics, and due to its widespread use and content quality, it has almost become a standard in computational biology. The algorithms described above are all available as Bioconductor R packages, thereby enabling the full SRP pipeline to be run in the R environment. An additional capability of Bioconductor is its software versioning and archiving policy put in place in the repository [29], which significantly contributes to the reproducibility of the processing pipeline.

Experimental reproducibility. Improving the reproducibility of microarray experiments has become an important goal for many years, as illustrated by initiatives such as the Microarray Quality Control Phase I (MAQC-I) project [30]. Several components of the data production instruments are not under the direct control of the experimentalist and as such quality must be guaranteed by the manufacturers. Technological progress in the field of bioanalytics has led to a fairly robust protocols that display acceptable reproducibility. In this context, the experimentalist focuses on following the established protocols and controlling the quality at every step. Insufficient quality introduces variability into the data, such that the extraction of the targeted signal becomes more difficult due to an increased signal-to-noise ratio. Another aspect relevant for reproducibility is the availability of detailed information describing the experiment. A satisfactory solution is provided by the MIAME-based exchange format MAGE-TAB [31], which has become the standard for data sets deposited in public microarray databases such as ArrayExpress [32].

In conclusion, complying with all quality aspects is essential to producing robust SRPs and to optimal acceptance of resulting compound assessment outcomes.

13.4
Step 3: Identify Perturbed Biological Networks

A substrate of *a priori* biological knowledge is required in order to evaluate the SRPs in a causal, mechanistic manner that can facilitate the determination of biological impact for the compound under assessment. In order to apply the strategy outlined in Section 13.1, highly comprehensive causal network models of the biological processes relevant to risk assessment must be constructed. This approach provides a more detailed molecular understanding of biological network perturbations compared to the gene lists used in more classical toxicogenomics studies [33] and enables a tighter mechanistic linkage between exposure and disease risk. We have developed a strategy to build such network models using Biological Expression Language (BEL), a semantic programming language that allows a flexible representation of biological processes in a computable format [34]. As shown in Figure 13.5, the design and construction of causal network models is an iterative, multistep process. Network construction is explained in this section, aided by two concrete examples from previously published network models describing cell proliferation and cellular stress in nondiseased pulmonary tissues [35,36].

In step 1, the "literature model" is constructed (Figure 13.5, upper panel). The biological boundaries of the network are defined by a team of discipline-specific field experts, guided by a literature survey of the relevant signaling pathways related to the process of interest (e.g., cell proliferation in lung). BEL-encoded causal relationships describing these pathways are extracted from the Selventa Knowledgebase, a comprehensive repository containing over 1.5 million nodes (biological concepts and entities) and over 7.5 million edges (connections between nodes). The Selventa Knowledgebase is manually curated from peer-reviewed scientific literature as well
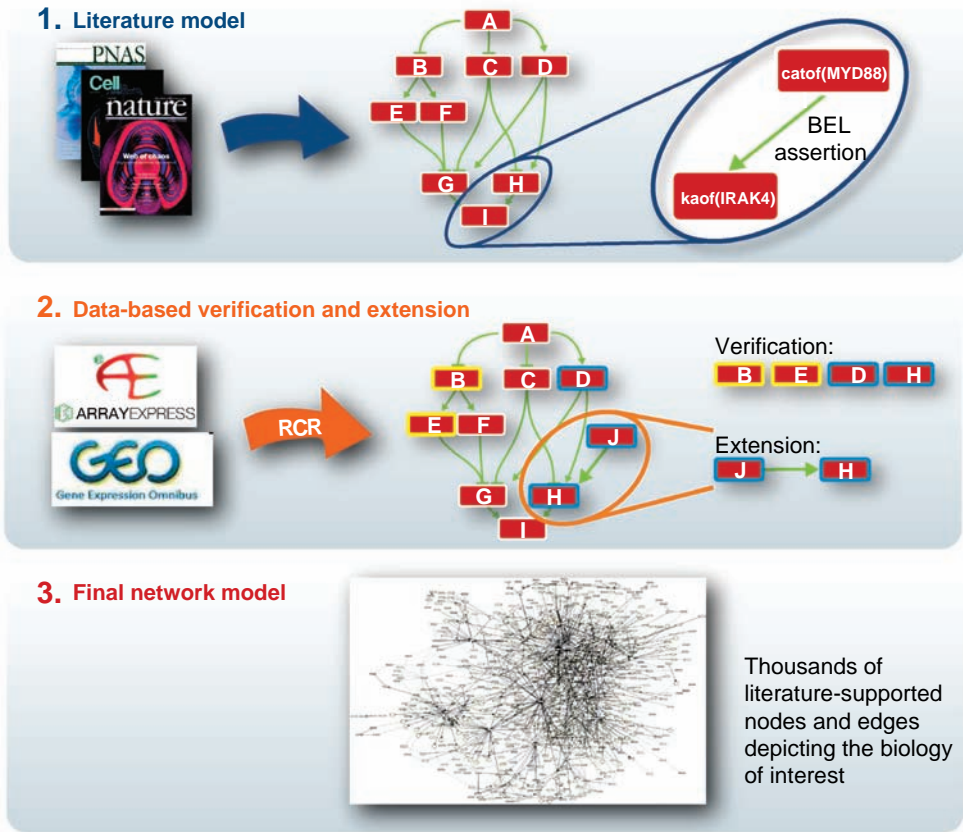
Figure 13.5 The three-step process for identifying and constructing perturbed networks.

as other public and proprietary databases [37]. These relationships are assembled to nucleate a network with causal relationships within the boundaries of tissue context and known biological mechanisms. Nodes in the networks are biological entities such as protein abundances, mRNA expressions, and protein activities. Network nodes can also represent biological processes (e.g., apoptosis), as well as chemicals or small molecules whose transcriptional signatures represent similar signaling to the one induced by the compound exposure. Edges are relationships between the nodes, and are categorized as either causal or noncausal. Causal edges are directional cause–effect relationships between nodes (e.g., catalytic activity of MYD88 directly increases kinase activity of IRAK4), whereas noncausal edges connect different forms of a biological entity, such as a particular protein abundance to its phosphorylated form (e.g., TP53 protein abundance to TP53 phosphorylated at serine 15). The literature-based model constructed in step 1 broadly covers biological signaling within each process of interest. It utilizes a modular design, where the content of each module is constrained to discrete areas of signaling. For example, in the network describing cell proliferation,

a module depicting cell cycle regulation was combined with modules describing the pathways known to influence cell cycle progression in the lung [e.g., Wnt, Hedgehog, and prostaglandin E2 (PGE2) signaling] [36].

In step 2, the "integrated model" is constructed (Figure 13.5, middle panel). The content of the literature model is first verified by performing reverse causal reasoning (RCR) on relevant high-throughput data sets available from public transcriptomic data repositories such as GEO or ArrayExpress. RCR is a method that takes SRPs obtained from gene expression profiles as input and uses statistical and biological criteria to make predictions about the activity states of biological entities termed HYPs as outputs [37–40]. The abbreviation HYP is derived from the word "hypothesis" and is an appropriate term to use since RCR determines which HYPs contained in the Selventa Knowledgebase can be considered as hypotheses for explaining the observed SRPs. HYPs are described in greater detail in Section 13.5. By design, a large fraction of the nodes of the literature model are HYPs and also involve causal relationships extracted from the Selventa Knowledgebase. Taking advantage of this feature, the verification of the literature model is performed as follows. First, a suitable data set is chosen for verification. Ideally, data sets used to verify network content are generated from an experiment perturbing a biological mechanism captured by the network model under investigation (e.g., a data set investigating the cellular response to oxidative stress would be used to verify the cellular stress network model). From the differentially expressed genes in a data set, RCR predictions are computed for all HYPs in the Selventa Knowledgebase, including the network nodes that are HYPs. Model HYPs that are predicted by RCR are mapped to the network model and analyzed for qualitative pathway activation, including directional consistency with the network edges. Finally, any additional phenotypic observations performed in parallel with the transcriptomic data (e.g., nuclear translocation of transcription factors, protein production/stabilization, or physiological endpoints) are compared with the RCR predictions to verify that the model is indeed competent to capture the relevant and expected biology. In addition, HYPs predicted by RCR, which were not already represented in the literature model, are used to extend the model, provided there is strong mechanistic connection to the biological processes underlying the network. Using this approach, a more comprehensive "integrated model" is generated, including nodes derived from existing literature as well as nodes derived from experimental data sets. In this way, the combined use of molecular profiling data and prior biological knowledge of cause–effect relationships is exploited to reinforce the content of a given network model.

The final network model is generated during step 3 of network construction (Figure 13.5, lower panel). In this last step, discipline-specific scientific experts conduct a terminal round of systematical manual review and refine the content and connectivity of the integrated model. Ultimately, this three-step methodology results in computationally optimized network models whose nodes (including HYPs) have inherent causal linkage supported by published literature [35,36,41,42].

Having outlined the three-step process to construct causal network models that describe relevant biological processes, the following key aspects are worth closer examination. A network model describing cell proliferation networks has been

completed and published, and is used here to provide concrete material to illustrate the process [36].

Applying RCR for verification: During step 2, RCR-derived HYPs from process-relevant data sets are mapped onto the model scaffold in order to verify the representation of biologically vetted mechanisms. During the construction of the cell proliferation network model, RCR analysis was applied to four transcriptomic profiling data sets characterizing proliferation in tissues and cell types relevant to the pulmonary system (e.g., fibroblasts *in vitro* and whole lung *in vivo*) in order to derive HYPs related to cell proliferation [43,44]. RCR-derived HYPs verified the central roles of known cell cycle regulators MYC, RB1, and CDKN1A in controlling lung cell proliferation. In addition, because RCR can predict the activity states of any HYP within the Selventa Knowledgebase, the computational approach taken in step 2 not only verifies that the initial model draft has the appropriate content but also serves to identify novel regulatory mechanisms that should be included in the network model.

Using phenotypic readouts: The data sets used in step 2 are derived from experiments with biological endpoints covering the specific biology represented in the network: an essential requirement for linking differences in gene expression with phenotypic outcomes. In the network model characterizing pulmonary cell proliferation, a molecular profiling data set from the hyperproliferative embryonic lung of a genetically modified mouse was used [45]. In addition to the transcriptomic data, a time-matched measurement of 5-bromo-2-deoxyuridine (BrdU) incorporation into DNA provided a well-established phenotypic readout of increased cell proliferation. Multiple data sets addressing discrete and yet overlapping aspects of the biology being modeled are used in step 2, ensuring broad coverage of the network model through computational analyses. In a recently published network model describing the cellular response to stress [35], transcriptomic data were obtained from experiments where an oxidative burden was induced by hyperoxic conditions, hypoxic conditions, and cell culture in the presence of biological entities known to induce oxidative stress.

Properties of network models: Network models preserve the topology of the network such that causal relationships (signaling pathways) can be traced from any point in the network to a measurable entity, enabling mechanistic linkage between the SRP and the upstream causal network that represents it. Furthermore, the network models are dynamic and the assumptions used to build them can be modified or restated as needed. This feature enables adaptability for experiments performed in different tissue contexts and species, thereby allowing iterative testing and improvement as new knowledge becomes available. In addition to published cell proliferation and cellular stress networks, additional network models are currently in construction. The networks (and the causal relationships describing their connectivity) will be made freely available to the scientific community through peer-reviewed publication [35,36,41].

In summary, the three-step process explained in this section enables the construction of causal network models describing biological processes perturbed by

exposure to the compound under assessment. Essentially, the models constitute substrates on which SRPs can be imposed and mechanistically interpreted. They provide the added-value input for the next two steps of the five-arrow strategy: the network perturbation amplitudes (NPAs) and the biological impact factor (BIF) calculations (Figure 13.1).

13.5
Step 4: Compute Network Perturbation Amplitudes

By this stage in the overall process, the SRPs have been obtained from an appropriately designed experiment and the application of high-throughput systems-wide tools to measure gene expression, protein expression, and posttranslational modifications. The network model(s) that encode the causal and noncausal relationships have been constructed from external and internal data sources. The next step combines these two elements to compute the NPAs. The purpose of this scoring scheme is to derive a "response profile" at the network level, which then allows a coarse-grained view of the effects of the applied treatment encompassed by the SRPs. As a consequence, the NPA scores are expected to correspond to the resulting changes in the activity of the cellular processes described by the network model. For instance, in the case of the proliferation network (see Section 13.4), a positive NPA score would correspond to an increase in the cell division rate. Therefore, comparisons of the NPA scores with independent measurements of the cellular processes described by the network models are necessary in order to validate the NPA approach. Successful validation of NPAs then paves the way for computing the BIF, which constitutes the last component of the overall strategy shown in Figure 13.1. In this section, a complete proof-of-principle NPA study is described [46]. It uses a reasoning process that first applies the NPA approach to a single network node, designated a HYP, and then extends it to complete network models. The perspectives opened by the NPA approach are then discussed, both at conceptual and methodological levels.

The proof-of-principle NPA study was based on a well-understood and controlled experimental system comprising three components: (i) the actual biological experiment, (ii) the related network models, and (iii) the assay measuring the perturbations in the system. The experiment consisted of cultures of normal human bronchial epithelial (NHBE) cells treated with the proinflammatory signaling mediator TNFα. The design of this experiment, its execution, and the subsequent computation of the SRPs were performed in accordance with the guidelines described in Sections 13.2 and 13.3. In the context of TNFα-treated NHBE cells, the stress and immune response transcription factor NF-κB is known to be a major mediator of the induced signaling response. Three network models describing NF-κB biology were therefore assembled using the information contained in the Selventa Knowledgebase, following the process described in Section 13.4. These three networks have slightly different structures, facilitating the reasoning used to derive the NPA scoring scheme, as will become clear. The last component of the NPA study was an assay to measure the perturbations in the network due to the
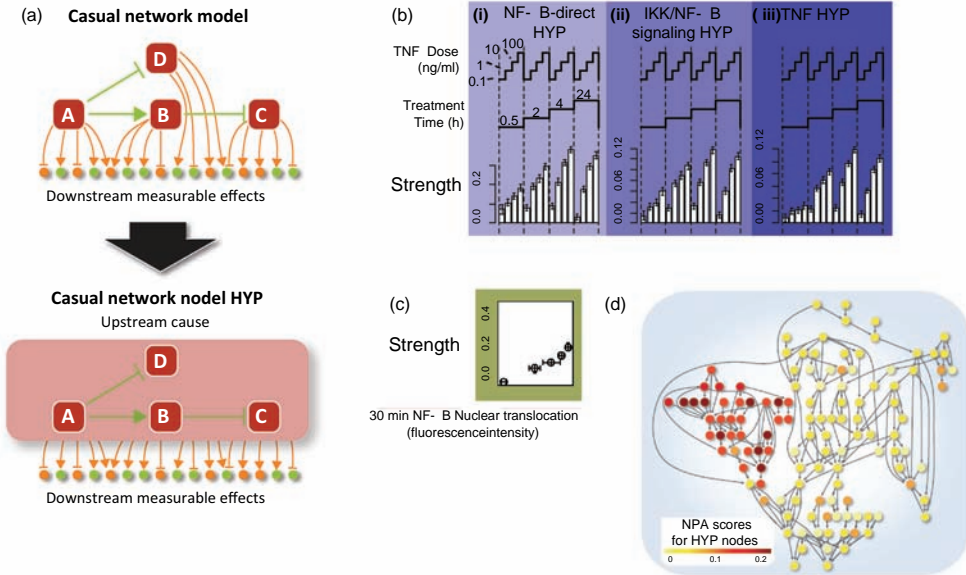
Figure 13.6 The computation of network perturbation amplitudes (NPAs). (a) Schematic representation of the aggregation of a network models into a single HYP. (b and c) NPA results for the TNFα-treated NHBE cells experiment.

(d) The graph theoretic approach for NPA calculation: the coloring scheme follows the NPA scores of the HYP nodes of the network model and reveals the structure of the perturbation inside the network.

biological activity of NF-κB. Since the main activity of NF-κB consists in regulating the transcription of genes, a quantitative detection of the NF-κB complex translocation into the cell nucleus, as measured *in situ* using fluorescence-based techniques [47], was deemed the most appropriate measure of biological activity.

The starting point for deriving the NPA scoring scheme is the HYP, which is the term used to describe the most basic constituent of the causal network models (see Section 13.4). As shown in Figure 13.6a, a HYP actually captures the causal regulatory relationships between one upstream controller and its downstream targets, where a transcription factor like NF-κB can be one such upstream controller. In this case, the downstream targets are the genes that are either up- or down-regulated upon activation of NF-κB, as indicated by the two arrow types "→" and "—|", respectively. Using "backward" HYPs is advantageous in that, for gene expression data, the results do not depend on the "forward" assumption that mRNA expression changes are always directly correlated with protein activity changes [48–50]. Often, this assumption does not take into account the effects of translational or posttranslational regulation on protein activity. In the RCR approach, a statistically significant match between a SRP and a HYP indicates that the biological process modeled by the HYP is very likely to have been activated in the experiment. This conclusion leads to a mechanistic hypothesis about the

underlying biology. Here, the "NF-κB-direct" HYP is first considered. It contains $N = 155$ downstream genes known to be directly regulated by NF-κB, that is, genes whose expression is controlled in an NF-κB-dependent manner and whose promoter sequences are directly bound by NF-κB.

The causal structure of a HYP implies that a perturbation of the upstream controller propagates to its downstream targets, modulated by the positive or negative signs $s_i = \pm 1$ associated with each connection i, as indicated by the two arrow types "→" and "—|" in Figure 13.6a. In the case of TNFα-treated NHBE cells, the "NF-κB-direct" HYP is expected to provide an adequate description of the activated biology. Therefore, the sign s_i of its downstream genes and the signs of the corresponding differential expressions β_i contained in the SRPs are very likely to be the same. This means that the majority of the products $\beta_i \cdot s_i$ are positive, so that their sum over all the HYP's downstream genes is also positive. This property is exploited in defining an NPA score called *strength* by the following expression:

$$strength(\text{HYP}) = \frac{1}{N} \sum_{i=1\cdots N} \beta_i \cdot s_i.$$

As suggested by its name, this quantity estimates the strength of the TNFα-induced NF-κB activation based on the "NF-κB-direct" HYP. In other words, it measures the amplitude of the TNFα-induced perturbation of the activity of the "NF-κB-direct" network model. Mathematically, *strength* is the mean treatment-induced differential expressions of the HYP's downstream genes, adjusted for the sign of their causal connection to the upstream controller of the HYP. It can be deduced from the general geometric considerations. In essence, *strength* constitutes an authentic quantitative measure of the network model perturbation. Indeed, it not only counts the number of genes whose differential expression agrees with the HYP signs s_i but also takes into account their magnitudes β_i, without any need of significance thresholds. In contrast, thresholds are a key requirement for the previously available measures of HYP significance, "richness" and "concordance," as used in the RCR framework [37].

The application of the *strength* NPA score to the 16 SRPs obtained from the TNFα-treated NHBE cells experiment is shown in Figure 13.6b. The *strength* results obtained with the "NF-κB-direct" HYP display a monotonically increasing dose dependency at all time points. This behavior is consistent with the fact that increasing TNFα dose amplifies the resulting NF-κB activation in the cell culture experiment. Figure 13.6c shows the results of the fluorescence intensity measurements of the differential nuclear translocation of the NF-κB complex. At 30 min, results confirm the monotonically increasing dose dependence of the NF-κB responses, attested by a correlation coefficient of 0.98. These results demonstrate that the NPA *strength* scores based on the "NF-κB-direct" HYP quantify NF-κB transcriptional activity. This process is an important component of the perturbations induced by the TNFα treatment on the NHBE cells. The fluorescence intensity measurements provide an independent experimental confirmation of the validity of NPA scoring scheme presented here.

Having validated the *strength* scoring scheme for the "NF-κB-direct" HYP, the next step in the NPA approach consists of improving the network model to more comprehensively cover the biology of the cellular TNFα response. A second network model, the "TNF" HYP, was constructed, consisting of 1741 downstream genes, all of which are known to be modulated in TNFα-treated cells. Even if the "TNF" HYP includes biological processes that are more distant from the measured NF-κB process, its *strength* results remain very similar to the ones obtained with the "NF-κB-direct" HYP (Figure 13.6b). This suggests that there is not a large difference between the global behavior of genes that are known to be directly regulated by NF-κB and the behavior of genes where knowledge of direct regulation is uncertain. In this situation, the "TNF" HYP can be seen as a "black box" masking the underlying (unknown) network structure and retaining only the regulatory signs of the downstream genes. The next step consists of using the "IKK/NF-κB signaling" network model, comprising 40 nodes that together regulate a total of 992 downstream genes. This model describes processes more directly connected to NF-κB than the "TNF" HYP and is more representative of the type of networks discussed in Section 13.4. As represented in Figure 13.6a, the "IKK/NF-κB signaling" network model is first transformed into a single "aggregated" HYP by collecting all the downstream genes of its HYP nodes and assigning them the suitable sign. This operation is possible only when the networks are causally consistent, which is the case for the "IKK/NF-κB signaling" network model. This condition ensures that no ambiguities about the signs of regulation appear in the model. The aggregated "IKK/NF-κB signaling" HYP resembles the "TNF" HYP already discussed, except for the fact that the details masked by the "black box" are known explicitly in this case. The results of the *strength* score computation for the aggregated "IKK/NF-κB signaling" HYP are shown in Figure 13.6b, displaying very similar patterns of response, which ultimately confirm the validity of *strength* as a genuine method for computing NPAs.

The proof-of-principle study reports many additional results that support the NPA approach presented here [46]. Three complementary NPA metrics besides *strength* were created and successfully tested, bringing new additional features to the process. For instance, the geometric perturbation index (GPI) metric was designed to reduce the noisy contribution of the statistically insignificant genes involved in a HYP. It reduces the probability of getting arbitrarily biased scores in the case of large HYPs with more than thousand downstream genes. Two statistics that complement the NPA scores and allow assessment of their significance were also derived. To further confirm the validity of the developed NPA methodology, the results from other data sets and networks were taken into consideration. The proof-of-principle study also showed that assessing the amplitude of network perturbations with four complementary NPA methods highlights those conclusions that are robust versus those that may be specific to a particular NPA method.

From a broader perspective, NPA scoring is an integrated approach that combines high-throughput experimental data with a knowledge-driven network model to provide measurable quantities causally affected by a targeted biological process. This allows the activity changes of that process to be quantified relative to a control (nonperturbed) state of the system. The utility of the NPA method lies in the synergy

of on-demand HYP generation from an extensive causal knowledgebase, with a continuous measure of its activity change. Today, four NPA scoring methods have been developed with complementary strengths and are individually providing distinct advantages for specific circumstances. When applied to the TNFα-treated NHBE cells experiment, NPA scores for NF-κB correlated with the expected dose–response relationships and specific measured pathway outputs. NPA scoring also suggested possible "cross talk" between NF-κB activation and the cell cycle that could be investigated experimentally.

With a broad spectrum of biology available to score within the Selventa Knowledgebase, NPA metrics and statistics can be used to assess amplitude of perturbation on many orders – from a single molecule to that of a complex, higher order causal network model representing complex biological processes. In the case of larger network models, feedback interactions may lead to inconsistent paths in the network and these inconsistencies may extend to existing NPA methods. One possibility would be to directly compute the NPA metrics for the network model nodes that are HYPs, as depicted in Figure 13.6d, to allow graph theoretic methods to be applied to the network to derive the response at a global level. This approach would not only replace the aggregation of a network model with a single HYP (Figure 13.6a) but also reveal the structure of the perturbation inside the network (Figure 13.6d) [51]. While the NPA approach has validated utility in its current form, such considerations show room for further methodological and algorithm development.

In summary, the NPA approach enables a quantitative, systems-wide understanding of the biological mechanisms leading to diseases. The described algorithms are the first step toward the development of computational tools designed to comparatively assess any perturbation in any biological systems.

13.6
Step 5: Compute the Biological Impact Factor

The final step of the strategy to quantitatively describe the effects of perturbations within biological networks is the computation of a BIF. This factor represents a holistic score that describes the systems-wide effect of all the processes captured in the underlying network models and their associated NPA scores. The compounds' attributed BIF values can then be quantitatively compared based on a high-level view of their biological effects. In summary, a well-defined framework (steps 1–4) has been established that enables transparent information agglomeration such that entire SRPs can be mathematically transformed into a small set of numbers (NPA scores and then a BIF value). If the validity requirements of the four steps are met, the BIF has the potential to provide a simple but scientifically sound measure of the biological impact of a compound on a system. The development of the BIF methodology is currently ongoing [52] and accordingly the content of this section focuses on outlining the strategic and scientific contexts rather than the actual algorithms. The concept of a BIF is demonstrated using an example to estimate rat nasal epithelium tumorigenesis in

response to formaldehyde exposure. The perspectives resulting from the BIF interpretation and utilization are also briefly discussed.

As indicated by its name, the BIF aims to quantify the biological impact resulting from the exposure of a biological system to one or several compounds. As suggested by Figure 13.1, its most direct application is in the explicit comparison between different compounds. The BIF scores provide quantitative measures of the impacts caused by each compound, which can be compared with each other. This relative approach is particularly useful in situations where one of the compounds is well characterized in terms of perturbed biological networks and long-term disease risk, while the others are much less studied. In this case, the BIF provides an explicit way of assessing the expected effect of the less-studied compounds, based on the existing knowledge available for the well-studied, or reference, one. Another appropriate application is in the situation where a disease phenotype of the exposed organism is available alongside the measured SRPs. In this case, and in direct line with the widely used concept of disease association, the BIF can be calibrated with a quantitative measure of health impact. If the calibration is done in a robust manner, it opens up broader perspectives in the context of personalized health and safety assessment. Even in the absence of an explicit disease phenotype, a BIF can still be amenable to calibration and thus be used to encompass information relevant to disease risk. This statement is based on the assumption that the mechanistic characterization of early biological effects, in terms of perturbations of the relevant biological networks, is strongly indicative of the long-term disease outcome. From this perspective, the perturbations of the biological networks are expected to collectively serve as prospective biomarkers for disease risk, similar to compound metabolites detected in body fluids [53,54]. As such, the BIF enables the identification of risk factors and allows the potential for "red flags" to identify their manifestations in the observations constituted here by the NPA scores and the SRPs.

In light of the initial observation regarding the limited utility of epidemiological studies to link short-term effects with long-term diseases, the usefulness of the BIF concept becomes obvious. The short-term quantification of perturbation caused by interventions such as drugs, diets, or environmental conditions can be linked to potential longer term risk through the identification of the BIF. As emphasized throughout this chapter, since the BIF is supported by the mechanistic information contained in all the underlying networks, it can be viewed as a "quantitative mechanistic meta-biomarker" of the effects associated with exposure to test compound. Since it aggregates NPA scores that have themselves already filtered a large fraction of the noise initially contained in the SRPs [45,46], the BIF is expected to produce results with increased robustness against technical and biological sources of variability. Although this aspect has not yet been concretely tested for the BIF, the MAQC-II study has clearly shown that results based on biomarkers involving multiple genes are much less sensitive to the variances inherent in the underlying technologies [55].

In Figure 13.1, the BIF is represented as a radar chart in which the multiple axes contain the NPA scores computed for each of the considered biological network models. Computing the surface of the polygon formed by the NPA scores obtained for a given SRP constitutes an intuitive BIF algorithm. Similarly, the fundamental

idea behind the BIF is to use the amplitudes of the perturbations induced by the exposure in an appropriate set of biological network models as the input of a simple scoring scheme, which provides a quantitative measure of their global effect. From this point of view, the BIF algorithm is first and foremost intended to detect and display trends in its input data set. As a consequence, the *a priori* selection of networks to be included in the BIF calculation, while it must be biologically sound, does not constitute its most critical aspect, since only the significantly perturbed ones will contribute to the BIF results. Ideally, even if the chosen networks do not exhaustively cover the underlying biology, they will still capture a significant portion of the systems response due to the strategy put in place in step 3 (see Section 13.4).

Having computed the NPA scores for the selected biological network models (step 4), the relative importance of each network model must be determined. While the BIF deduced from the radar chart in Figure 13.1 weights every axis equally, other choices are possible. Network preference based on *a priori* qualitative knowledge is not easily translatable into objective and reproducible weights. Data-driven weighting schemes, such as multivariate dimension reduction methods may be more appropriate [56]. The final step of the BIF calculation consists of aggregating the weighted NPA scores. As illustrated by the surface-based BIF from the radar chart in Figure 13.1, a simple sum of the weighted NPA scores is not necessarily the most meaningful solution. Methods based on more advanced geometric considerations may be more appropriate. The aggregation process is also expected to determine the contribution to the BIF of nodes belonging simultaneously to several network models, such as the highly connected NF-κB transcription factor. Additional methods are being developed to avoid overweighting these contributions.

To illustrate the concept of a BIF, an example showing the estimation of nasal epithelium tumorigenesis in rats after exposure to formaldehyde is presented [3]. For a simple BIF, the proliferation and the inflammatory networks were identified as underlying processes relevant for tumorigenesis. Both networks were naively assumed to contribute equally to tumorigenesis, and thus were weighted equally. The nasal epithelium tumorigenesis BIF in rats was evaluated using transcriptomic data following exposure to multiple doses of formaldehyde for 13 weeks [57]. NPA *strength* scores were normalized for each network to their highest values across the different doses. Figure 13.7 shows that significant correlation was observed between the BIF derived at an early stage following the 13 week exposure to formaldehyde and the tumorigenesis rates for rats exposed to the same doses of formaldehyde for 2 years [58]. This demonstrates that even a simple BIF, derived from systems-wide data obtained in short-term experiments, can be a good predictor of long-term disease outcome. Figure 13.7 also suggests a threshold effect with tumorigenesis only becoming significant above a BIF of 0.4. This observation can be exploited to provide a concrete estimate of the tumorigenesis risk, based on the measureable NPA values and BIF. Even if a BIF is not calibrated, because the long-term disease outcome data are not available, it can be used to rank biological network perturbations based on their expected biological outcomes.

The calibrated BIF has been thus presented as a means to correlate late disease onset (tumorigenesis rate after a 2 year exposure to formaldehyde in rats) based on early
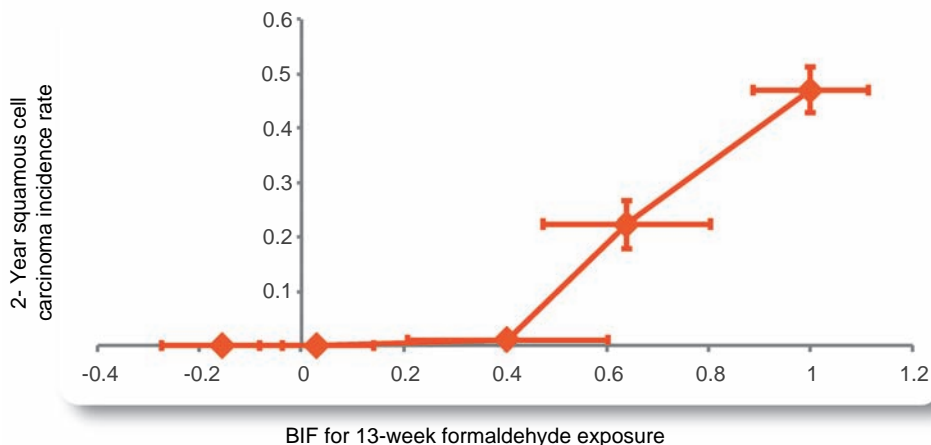
Figure 13.7 The biological impact factor (BIF) for compound testing in the example case of formaldehyde exposure. Early effects, that is, perturbations of relevant biological networks, correlate with the long-term health impact, that is, tumor incidence rate after 2 years (given by fraction of rats with squamous cell carcinoma in the nasal epithelium).

perturbations of the proliferation and inflammation networks (due to a 13 week exposure to formaldehyde in rats). It could be also used to predict the long-term effects. In essence, the BIF offers the potential to quantitatively describe the long-term impact of short-term network perturbations. It can be used as a scale for comparison or for threshold establishment, based on an associated outcome calibrated with the computed BIF values. Furthermore, whereas today it is necessary to correlate defined exposure modalities (time and dose) of a specified compound with the rate of disease onset [55,58], such a mechanism-based BIF allows the explicit association of biological network perturbations with disease onset as a function of the exposure regimen. This would allow the mechanism-based estimation of the risks of long-term disease caused by compounds for which no long-term epidemiology data are available. In addition, the process of computing a BIF from systems-wide measurements mapped to contributing biological networks enables the simultaneous identification of mechanistic biomarkers, which can be used as assessment tools for testing compounds.

13.7
Conclusions

Our systems biology-based approach to quantifying the biological impact caused by exposure to compounds is based on the five-step strategy illustrated in Figure 13.1. It consists of systematically exploring the "cubic" design space depicted in Figure 13.2 in order to deduce the biological mechanisms that translate from preclinical experimental systems to humans and their populations. Steps 1–4 provide a

well-defined framework for the identification of biological networks that are perturbed by short-term exposure to compounds. In step 5, these results are summarized into a BIF that enables the linking of the observations of early effects with long-term health impacts. An example is shown in Figure 13.7 for the particular case of formaldehyde exposure and long-term tumorigenesis in rats. Fundamentally, the computed BIF can be viewed as a prospective biomarker for disease risk, supplemented by mechanistic attributes that enable its potential translation to humans.

We thus propose that experiments performed over hours, days, or weeks can be used to measure the degree of perturbation of individual networks that can then be aggregated into an estimate of risk for disease onset, or prognosis for disease progression. Furthermore, time- and exposure-dependent changes of this risk estimate can be readily derived from appropriate experimental data to further provide an indication about risk modification as a function of time and exposure. Applications of this framework include the evaluation of the degree of unwanted biological impact caused by (i) different manufactured products for safety comparisons, (ii) therapeutics (especially those for chronic use), and (iii) environmentally active substances to predict safety of long-term exposure and the relationship to adverse effect and onset of disease.

The systems biology approaches to compound testing described in this chapter show novel applications of data mining, which can become pertinent in the context of drug discovery. They consist in a five-step strategy using biological network models to mine unstructured high-throughput data generated during well-designed experiments. These processes involve the calculation of Network Perturbation Amplitudes (NPA) and Biological Impact Factors (BIF). These two quantities provide a quantitative, mechanism-based, and, therefore, interpretable assessment of the systems-wide biological impact of exposures to the tested compounds.

References

1 Waters, M.D. and Fostel, J.M. (2004) Toxicogenomics and systems toxicology: aims and prospects. *Nature Reviews. Genetics*, **5**, 936–948.

2 Harrill, A.H. and Rusyn, I. (2008) Systems biology and functional genomics approaches for the identification of cellular responses to drug toxicity. *Expert Opinion on Drug Metabolism and Toxicology*, **4**, 1379–1389.

3 Hoeng, J., Deehan, R., Pratt, D., Martin, F., Sewer, A., Thomson, T.M., Drubin, D.A., Waters, C.A., De Graaf, D., and Peitsch, M.C. (2012) A network-based approach to quantifying the impact of biologically active

substances. *Drug Discovery Today*, **17**, 413–418.

4 FDA's Critical Path Initiative, http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/ucm076689.htm

5 Ekins, S., Nikolsky, Y., and Nikolskaya, T. (2005) Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. *Trends in Pharmacological Sciences*, **26**, 202–209.

6 Krewski, D., Westphal, M., Al-Zoughool, M., Croteau, M.C., and Andersen, M.E. (2011) New directions in toxicity testing. *Annual Review of Public Health*, **32**, 161–178.

7 Pleil, J.D. and Sheldon, L.S. (2011) Adapting concepts from systems biology to develop systems exposure event networks for exposure science research. *Biomarkers*, **16**, 99–105.

8 Edwards, S.W. and Preston, R.J. (2008) Systems biology and mode of action based risk assessment. *Toxicological Sciences*, **106**, 312–318.

9 Ideker, T. and Sharan, R. (2008) Protein networks in disease. *Genome Research*, **18**, 644–652.

10 Schadt, E.E. (2009) Molecular networks as sensors and drivers of common human diseases. *Nature*, **461**, 218–223.

11 Barabasi, A.L., Gulbahce, N., and Loscalzo, J. (2011) Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, **12**, 56–68.

12 del Sol, A., Balling, R., Hood, L., and Galas, D. (2010) Diseases as network perturbations. *Current Opinion in Biotechnology*, **21**, 566–571.

13 Scott, D.J., Devonshire, A.S., Adeleye, Y.A., Schutte, M.E., Rodrigues, M.R., Wilkes, T.M., Sacco, M.G., Gribaldo, L., Fabbri, M., Coecke, S., Whelan, M., Skinner, N., Bennett, A., White, A., and Foy, C.A. (2011) Inter- and intra-laboratory study to determine the reproducibility of toxicogenomics datasets. *Toxicology*, **290**, 50–58.

14 Reimers, M. (2010) Making informed choices about microarray data analysis. *PLoS Computational Biology*, **6**, e1000786.

15 Slonim, D.K. and Yanai, I. (2009) Getting started in gene expression microarray analysis. *PLoS Computational Biology*, **5**, e1000543.

16 Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

17 Wu, Z., Irizarry, R.A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004) A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Department of Biostatistics Working Papers, Johns Hopkins University.

18 Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., and Speed, T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, **31**, e15.

19 Smyth, G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article3.

20 Tusher, V.G., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116–5121.

21 Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, **57**, 289–300

22 Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., and Liu, C. (2011) Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One*, **6**, e17238.

23 Ioannidis, J.P., Allison, D.B., Ball, C.A., Coulibaly, I., Cui, X., Culhane, A.C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G.P., Petretto, E., and van Noort, V. (2009) Repeatability of published microarray gene expression analyses. *Nature Genetics*, **41**149–155.

24 McCall, M.N., Bolstad, B.M., and Irizarry, R.A. (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics*, **11**, 242–253.

25 Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., and Vert, J.P. (2007) Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**, 35.

26 Choe, S.E., Boutros, M., Michelson, A.M., Church, G.M., and Halfon, M.S. (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, **6**, R16.

27 Wang, D., Cheng, L., Wang, M., Wu, R., Li, P., Li, B., Zhang, Y., Gu, Y., Zhao, W., Wang, C., and Guo, Z. (2011) Extensive increase of microarray signals in cancers calls for novel normalization assumptions. *Computational Biology and Chemistry*, **35**, 126–130.

28 Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis,

B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., and Zhang, J. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80.

29 Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., and Zhang, J. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5** (10), R80.

30 Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., de Longueville, F., Kawasaki, E.S., Lee, K.Y., Luo, Y., Sun, Y.A., Willey, J.C., Setterquist, R.A., Fischer, G.M., Tong, W., Dragan, Y.P., Dix, D.J., Frueh, F.W., Goodsaid, F.M., Herman, D., Jensen, R.V., Johnson, C.D., Lobenhofer, E.K., Puri, R.K., Schrf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P.K., Zhang, L., Amur, S., Bao, W., Barbacioru, C.C., Lucas, A.B., Bertholet, V., Boysen, C., Bromley, B., Brown, D., Brunner, A., Canales, R., Cao, X.M., Cebula, T.A., Chen, J.J., Cheng, J., Chu, T.M., Chudin, E., Corson, J., Corton, J.C., Croner, L.J., Davies, C., Davison, T.S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A.C., Fan, X.H., Fang, H., Fulmer-Smentek, S., Fuscoe, J.C., Gallagher, K., Ge, W., Guo, L., Guo, X., Hager, J., Haje, P.K., Han, J., Han, T., Harbottle, H.C., Harris, S.C., Hatchwell, E., Hauser, C.A., Hester, S., Hong, H., Hurban, P., Jackson, S.A., Ji, H., Knight, C.R., Kuo, W.P., LeClerc, J.E., Levy, S., Li, Q.Z., Liu, C., Liu, Y., Lombardi, M.J., Ma, Y., Magnuson, S.R., Maqsodi, B., McDaniel, T., Mei, N., Myklebost, O., Ning, B., Novoradovskaya, N., Orr, M.S., Osborn, T.W., Papallo, A., Patterson, T.A., Perkins, R.G., Peters, E.H., Peterson, R., Philips, K.L., Pine, P.S., Pusztai, L., Qian, F., Ren, H., Rosen, M., Rosenzweig, B.A., Samaha, R.R., Schena, M., Schroth, G.P., Shchegrova, S., Smith, D.D., Staedtler, F., Su, Z., Sun, H., Szallasi, Z., Tezak, Z., Thierry-Mieg, D., Thompson, K.L., Tikhonova, I., Turpaz, Y., Vallanat, B., Van, C., Walker, S.J., Wang, S.J., Wang, Y.,

Wolfinger, R., Wong, A., Wu, J., Xiao, C., Xie, Q., Xu, J., Yang, W., Zhong, S., Zong, Y., and Slikker, W., Jr. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, **24**, 1151–1161.

31 Rayner, T.F., Rocca-Serra, P., Spellman, P.T., Causton, H.C., Farne, A., Holloway, E., Irizarry, R.A., Liu, J., Maier, D.S., Miller, M., Petersen, K., Quackenbush, J., Sherlock, G., Stoeckert, C.J., Jr., White, J., Whetzel, P.L., Wymore, F., Parkinson, H., Sarkans, U., Ball, C.A., and Brazma, A. (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, **7489**.

32 Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., and Brazma, A. (2007) ArrayExpress: a public database of microarray experiments and gene expression profiles. *Nucleic Acids Research*, **35**, D747–D750.

33 Kiyosawa, N., Manabe, S., Yamoto, T., and Sanbuissho, A. (2010) Practical application of toxicogenomics for profiling toxicant-induced biological perturbations. *International Journal of Molecular Sciences*, **11**, 3397–3412.

34 Selventa, The openBEL portal, http://www.openbel.org/, 2012.

35 Schlage, W.K., Westra, J.W., Gebel, S., Catlett, N.L., Mathis, C., Frushour, B.P., Hengstermann, A., Van Hooser, A., Poussin, C., Wong, B., Lietz, M., Park, J., Drubin, D., Veljkovic, E., Peitsch, M.C., Hoeng, J., and Deehan, R. (2011) A computable cellular stress network model for non-diseased pulmonary and cardiovascular tissue. *BMC Systems Biology*, **5**, 168.

36 Westra, J.W., Schlage, W.K., Frushour, B.P., Gebel, S., Catlett, N.L., Han, W., Eddy, S.F., Hengstermann, A., Matthews, A.L., Mathis, C., Lichtner, R.B., Poussin, C., Talikka, M., Veljkovic, E., Van Hooser, A.A., Wong, B., Maria, M.J., Peitsch, M.C., Deehan, R., and Hoeng, J. (2011)

Construction of a computable cell proliferation network focused on non-diseased lung cells. *BMC Systems Biology*, **5** 105.

37 Selventa. Reverse Causal Reasoning Methods Whitepaper. http://www.selventa .com/technology/white-papers.

38 Kumar, R., Blakemore, S.J., Ellis, C.E., Petricoin, E.F., 3rd, Pratt, D., Macoritto, M., Matthews, A.L., Loureiro, J.J., and Elliston, K. (2011) Causal reasoning identifies mechanisms of sensitivity for a novel AKT kinase inhibitor, GSK690693. *BMC Genomics*, **11**, 419.

39 Smith, J.J., Kenney, R.D., Gagne, D.J., Frushour, B.P., Ladd, W., Galonek, H.L., Israelian, K., Song, J., Razvadauskaite, G., Lynch, A.V., Carney, D.P., Johnson, R.J., Lavu, S., Iffland, A., Elliott, P.J., Lambert, P.D., Elliston, K.O., Jirousek, M.R., Milne, J.C., and Boss, O. (2009) Small molecule activators of SIRT1 replicate signaling pathways triggered by calorie restriction *in vivo*. *BMC Systems Biology*, **3**, 31.

40 Laifenfeld, D., Gilchrist, A., Drubin, D., Jorge, M., Eddy, S.F., Frushour, B.P., Ladd, B., Obert, L.A., Gosink, M.M., Cook, J.C., Criswell, K., Somps, C.J., Koza-Taylor, P., Elliston, K.O., and Lawton, M.P. (2010) The role of hypoxia in 2-butoxyethanol-induced hemangiosarcoma. *Toxicological Sciences*, **113**, 254–266.

41 Westra, J.W., Schlage, W.K., Hengstermann, A., Gebel, S., Mathis, C., Thomson, T.M., Wong, B., Hoang, V., Veljkovic, V., Peck, M., Lichtner, R.B., Weisensee, D., Talikka, M., Deehan, R., Hoeng, J., Peitsch, M.C. (2013) A modular cell-type focused inflammatory process network model for non-diseased pulmonary tissue. *Bioinformatics and Biology Insights*, **7**, 167–192.

42 Gebel, S., Lichtner, R.B., Frushour, B.P., Schlage, W.K., Hoang, V., Talikka, M., Hengstermann, A., Mathis, C., Veljkovic, E., Peck, M., Peitsch, M.C., Deehan, R., Hoeng, J., and Westra, J.W. (2013) Construction of a computable network model for DNA damage, autophagy, cell death, and senescence. *Bioinformatics and Biology Insights*, **7**, 1–21.

43 Berenjeno, I.M., Nunez, F., and Bustelo, X.R. (2007) Transcriptomal profiling of the cellular

transformation induced by Rho subfamily GTPases. *Oncogene*, **26**, 4295–4305.

44 Ramirez-Valle, F., Braunstein, S., Zavadil, J., Formenti, S.C., and Schneider, R.J. (2008) eIF4GI links nutrient sensing by mTOR to cell proliferation and inhibition of autophagy. *The Journal of Cell Biology*, **181**, 293–307.

45 Okubo, T. and Hogan, B.L. (2004) Hyperactive Wnt signaling changes the developmental potential of embryonic lung endoderm. *Journal of Biology*, **3**, 11.

46 Martin, F., Thomson, T.M., Sewer, A., Drubin, D.A., Mathis, C., Weisensee, D., Pratt, D., Hoeng, J., and Peitsch, M.C. (2012) Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. *BMC Systems Biology*, **6**, 54.

47 Ding, G.J., Fischer, P.A., Boltz, R.C., Schmidt, J.A., Colaianne, J.J., Gough, A., Rubin, R.A., and Miller, D.K. (1998) Characterization and quantitation of NF-kappaB nuclear translocation induced by interleukin-1 and tumor necrosis factor-alpha: development and use of a high capacity fluorescence cytometric system. *Journal of Biological Chemistry*, **273**, 28897–2905.

48 Chen, G., Gharib, T.G., Huang, C.C., Taylor, J.M., Misek, D.E., Kardia, S.L., Giordano, T.J., Iannettoni, M.D., Orringer, M.B., Hanash, S.M., and Beer, D.G. (2002) Discordant protein and mRNA expression in lung adenocarcinomas. *Molecular & Cell Proteomics*, **1**, 304–313.

49 Guo, Y., Xiao, P., Lei, S., Deng, F., Xiao, G.G., Liu, Y., Chen, X., Li, L., Wu, S., Chen, Y., Jiang, H., Tan, L., Xie, J., Zhu, X., Liang, S., and Deng, H. (2008) How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. *Acta Biochimica et Biophysica Sinica (Shanghai)*, **40** 426–436.

50 Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology*, **4**, 117.

51 Martin, F., Sewer, A., Talikka, M., Xiang, Y., Hoeng, J., and Peitsch, M.C. (2013) Quantification of biological network perturbations: Impact assessment and

diagnosis using causal biological networks, submitted to "Bioinformatics".

52 Thomson, T.M., Sewer, A., Martin, F., Belcastro, V., Frushour, B., Gebel, S., Park, J., Schlage, W.K., Talikka, M., Vasilyev, D., Westra, J.W., Hoeng, J., and Peitsch, M.C. (2013) Quantitative assessment of biological impact using transcriptomic data and mechanistic network models, submitted to "Toxicology and Applied Pharmacology".

53 Church, T.R., Anderson, K.E., Caporaso, N.E., Geisser, M.S., Le, C.T., Zhang, Y., Benoit, A.R., Carmella, S.G., and Hecht, S.S. (2009) A prospectively measured serum biomarker for a tobacco-specific carcinogen and lung cancer in smokers. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, **18**, 260–266.

54 Yuan, J.M., Koh, W.P., Murphy, S.E., Fan, Y., Wang, R., Carmella, S.G., Han, S., Wickham, K., Gao, Y.T., Yu, M.C., and Hecht, S.S. (2009) Urinary levels of tobacco-specific nitrosamine metabolites in relation to lung cancer development in two prospective cohorts of cigarette smokers. *Cancer Research*, **69**, 2990–2995.

55 Fan, X., Lobenhofer, E.K., Chen, M., Shi, W., Huang, J., Luo, J., Zhang, J., Walker, S.J., Chu, T.M., Li, L., Wolfinger, R., Bao, W., Paules, R.S., Bushel, P.R., Li, J., Shi, T., Nikolskaya, T., Nikolsky, Y., Hong, H., Deng, Y., Cheng, Y., Fang, H., Shi, L., and Tong, W. (2010) Consistency of predictive signature genes and classifiers generated using different microarray platforms. *The Pharmacogenomics Journal*, **10**, 247–257.

56 Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

57 Andersen, M.E., Clewell, H.J., 3rd, Bermudez, E., Dodd, D.E., Willson, G.A., Campbell, J.L., and Thomas, R.S. (2010) Formaldehyde: integrating dosimetry, cytotoxicity, and genomics to understand dose-dependent transitions for an endogenous compound. *Toxicological Sciences*, **118**, 716–731.

58 Monticello, T.M., Swenberg, J.A., Gross, E.A., Leininger, J.R., Kimbell, J.S., Seilkop, S., Starr, T.B., Gibson, J.E., and Morgan, K.T. (1996) Correlation of regional and nonlinear formaldehyde-induced nasal cancer with proliferating populations of cells. *Cancer Research*, **56**, 1012–1022.

Index