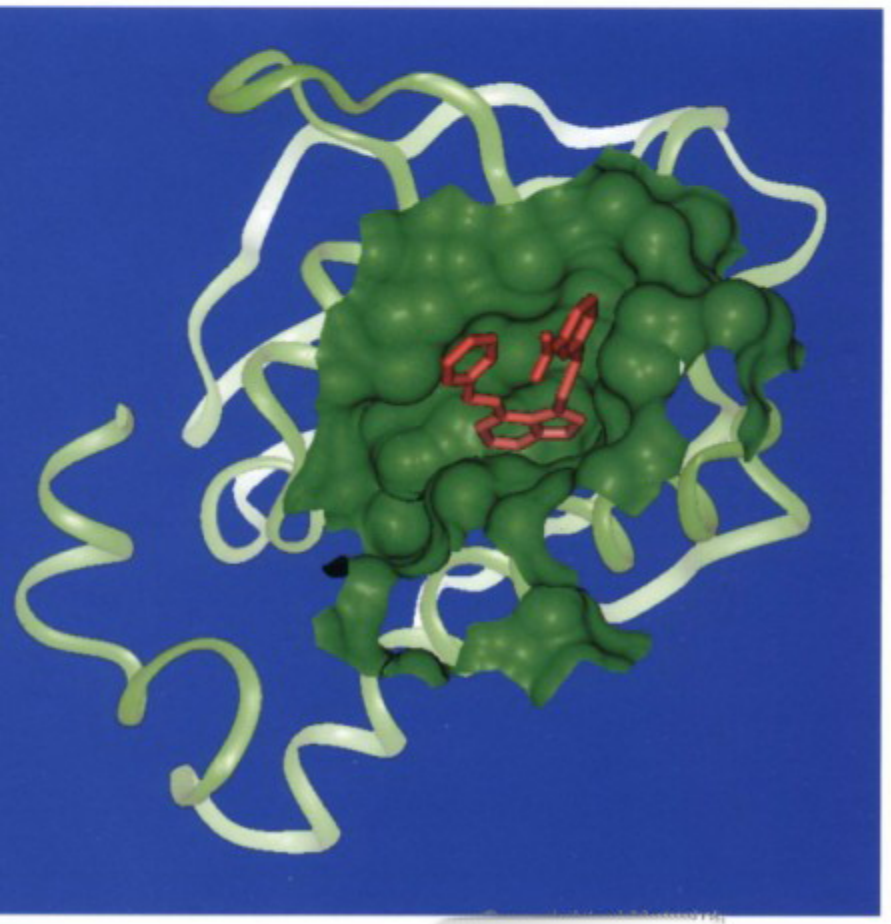
WILEY-VCH

Virtual Screening for Bioactive Molecules

Edited by H.-J. Böhm and G. Schneider



Methods and Principles in Medicinal Chemistry

Volume 10

Edited by
R. Mannhold,
H. Kubinyi,
H. Timmerman

Virtual Screening for Bioactive Molecules

Edited by Hans-Joachim Böhm and Gisbert Schneider

WILEY-VCH

Methods and Principles in Medicinal Chemistry

Edited by
R. Mannhold
H. Kubinyi
H. Timmerman

Virtual Screening for Bioactive Molecules

Edited by
Hans-Joachim Böhm and Gisbert Schneider

WILEY-VCH

Weinheim · New York · Chichester · Brisbane · Singapore · Toronto

Series Editors:
Prof. Dr. Raimund Mannhold
Biomedical Research Center
Molecular Drug Research Group
Heinrich-Heine-Universität
Universitätsstraße 1
D-40225 Düsseldorf
Germany

Prof. Dr. Hugo Kubinyi
Combinatorial Chemistry
and Molecular Modelling
ZHF/G, A 30
BASF AG
D-67056 Ludwigshafen
Germany

Prof. Dr. Hendrik Timmerman
Faculty of Chemistry
Dept. of Pharmacochemistry
Free University of Amsterdam
De Boelelaan 1083
NL-1081 HV Amsterdam
The Netherlands

Volume Editors:
Dr. Hans-Joachim Böhm,
Dr. Gisbert Schneider
Division Pharmaceuticals
F. Hoffmann-La Roche Ltd.
CH-4070 Basel
Switzerland

Cover illustration: A novel DNA gyrase inhibitor, which was designed with the help of Virtual Screening methods at F. Hoffmann-La Roche Ltd., Basel. For details, see H.-J. Böhm, M. Boehringer, D. Bur, H. Gmuender, W. Huber, W. Klaus, D. Kostrewa, H. Kuehne, T. Luebbers, N. Meunier-Keller, F. Mueller, *J. Med Chem.* 2000, *43*, 2664-2674. Graphics created using Insight II (Molecular Simulations Inc., San Diego).

Library of Congress Card No. applied for.

British Library Cataloguing-in-Publication Data: A catalogue record for this book is available from the British Library.

Die Deutsche Bibliothek – CIP Cataloguing-in-Publication-Data
A catalogue record for this publication is available from Die Deutsche Bibliothek

ISBN 3-527-30153-4

Composition: Datascan GmbH, D-64295 Darmstadt
Printing: Druckhaus Darmstadt GmbH, D-64295 Darmstadt
Bookbinding: Osswald & Co., D-67433 Neustadt (Weinstraße)

Printed in the Federal Republik of Germany.

Preface

The present volume of the series "Methods and Principles in Medicinal Chemistry" focuses on a timely topic: Virtual Screening. This new branch of Medicinal Chemistry represents an attractive alternative to high-throughput screening techniques.

Virtual Screening involves several techniques. The handling and screening of large databases with clustering and similarity searching deserves mention. Once a reduced selection has been obtained, docking or alignment techniques gain prime impact. Alternatively, libraries based on principles from combinatorial chemistry or NMR evidence can be screened for complementarity with binding sites, or for similarity with lead structures. Computational speed and reliable scoring functions are essential in Virtual Screening, both for docking and for alignment. Structure-based design and molecular modelling were introduced to Medicinal Chemistry in the 1980s. Presently, a second wave of computational chemistry is rapidly gaining impact; and Virtual Screening forms the centerpiece of this development.

The book begins with a general introduction to the basics, including experimental screening techniques. Subsequent chapters highlight the crucial concepts in chemical library profiling, similarity searching and diversity assessment, property prediction, drug-likeness assessment, docking, and SAR modelling. Many novel techniques are described and several successful applications are presented to highlight their usefulness. The twelve chapters, written by leading experts in the field, cover the major aspects of an important discipline in Medicinal Chemistry in an authoritative and easy-to-read fashion.

This book on Virtual Screening is at the forefront of science; and all techniques have been developed in very recent years. Research groups that are able to implement these methods for their drug and agro research, as well as for the prediction of bioavailability and toxicity, will have a competitive advantage. The final goal of a fast and reliable estimation of affinity constants and biological properties, in general, depends heavily on our ability to improve the underlying scoring functions. This will be the next step and the next problem to be solved. Definitely, research on the rational selection and evaluation of compounds from huge series will be further stimulated by this book. The editors of the series are grateful to the contributors to this volume, in particular Hans-Joachim Böhm and Gisbert Schneider as well as Wiley-VCH publishers, for this extremely fruitful collaboration.

August 2000

Raimund Mannhold, Düsseldorf
Hugo Kubinyi, Ludwigshafen
Henk Timmerman, Amsterdam

A Personal Foreword

This book gives an introduction to a broad collection of Virtual Screening methods. What is Virtual Screening? Basically, this term summarizes various computer-based methods for the systematic selection of potential drug candidates from a large set of molecular structures. Molecular properties taken into account include binding to the protein target, physicochemical properties and – whenever possible – pharmacokinetic attributes. We are convinced that Virtual Screening is an indispensable tool for medicinal chemists. It can be used to analyze large collections of chemical and biological data, and to offer novel suggestions on how to move forward in a drug design project. Virtual Screening methods are decision support systems. It is evident that such approaches will have a continually increasing impact on the Drug Discovery process.

The basic concepts of Virtual Screening have already been outlined more than a decade ago. Since then, we have witnessed the introduction of high throughput screening and combinatorial chemistry. Combined with new computational concepts and algorithms, we are now perfectly positioned to fully capitalize on a synergistic use of in silico tools and "wet bench chemistry". The book covers both established and more recent techniques that are still in a more exploratory status. It should be stressed that currently Virtual Screening is in a rapid development process leading to many new ideas and applications with several scientific disciplines being involved. Therefore, it is impossible to fully cover all facets of this exciting discipline in the present volume. Nevertheless, we trust that the book will be useful to a broad range of researchers in pharmaceutical industry and academia.

All authors are very much thanked for their great enthusiasm and excellent contributions. Without their willingness and support editing this book would not have been possible. We equally wish to thank Hugo Kubinyi for many helpful comments and encouragement, Gudrun Walter from Wiley-VCH for great engagement and careful editing, and our colleagues at F. Hoffmann-La Roche for many stimulating discussions and valuable support.

Basel, July 2000

Hans-Joachim Böhm
Gisbert Schneider

Contents

List of Contributors

Dr. Dimitris K. Agrafiotis

3-Dimensional Pharmaceuticals, Inc.
665 Stockton Drive, Suite 104
Exton, PA 19341
USA

Dr. John M. Barnard

Barnard Chemical Information Ltd.
46 Uppergate Road
Stannington
Sheffield, S6 6BX
UK

Dr. Hans-Joachim Böhm

F. Hoffmann – La Roche Ltd.
Pharmaceuticals Division
CH-4070 Basel
Switzerland

Dr. Gianpaolo Bravi

Computational Chemistry and Informatics
Glaxo Wellcome R&D
Gunnels Wood Road
Stevenage, Hertfordshire, SG1 2NY
UK

Dr. Pierre Bruneau

AstraZeneca Ltd.
8AF15, Mereside
Alderley Park, Macclesfield
Cheshire, SK10 4TG
UK

Dr. Geoffrey M. Downs

Barnard Chemical Information Ltd.
46 Uppergate Road
Stannington
Sheffield, S6 6BX
UK

Dr. Richard M. Eglen

LJL BioSystems, Inc.
404 Tasman Drive
Sunnyvale, CA 94089
USA

Dr. Emanuela Gancia

Computer-Aided Drug Design
Celltech Chiroscience
Cambridge Science Park, Milton Road
Cambridge, CV4 4WE
UK

Dr. Holger Gohlke

Universität Marburg
Institut für Pharmazeutische Chemie
Marbacher Weg 6
D-35032 Marburg
Germany

Dr. Andrew C. Good

Bristol-Myers Squibb
5 Research Parkway
P.O. Box 5100
Wallingford, CT 06492-7660
USA

Dr. Ulrich Grädler

Universität Marburg
Institut für Pharmazeutische Chemie
Marbacher Weg 6
D-35032 Marburg
Germany

Dr. Darren V. S. Green

Computational Chemistry & Informatics
GlaxoWellcome R&D
Gunnels Wood Road
Stevenage, Hertfordshire, SG1 2NY
UK

Dr. Sven Grüneberg

Universität Marburg
Institut für Pharmazeutische Chemie
Marbacher Weg 6
D-35032 Marburg
Germany

Dr. Mike M. Hann

Computational Chemistry & Informatics
GlaxoWellcome R&D
Gunnels Wood Road
Stevenage, Hertfordshire, SG1 2NY
UK

Dr. Sergei Izrailev

3-Dimensional Pharmaceuticals, Inc.
665 Stockton Drive
Exton, PA 19341
USA

Prof. Gerhard Klebe

Universität Marburg
Institut für Pharmazeutische Chemie
Marbacher Weg 6
D-35037 Marburg
Germany

Dr. Jonathan Knowles

F. Hoffmann – La Roche Ltd.
Pharmaceuticals Division
CH-4070 Basel
Switzerland

Dr. Oliver Krämer

Universität Marburg
Institut für Pharmazeutische Chemie
Marbacher Weg 6
D-35032 Marburg
Germany

Dr. Victor S. Lobanov

3-Dimensional Pharmaceuticals, Inc.
665 Stockton Drive
Exton, PA 19341
USA

Dr. Jonathan S. Mason

Bristol Myers Squibb
Mailstop H23-07
P.O. Box 4000
Princeton, NJ 08540
USA

Dr. Jeff Morris

AstraZeneca Ltd.
8AF15, Mereside
Alderley Park, Macclesfield
Cheshire, SK10 4TG
UK

Dr. Stephen D. Pickett

Aventis
Rainham Road South
Dagenham, Essex, RM10 7XS
UK

Dr. Dmitrii N. Rassokhin

3-Dimensional Pharmaceuticals, Inc.
665 Stockton Drive
Exton, PA 19341
USA

Dr. Jens Sadowski

Structural Chemistry Laboratory
AstraZeneca R&D Mölndal
S-43183 Mölndal
Sweden

Dr. Gisbert Schneider

F. Hoffmann – La Roche Ltd.
Pharmaceuticals Division
CH-4070 Basel
Switzerland

Dr. Martin Stahl

F. Hoffmann – La Roche Ltd.
Pharmaceuticals Division
CH-4070 Basel
Switzerland

Dr. Patrick Walters

Vertex Pharmaceuticals, Inc.
140 Waverly Street
Cambridge, MA 02139
USA

Dr. Lutz Weber

Morphochem AG
Am Klopferspitz 19
D-82152 Martinsried
Germany

Prof. Peter Willett

Krebs Institute for Biomolecular Research
and Department of Information Studies
University of Sheffield
Sheffield, S10 2TN
UK

Why is Virtual Screening Important?

Jonathan Knowles

One of the important abilities that mankind acquired in the twentieth century was the ability to discover highly active, small, organic molecules that are used to treat people who are sick. This ability was developed by combining the amazing discoveries in synthetic organic chemistry with information from clinical chemistry to give us the powerful medicines available today. Much of the progress to date has been made by creating molecules based on the structures of natural signalling molecules such as histamine and serotonin. The derivative molecules were then used to investigate the specificity of the biology through the science of pharmacology. Additionally, natural products have been identified and modified to provide a number of important therapeutic advances.

The rapid integration of computer systems in almost every aspect of human existence, and the probability that – despite the difficulties – this process increases the wellbeing of individuals and societies that embrace this technology suggests that computers will play an increasingly important role also in the practise and evolution of medicine in the future. This will be particularly true for the Drug Discovery process.

Virtual screening has its roots in computational chemistry and in structural biology. In the 1970s, the development of structural biology and the growing availability of atomic structures of diverse proteins, led to the hope that it would be possible to identify new medicines by first solving the structure of the potential drug target at the atomic level and then using this information to design small molecules that had the required effect. However, while it is true to say that protein structures have been sometimes very useful in guiding the thinking of creative medicinal chemists, the number of cases where a drug has been actually designed *de novo* is small. This is due in part to the relatively low resolution of the structures and the earlier limitation computing power which made realistic simulations difficult in practise.

Over the past ten years, the three technical revolutions of molecular biology, automation and informatics occurred simultaneously, and these are now bringing about major changes in biology and chemistry. These three revolutions are dramatically altering the way in which we look for new medicines and more importantly, the opportunities to greatly improve the practise of medicine.

Combining molecular biology, informatics and automation has brought us *Genomics*, i.e. an increasingly complete list of all the genes and all the proteins that make up the biology of man, with a growing understanding of the primary function of each gene and protein. This list, in itself, does not tell us which proteins to modulate in order to treat a specific disease but it shows us the staggering number of new unexplored opportunities. The likely publication of the full list of human gene before readers read this book is evidence of the dramatic rate of progress

in this area. The same three technologies are also being applied to identify the genetic risk factors which help us to identify the mechanisms most likely to be useful in treating disease. So the number of potential and partially validated drug targets is increasing dramatically.

At the same time there has been an important change in the philosophy of medicinal chemistry. Formerly, chemists only synthesized large quantities of pure substance, which was then tested in complex and sophisticated biological systems. Bringing together synthetic chemistry, automation and informatics has given rise to a way of thinking about chemistry where the diversity of the compounds synthesized is more important than the quantity and to some extent quality. This is known variously as parallel synthesis or combinatorial chemistry and has lead to the creation of new groups of chemists in companies both large and small, and in universities around the world. The design of new chemical libraries, some very large and some much smaller and more focused around particular pharmacophores is now seen as a critical activity as part of the discovery of new medicines.

The application of automation and informatics to biology has lead to the development of *ultra*-high-throughput screening systems using miniaturized biology, often with fluorescent detection of relevant changes. A few of the leading pharmaceutical companies now have the ability to carry out over a million assays per day and, through globally integrated informatics systems, to automatically capture data from all these experiments no matter where they are carried out in the world. These systems are now being used to bring together the new targets from Genomics and the new chemical diversity from combinatorial chemistry. In addition, in many pharmaceutical companies, there is a growing body of information that relates to the currently less predictable properties of molecules like toxicity and oral bioavailability in animals and man. Thus a very large amount of information is now being generated and stored. This infomation could be used to predict the classes of molecule more likely to be medicines much faster than by carrying out the physical experiment. In addition, the rapid growth of the number of compounds in chemical libraries and the number of potential targets from Genomics makes virtual screening critically important for the future.

Better decisions come from having access to all the relevant information and the ability to analyze these data in such a way that real knowledge is created. We stand at the threshold of a new century in which informatics will become as essential a tool for research groups in biology and chemistry as it is today for physics. The creation of new knowledge from vast assemblies of disparate data using novel informatics approaches is one of the most exciting scientific challenges of our age. This will be nowhere more true than for those of us who wish to discover effective new medicines for untreated diseases. Those groups that can collect, analyze, and interpret the dramatically increasing quantities of relevant information to allow better decisions will clearly be more successful at identifying the effective medicines of the future.

In summary, the role of informatics will be critical as a primary interface between biology, chemistry, and medicine. Discovering new medicines absolutely requires the integration of diverse information from medicine, fundamental biology, genetics and Genomics, and chemistry. The generation of "real" knowledge from this diverse information will give us the ability to understand and predict the relationship between biology and chemistry – and this is the centerpiece of virtual screening. It is also one of the most exciting areas of science. The virtual screening of molecules to identify new medicines as described in the following Chapters is today already an important issue for those who wish to be successful in this art. In the near future it will become absolutely central to the whole process of effective Drug Discovery.

Virtual Screening for Bioactive Molecules
Edited by Hans-Joachim Böhm and Gisbert Schneider

1 High-Throughput Screening and Virtual Screening: Entry Points to Drug Discovery

Richard M. Eglen, Gisbert Schneider, Hans-Joachim Böhm

1.1 Introduction

The goal of pharmaceutical research is to discover new molecules with a desired biological activity that are useful in the efficient and safe treatment of human diseases. The discovery process is quite complex and can be divided into several steps. The first step is normally the selection of a molecular target, e.g. an enzyme or a receptor that is associated with the disease. This selection process is still primarily driven by searching publications, although bioinformatics in concert with genetics and genomics/proteomics play an increasingly important role in the target selection process. One may attempt an early target validation, e.g. by "knock-out" experiments monitoring the effects of corresponding gene loss to an organism. The next step is the identification of lead molecules (*lead identification phase*). In most discovery programs, once the biological target is validated a series of robust and miniaturized biological assays is set up, and several hundred of thousands of compounds are tested in this "primary screening" round. High-throughput screening (HTS) is a routine aspect of drug discovery in almost every large, fully integrated pharmaceutical company. All major pharmaceutical companies have invested heavily in the process of HTS by:

- setting up biological assays that can be processed rapidly using small amounts of material,
- building large collections of chemical compounds typically in the range of 100000–1000000 molecules, and
- storing the compounds in a way that is suitable for rapid access and retrieval.

Once active compounds (often termed "hits") are identified, the potency is estimated in a second screening run, and their chemical structure and mechanism of action are determined. These are generally referred to as "validated hits". If both structure and biological activity can be confirmed and the compound is considered chemically tractable, a further exploration of validated hits is started: the *lead development phase*. The goal is to further characterize the compound class, establish a structure–activity relationship for the initial hit including closely related molecules. If compounds are identified that bind tightly to the molecular target (typically, for receptor binding, a K_i in the nanomolar range is considered to indicate sufficiently tight binding), these molecules are subsequently characterized more carefully. For this purpose, a number of other properties besides the K_i are also taken into account, such as bioavailability, metabolic and chemical stability, physicochemical properties (solubility, lipophilicity), selectivity, etc.

In its broadest sense, HTS is an automated process that rapidly assays large numbers of compounds (10^4–10^5 and above) against a target and subsequently analyses the data to iden-

tify novel chemical leads [1]. *Ultra* HTS (uHTS) is a technical extension of HTS, in which more than 100000 compounds can be screened daily, notably with minute assay volumes. Consequently, uHTS is associated with the highly sophisticated handling of small fluid volumes and highly sensitive assay detection systems. Therefore, of all the innovative technological initiatives applied to pharmaceutical research in the 1990s, HTS is probably the best integrated. The speed of this integration of HTS in drug discovery efforts clearly arises from the plethora of biological targets and the notable lack of chemical leads from which medicinal chemistry programs can be initiated.

The growth of HTS in the pharmaceutical industry can be illustrated in the following numbers: in 1998, HTS laboratories read on average 55000 wells per week and by 2003 will be reading an average 350000 wells per week, an increase of more than 500%. HTS is thus evolving into a technology-driven process, that in the future will be expected to deliver two to five series of lead compounds per project as a starting point for medicinal chemistry and lead optimization [2,3]. Several years ago it was recognized that without a dramatic expansion of random screening libraries, HTS processes would quickly exhaust the number of available compounds. In the last decade or so, consequently, the need to develop chemical parallel synthesis technologies largely stems from the speed at which compounds could be screened [3]. Furthermore, as occurred in genomic and genetic research, the very large volume of data arising from HTS drove the need for sophisticated database archiving and searching methods [4]. Related to this has arisen the concept of virtual screening, in which the activity space of compounds for novel targets can be assessed *in silico*, and predictions of pharmacokinetic properties are considered in the very early phases of the drug discovery process.

Drug discovery has thus seen a rebirth of random screening, as practiced in the early days of pharmaceutical research, although admittedly at a higher level of sophistication [2]. Due to the range and novelty of the targets under examination, the dialogue between chemistry and biology in the rational design of chemical leads generally occurs after an HTS screening campaign. Currently HTS is therefore a field under constant pressure to accelerate assay throughput and reduce assay costs while maintaining flexible platforms with which to screen

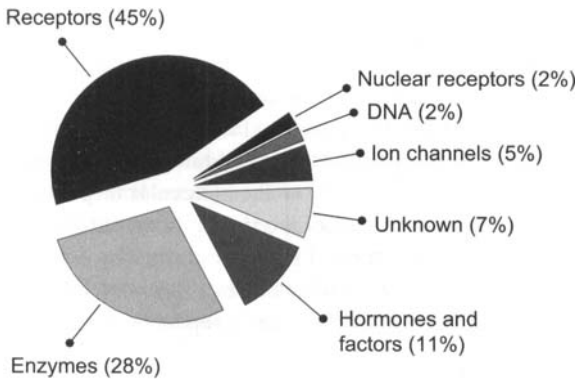


Figure 1.1. Classes of 485 current drug targets. Adapted from reference [20].

several target types. Notable in HTS innovation is the development of highly sensitive detection strategies allowing assay miniaturization with consequent reduction in both the costs per data point and amounts of compound or reagents consumed.

The dominant target classes at which HTS assays are directed are listed in Figure 1.1. This reflects the historical success with which leads have been found and the commercial success of drugs directed at these targets. Currently, there are about 500 targets for which marketed drugs are available. This list is composed of the total number of receptors, ion channels and enzymes in the human that can be modulated to alleviate disease [2]. In the future, of course, the diversity of targets will increase, given the size of the human genome and the genetic analysis underway to identify targets associated with specific diseases.

1.2 Miniaturization and Detection Strategies

To understand and assess the various virtual screening approaches presented and discussed in this book, it is necessary to comprehend some more details associated with HTS. Two major aspects, advanced miniaturization and detection strategies, made possible the technological breakthrough in HTS. These developments and their relation to virtual screening are briefly discussed in the following.

1.2.1 Screening Plate Format and Fluidics

The dominant drive to identify novel detection strategies is to continuously move to greater and greater assay miniaturization [3,4]. The 96-well microtiter plate was originally planned as a tool for increasing the throughput of viral titer assays. Later, ELISA assays using this format were developed, and were associated with liquid handling devices and plate readers, all predicated on the 96-well format. HTS devices were developed from these technologies and all subsequent plate arrays (384, 1536, 3456 wells, etc.) and associated devices have maintained the 96-well plate aspect ratio [4].

Miniaturization of the assay, initiated several years ago by the introduction of the 384-well screening plate, immediately provided two key benefits. First, decreased cost per datapoint and second decreased consumption of the screening library. The latter resulted in a reduced need for amount of synthesized compound; a critical issue, given the small amounts of compound generally synthesized combinatorially. Although the additional benefit of increasing the assay throughput clearly occurs, this is not usually the principal driver.

Critical milestones in the miniaturization efforts have emanated from developments in the microliter volume control of diverse fluids. Two problems have been solved in the last three years or so in this regard. First, reformatting (moving compounds into miniaturized formats) and dispensing (placing of the reagents into the microwell). The dispenser must therefore place identical fluid volumes accurately with each test sample with minimal contamination – a requirement met by exploiting the ink jet technology for high-speed printing. Reformatting requires the handling of stock solutions, usually in DMSO, without contamination. In most systems, the rate-limiting step of the screening operation lies in the reformatting, rather than in the assay *per se* [4].

1.2.2 Detection Strategies

Historically, one of the most widely used and successful screening strategies is radioligand binding, particularly at cell-surface receptors. Although several technical innovations have been introduced that have accelerated the throughput of the technique, the inherent disadvantages of separation and multi-step separation methods are maintained. New, non-separation approaches include the development by Amersham International [6] of the scintillation proximity assay (SPA) and depend upon the radioligand being brought into sufficiently close proximity to a fluorophore-containing microsphere for energy transfer to occur. This is accomplished by immobilizing the target molecule for the ligand on a microsphere. Light is emitted as the assay signal; and the amount emitted is proportional to the quantity of radioligand bound. The radioligand in free solution is undetected. Although ideally suited to surface bound receptors, the technique has been applied to several protein–protein interactions assays [5].

Many isotope-based assays have analogous fluorescent-based approaches, inherently amenable to miniaturization techniques [4]. The trend away from isotopic means of detection has accelerated and over 90% of screening assays will be amenable to the use of fluorescence analysis within two years. The attractiveness of the approach lies with ratiometric dyes, in which the signal to noise ratio of the assay is markedly enhanced and thus assay volumes of minute proportions (moving from 50–200 μl to 1–2 μl) become highly feasible. Equally important has been the development of robust, "addition only" assay techniques [6,7].

Fluorescence approaches as detection strategies for HTS comprise four main categories: fluorescence resonance energy transfer (FRET), time-resolved fluorescence (TRF), fluorescence correlation spectroscopy (FCS) and fluorescence polarization (FP). FRET occurs between two fluorophores in close proximity with suitably overlapping spectra. In general, excitation of a donor fluorophore can result in a transfer of energy from the donor to an acceptor with a longer wavelength of emission. An event causing an increase in the distance between these two fluorophores can be measured by disruption in the FRET through re-establishment of fluorescence emission from the donor. In HTS, the two fluorophores can be brought together by a linker moiety, or by being confined by the cell membrane, as occurs in the case of novel voltage sensitive dyes. A major advantage of the technique is that the results produced are in the format of a ratio between the two wavelengths, thus reducing background artifacts, such as differences in cell number, probe concentration or light fluctuations. For these reasons, FRET assay systems have been extensively employed in assay miniaturization [6].

Time-resolved fluorescence (TRF) intensity is based principally upon the use of fluorescent lanthanide ions in the context of homogeneous assay systems. These assays are highly sensitive and non-isotopic [7]. Homogeneous TRF (HTRF) has been developed by Canberra–Packard and uses a Europium chelate ion caged in a proprietary macropolycyclic ligand containing 2,2′ bipyridines as light absorbers. Energy is transmitted form the Eu cryptate when excited at 337 nm to a fluorescence acceptor molecule: a modified proprietary allophycocyanin, XL665. In the presence of pulsed laser light, energy is transferred from the Eu cryptate to the XL665, resulting in the emission of light at 665 nm over a prolonged time scale (microseconds). The signal is thus distinguished from natural fluorescence occurring in the absence of the cryptate. The energy transfer efficiency is thus 75% over 7.5 nm and

rapidly diminishes over greater distances. Consequently the technique is ideal for homogenous assays since molecules in free solution are unlikely to emit a signal and are rarely in close enough proximity. HTRF is commercially available, using Discovery instrumentation, and is amenable to 96- and 384-well formats. It works optimally with *in vitro* biochemical assays, although cell-based assays are potentially under development. LANCE, alternatively, employs technology developed by Wallac and is based around lanthanide rather than Europium chelates. The technology has been used for receptor ligand binding assays, as well as protein–protein interactions. A multiplate reader, the Victor, has been commercially developed and can read both 96- and 384-well formats, again with various assay formats [7].

1.2.3 Cell-Based Reporter Gene Assays

In the context of cell-based assays, FRET has been extensively used in the development of sensitive cell-based reporter gene assays [8]. The use of cell systems in HTS and uHTS is increasing. CHO cells, for example, can be dispensed into plate formats as high as 3456 (approx. 250 cells per well). Smaller assays, using single cell systems, may pose insuperable problems for HTS due to stochastic variations in cell populations and the variability in response [3,6]. A reporter gene construct consists of an inducible transcriptional control element driving the expression of a reporter gene. In mammalian cells that stably express a reporter gene construct, the functional reporter response can be used to select appropriate cell lines for HTS. The first reporter gene to be widely used was chloramphenicol acetyltransferase. However, as its use was laborious and required radioisotopes, agents such as luciferase from *Photinus pyralis* and principally green fluorescent protein (GFP) have superceded its use in HTS [6,9].

GFP from marine animals such as *Aequorea victoria*, and most recently from soft sea coral in the phylum *Cnidaria*, emit light when energy is transferred from the calcium activated photoprotein aequorin [9]. The cloning of wild type GFP and several mutants have established the protein as a powerful reporter for several research applications including HTS. When the GFP gene is expressed in mammalian cells and illuminated by bright light, GFP emits a bright green fluorescent light that is easily detected. Enhanced GFP (EGFP) has mutated changes of Ser65 to Thr65, as well as 190 silent base changes to contain residues preferentially expressed in human proteins. EGFP has the advantages of enhanced detection, improved solubility, more efficient protein folding, and faster chromophore oxidation to produce the fluorescent form of the protein. Other variants of EGFP include cyan, red, and yellow that with different detection filters, permit simultaneous analysis of multiple gene expression cascades, including protein translocation episodes [9].

Most important in the context of HTS has been the appropriate combination of GFP color variants to develop cell-based assays for molecular proximity based on FRET techniques. The compact β can structure of GFP renders it extremely stable; a serious disadvantage when using GFP to monitor changes in gene expression. In HTS this leads to high backgrounds in the assay readout. Consequently, companies such as Clontech have identified several destabilized forms of GFP (dGFP) for use in HTS. Since screening assays based on dGFP engineered cell lines are inexpensive and require no washing or substrate addition steps, HTS can be undertaken in real time using living cells [9].

Aurora Biosciences have designed a novel strategy using the bacterial enzyme β lactamase for use in the reporter. When combined with a ratiometric β lactamase substrate that localizes in the cell cytoplasm, a fluorogenic reporter gene assay is possible. An advantage of the system is that individual cells loaded with the fluorescent substrate can be sorted by FACS (fluorescence activated cell sorter), and selection of optimized cell lines for assays development becomes possible. Cellomics have developed the concept of high content screening using Arrayscan technology in which detailed temporal spatial relationships of cellular proteins can be assessed [9].

1.2.4 Fluorescence Correlation Spectroscopy

The ultimate assay miniaturization lies in the detection of single molecules. This is feasible using the technique of fluorescence correlation spectroscopy (FCS). This measures temporal fluctuations in the fluorescence signal detected from the diffusion of individual fluorescent molecules in and out of a focused confocal element, usually in volumes of less than a femtoliter [10]. Interactions of single molecules can theoretically be studied by this technique, allowing nanoscale detection. The approach combines homogeneous mixtures of reagents, high sensitivity, true equilibration in complexation reactions, and a wide range of solution conditions. By choosing the appropriate fluorescence label, the readout can provide information on the size, distances, ligation state, conformational rearrangements, and sample heterogeneity. Evotec systems have exploited the approach most extensively and have commercialized the EVOscreen platform, using FCS and a proprietary single-molecule detection strategy [10].

1.2.5 Microchip Fabrication

Just as the strategies for synthesizing complex DNA arrays on small glass surfaces have greatly impacted the diagnosis of genetic diseases, similar approaches will influence HTS. Indeed, it is now possible to test the effect of a compound on the expression of a single gene, a large family of genes or segments of an organism's entire genome. Nonetheless, transcriptional assays are generally difficult to translate to the HTS format although, as commented above, reporter gene assays have found utility. More likely to emerge as an HTS trend is the use of microplate technology, in which credit card-sized glass chips are engineered to possess integrated synthesis and detection devices. The fluid is precisely moved by changes in the electrostatic forces across the fabricated channels. Although not extensively validated in HTS as yet, the potential of systems such as those being developed by Caliper and Orchid will hopefully allow the intimate association of compound synthesis and screening [4].

1.2.6 Remarks and Summary

Assay miniaturization is a continuous process and the question can reasonably be posed "how much is enough?" It has been suggested that the assay volumes of 1 μl in a 1536 for-

mat may represent a format sufficient for the foreseeable future of HTS. This reasoning comes from speculation based on the number of targets and the number of compounds. One analysis postulates that there will be approximately 10000 targets and about 10^7 compounds to be screened. This suggests that the total database for structure activity screening is 10^{12} or a relatively modest 650 million 1536 plates [4].

Although much precision has been gained through the development of advanced detection methods, a problem still remains in the reliability estimates of raw data. Success in several virtual screening techniques heavily depends on reliable and "sound" measurements in HTS. One cannot expect highly accurate predictions from computer-based experiments, which are performed using noisy input or reference data. This fact must not escape our attention when discussing and assessing virtual screening results. An additional source of noise arises from cross-experiment data collections, i.e. HTS screening data that were obtained by different detection techniques. The virtual screener is often confronted with such data, which is in part a result of the rapid technology development whenever a new detection technique enters the HTS scene.

An unknown factor in this speculation lies in potential future applications of HTS, notably into the arena of surrogate assays for ADME/toxicological screening. For example, HTS using a colorectal adenocarcinoma cell line, Caco-2 cells, may have a predictable value in the oral absorption of a compound across the intestinal lumen. Similarly, the screening against a panel of cytochrome P450 enzymes may have a predicative value in the metabolism of compounds. These considerations extend into discussions of compound optimization and are covered further below.

1.3 Compound Libraries

How successful is HTS and thus the return on the high level of investment in the technology? Some literature reports are very promising [11]. An alternative view has been posed by Drews [2] in which the success of HTS was claimed to be arguable, since the introduction of new compounds committed to full development has increased only moderately, if at all. The failure to provide higher quality leads stems in part from the non-validated nature of the biological target at the initiation of the program. However, the speed and cost effectiveness of modern HTS permits the screening of several targets to be conducted in parallel to traditional target validation procedures. Since it is unlikely that a general solution will be found to accelerate and increase the accuracy of target validation, screening of such targets that may fail as drug discovery programs will probably continue [2].

The major reason for the perceived "failure" of HTS lies in the quality of HTS screening libraries, specifically the diversity of the structural themes [12,13]. Combinatorial libraries usually provide small amounts of uncharacterized compounds for screening. Once these samples are further characterized, the data are of interest for structure–activity purposes. In most companies, these compounds are also present with the historical collection of compounds, generally derived from classical medicinal chemistry programs, most of which have very well defined chemical characteristics. Commercial compound collections can also be purchased which fall between these two extremes. Collectively, therefore, the information used to relate

biological activity and chemical structure must clearly integrate all of these types of compound, since all will be used for HTS purposes.

Although assessment of the diversity of a compound library is covered elsewhere in his volume (see Chapter 12), there are least two approaches to address the issue in the context of HTS. The first is clearly to assess the diversity space using chemical criteria and several algorithms are now available to do that. The second approach is to assess the diversity space, based on HTS operational experience. It is likely that, after extensive screening of the library at several targets and target classes, the structure–activity database will point to areas of success or failure in terms of identifying leads. Thus the library may be said to be "GPCR-rich", "kinase-rich", etc. Importantly, the operational structure–activity relationships should also facilitate design of other compound arrays.

An experiment-based understanding of the screening library diversity should also provide compounds that are "frequent hitters", i.e. compounds that are not necessarily chemically reactive, but have structures that repeatedly bind to a range of targets via unspecific interactions, or cause a false positive signal for other assay-inherent reasons. Clearly removal of these compounds from the library is an advantage in HTS, as is understanding the reason for their promiscuity of interaction.

A further HTS issue (in the context of the screening library) relates to identifying a screening library subset, ostensibly representative of the diversity of the whole library, that is screened at all targets, usually as a priority in the screening campaign. Assessment of chemical versus operational understanding of diversity is critical in the design of the library subset. Moreover, there are advantages at screening the whole library. First, since HTS or uHTS is generally unconstrained by cost or compound usage, it is as easy to screen 250000 compounds as it is to screen 12000. Second, the screening campaign increases the likelihood of finding actives, especially for difficult targets, as well as finding multiple structurally distinct leads. Indeed, a direct comparison of the approach of screening a representative library has been reported from Pfizer, in which it was noted that 32 out of the 39 leads were missed in comparison to those found by screening the whole library [14]. Alternatively, Pharmacopeia have reported that receptor antagonists for the CXCR2 receptor and the human bradykinin B1 receptor were derived from the same 150000-compound library, made using the same four combinatorial steps. Notably, this library was neither based on known leads in the GPCR field nor specifically targeted towards GPCRs. On the other hand, researchers at Organon reported that it is possible to rationally select various "actives" from large databases using appropriate "diversity" selection and "representativity" methods [15]. In Chapters 6 and 12 such virtual screening methods will be treated in detail.

The main aim of virtual screening is to select activity-enriched sets of molecules – or single molecules exhibiting desired activity – from the space of all synthetically tractable structures. How big is this space? It has been estimated that the medicinal chemist – and thus the virtual screener – is confronted with approximately 10^{100} feasible organic compounds [16]. Currently the most advanced uHTS techniques allow for testing ~10^5 compounds per day, and a typical corporate screening collection contains several hundred thousand samples. Although these facts alone represent a technological revolution, the turnover numbers still are vanishingly small compared to the size of total chemical space. As a consequence of this conclusion, even uHTS combined with fast, parallel combinatorial chemistry can only be successful if a reasonable pre-selection of molecules (or building blocks) for screening was done.

Otherwise this approach will essentially represent a random search with extremely long odds. Surprisingly, some hits do occur in the majority of current screening assays. This observation supports the assumption that most screening libraries and historical corporate compound databases are enriched in "drug-like" molecules already. This makes perfect sense because, ever since the beginning of pharmaceutical drug discovery, medicinal chemists have stored their knowledge about what they think makes a molecule inherently a drug in such libraries. Nevertheless, much improvement of both general-purpose and focused screening libraries is conceivable by the use of virtual screening techniques [16,17].

As we have learned from many years of "artificial intelligence" research, it is extremely difficult (if not impossible) to develop virtual screening algorithms mimicking the medicinal chemists' gut feeling. Furthermore, there is no common "gut feeling" as different chemists have different educational background, skills and experience. Despite such limitations there is, however, substantial evidence that it is possible to support drug discovery in various ways by help of computer-assisted library design and selection strategies. There are two specific properties of computers, which make them very attractive for virtual screening applications:

1. By help of virtual synthesis hitherto unknown parts of chemical space can easily be explored, and

Table 1.1. Some chemical structure databases frequently used as compound sources or reference data collections in virtual screening.

Database	No. of structures	Description
ACD[1]	> 250000	*A*vailable *C*hemicals *D*irectory; catalogue of commercially available specialty and bulk chemicals from over 225 international suppliers
Beilstein[2]	> 7000000	Covers organic chemistry from 1779
CSD[3]	> 200000	*C*ambridge *S*tructural *D*atabase; experimentally determined three-dimensional structures of small molecules
CMC[1]	> 7000	*C*omprehensive *M*edicinal *C*hemistry database; structures and activities of drugs having generic names (on the market)
MDDR[1]	> 85000	*MACCS-II Drug Data Report*; structures and activity data of compounds in the early stages of drug development
MedChem[4]	> 35000	*Med*icinal *Chem*istry database; pharmaceutical compounds
SPRESI[4]	> 3400000	Substances and bibliographic data abstracted from the world's chemical literature
WDI[5]	> 50000	*W*orld *D*rug *I*ndex; pharmaceutical compounds from all stages of development

[1] Molecular Design, San Leandro, CA, USA
[2] Beilstein Informationssysteme, Frankfurt, Germany
[3] CSD Systems, Cambridge, UK
[4] Daylight Chemical Information Systems, Claremont, CA, USA
[5] Derwent Information, London, UK

2. the speed and throughput of virtual testing can be far ahead of what is possible by means of "wet bench" experimental systems.

Once a reliable virtual screening process has been defined, it can help to save resources and limit experimental efforts by suggesting defined sets of molecules. Several such applications will be presented in the following Chapters of this book.

Two complementary compound sources are accessible for virtual screening, databases of known structures and *de novo* designs (including enumerated combinatorial libraries). Some major databases frequently employed for virtual screening experiments are listed in Table 1.1. In addition, several companies offer large libraries of both combinatorial and historical collections on a commercial basis. Usually the combinatorial collections contain 100000–500000 structures, whereas historical collections rarely exceed 100000.

Most pharmaceutical companies now have access to HTS and uHTS technologies. They also use strategies for target identification and validation including genetic, genomic, and proteomic approaches. They may also have similarly sized and similarly diverse screening libraries. Where will lie their competitive advantage? One answer is the ability to mine the increasing amounts of data. A robust bioinformatics and cheminformatics capability is thus a vital strategy to seamlessly link HTS screening data to medicinal chemistry. Consequently, the proprietary corporate database should grow in value as an asset of the organization. Open, easy access to the biology/chemistry database, particularly when augmented by "drugable characteristics" from HTS pharmacokinetic and toxicologic studies, provide a pivotal tool in drug discovery [17].

1.4 Multi-Dimensional Optimization: Qualifying HTS Lead Candidates

It has been pointed out that while enormous investments have been made in HTS and related techniques, the output of pharmaceutical research in terms of new medicines has not increased overall [2]. It appears that a number of factors might contribute to this dilemma:

- While the number of compounds available to pharmaceutical research organizations has increased strongly over the past few years, one should keep in mind that this still represents only a tiny fraction of all possible molecules or even all molecules that can be made by today's methods of organic chemistry. Thus, even the largest compound collections represent just a small part of the molecules that could be made (see previous Section).
- In-house compound collections usually have a high historical bias reflecting the therapeutic areas that the company has worked on in the past and the type of targets. Therefore, a research organization that has, e.g. worked on the dopamine receptor subtypes D1, D2, and D3 has a high probability that some of the molecules synthesized for these projects will also show binding affinity for the D4 receptor. On the other hand, completely new targets will less likely yield hits from HTS. In many cases, HTS simply produces no suitable hits.
- It is obvious that tight binding to the target is not sufficient to qualify a molecule as a drug. Other properties such as physicochemical properties also have to be taken into account.

For example, a poor aqueous solubility may prevent a potent ligand reaching its target within the organism. It has been observed that the introduction of combinatorial chemistry has led to an increased fraction of molecules with undesirably high molecular weight above 500, high lipophilicity, and thus poor water solubility.

The whole drug discovery process can be compared with a pipeline: Pharmaceutical companies are considered to be healthy if they have a sufficient number of drug candidates with substantial economic potential "in the pipeline". Similar to an oil pipeline, the throughput in pharmaceutical research is limited by bottlenecks in the pipeline. The present book aims to describe the use of computational chemistry techniques to improve the drug discovery process by addressing the bottlenecks indicated above. The first step is to use computational techniques in the selection process to add new molecules to the general screening library. Computational tools that can be used for this purpose include (Figure 1.2):

- Similarity searching (see Chapters 4, 5, 8, and 12 in this book): one may want to exclude from synthesis or purchase molecules that have a very high similarity to compounds already present in the compound collection.
- Calculation of physicochemical properties (see Chapters 2 and 3).
- Exclusion of compounds with undesirable functional groups (see Chapters 6, 7, and 12).

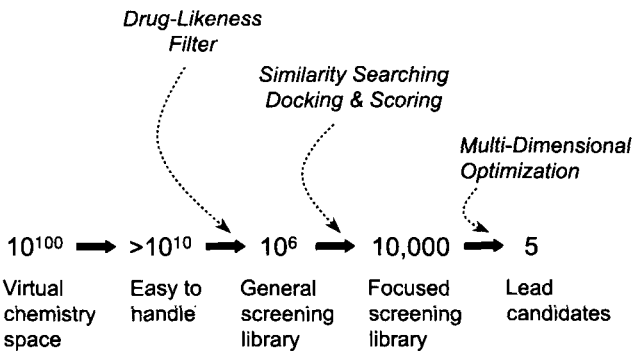


Figure 1.2. Virtual screening influences candidate selection from a large collection of feasible molecules to promising lead candidates at various stages of the Drug Discovery pipeline. Some examples are shown (linked via dotted lines to the pipeline). These virtual screening approaches are paralleled and fed with data by HTS/uHTS and other *in vitro* and *in vivo* assays.

The second step is to look at receptor-specific interactions. Again, a number of approaches will be discussed in this book:

- Use the known 3-D structure of the target in the design of new molecules, e.g. from virtual combinatorial libraries (see Chapters 10 and 11).
- Pharmacophore patterns can be used to select molecules with a higher probability than random of binding to a certain molecular target (see Chapter 7).

Finally, the last challenge is to focus on those hits with the best physicochemical and pharmacokinetic properties. This is explained in a bit more detail in the following.

The introduction of combinatorial chemistry, HTS and the presence of large compound selections have put us in the comfortable position (at least for certain classes of molecular targets) that there is a large number of hits to choose from for lead optimization. Assuming a hit rate of 0.1–1% and a compound collection size of 10^6 compounds, we have (or will have) about 1000–10000 hits that are potential starting points for further work. It is important to realize that while the screening throughput has increased significantly, the throughput of a traditional chemistry laboratory has not. While it is true that automated and/or parallel chemistry is now routinely used, there are still many molecules that are not amenable to these more automated approaches. Therefore the question is: "How can subsequent lead optimization fully exploit this increased amount of information?" Computational techniques can be used to address the question in a variety of fashions:

- Many of the hits are false positives, which means that the observed effect in the assay is not due to the specific binding to a certain pocket in the molecular target. Docking techniques can be used to place all compounds from the chemical library in the binding pocket and score them. If HTS data are available, a comparison of the *in silico* docking results and the HTS data can be used to prioritize the hits and focus the subsequent work on the more promising candidates. If no HTS data are available (e.g. when no assay amenable to HTS is possible or if no hits were obtained) then docking can be used to select compounds for biological testing.
- Many of the compounds have undesirable properties such as a low solubility, or high lipophilicity. *In silico* prediction tools can be used to rank the HTS hits. While it is generally true that insufficient PC properties can be remedied in the lead optimization process (e.g. large, lipophilic sidechains can be removed, or a "pro-drug" can mask highly acidic groups), it may be advisable if possible to focus on compounds without obvious liabilities.
- Toxicity and metabolic stability are extremely important parameters in the process of evaluating a compound for further development. In the past, these parameters were only taken into account at the later stages of the drug discovery process, partly because these parameters are time-consuming to establish and partly because small modifications to a molecule are known to have dramatic effects on these parameters. However, the availability of large databases on toxicity and metabolism has now increased the chance of sensibly relating chemical structures to these effects and developing alert systems that again can be used to prioritize the hits.
- In the past, the analysis of HTS data was primarily performed by medicinal chemists, looking at the active compounds and then deciding which hits their efforts should focus on. First, with the increase in the number of hits, this approach becomes increasingly ineffective and computational techniques are increasingly used to classify the hits and derive hypotheses. Second, one should keep in mind that is basically impossible for a human being to take into account the large number of inactive compounds. However, the development of a pharmacophore hypothesis, for example, requires the incorporation of information on inactive compounds. Again, computational techniques are important for taking all data from HTS into account. Several developments are ongoing in this area [18].

1.5 Conclusions

The development of random or blind screening, admittedly at a high level of technical sophistication, mirrors the early days of drug discovery and provides a solid basis for the rational design of novel therapeutics. Given access by the HTS laboratory to a large, diverse compound screening library, a stream of hits can generally be anticipated. These should flow at a predetermined rate to reduce the false-positive hit-rate and thereby reduce the burden on follow-up screening ("cherry picking"). Rarely, if ever, does HTS yield compounds of adequate potency and selectivity that can be used to validate the biological target [13,19].

The situation could be improved by screening a subset of the compound library, implicitly designed for a particular biological target. Generally, however, significant chemical effort, combinatorial or otherwise, is required to optimize the compound for target validation purposes. The chemical effort will be augmented if the target needs to be validated *in vivo*, since pharmacokinetics issues become critical. Several companies have concluded that HTS, and to some extent high throughput chemistry, has moved the bottleneck of drug discovery from screening to lead optimization. Viewed in this light, the outcomes of HTS are critical in determining the placing of restricted chemistry resources. Indeed, if no chemically tractable leads are found, then it might be reasonable to terminate the discovery program, even if the biological hypothesis remains highly attractive. This clearly is a practical viewpoint in the pharmaceutical industry. Sometimes *de novo* design in combination with virtual screening strategies can provide a means to overcome such limitations.

Many compounds fail as clinical candidates due to non-target related deficiencies in adsorption, metabolism, excretion and toxicity. To date, relatively little effort has been expended in addressing this bottleneck in drug discovery. Future technologies will undoubtedly address the problem and should have a predictive value for clinical efficacy and safety. Taken together, the "gene-to-screen" approach, frequently applied to HTS, seems naïve. The real value of HTS could be to enable a pharmaceutical company to make better decisions as to resourcing its drug discovery programs. Underpinning this effort is the bioinformatics/cheminformatics matrix for the HTS groups. If executed properly, the era of large-scale virtual screening will become a reality.

We anticipate that while the size of the compound libraries and the number of HTS will continue to increase, leading to a larger number of hits, the number of leads actually being followed up per project will roughly remain the same. The challenge is to select the most promising candidates for further exploration and computational techniques will play a very important role in this process. Many of the available techniques are highlighted in the following Chapters of this book.

Acknowledgements

Michael Schultz and Petra Schneider are thanked for helpful comments on the manuscript.

References

[1] MORPACE Pharma Group Ltd., *From data to drugs: strategies for benefiting from the new drug discovery technologies.* Scrip Reports, July **1999.**

[2] J. Drews, *Rethinking the role of high-throughput screening in drug research*, Decision Resources **1999.**

[3] K. R. Oldenburg, *Ann. Reports Med. Chem.* **1998**, *33*, 301–311.

[4] J. J. Burbaum, *Drug Discovery Today* **1998**, *3*, 313–322.

[5] N. Bosworth, P. Towers, *Nature* **1989**, *341*, 167–168.

[6] L. Mere, T. Bennett, P. Coassin, P. England, B. Hamman, T. Rink, S. Zimmerman, P. Negulescu, *Drug Discovery Today* **1999**, *4*, 363–369.

[7] I. Hemmila, S. Webb, *Drug Discovery Today* **1997**, *2*, 373–381.

[8] C. M. Suto, D. M. Ignar, *J. Biomolec. Screening* **1997**, *2*, 7–9.

[9] S. R. Kain, *Drug Discovery Today* **1999**, *4*, 304–312.

[10] M. Auer, K. J. Moore, F. J. Meyer-Almes, R. Guenther, A. J. Pope, K. A. Stoeckli, *Drug Discovery Today* **1998**, *3*, 457–465.

[11] P. Gund, N. H. Sigal (1999) *Pharmainformatics* **1999**, Suppl. S25–S29.

[12] J. H. Wikel, R. E. Higgs, *J. Biomolec. Screening* **1997**, *2*, 65–67.

[13] S. Fox, S. Farr-Jones, M. A. Yund, *J. Biomolec. Screening* **1999**, *4*, 183–186.

[14] R. W. Spencer, *Biotechnol. Bioeng.* **1998**, *61*, 61–67.

[15] D. M. Bayada, H. Hamersma, V. J. van Geerestein, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1–10.

[16] W. P. Walters, M. T. Stahl, M. A. Murcko, *Drug Discovery Today* **1998**, *3*, 160–178.

[17] D. E. Clark, S. D. Pickett, *Drug Discovery Today* **2000**, *5*, 49–58.

[18] M. Hann, R. Green, *Curr. Opin. Chem. Biol.* **1999**, *3*, 379–383.

[19] J. Major, *J. Biomolec. Screening* **1998**, *3*, 13–17.

[20] Goodman & Gilman's The Pharmacological Basis of Therapeutics, Ninth Edition **1996**. Adapted from J. Drews, S. Ryser (Eds.) *Human disease – from genetic causes to biochemical effects*, Ex Libris Roche, Basel **1997**.

Virtual Screening for Bioactive Molecules
Edited by Hans-Joachim Böhm and Gisbert Schneider

2 Library Filtering Systems and Prediction of Drug-Like Properties

W. Patrick Walters, Mark A. Murcko

2.1 Introduction

Recent advances in combinatorial chemistry and high throughput screening have dramatically increased the scale on which drug discovery efforts are carried out. Drug companies now routinely assay several hundred thousand compounds against each new drug target and the size of the typical screening library is soon expected to approach a million compounds. Likewise, the number of compounds, which can be synthesized in one year by a dedicated combinatorial chemist, is typically 10000 to 100000 [1–3]. In the early 1990s, many researchers believed this increase in the number of compounds synthesized and screened would lead to a concomitant increase in the number of clinical candidates produced. However, anecdotal evidence from several research laboratories suggests that raw speed and sheer numbers are not sufficient to make significant contributions to the drug discovery process [4,5].

The utility of first-generation combinatorial libraries has generally been considered to be quite low because these libraries tend to be populated with large, lipophilic, highly flexible molecules. Support for this thesis comes from Lipinski [6], who analyzed the compounds synthesized at Pfizer between 1984 and 1994. He showed that the percentage of compounds with a molecular weight greater than 500 doubled over the 10-year period. We also should remember that the number of high-quality lead molecules derived from HTS is typically quite low. Spencer [7] reported the number of quality hits is in the range of 1 per 100000 compounds screened for "easier" targets such as enzymes and receptors, but much worse for "harder" targets such as cytokines and growth factors.

As a consequence, many researchers have begun to pay closer attention to the nature of the compounds synthesized and screened. The term "drug-like" is often employed in presentations, although it is seldom clearly defined. Many researchers seem to mean some combination of overall size, charge, and lipophilicity, while others suggest that the presence or absence of certain functional groups may be associated with "drug-like" properties. It is clear that all these ideas are fairly qualitative and require careful scrutiny.

2.2 Simple Counting Methods to Predict Drug-Likeness

Many researchers over the years have attempted to show that drug-like molecules tend to have certain properties. For example, log*P*, molecular weight, and the number of hydrogen-bonding groups has been each correlated with oral bioavailability. In principle, then, one

should be able to very simply improve the "odds of success" by biasing selection towards compounds that have certain properties.

Recently, researchers at Pfizer have extended this idea with the establishment of the "rule of five", which provides a heuristic guide for determining if a compound will be orally bioavailable. The rules were derived from analysis of 2245 compounds from the World Drug Index [8]. Only those compounds with a USAN (United States Adopted Name) or INN (International Nonproprietary Name) and an entry in the "indications and usage field" of the database were included in the analysis. The assumption is that compounds meeting these criteria have entered human clinical trials and therefore must possess many of the desirable characteristics of drugs. It was found that in a high percentage of compounds, the following rules were true: hydrogen bond donors ≤ 5, hydrogen bond acceptors ≤ 10, molecular weight ≤ 500, and log$P \leq 5$. The majority of the violations came from antibiotics, antifungals, vitamins and cardiac glycosides. The authors suggest that, despite their violations of the "rule of five", these classes of compounds are orally bioavailable because they possess groups which act as substrates for transporters.

Ghose and co-workers [9] extended this work by characterizing 6304 compounds (taken from the Comprehensive Medicinal Chemistry Database [10]) based on computed physicochemical properties. They established qualifying ranges, which cover more than 80% of the compounds in the set. Ranges were established for *A*log*P* [11], molar refractivity, molecular weight, and number of atoms. These ranges, which agree well with those determined by Lipinski, are shown in Table 2.1.

Table 2.1. Quantifying ranges for predicted drug properties.

Predicted property	**Min.**	**Max.**	**Ave.**
Log*P*	–0.4	5.6	2.3
Molar refractivity	40	130	97
Molecular weight	160	480	360
Heavy atoms	20	70	48

As mentioned in the introduction, many first generation combinatorial libraries were of limited utility, presumably because they contained large numbers of compounds with undesirable physical properties. A great deal of recent effort has gone into the development of computational techniques for optimizing the calculated properties of a combinatorial library. One method for designing combinatorial libraries with reasonable physical properties was presented by Fecik [12]. The method involves calculating the molecular weight of the library scaffold and estimating the optimal weight of the sidechains to be added. The method is illustrated for a benzodiazepine library in Figure 2.1. The core structure has a molecular weight of 156. In order to produce compounds with a molecular weight less than 500, the sum of the side chain masses must be less than 344. With four points of diversity, this suggests that each sidechain should have fewer than six heavy atoms. Obviously other combinations that put a large sidechain at one position and smaller sidechains at other positions can also be used. Simple guidelines such as these can produce a dramatic reduction in the number of reagents that must be considered by the medicinal chemist.

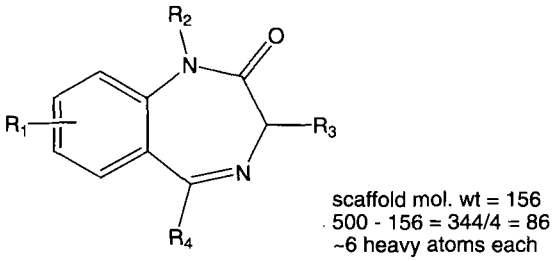


Figure 2.1. Determining the optimal sidechain masses for a benzodiazepine library.

Another approach is to treat the library design as a multivariate optimization problem and apply algorithms designed for global optimization. One example of such a method is the SELECT program developed by Gillet and coworkers [13]. SELECT uses a genetic algorithm to optimize both the diversity and calculated physical properties of a combinatorial library (for more information on genetic algorithms, see Chapters 8 and 9 in this volume). SELECT begins with a population of chromosomes, each of which represents a single combinatorial library. For a two-component library being constructed from ten reagents of type R1 and ten reagents of type R2, the chromosome would have 20 elements. The fitness of the chromosome is evaluated by enumerating the library (for our 10 x 10 example, 100 products would be enumerated for each chromosome) and calculating a set of physical properties. The distribution of properties is then compared with a distribution obtained from a drug database. A fitness score is assigned, based on the diversity of the library and the similarity of the library properties to those of the drug database. The most "fit" members of the population are then chosen to produce the next generation. In the crossover phase, two of the selected parents are paired and reagents are exchanged. During the mutation phase, a reagent is replaced by another random reagent of the same type. The new population created by crossover and mutation then replaces the current population and the cycle is repeated for a predetermined number of generations. The advantage to this approach is that the optimization method is independent of the method for assessing drug-like character. Any of the techniques described in this chapter could have been used as a fitness function.

2.3 Functional Group Filters

Medicinal chemists have known for quite some time that certain functional groups tend to be toxic or unstable under physiological conditions [14]. Unfortunately, very little of this information has been collected or formalized. This was less of a problem in the past when medicinal chemists only had to consider tens of molecules. However, today's chemist is presented with the daunting task of choosing between thousands or even millions of molecules. These challenges have led several groups to create automated methods of identifying problematic molecules.

One attempt to formalize the collective knowledge of medicinal chemists, enzymologists, pharamacologists and molecular modelers is the REOS (Rapid Elimination of Swill) pro-

gram developed at Vertex [15]. The program's primary function is to analyze databases of screening compounds or proposed chemical libraries and "filter out" molecules which may be problematic (swill). REOS is a hybrid method, which combines a set of functional group filters with some simple counting schemes similar to those in the rule of five. The functional group filters implemented in REOS identify reactive, toxic, and otherwise undesirable moieties. Initial filtering is based on a set of seven property filters. The default values for these filters are shown in Table 2.2. Hydrogen bond donors, acceptors and charged groups are determined using a set of rules similar to those used in the PATTY program developed at Merck [16]. Log*P* can be calculated based on a variety of schemes [11,17,18]. A web-based interface makes it trivial to modify parameters to suit the needs of a particular drug discovery project.

Table 2.2. The default property filters employed by REOS.

REOS property filter	Min.	Max.
Number of hydrogen bond donors	0	5
Number of hydrogen bond acceptors	0	10
Formal charge	–2	+2
Number of rotatable bonds	0	8
Molecular weight	200	500
Number of heavy atoms	20	50
Log*P*	–2	5

One may initially question the validity of setting a minimum value for the number of heavy atoms and molecular weight. In many cases, the final product of a lead optimization effort may be considerably larger and more hydrophobic than the initial lead. Thus, it would be better to start with a small lead. However, while there are cases where very low molecular weight compounds have been found to be highly potent [19], these tend to be rare. In order to improve the odds of finding an inhibitor, REOS typically employs the molecular weight and heavy atom minima specified in Table 2.3.

Table 2.3. Some of the functional group filters employed by REOS and their SMARTS representations.

REOS filter	SMARTS representation
Nitro groups	[N;+0,+1;$(N(=O)~[O;H0;-0,-1])]
Long aliphatic chains	[CH2][CH2][CH2][CH2][CH2][CH2][CH2][CH2]
Primary alkyl halides	[Cl,Br,I][CH2]
Acid or sulfonyl halides	[S,O]=C-[Cl,Br,I]
Aldehydes	[HC]=O
Anhydrides	O=[C,S,P]-O-[C,S,P]=O
Diimides,isocyanates, isothiocyanates	N=C=[N,O,S]
Acetals	O-&!@[CH2]-&!@O 0
Alpha halo ketones	[F,Cl,Br,I]-[CH]-C=O
Triflates	OS(=O)(=O)C(F)(F)F
Halogen-N,O	[F,Cl,Br,I]-[N,O]
N and S mustards	[N,S][CH2][CH2][Cl,Br,I]
Peroxides	OO
Epoxides and aziridines	C1[O,N]C1
1,2 Dicarbonyls	C(=O)C(=O)

In addition to the property filters, REOS also allows the user to remove compounds using a set of more than 200 functional group filters. Rather than providing a simple "accept/reject" facility, REOS allows the user to specify a maximum allowed quantity for each functional group. For instance, if one is selecting screening compounds, it is usually desirable to eliminate aldehydes due to their reactivity with biological nucleophiles. In this case, the maximum number of allowed aldehydes would be set to 0. However, if one is using aldehydes in the synthesis of a combinatorial library, it may be desirable to select reagents containing only one aldehyde. This can be easily accomplished by setting the maximum allowed value to 1. Examples of the functional group filters employed by REOS are listed in Table 2.2. Further examples have been published elsewhere [14,20,21].

In REOS, functional group filters are specified using the SMARTS [22] pattern-matching language developed at Daylight Chemical Information Systems. SMARTS is an extended version of the SMILES [23,24] notation developed specifically for substructure searching. For instance, an acid halide can be specified by the pattern C(=O)[F,Cl,Br,I]. In this example, the parentheses are used to indicate a branch while the square brackets are used to specify a set of options. With a simple modification, this pattern could be used to specify sulfonyl halides as well [C,S](=O)[F,Cl,Br,I]. More complex rules can also be defined for abstract atom types such as hydrogen bond donor or lipophilic atom. The major advantage of SMARTS patterns is that they are simple ASCII text, which can be easily modified and used by a variety of applications. The primary disadvantage is that writing such patterns takes a bit of practice and the notation may not be immediately accessible to medicinal chemists. This situation has been somewhat remedied by the recent appearance of a number of tools which are capable of converting a graphical representation of a structure query into a SMARTS pattern [25,26]. Variations of the SMARTS language are also available in Unity suite of programs from Tripos [27] and the Molecular Operating Environment (MOE) from The Chemical Computing Group [28].

Figure 2.2 shows the results of a REOS analysis of seven databases. The first database consists of 5120 compounds from the CMC [10] database. The other six databases consist of commercially available screening compounds. The screening databases contain between 10000 and 100000 compounds. Figure 2.2a shows the percentage of each of the seven databases which passed the rule of five filters. Parameters for this analysis were set to the values defined by Lipinski and as described earlier. Some readers may find it surprising that approximately 25% of the compounds in the CMC database do not meet the rule of five criteria. Failures can primarily be attributed to compounds with molecular weight > 500 (12%) and log$P$ > 5 (13%). A large percentage of the high molecular weight compounds were either antibacterials or antineoplastics. This reflects the history of drug discovery efforts in these areas, which have traditionally been based on screening large collections of natural products. Many of the compounds with high log$P$ values were from classes considered to be CNS-active (antiparkinsonian, antipsychotic, antidepressant). This is consistent with the fact the CNS compounds tend to be more lipophilic than other biologically active molecules [29]. Examples of drugs which violate the rule of five filters are shown in Figure 2.3.

Figure 2.2b shows the number of compounds from each database that passed the functional group filters. Although more of the CMC compounds passed the filters than most of the screening databases it is surprising that approximately 40% of the CMC compounds were eliminated. The largest number of rejections (185) was due to the presence of nitro groups.

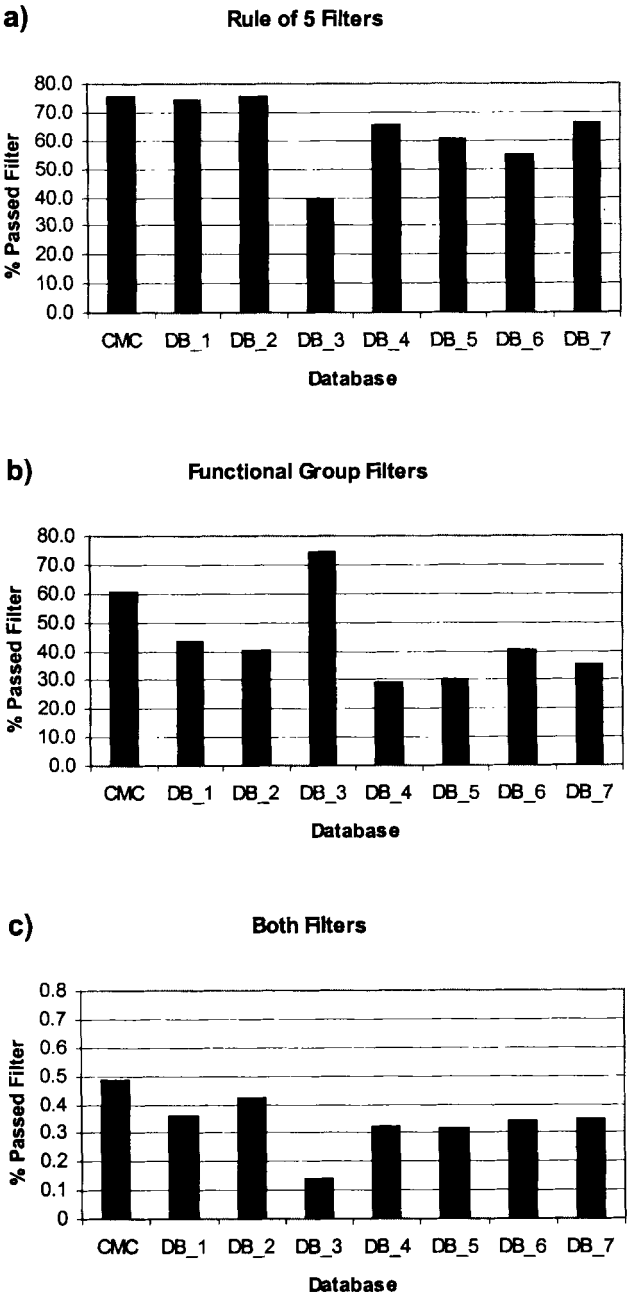
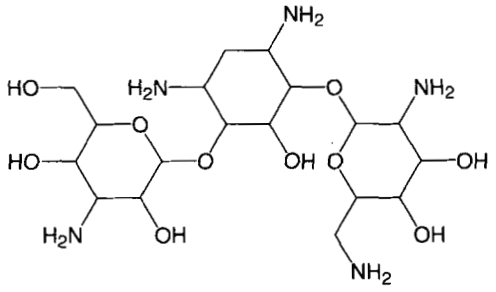
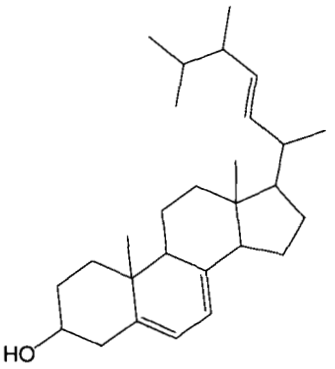


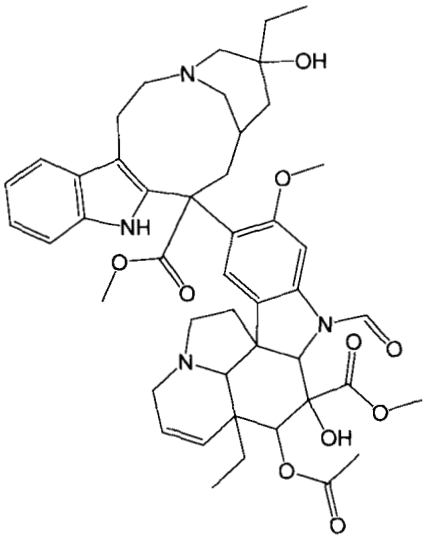
Figure 2.2. A REOS analysis of seven databases. a The percentage of each database which passed the rule of five filters. b The percentage of each database which passed the functional group filters. c The percentage of each database which passed both sets of filters.



HB Donors = 16



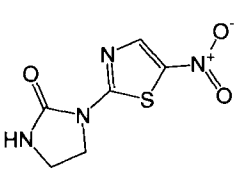
clogP = 9



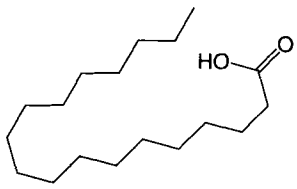
mol. wt. = 825

Figure 2.3. Drugs rejected by the rule of five filters.

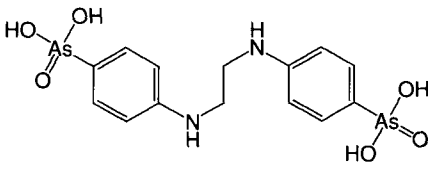
Nitro groups tend to activate aromatic rings [30] and may increase a molecule's tendency to generate false positives under assay conditions. In addition, nitro compounds tend to be colored and may interfere with assays which employ a spectrophotometric readout. Another 125 compounds were rejected because they contained atom types other than C, O, N, S, P, F, Cl, Br, I, Na, K, Mg, Ca, and Li. The majority of these rejections were due to antineoplastics containing Pt and As, antacids containing Al or Si, or vitamins containing Fe and Co. Further examples of drugs rejected on the basis of functional group filters are shown in Figure 2.4.



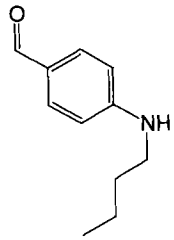
nitro group



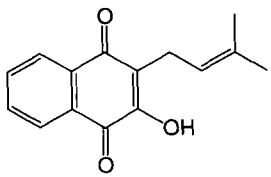
long alkyl chain



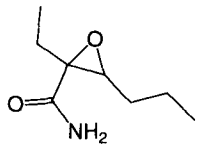
unusual atom type



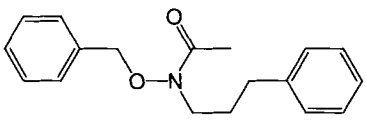
aldehyde



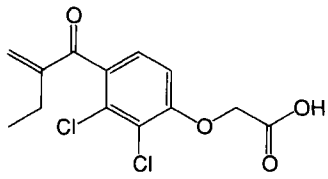
quinone



epoxide



N-O single bond



Michael acceptor

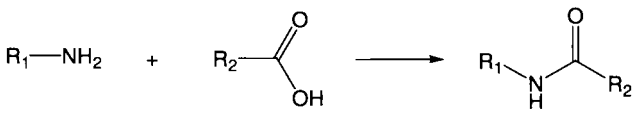
Figure 2.4. Drugs rejected by functional group filters.

Figure 2.2c shows the intersection of the two lists described above, i.e. the compounds which passed both the rule of five filters and the functional group filters. It is quite striking that in every case, more than half the molecules are eliminated. Tools such as REOS can provide a significant reduction in the number of compounds which must be considered. Careful filtering can also decrease the amount of time required for follow-up screens by reducing the

number of false positives. While functional group filtering is highly effective, one must avoid the temptation to become overzealous. It is important to periodically examine the results of a filtering analysis to insure that "interesting" molecules are not being inadvertently eliminated.

The results of functional group filtering are even more dramatic when considering the synthesis of a combinatorial library. The REOS analysis of combinatorial libraries is typically carried out in three steps:

1. Filter the reagents. In this step, reactive and toxic reagents are removed. Reagents that will clearly create products that violate the molecular weight limits are also removed.
2. Check the reagents for compatibility with the chemistry. For example, when synthesizing amides, one can simplify the chemistry by removing acids containing basic amines and amines containing acidic functionality.
3. Filter the products. This allows a final consideration of properties such as log*P*, which cannot be calculated from reagents in a straightforward fashion. In addition, this step is necessary for functional group filters which specify a maximum count. For instance, let us suppose we have a rule that only permits molecules with fewer than three bromine atoms. Two reagents having two bromine atoms each would pass the filter, but their product having four bromine atoms would fail.



Full Library	6,681	4,288	28.6 M
REOS Library	2,578	1,796	4.6 M
Chem. Comp.	2,054	1,666	3.4 M
Product Filtering			1.0 M

Figure 2.5. A REOS analysis of combinatorial library.

Figure 2.5 shows the number of reagents that could be used in the synthesis of an aryl amide library from benzoic acids and anilines. There are more than 28 million possible combinations for this two-component library. The three filtering steps described above reduce the size of the library almost 30-fold.

2.4 "Chemistry Space" Methods

The idea that similar compounds will have similar properties is one of the fundamental underpinnings of QSAR and molecular similarity analysis. One recent extension of this idea is the concept of a "chemistry space". A chemistry space is typically defined by calculating a number of descriptors for each molecule and using the descriptor values as points in a multi-

dimensional space. As an example, let us assume that we have calculated molecular weight, log*P* and the number of hydrogen bond donors for a set of molecules. These three descriptor values can then be used to define a point in a three-dimensional space that represents each molecule. In practice, large numbers of descriptors are calculated and statistical techniques such as principal components analysis or factor analysis are used to reduce the dimensionality of the descriptor space.

Several research groups have attempted to define the "chemistry space" which is occupied by drug-like molecules. The basic idea is that drugs will tend to possess distinct values for certain properties and, as a result, will be shown to be distinct from non-drugs when analyzed in a multi-dimensional space. Cummins *et al.* [31] compared five databases – CMC [10], MDDR [32], ACD [33], SPECS [34], and their in-house Wellcome registry. They calculated 28 topological indices as well as an estimate of the free energy of solvation for 300000 compounds. Factor analysis was used to reduce the 28-dimensional descriptor space to four dimensions. The descriptor space was then partitioned and the occupancy of the resulting sub-hypercubes was examined. The percentages of the total volume occupied by the databases were: CMC 27%, Wellcome registry 72%, MDDR 69%, SPECS 46%, and ACD 72%. The authors also found a 92% overlap between CMC and ACD. Thus, while the method may be used to identify interesting regions of space, it may not by itself be an effective discriminator between "drugs" and "non-drugs".

Gillet *et al.* [35] used profiles of calculated properties (numbers of hydrogen bond donors and acceptors, molecular weight, rotatable bonds, aromatic rings, and a shape descriptor) to differentiate between a set of drugs represented by 14861 compounds from the World Drug Index and set of non-drugs represented by a set of 16807 compounds from the SPRESI database. A genetic algorithm was used to derive a set of optimal weights for the properties. The best weighting schemes were able to provide a five- to six-fold enhancement over random selection. The authors were also able to achieve similar results in using property profiles to identify drugs belonging to a specific therapeutic class from a larger drug database.

2.5 Examination of Building Blocks in Known Drugs

A very different approach is to analyze the building blocks commonly found in drugs to see whether non-random patterns can be unearthed. This work does not directly confront the problem of distinguishing drugs from non-drugs, but it helps to define what drugs are and thereby helps chemists to think about preferred moieties for library design.

One of the first studies of this type was carried out by Bemis and Murcko [36], who developed a method of organizing drugs by decomposing molecules into frameworks, sidechains and linkers. Their method, which is outlined in Figure 2.6, begins by successively removing monovalent atoms until only rings and acyclic linkers between rings remain. The "scaffolds" defined by these rings and linkers can then be further abstracted by considering all atom and bond types as equivalent. An examination of 5120 compounds from the CMC [10] yielded 1170 scaffolds. This suggests that drugs are rather diverse. However, when atoms and bonds were considered equivalent, only 32 frameworks described the shapes of half the drugs in the set. These frameworks are shown in Figure 2.7. Even when atom types and hybridization are

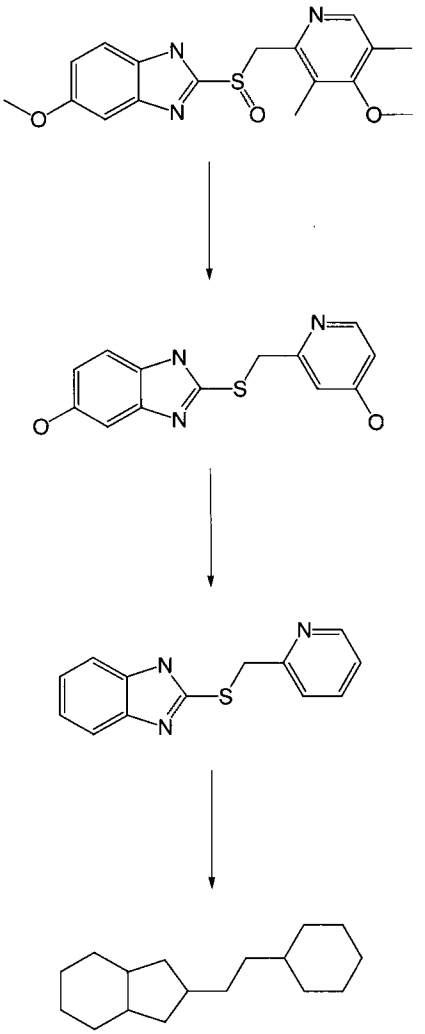


Figure 2.6. The process for reducing a drug molecule to a framework.

considered, 25% of all drugs are found to utilize only 42 frameworks. These surprising results suggest that a small number of common "shape themes" can be re-used in widely divergent drug design situations.

A second study carried out by the same authors [37] examined the sidechains found in the same set of drug molecules. Using an atom-pair shape descriptor, they found that there were 1246 sidechains among 5090 compounds analyzed. The mean number of sidechains was found to be four and the average number of heavy atoms per sidechain was two. As with the frameworks analysis, a small number of themes predominated. A set of 20 sidechains was sufficient to cover more than 11000 of the 15000 occurrences in the database.

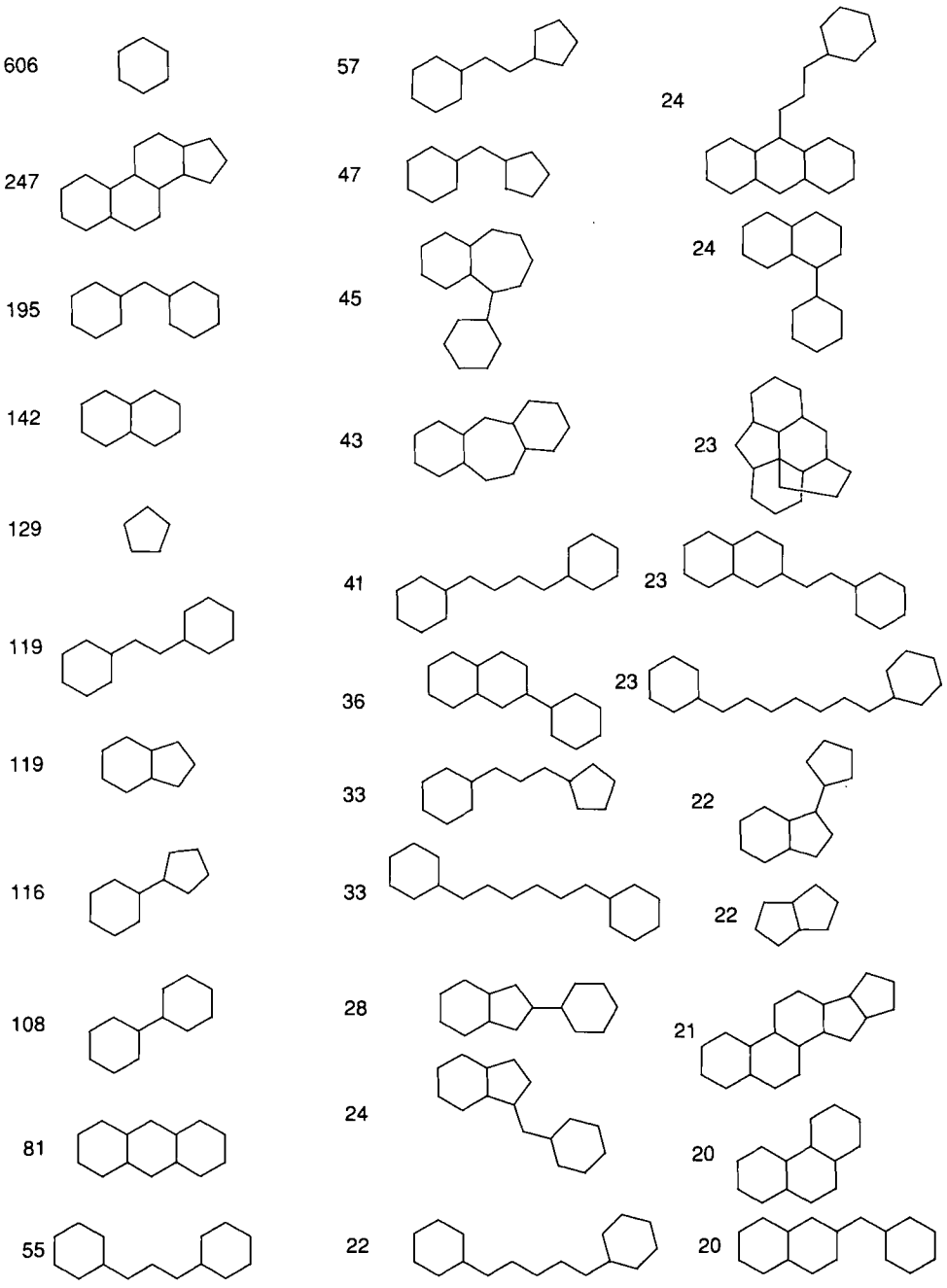


Figure 2.7. The top 32 frameworks found in drugs. Numbers indicate the frequency of occurrence in the drug database.

A number of interesting applications of this analysis have appeared in the recent literature. Fejzo and coworkers [38] have used the frameworks and sidechains to create a library of "drug fragments" which can be screened using NMR techniques to assess binding. This "SHAPES" library can be used to identify weakly binding hits (up to 10 mM) which would be missed by standard enzymatic assays. These weak binders can then be used as the basis for pharmacophore hypotheses, combinatorial library design, database searches, or a number of other techniques.

Wang [39] has utilized the frameworks and sidechains approach described above to study toxic molecules and identify structural features associated with specific toxicity (abnormality, tumor, mutation, etc.). The analysis was carried out on a set of 56862 structures from the Registry of Toxic Effects of Chemical Species (RTECS). Each structure was first reduced to a set of frameworks and sidechains. The authors then compared the frequency of occurrence of a particular framework in the entire database to its frequency in a specific toxicity subset. This allows the discrimination between "composition frameworks" which occur in a variety of molecules and "toxicity-conferring frameworks" which occur primarily in molecules with a specific toxicity. The toxicity-conferring frameworks can then be used to screen a database and identify potentially toxic molecules. Toxicity is a major reason for the failure of drugs in clinical trials [40] and this will undoubtedly continue to be an area of active research.

Standard Tanimoto Coefficient = (A & B)/(A + B − A & B)
Substructure Similary Tanimoto Coefficient = (A & B)/A

where
A = the number of bits turned on in fingerprint A
B = the number of bits turned on in fingerprint B
A & B = the number of bits in common to fingerprints A and B

Figure 2.8. The Tanimoto coefficient and alternate form used to identify drug-like reagents.

The Glaxo group developed two different approaches to identifying drug-like reagents that can be used in the synthesis of combinatorial libraries. The first method [41] involved clustering sets of monomers and calculating the similarity of the cluster centroid to 30000 compounds from the Derwent Standard Drug File. Similarities were calculated using Daylight fingerprints in two ways. Since the monomers were typically much smaller than the drug molecules, an alternate formulation of the Tanimoto coefficient was used to calculate a "substructure similarity". The alternate form is given in Figure 2.8. This analysis led to the identification of a series of monomers which occur frequently in drugs and can then be used as the basis of a library design. The major drawback to this approach is that it cannot be used to suggest reagents which are not already in the initial pool. This deficiency was addressed though the development of a second program know as RECAP [42] (Retrosynthetic Combinatorial Analysis Procedure). RECAP begins with a collection of drug molecules and then fragments these molecules using any of the 11 retrosynthetic "reactions" shown in Figure 2.9. The resulting fragments are then clustered and transformed into sets of monomers that can form the basis of library design effort. Because the monomers come from known drugs, there

is a high likelihood that the molecules constructed from them will contain biologically interesting motifs.

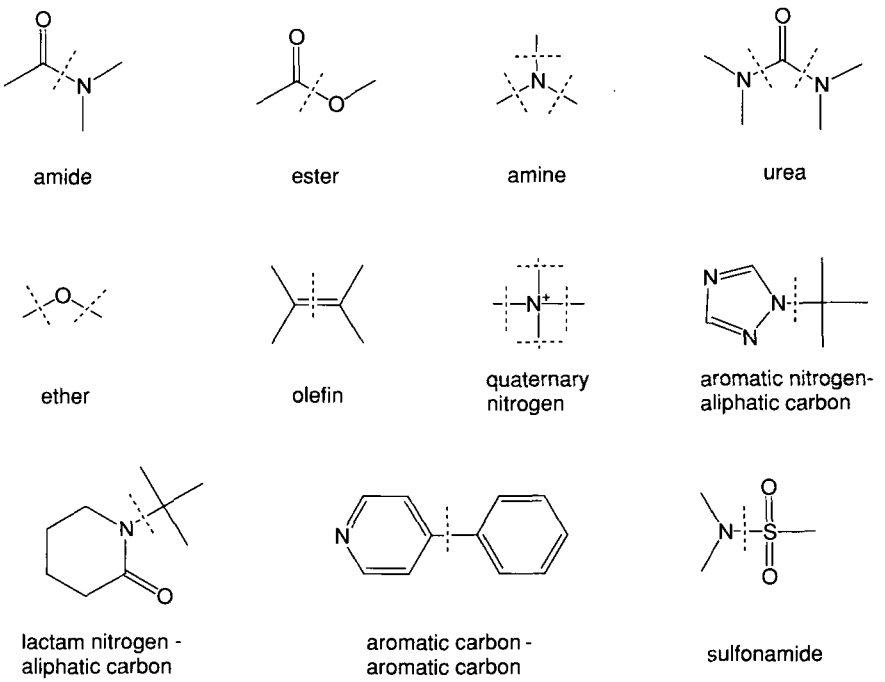


Figure 2.9. The retrosynthetic reactions used by RECAP.

2.6 Other Methods

Both Ajay [43] and Wagener [44] have separately reported methods that use machine learning to distinguish drugs from non-drugs. Machine learning programs operate by examining a set of training examples, each of which is assigned to belong to a particular class. They then derive a rule or set of rules that assign new examples to these classes. The primary advantage of machine learning methods is their potential to express the rules in a form that can be easily understood by humans. In the work published by Ajay, the training and test sets consisted of 3500 compounds each from the CMC (drugs) and the ACD (non-drugs). The machine learning program C4.5 [45] was used with a set of seven one-dimensional descriptors to produce a decision tree. A portion of this decision tree is shown in Figure 2.10. The decision tree was able to correctly classify approximately 80% of the CMC compounds and approximately 70% of the ACD compounds. Rules can be constructed by walking up the tree from bottom to top. For instance, the rule

IF mw <= 388.7 AND
kap <= 10.924 AND
don > 1 AND
acc > 3 AND
acc <= 8 AND
don <= 3 THEN
class = drugs

can be obtained from the decision tree in Figure 2.10. The primary disadvantage of this method is its tendency to "overtrain" and produce rules based on chance correlations in the data.

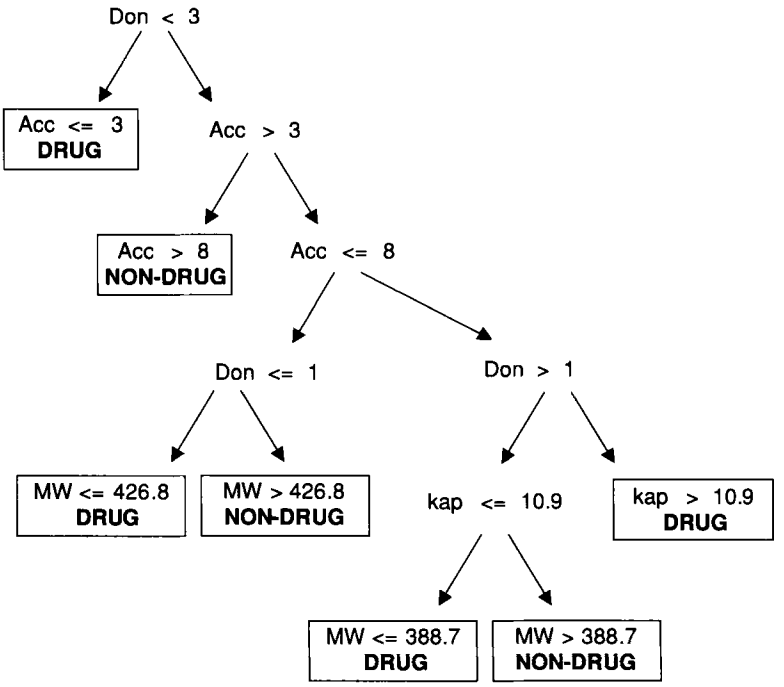


Figure 2.10. A portion of a decision tree for distinguishing drugs from non-drugs.

Neural networks are another automated classification method that has received a great deal of recent attention. Two recent papers [43,46] have discussed the use of neural networks in distinguishing drugs from non-drugs. This topic is discussed elsewhere in this volume (see Chapter 6) and will not be discussed in detail here. Suffice it to say that both methods were highly predictive and show great promise.

Another approach to identifying drug-like molecules was recently published by Wang [47], who introduced the concept of multilevel chemical compatibility (MLCC). In MLCC, a mol-

ecule is characterized according to a set of local atom environments. These atom environments are defined using single atoms as well as dicentered, tricentered, and tetracentered groups. Each unique atom environment is assigned an occupancy based on the number of times that atom environment occurs in the molecule. These occupancies are then compared with those calculated for a set of 11704 drugs taken from the CMC and MDDR databases. The method was tested using four sets of compounds: top selling drugs (100), compounds currently undergoing biological testing (68017), anticancer agents (461), and a set of reactive molecules which are considered problematic for biological testing (57). The results of the analysis using tetracentered atom environments are shown in Table 2.4. While the program was able to completely eliminate the problematic compounds, it was unable to recognize 24% of the drugs. The authors point out that most of unrecognized drugs contain unusual building blocks and highlight deficiencies in their drug database.

Table 2.4. The results of an MLCC analysis of four sets of compounds.

Category	% Considered drug-like
Drugs	76
Bio-testing	27
Anticancer	19
Problematic	0

2.7 Conclusions and Future Directions

As we have shown, a wide variety of methods have already been applied to the problem of identifying molecules with desirable or "drug-like" properties. These methods appear to be meeting with some success. A key issue is whether general ("global") rules can be formulated, or whether rules will always need to be "local" and situation-specific.

Another trend we may witness in coming years might be attempts to predict the various properties which contribute to a drug's success rather than the more complex problem of "drug-likeness" itself. These might include oral absorption, blood-brain barrier penetration, toxicity, metabolism, aqueous solubility, log*P*, p*K*a, half-life, and plasma protein binding. Some of these properties are themselves rather complex and are likely to be extremely difficult to model. In our view, however, it should be possible for most properties to be predicted with better-than-random accuracy.

Future work is likely to include additional approaches and more robust attempts at validation of these methods. Also, one hopes that the judicious use of these predictions may lead to increased efficiency in the selection of combinatorial and HTS screening libraries. However, we are probably still several years away from a definitive experiment proving this point. Further off, in all likelihood, will be the ability to predict "downstream" issues pertaining to formulation, manufacturing, shelf-life, chemical stability, and so forth. These too are critical for the success of a drug.

References

[1] R. E. Dolle, *Mol. Divers.* **1997**, *2*, 223–226.
[2] E. M. Gordon, *Curr. Opin. Biotechnol.* **1995**, *6*, 624–631.
[3] D. Brown, *Mol. Diversity* **1997**, *2*, 217–222.
[4] M. A. Gallop, The Second Lake Tahoe Symposium on Molecular Diversity: Lake Tahoe, CA, Jan 27, **1998**.
[5] C. B. Cooper, National Managed Healthcare Conference: Boston, MA, May 20, **1998**.
[6] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Delivery Rev.* **1997**, 23, 3–25.
[7] R. W. Spencer, *Biotechnol. Bioeng.* **1998**, *61*, 61–67.
[8] *World Drug Index (WDI)*; Derwent Information: London, UK, http://www.derwent.com.
[9] A. K. Ghose, V. N. Viswanadhan, J. J. Wendelowski, *J. Comb. Chem.* **1999**, *1*, 55–67.
[10] *Comprehensive Medicinal Chemistry (CMC) Database*; Molecular Design Limited, San Leandro, CA, http://www.mdli.com.
[11] A. K. Ghose, V. N. Viswanadhan, J. J. Wendoloski, *J. Phys. Chem. A.* **1998**, *102*, 3762–3772.
[12] R. A. Fecik, K. E. Frank, E. J. Gentry, S. R. Menon, L. A. Mitscher, H. Telikepalli, *Med. Res. Rev.* **1998**, *18*, 149–185.
[13] V. J. Gillet, P. Willett, J. Bradshaw, D. V. S. Green, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169–177.
[14] G. M. Rishton, *Drug Discovery Today* **1997**, *2*, 382–385.
[15] W. P. Walters, M. T. Stahl, M. A. Murcko, *Drug Discov. Today* **1998**, *3*, 160–178.
[16] B. L. Bush, R. P. Sheridan, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756–782.
[17] *CLOGP3*; Daylight Chemical Information Systems: Santa Fe, NM, http://www.daylight.com.
[18] I. Moriguchi, S. Hirono, Q. Liu, Y. Nakagome, Y. Matsushita, Chem. Pharm. Bull. 1992, 42, 976-978.
[19] I. D. Kuntz, K. Chen, K. A. Sharp, P. A. Kollman, Proc. Natl Acad. Sci. USA 1999, 96, 9997-10002.
[20] M. Hann, B. Hudson, X. Lewell, R. Lifely, L. Miller, N. Ramsden, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897–902.
[21] A. R. Leach, J. Bradshaw, D. V. S. Green, M. M. Hann, J. J. Delany, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1161–1172.
[22] C. A. James, D. Weininger, J. Delany, *Daylight Theory Manual* **1997**, http://www.daylight.com/day-html/doc/theory/theory.toc.html.
[23] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
[24] D. Weininger, A. Weininger, J. L. Weininger, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
[25] *JavaGrins*; Daylight Chemical Information Systems: Santa Fe, NM, http://www.daylight.com.
[26] *MolSMART*; Barnard Chemical Information: Sheffield, UK, http://www.bci1.demon.co.uk.
[27] *Unity*; Tripos: St. Louis, MO, http://www.tripos.com.
[28] *Molecular Operating Environment*; Chemical Computing Group: Montreal, Quebec, Canada, http://www.chemcomp.com.
[29] Ajay, G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1999**, *42*, 4942–4951.
[30] J. March, *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure*, John Wiley & Sons **1992** p. 599.
[31] D. J. Cummins, C. W. Andrews, J. A. Bentley, M. Cory, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750–763.
[32] *MACCS-II Drug Data Report (MDDR)*; 98.2 ed.; Molecular Design Limited, San Leandro, CA, http://www.mdli.com.
[33] *Available Chemicals Directory (ACD)*; 98.2 ed.; Molecular Design Limited: San Leandro, CA, http://www.mdli.com.
[34] *SPECS/BioSPECS Database*; Specs and BioSPECS: Rijswijk, The Netherlands, http://www.biospecs.com.
[35] V. J. Gillet, P. Willett, J. Bradshaw, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165–179.
[36] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887–2893.
[37] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1999**, *42*, 5095–5099.
[38] J. Fejzo, C. A. Lepre, J. W. Peng, G. W. Bemis, Ajay, M. A. Murcko, J. M. Moore, *Chem. Biol.* **1999**, *6*, 755–769.
[39] J. Wang, L. Lai, Y. Tang, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1173–1189.
[40] T. Studt, *Drug Discov. Devel.* **1999**, *Jan*, 40–41.
[41] X. Q. Lewell, R. Smith, *J. Mol. Graph Model.* **1997**, *15*, 43–48.
[42] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
[43] Ajay, W. P. Walters, M. A. Murcko, *J. Med. Chem.* **1998**, *41*, 3314–3324.

[44] M. Wagener, V. J. van Geerestein, *Analysing Large Datasets with Decsion Trees: Discriminating Between Potential Drugs and Non-drugs:* 217th American Chemical Society National Meeting, Anaheim, CA, **1999**.
[45] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, **1993**.
[46] J. Sadowski, H. Kubinyi, *J. Med. Chem.* **1998**, *41*, 3325–3329.
[47] J. Wang, K. Ramanarayan, *J. Comb. Chem.* **1999**, *1*, 524–533.

Virtual Screening for Bioactive Molecules
Edited by Hans-Joachim Böhm and Gisbert Schneider

3 Prediction of Physicochemical Properties

Jeff J. Morris, Pierre P. Bruneau

3.1 Introduction

Physical properties are becoming recognized as important factors, which govern the ability of drugs to show an effect *in vivo*. There is little point in making very potent compounds which bind tightly to the receptor, if they cannot show their effect due to poor solubility, or poor bioavailability. Alongside this has come the realization that if the physical properties are not optimal, it is often difficult and time consuming to "fix" them later. Therefore, it is far better to start with something that looks reasonable in the first place. This is particularly true now that automated synthesis can easily make hundreds or thousands of these rather unsatisfactory structures at one go. Consequently it would be really helpful if we could guide the selection of candidate compounds, both for the output of high-throughput screening (HTS), and for the input of combinatorial chemistry by the use of calculated physical properties.

The purpose of this Chapter then is to look at the methods, which are available to estimate some of the common physical properties held to be important for *in vivo* activity. The properties, we will consider are lipophilicity as represented by log$P$ in *n*-octanol, aqueous pKa, water solubility, and binding to serum albumin. In particular we focus our attention on how well these methods can be expected to predict, and where they might fall down. As they become more widely available, they are increasingly used by non-specialists who may not have a clear idea of their limitations. It will be for the reader to judge whether some or all of them are suitable for their particular purpose.

3.2 Prediction of Lipophilicity

The partition coefficient (P) of a compound is defined simply by Eq. (3.1)

$$P = C_{\text{solv}}/C_{\text{aq}} \tag{3.1}$$

where C_{aq} is the concentration in an aqueous phase, and C_{solv} is the concentration in a non-aqueous phase. This definition is quite general and applies to neutral compounds. It should be carefully distinguished from the distribution coefficient, D, which applies to solutes with an ionizable group (see Section 3.2.5). Lipophilicity is one of the most fundamental properties of organic molecules. It is implicated in distribution processes across the intestinal mucosa [1], across the blood-brain barrier [2], and is related to the volume of distribution in man, and thereby tissue levels [3]. The measured lipophilicity depends on the non-aqueous solvent chosen, and *n*-octanol is the most commonly used. However there is no reason why it

should perform better than any other solvent in relating to processes *in vivo* [4]. The original decision to use *n*-octanol was based primarily on experimental considerations (see Section 3.1.6). Over the years many hundreds of relationships have been discovered between measured lipophilicity in *n*-octanol and biological activity [5]. This fueled interest in the field, and there are now more than 30000 measurements available in the Pomona database [6] for the octanol–water system, many of them referring to neutral compounds. These measurements are usually of high quality, and the most reliable are given a "star". For these reasons there is a very large data set to enable the calibration of predictive methods.

Over the years there have been a large number of methods developed which attempt to calculate partition coefficients for this particular solvent system, and these have been reviewed and described [7–11]. The present Chapter will concentrate only on any improvements or changes which have been published since a recent review [11] in 1997, or on any new methodology which has appeared subsequently. The methodologies used to calculate log*P* can be grouped into two broad families:

1. Fragment-Based Methods, and
2. Methods Based on Molecular Properties.

3.2.1 Fragment-Based Methods

In these methods each structure is broken down into a set of sub-structural fragments. The contribution of each fragment in each structure to the log*P* of the entire set is then assessed. The methods differ in the way the fragments are produced, the number and size of the fragments, and the way in which they are related to the log*P*. All of these fragment-based approaches are dependent on several critical factors for success. These are fairly self-evident, but it is surprising how many of the newer methods do not explicitly state that they are fulfilled:

1. Every atom in every molecule should be accounted for by the fragment definition. If this criterion is not met, then essentially there are missing values.
2. Each fragment definition should hit only one atom in the structure. If the same atom is hit multiple times by different definitions, then the results are difficult to interpret.
3. Some of the fragment definitions in the training set will have few representatives, so in the database there may be only one or two examples for a particular fragment type. In these instances there will be more uncertainty in the fragment contribution to the log*P*.

Another advantage often quoted for fragment-based methods is that they are fast to calculate. Given the large number of measured values available, it is sometimes difficult to understand why the training set used to develop some of these methods is so small.

CLOGP [12] differs from most fragment methods in that it relies on accurate measurement of a set of fundamental fragments. The log*P* is then built up by adding the contribution of these fragments together. However an additional complication is that the method attempts to take account of different molecular environments using a series of correction factors, which can get very complex. However, the program reports exactly how it did the calculation, and whether it had all the fragments in the structure, or had to guess at some. In addition its

rules are derived from a very large training set. Recently, Leo [13] reported that CLOGP had been modified so that fewer missing fragments were reported. He argued that fragment values for most functional groups have been measured but not necessarily in the appropriate molecular environment. So another set of rules has been developed to estimate functional groups in their new environment. As yet there are no reports of the program but Leo states that the precision will be lower.

Rekkers method [10] along with those of Ghose and Crippen [14], and Moreau and Broto [15] have all been previously reviewed [7,11], and are incorporated into PROLOGP [16]. This gives the user the option to calculate using either a fragment-based or atom-based method.

ACDLABS [17] have introduced a suite of software to calculate physical properties, including ACDlog*P*, which calculates log*P* octanol. The methodology has been described in a recent software review [18], and is based on a set of fragments and atomic contributions, together with a set of correction factors which apply to certain molecular features. One key advantage is the ability to estimate missing fragments from a rule base. The prediction accuracy (reported on the ACD website) has been estimated as 0.21 log units based on the standard deviation of a validation set of 3600 compounds. There is no doubt that ACDlog*P* goes further than any other program in terms of integration with other software, the user interface, and its handling of missing fragments. Future rigorous assessment will lead to a more precise estimation of its prediction performance (see Section 3.2.3 for a discussion).

ALOGP [19] is basically a refined method of the original Ghose and Crippen work [14] and it works by breaking the structure down into atomic fragments. Multiple regression is then used to obtain the average contribution of each atom across the set. The method was criticized on two counts. Firstly, because it led to atomic contributions which did not seem to make chemical sense; for example in diethyl ether the oxygen appears to be more hydrophobic than the carbon atoms flanking it [7]. Secondly, an atom-based method cannot take account of the longer-range interactions between functional groups which are known to exist and contribute to log*P* [7]. Recently, ALOGP has been revised to take account of some of the criticisms. In the new method [20], there are 68 atomic definitions, developed using SMARTS [21], and the authors explicitly state that their atomic definitions hit each atom in the structure once only. The chemical interpretation is improved by constraining several carbon atom types to have a positive contribution to log*P* in the fitting process. Finally the training set was expanded so that the coefficients for each atom type were determined by fitting to the 9000 compounds in the Pomona Database. The standard deviation is given as 0.67 log units. We believe that these improvements to the method make it the best of the atom-based approaches, and the results are probably as good as is possible without including interactions between functional groups.

XLOGP is another atom-based method [22]. In this case there are 80 atom descriptors which were originally based on the Tripos atom types. The coefficients for these atom types were determined by regression using the log*P* values for 1831 compounds. Overall, these coefficients look chemically sensible. In addition, correction factors were applied for hydrocarbons, amino acids, intramolecular hydrogen bonds, halogen–halogen 1–3 interactions. The standard deviation on the fitting for these 1831 compounds was 0.37 log units. For a test set of 19 compounds the standard deviation was 0.52 log units.

It is interesting that in both ALOGP and this XLOGP certain carbon atoms were found to be too hydrophilic. In both cases corrections were applied, albeit in rather different ways.

3.2.2 Methods Based on Molecular Properties

In these methods molecular properties are calculated for each structure and then related to the log*P*. Again these methods were reviewed recently [11], so only new approaches or improvements to existing methods published since 1997 are covered here.

AUTOLOGP [23,24] uses four types of descriptor for each structure:

- A lipophilicity descriptor derived from Rekker's fragmentation scheme.
- A molar refractivity descriptor derived from the Hansch–Leo fragments [5] or from the Lorentz–Lorentz equation (Eq. 3.2)

$$MR = (n^2-1)/(n^2+1) \times (MW/d) \tag{3.2}$$

  where n is the refractive index, MW is the molecular weight, d is the density.
- Two descriptors encoding the hydrogen bond donor and acceptor ability of the compound.

These descriptors were then converted into four autocorrelation vectors, scaled and fed into a conventional feed-forward back-propagation neural network (for details on neural networks, see Chapters 6 and 8). Trial and error was used to establish the number of autocorrelation components. Four configurations of a 35:32:1 network architecture were used for the final predictions. The training set consisted of 7200 compounds, with 200 structures held out to control training. Predictive performance was judged using a test set of 519 compounds, and the rms error was quoted as 0.39 log units. Autocorrelation vectors are well established now, and the approach looks promising. However the use of the neural network approach means that chemical interpretation is not as straightforward as it is with the fragment-based methods.

32 atom-type E-state topological indices originally derived by Kier and Hall [25] were calculated for 326 drug-like molecules extracted from the Pomona Database, and a feed-forward back-propagation neural network with five neurons in the hidden layer was used to define the relationship between the topological descriptors and log*P* [26]. In this network the authors use a control set to determine when to stop training, and avoid overfitting. 19 compounds were used as a test set and the standard error of prediction was 0.57 using all descriptors and compounds.

Another approach [27] tried to use the connection table directly as an input to a neural network. Better results were obtained when the connection table was replaced by the atom type and bond type as dummy variables. The position of each atom was enumerated by an atom-numbering scheme. There were 268 compounds in the training set, and 147 variables. Using the architecture described, there is a possibility of overfitting, unless great care is taken. Accordingly, the error was 0.28 log units for the fit, and 0.66 log units for the test set.

One of the best known methods for the calculation of log*P* is that based on the LSER approach [28]. In this treatment [29], which was originally applied to monofunctional solutes, log*P* is regarded as a combination of the solute hydrogen bond donor ($\Sigma\beta\alpha_2^H$) and acceptor ability ($\Sigma\beta_2^H$), a dipolarity term (π_2^H), a cavity term represented by the volume of the solute (Vx), and an excess molar refraction term (R_2) in log units, as shown in Eq. (3.3)

$$\log P(\text{octanol}) = 0.088 - 0.562R_2 + 1.054\,\pi^2{}_H + 0.03\,\Sigma\alpha_2^H - 3.46\,\Sigma\beta^2{}_H + 3.81V\text{x}. \tag{3.3}$$

This approach is appealing because it is easy to interpret in a chemically meaningful way. So in Eq. (3.3) for example, it is easy to see that the solute hydrogen bond acceptor ability tends to drag the compound into water rather than octanol, whereas hydrogen bond donors are happy in both. In addition, there is the added advantage that no special schemes are needed to extend the approach to other solvent systems. All that is needed are the log*P* values, and the differences between the solvents are found in the relative size of the coefficients (see Section 3.1.5). However, there are a number of caveats on the practical use of this approach. The first is that all the parameters in Eq. (3.3) have to be available for all the structures. Clearly there is no problem with volume, but despite large compilations there are still significant gaps in the database for the other descriptors. Secondly, Abraham has adapted his original approach to encompass multifunctional solutes, but again the practical problem remains of how to assign hydrogen bonding numbers to complex structures.

Interestingly, there is now quite a lot of work going on to replace the experimentally measured parameters of the LSER with calculated properties. In one approach a set of descriptors has been developed based on surface area, and several parameters which are derived from the electrostatic potential on the surface [30,31], and encouraging results has been obtained with these descriptors.

In another quite different approach [32], a set of fragments was defined based on those first used by Klopman [33] in his earlier group contribution approach for calculating log*P*. These fragments were coded using SMARTS, and the occurrences were counted for each structure in the database of Abraham. The counts were then regressed against the experimental measures of logR_2, π^2_H, $\Sigma\beta^2_H$, and log$L$, although new structural definitions were needed for $\alpha_2{}^H$. Good relationships were found for all these parameters with the counts. LSER has always been a very nice approach but, because of the problems of assigning descriptors to complex structures, it is really only a practical proposition for experts in the field. Perhaps the approaches described are the first steps towards broadening the appeal of this method by producing these descriptors automatically from any database of structures.

Another approach to the calculation of log*P* octanol uses properties derived from semi-empirical molecular orbital calculations to feed into a neural network [34]. In this work, the semi-empirical methods AM1 or PM3 were used to optimize the structures and properties were calculated. Both regression and neural networks were used to fit the properties to log*P* measured on 1 005 compounds. The regression model ended with 12 terms, and a standard deviation of 0.8 log units. The network was developed with a architecture of 16:25:1 using a training set of 980 compounds, and the standard deviation was 0.41 log units. The standard error on the test set was 0.53 log units.

Work continues to be reported using theoretical methods to describe desolvation. Much of this work is very intensive computationally. In addition the octanol–water system is a rather complex system to model because experimentally the two phases are not pure liquids but are mutually saturated. So the *n*-octanol contains around 5M water, and the water contains around 0.01M *n*-octanol.

The parameterization of the AMSOL [35] suite has been extended to include octanol as one of the solvents [36]. The electrostatics are represented by the generalized Born treatment, and the rest by the solvent accessible surface area in combination with empirical atom surface tensions. The authors observed that when the empirical terms are added the contribution of the electrostatic term becomes very small. The implementation described for oc-

tanol–water partition coefficients SM5.0R offers no increased accuracy over previous AMSOL models, but is much faster to compute. This speed is achieved by neglecting the electrostatic term, which means that no explicit calculation is needed. All contributions to the solvation free energy are related to the solvent-accessible surface area. In addition the model has also been parameterized for solvents other than octanol. The authors warn that they are not sure how well the model will cope with structures that are very different from their parameterization set. Nevertheless for the ten simple solutes described, the agreement between calculated and measured log$P$ is very good.

3.2.3 Predictive Ability of Existing Techniques

With so many methods available, one key question is how well do they really work in practice? In 1997 Carrupt and co-workers noted [11] that *"because of the limited validation domain of some techniques, these comparisons should be re-evaluated whenever a new model is proposed. In addition the number of solutes for which full comparison is made is so limited that the reliability of such studies is highly questionable."* We believe that the situation is not much better now. Another potential problem with most comparative studies is that they all draw on published data and are therefore by nature retrospective. So it is not easy to tell whether the compounds being used to test predictive ability were in the set used to derive the method in the first place. We believe this is another confounding factor in the apparent variations in predictive ability between some of the methods. Consequently these studies, valuable though they are, may be unintentionally biased.

Some independent studies have been carried out, but on rather small sets of compounds, at least by today's standards. Van de Waterbeemd and colleagues [37] looked at around 180 compounds and 14 different methods, and concluded that fragment-based methods such as those described by Leo or Rekker were best. In this comparison, regression coefficients were given between the measured values of the test set and the predicted values. The correlation coefficients ranged from 0.86 to 0.96, although the prediction error was not reported. Another nice comparison [38] used principal component analysis to highlight the similarities and differences between the different methods of calculation. In this study 149 compounds were used in total, including 111 simple compounds and 48 drugs. The methods were grouped on their ability to predict both HPLC retention data and log$P$ octanol. Again the fragment methods performed well. Other head-to-head comparisons of various methods have also been done, usually by the authors of the methods. For example, a comparison between AUTOLOGP, and the method of Klopman [39] showed for 1400 or so compounds a rms residual of 0.33 and 0.37 respectively. The authors conclude that both models work well but AUTOLOGP is superior. To be honest, from these results and from the inspection of the plots we are hard pressed to tell the methods apart.

For all of the reasons describe above we set up our own prediction test [40] which will be reported separately. The test was to see how well the methods could predict the log$P$ of around 1200 neutral compounds, which had been measured in-house over the past 20 years. These compounds do not appear in any of the public databases, and we believe the log$P$ values to be reliable. In this way we should have a true test of predictive ability for the calculation methods, that is to say, can they predict the log$P$ of compounds they have never "seen"

before? The compounds themselves are representative of our drug-hunting programs, and are derived from over 500 different chemical families, with the tautomers fixed to the appropriate form.

We have picked representative methods from those available, and they are ACDLOGP, CLOGP, ALOGP, MLOGP, and CLIP.

MLOGP [41] uses 13 descriptors, and non-linear regression is applied to relate these final 13 descriptors to log*P* measured on 1230 compounds. It was validated on a test set of only 22 compounds, and because of this, the conclusions regarding its superiority over other methods have been criticized [42], although since then a more extensive comparison has been carried out [43]. In this instance we have used coefficients obtained by applying the method to the 9000 or so Starlist compounds. These were then applied to the test set, and are reported in Table 3.1.

Table 3.1. Predicted log*P* of 1357 AstraZeneca compounds.

Method[a]	*n*[b]	Slope[c]	Intercept[d]	*r*2[e]	sd[f]	sd resid[g]
CLOGP	1357	0.75	0.4	0.64	0.86	0.94
ACDLOG	1345	0.70	0.65	0.67	0.80	0.94
MLOGP	1357	0.46	1.18	0.26	1.2	1.6
ALOGP	1279	0.82	0.41	0.75	0.71	0.77
CLIP	1224	0.76	0.64	0.66	0.84	0.92
CLOGP 461[h] ALLFRAG	1283	0.80	0.38	0.72	0.73	0.94
MLOGP[j] 10 OUT	1348	0.71	0.66	0.42	1.1	1.16

[a] Method (see text),
[b] no. of observations,
[c] fitted slope,
[d] fitted intercept,
[e] squared correlation coefficient,
[f] standard deviation,
[g] standard deviation of the difference between the measured and observed,
[h] CLOGP v4.61 for compounds with missing fragments excluded (see text),
[j] MLOGP with ten extreme outliers removed (see text).

CLIP uses the 3-D structure and is based on the molecular lipophilicity potential (MLP) [44]. The MLP at a fixed point can be expressed by an equation involving atomic contributions, and a distance function. If the MLP values are summed over the surface area of the molecule two parameters can be generated, one for the total positive potential, and one for the total negative potential, representing the hydrophobic and hydrophilic parts of the molecule respectively. These two parameters are used as descriptors for log*P* in a regression equation. The other methods, ALOGP, CLOGP and ACDLOGP are described in Section 3.1.2.

Reported in Table 3.1 are the details of the fit for the relationship between the observed and predicted log$P$ of the test compounds for each of these methods. In addition we also report the standard deviation of the difference between the log$P$ calculated by the different methods and the measured values. We believe this difference is the best measure of a method's ability to predict. It is quite clear from Table 3.1 that the errors are much larger than expected, as judged from some of the claims made in the literature. However, none of these methods had seen these actual structures before, so we believe this is a more realistic test of how well they can be expected to predict in practice. CLOGP for all compounds has a standard deviation of prediction of 0.94 log units, however around 74 compounds had a missing fragment, and if we eliminate these (CLOGP461ALLFRAG in Table 3.1) the standard deviation drops to 0.79 log units, and the fit between observed and measured improves. We believe the same would be true for ACDLOGP, a similar method, but we have not tested it here. There was no discernible difference in overall performance between CLOGP and ACDLOGP, two similar methods. If we define failure as wrong by more than one log unit, then ACD failed on 234 compounds and CLOGP on 228, but only 40% of the compounds were failures in both methods. Within chemical families, even if the absolute values were incorrect, the relative rankings were correct with both methods. In other words, both methods "failed" on an equal proportion of chemical families, but the families were quite different.

In this set of compounds ALOGP seems to do well, but only compounds and fragments which obeyed the rules we set out at the start of Section 3.1.1 were included in the analysis, so not all compounds were included. The only 3-D method here, CLIP, did no better on these structures than the simpler fragment-based methods, given the complications of conformational flexibility which were not addressed here. MLOGP had ten huge outliers in our dataset, and when these were removed the method did better, but still was quite poor compared with the other methods used. Perhaps the main point to take is that errors of the magnitude shown in Table 3.1 are to be expected when methods are applied to a disparate set of structures which they have never seen before. Consequently there is no one method that is superior to any other; they all calculate some chemical families well, and some badly. Unless the methods are well understood, it is difficult to decide ahead of time which ones are likely to work best in any given situation. Clearly the question will decide the approach to be taken. If all that is needed is an indication that, say, log$P > 3$ then the calculation methods described may be quite adequate. If, on the other hand, log$P$ values are needed for correlation or to calculate log$D$, then any calculation procedure will need to be calibrated with measurements.

Tautomers provide a real problem for the calculation of log$P$ *de novo* [5]. If the structure present in the databases is in the incorrect tautomeric form, then most of the methods described in the previous section will fail. In ACDLOGP this is apparently addressed.

3.2.4 Other Solvent Systems

The *n*-octanol–water system has some appealing experimental properties which make it attractive: it is practically insoluble in water, it has very low vapor pressure, and it is transparent in the UV region. However, it is arguable whether octanol–water is the most appropriate solvent system to model a membrane, and from time to time over the years other solvent sys-

tems have been proposed and compared [4]. For example, for blood–brain distribution the difference between log*P* measured in cyclohexane and log*P* measured in octanol is relevant [45]. In an earlier study Davies and co-workers argued that anaesthetic activity was better modeled by gas–ester partition coefficients [46]. Two related factors have conspired to hold back development of these other solvent systems. The first is that the measurements are more difficult than in octanol, so there are far fewer reliable measurements. Secondly, the paucity of measurements has held back the development of methods to calculate log*P* from structure, again in direct contrast to the situation in *n*-octanol.

The LSER approach (described in Section 3.2.2) is perhaps the only method which can be used to reliably predict partition coefficients in solvent systems other than octanol, but all the caveats described in Section 3.2.2 still apply. As an example, Eq. (3.4) shows the situation found for partitioning between vapor–$CHCl_3$ [47]. Here the hydrogen bond acceptor coefficient (for β) is positive which suggests that acceptors prefer $CHCl_3$ to the vapor phase.

$$\log P(CHCl_3) = 0.116 - 0.467R_2 + 1.023\,\pi^2{}_H + 0.138\,\Sigma\alpha_2{}^H + 1.432\,\Sigma\beta^2{}_H + 0.994\log L^{16} \quad (3.4)$$

3.2.5 Effect of Ionization

All of the methods described in the previous sections refer to neutral compounds. For ionized compounds, the distribution coefficient *D* for basic compounds, is given by Eq. (3.5)

$$\log D = \log P - \log_{10}(1 + 10^{(pKa-pH)}). \quad (3.5)$$

Eq. (3.5) shows that *D* depends upon the log*P* and p*K*a of the solute and the pH of the experiment, and the assumption is that only the neutral form partitions into the organic layer: the so called pH-partition hypothesis. For low dielectric solvents this is a reasonable assumption. However, octanol contains roughly 5M water and it is not unknown for ionized species to partition into the organic layer as well. Under these conditions log*D* depends on the experimental conditions of pH, concentration of counter-ion, and the nature of the counter-ion. In order to calculate log*D* from structure, two pieces of information are required: the p*K*a, and log*P* of the solute, which are then combined using Eq. (3.5) to calculate log*D*. Clearly, predictions can only be as good as the individual estimates of log*P* (see Section 3.2.3), and p*K*a (see Section 3.4). There are two commercial systems available to calculate log*D*.

The PROLOGD [48,49] method has been validated using log*D* measured on a set of 84 compounds at a variety of pH values. PROLOGP contains three ways to estimate log*P*, and the method used to estimate p*K*a is described in Section 3.4.2. The standard deviation of the log*D* estimate is around 0.7–0.8 log units for these compounds.

Again with ACDlog*D*, to our knowledge, there is no published validation of the log*D* calculation, apart from that on the ACD web site. To us, it seems almost impossible to have any idea of the prediction error from the data provided. However it cannot be expected to perform well on structures which are not well represented in the databases.

In our opinion they may find some use in well defined and well behaved systems, or in combination with measurements to validate the approaches.

3.3 Prediction of Solubility

Aqueous solubility is an important property in governing the ability of a drug to be absorbed. In addition it represents the maximum free concentration which the drug can achieve in the biophase. Solubility remains one of the most challenging properties to predict from chemical structure because there are two processes to consider:

1. The disruption of the crystal, which is related to the lattice energy.
2. The aqueous solvation of the drug.

The two main approaches used are those based on molecular properties and those based on group or fragment contributions. There have been many attempts to predict solubility, and these have been nicely reviewed recently [50,51].

3.3.1 Fragmental Approaches

Since the solubility of a solute in water includes a solvation term, and fragment methods have worked well for the prediction of other solvation-based phenomena such as partition coefficient (see Section 3.2.1), there was a natural tendency to see whether these fragment-based methods could be successful in predicting solubility. Just as in the fragment contributions approach log$P$, so the solubility is defined as an additive property of the atoms or the fragments of the molecule. However there are two key differences here, which we believe may hamper the success of these approaches:

1. Because solubility contains contributions from both the lattice energy and solvation, the group contribution to both these processes must be constant across a wide variety of chemical types and structural environments. To compensate for this effect, many methods use a term which involves the melting point to specifically account for effects in the crystal.
2. The fragment approaches rely for their success on having a wide representation of structural fragments in a wide variety of molecular environments. In contrast to the situation with log$P$ where there are some 30000 measurements in the public domain across huge varieties of structures, the database of measurements for solubility does not cover the same richness of functionality.

Irmann [52] pioneered the fragment approach by proposing Eq. (3.6), which is a hybrid of series, atoms and fragments contributions.

$$\log S_w = a + \Sigma n_i b_i + \Sigma n_j c_j + 0.0095(M_p - 25) \tag{3.6}$$

where a is the contribution of the compound type, b_i is the contribution of the ith atom type which occur n_j times and c_j is the contribution of the jth fragment which occurs n_j times. The final term in his scheme is derived from the entropy of fusion at the melting point, which he took to be a constant 13 cal Deg^{-1} mol^{-1}.

Another example of this approach has been published by Yoshimoto et al. [53] who defined six fundamental fragment solubility constants using a restricted set of 46 liquid aliphatic hydrocarbons. Another set of 19 fragment constants were determined from 249 liquid aliphatic compounds with diverse functional groups followed by 15 fragment constants determined from 58 aromatic liquid. Obviously the overall training data set is neither "drug-like" nor even multi-functional. The Irmann and Yoshimoto approaches have been reviewed and discussed by Suzuky [54]. Other group contribution approaches such as UNIFAC [55] have been proposed, and used [56] but, because they refer to the liquid phase, they too require a heat of fusion or a melting point.

The AQUAFAC (AQUeous Functional group Activity Coefficient) method is one of the most elaborated and most successful fragmental techniques to calculate aqueous solubility. In this method, the aqueous activity coefficient is estimated through fragment contributions, and the packing forces are accounted for by a term incorporating the melting temperature. Myrdal *et al.* [57] defined a set of 24 possible groups from a selected data set of 2400 mainly mono-functional solutes. The different molecular environments led to 52 fragments, of which only 35 could be used. Regression analysis was then used to find the relative weights for each of the fragments. These fragments give good solubility predictions with rmse = 0.4 log units of molar solubility. It has to be noted that on this data set AQUAFAC does not give a significant improvement compared to the so-called general solubility equation which uses calculated $\log P_{oct}$ values (see Section 3.2.2). Furthermore the method still needs measured *Mp* values, and a measured or estimated entropy of melting. Later, the method was extended [58] by adding a further 168 compounds which were much more diverse, and contained multi-functional solutes. This led to a further 26 fragments. However, it is striking that around 30% of these new compounds are barbiturate derivatives. Consequently it is not clear whether this introduces a bias into the training set. Nevertheless the group contributions to the solubility themselves are interesting, and seem to make sense. For example $CONH_2$ attached to sp^3 C is more soluble by 1.5 log units on average than when attached to an sp^2 C. Interestingly, a direct comparison between the fragment values derived from solubility and those derived from partitioning have been carried out [59]. For non-hydrogen bonding groups the fragment values are comparable, but hydrogen bond donor groups favor partitioning into octanol more than they favor increasing solubility. This is also confirmed by the coefficients obtained from the approach of Abraham (see Section 3.3.2).

An approach identical in concept to one derived for partition coefficients has been proposed by Klopman and co-workers [60]. They used a final set of 469 mainly mono-functional compounds from the literature, and 45 fragments, which occurred at least three times in the data set. They obtained a sd of 0.46 log units on the fit, and a sd of 0.5 log units cross-validated. In this instance no specific term was used to account for the solid state. Although the data set is rather small again, the fragment contributions to solubility do seem to make sense, so for example an aromatic ether is less soluble than an aliphatic ether by nearly 1 log unit. It is rather difficult to cross-compare the fragment values obtained from different methods, because all the fragmentation schemes and data sets are slightly different.

3.3.2 Property-Based Methods

Probably the best known approach is that based on log$P$, which started from an empirical observation by Hansch [61] that the aqueous solubility (logS_w) can be correlated with octanol–water partition coefficient (logP_{oct}), for a series of liquid non-electrolytes. Yalkowsky [50] combined the approach of Hansch with that of Irmann (see Section 3.3.1), and added a thermodynamic rationalization to produce what has come to be known as Yalkowsky s equation (Eq. 3.7). This equation relates logS_w to logP_{oct}, the entropy of fusion (ΔS_f, taken to be constant) and the melting point (Mp) of the compound.

$$\log S_w = -1.00\log P_{oct} - 1.11\,\frac{\Delta S_f\,(Mp - 25)}{1364} + 0.54 \tag{3.7}$$

Given the widespread use of this approach we felt it was worth just outlining again the assumptions and approximations made when using this equation. The relationship linking the mole fraction aqueous solubility, X_w, to the ideal solubility, X_i, and the activity coefficient γ_w is shown in Eq. (3.8)

$$\log X_w = \log X_i - \log\gamma_w \tag{3.8}$$

where in the case of liquid solutes $X_i = 1$, and the mole fraction aqueous solubility is expressed solely by the activity coefficient γ_w. The octanol-water partition coefficient, XP_{oct} can be expressed in mole fraction terms by the activity coefficients as shown in Eq. (3.9)

$$\log XP_{oct} = \log\gamma_w - \log\gamma_o \tag{3.9}$$

Yalkowsky noted that most liquid solutes have similar polarity to octanol and so will be infinitely miscible in this solvent, thus the activity coefficient, $\gamma_o = 1$ (Eq. 3.10), in other words all organic compounds will form an ideal mixture with n-octanol.

$$\log X_w = -\log XP_{oct} \tag{3.10}$$

Converting from mole fraction aqueous solubility, X_w, back to molar units gives Eq. (3.11)

$$\log S_w = -\log P_{oct} + 0.87 \tag{3.11}$$

Hildebrand and Scott [62] demonstrated that the ideal solubility of solid solutes is less than 1 as shown by Eq. (3.12)

$$\log X_i = -\frac{\Delta S_f\,(Mp - 25)}{2.3RT} \tag{3.12}$$

[Eq. (3.12) is related only to the solute in the crystal]. The entropy of fusion, ΔS_f, is dependent on the molecular shape of the compound such that for spherical molecules $\Delta S_f = 3.5$eu, for rigid molecules $\Delta S_f = 13.5$eu and for molecules having $n > 5$ non-hydrogen atoms in a flexible chain $\Delta S_f = 13.5 + 2.5(n - 5)$eu. If we make the assumption that most drugs are rigid molecules, then at 25°C the ideal solubility is defined by Eq. (3.13)

$$\log X_i = 0.012(Mp - 25) \tag{3.13}$$

Combining Eq. (3.13 and 3.11) via Eq. (3.8) gives the simplified version of the Yalkoswsky's equation shown in Eq. (3.14)

$$\log S_w = -\log P_{oct} - 0.012\,(\mathrm{Mp} - 25) + 0.87 \tag{3.14}$$

Considering the approximations involved in generating Eq. (3.14), it would be surprising if it held generally for all compounds. Indeed it is common to use multiple regression to optimize the coefficients for $\log P_{oct}$ and Mp. Yalkowsky [113] obtained remarkable results on a diverse set of 873 molecules using CLOGP, ΔS_f having been estimated from the rotational symmetry of the structure and a measured Mp. The limitations of this approach are clear from the work of Meylan and co-workers [63] who collected a dataset of 1450 compounds, around one third of them liquids, from the literature. Using log$P$ and melting point as the descriptors they found the standard deviation to be 0.60 log units. Interestingly, even with these relatively simple compounds they identified 15 extra structural features, which appeared to introduce systematic bias into the model. When these structural features were included as dummy variables in the equation, the sd was lowered to 0.45 log units. In another study, 360 fairly simple compounds were clustered on the basis of the structural keys from Chemical Abstracts [64]. Models were generated for each cluster separately using log$P$ and melting point. Interestingly, this approach identified some of the same structural features that were identified by Meylan in his study. However, as both these authors point out they combine literature data, and it may be that some of the bias they see can be related to the variability in the measurements between the laboratories.

In our laboratories over the years, we also have found that this simple approach often does not produce general models. For example, the results shown in Eq. (3.15) are from a non-congeneric set of 22 compounds [65]. The model is quite poor despite the use of measured values.

$$\begin{aligned} &\log S_w = -0.0095\,(Mp - 25) - 1.214 \log P_{oct} + 0.85 \\ &n = 22,\ \ r^2 = 0.36,\ \ rmse = 0.95 \end{aligned} \tag{3.15}$$

Shown in Table 3.2 are results from medicinal projects within AstraZeneca where the approach has worked well [66].

There are four chemical series, which are all pure well-characterized crystalline compounds. In addition there are two other series which, with the introduction of a dummy variable to account for structural features (not shown in Table 3.2), also gave excellent relationships. The data in Table 3.2 were fitted using multiple regression to optimize the coefficients for log$P$ and melting point. The intercepts show a range of over 4 log units in solubility, rather than the theoretical value shown in Eq. (3.14). In our experience though, such relationships have been the exception rather than the rule, and it is difficult to predict ahead of time which ones are likely to work. It may of course be that our experience is not general, and others have been much more successful. However, packages which offer this treatment as a general solution for the estimation of solubility should be treated with caution.

The situation may not get much better either, particularly with the advent of automated chemistry where samples may be impure, not crystalline, or a melting point is simply not available due to decomposition. Consequently another limitation with this approach is the

Table 3.2. Successful relationships using log*P* and melting point to predict solubility for Medicinal Series.

Series No	No of compounds	$r2^a$	Intercept[b]	Coefficients log*P*[c]	*mpt*[d]	sd
1	10	0.72	–3.3	–0.7	–0.007	0.47
2	25	0.67	–3.1	–0.6	–0.004	0.41
3	12	0.55	–1.3	–0.07	–0.006	0.65
4	19	0.96	+0.9	–1.2	–0.017	0.24

[a] Squared correlation coefficient,
[b] intercept from fitting Eq. (3.14),
[c] coefficient of log*P* from fitting Eq. (3.14),
[d] coefficient of melting point from fitting Eq. (3.14).

requirement for melting points. So it would be very nice if there were methods around which might be used to estimate melting point. There are some interesting developments in this area. For example Katritzky and co-workers [67] have used calculated properties within the framework of their CODESSA program to produce models which relate these properties to the melting point for 443 substituted benzenes. A model was developed with between six and nine properties; however even in these simple cases, separate models were required for the different substitution patterns. The standard error was around 30° C. In another approach, Eq. (3.16) has been derived as a predictive model of the melting point [68]:

$$Mp = -\frac{\Sigma n_i m_i}{56.5 - 19.2 \log\sigma + 9.2\tau} \tag{3.16}$$

in which $\Sigma n_i m_i$ is a fragmental approach to calculate the enthalpy of melting, and the lower term evaluates the entropy of melting where s indicates the number of identical images that can be produced by rigid rotation of a molecule, and τ indicates the effective number of torsional angles in the structure [68,115].

We believe this is a very promising advance but it remains to be seen whether it will make the estimation of solubility from structure successful more consistently. Abraham *et al.* [51] have used their LSER approach to predict aqueous solubility. For 594 solutes, the sd on this set was 0.56 log units, and 0.5 log units for 65 compounds held out as a test set. For all 659 solutes the results are shown in Eq. (3.17)

$$\log S_w = 0.52 - 1.0R_2 + 0.77\,\pi_2^H + 2.17\Sigma\alpha_2^H + 4.24\Sigma\beta_2^H - 3.36\Sigma\alpha_2^H\Sigma\beta_2^H - 3.99V_x \cdot \quad (\text{with } n = 659, \text{sd} = 0.56, r^2 = 0.92) \tag{3.17}$$

This equation is very similar in form to the one they have used for partitioning processes, and is described in Section 3.2.2. The only difference is that a cross-term, $\Sigma\alpha_2{}^H\Sigma\beta_2{}^H$ has been introduced to partly account for interactions in the solid state. The logic is that in the solid state it will be more difficult to disrupt compounds with both hydrogen bond donor and hydrogen bond acceptor capabilities, and in line with this interpretation the coefficient for this

term opposes solubility. The strengths and weakness of using this approach in practice have been discussed in Section 3.2.2, and will not be elaborated further here.

The largely accepted explanations of the solubilization process, namely the enthalpic cost of creating a cavity in the bulk of water to receive the solute molecule and the entropic cost of rearranging the water molecules around the solute molecule have been challenged by Ruelle *et al.* [69]. In their mobile order theory "*the motions in the liquid state make needless the creation of a cavity to dissolve a foreign substance*" and the solvophobic effect is due to the reduced mobile order of the solvent molecules when solvent is diluted by a solute. The quantitative treatment of this theory leads to the universal solubility Eq. (3.18) [70], in which ln Φ_B is the aqueous solubility in volume fraction, A is the fluidization of the solute or its ideal solubility which can be evaluated with Eqs. (3.12 and 3.13), B is the combinatorial entropy originating from the exchange possibilities between the solute and the solvent molecules in solution, F is the mobile order entropy decrease (solvophobic effect) of the self-associated solvent molecules following the dissolution of the solute substance, D is the change in non specific forces upon mixing, O is the effect of the H-bonds between the proton-acceptor sites of the solute and the proton-donor solvents, and OH is the effect of the H-bonds formed by the amphiphilic groups of the solute including its self-association in solution.

$$\ln \Phi_B = -A + B - F - D + O + OH \tag{3.18}$$

In Eq. (3.18) the mobile order entropy decrease (F) is the most important term for compounds with poor aqueous solubilities, though this effect still remains important in determining the solubility of drugs. Group contributions of the six terms are compiled in tables and the universal solubility equation allows the calculation of the solubility of non-electrolytes not only in water but also in various alcohols. The approach is quite successful since it gives a sd = 0.45 ln Φ_B units, given that the ln Φ_B (in water) for a data set containing numerous drugs spans over more than 30 log units. Following earlier attempts to correlate volumes and surface areas with aqueous solubility reviewed by Yalkowsky and Banerjee [50], Silla *et al.* [71] published interesting results on series of hydrocarbons and oxygen containing aliphatics. Unfortunately the results they obtained on simple series cannot be extended to more complex structures often encountered in drug research. Furthermore the predictive ability of the model was not tested on compounds outside the training sets.

Bodor *et al.* [72] mixed dummy variables indicating the presence or absence of some chemical families, 3-D molecular descriptors like surface area or volume of the molecule, and descriptors issued from semi-empirical calculations like the dipole moment or different electrical charges attached to various atoms. In total they use 18 variables, which after linear regression led to a model having sd = 0.30 [72]. A corresponding neural networks regression [73] gave a slightly worse model with a sd = 0.43 which after optimization of the neural net architecture, yielded a better model having a sd = 0.27 [74]. The approach of Bodor in estimating log$P$, using similar methods, has been criticized [11] because it was not easy to interpret the physical relevance of some descriptors, and also because of the of the number of parameters considered, compared to the number of compounds in the training set. We believe the same criticisms can be applied here.

Jurs *et al.* [75] produced 144 chemical descriptors for each of 140 fairly simple compounds. Both regression and neural networks were used to produce a model after redundant vari-

ables were removed. They obtained a model using nine descriptors with a rms error of 0.32 log units for 123 compounds. The same nine descriptors were fed into a neural network in a 9:3:1 architecture to see if better models resulted. In both cases, polychlorinated biphenyls were not well predicted. The approach was further developed [76] using a more diverse set of 332 structures. In addition a quite sophisticated method was used to solve the problem of variable selection which is often a difficulty with this kind of approach. It is described here as an example of the problem that can be met. Initially Jurs calculated 210 descriptors. He removed redundant descriptors or ones that had little variation by examination of the pair-wise correlation between descriptors or by a Principal Component Analysis to leave 122 descriptors. A genetic algorithm and simulated annealing coupled with multiple linear regression gave an initial descriptor set, which was used to select an optimal neural network architecture. In the final step, the optimized network architecture is used to select a model from the 122 descriptors. The overall procedure is computationally very heavy but leads to a good model with nine descriptors and a rmse = 0.39 on a test set.

Katritzky *et al.* [77,80] use their CODESSA program to calculate approximately 1000 descriptors and to extract from them a model based on linear regression technique having six descriptors and a standard deviation error of 0.57.

Huuskonen [78] and co-workers carried out a very nice study where they did not attempt to find a universal model for solubility. Instead they studied three classes of drugs and the aim was to see if solubility could be predicted within the structural families. They used 28 steroids, 31 barbiturates, and 24 reverse transcriptase inhibitors. They used topological indices and established the models using a 5:3:1 feed-forward neural network architecture, and used cross-validation to assess the models. Models of varying predictive ability were found for all sets, but no common model. More recently they unified their approach by merging their three training sets plus other drugs into a single data set containing 211 examples [79]. By extending the descriptor set with Kier and Hall's atom type E-state index, they obtained promising results with a sd of the error of prediction equal to 0.53 log units on 51 randomly chosen test compounds. However we are bound to agree with their conclusion which states that *"The development of universal empirical models for predicting aqueous solubility from structure is still a challenge."*

3.3.3 Conclusions

The authors' experience in this field has shown that it is relatively easy to find a predictive model of solubility using the published data, and there are many successful reports of it in the literature, but that it is far more difficult to predict proprietary research compounds. To illustrate this point some data issued from the databases published in recent and successful papers discussed in this review, along with in-house data, are summarized in Table 3.3. The AstraZeneca data set is taken as an example of a set of compounds of interest to pharmaceutical research. These are thermodynamic solubility measurements done over the past ten years at Alderley Park, so they represent a selection from all our medicinally relevant series over that time period. There are some striking differences compared with the public domain databases. Compared with these databases, the AstraZeneca compounds are larger, by around 100 Da although still in an acceptable range. They are more lipophilic by around 1–1.5 units,

they are more likely to be multi-functional, and perhaps most strikingly they are less soluble by between 100- and 10000-fold. If our database is typical of other pharmaceutical companies, then maybe it is not surprising that the methods for solubility estimation do not work as well as we would like because the techniques have all been developed on databases which appear to have quite different characteristics to ours. It is also striking that the simple linear correlation with $\log P_{oct}$ gives at least a broad trend with all the databases except the AstraZeneca database. This is not to say that the derived models are not superior to those using CLOGP alone, but it does perhaps give another indication that the compounds available in these databases are not representative of our database.

Table 3.3. Comparative characteristics of some published databases used for training solubility models.

Database[a]	n[b]	MW (sd)[c]	$ClogP_{oct}$ (sd)[d]	$\log S_w$ (sd)[e]	r^2 (no descr.)[f]	r^2 with $ClogP_{oct}$[g]
Bodor [72]	331	124 (59)	2.39 (1.29)	−2.05 (1.59)	0.96 (19)	0.89
Katritzky [77]	411	111 (42)	2.00 (1.3)	−1.60 (1.53)	0.88 (6)	0.86
Abraham [51]	648	147 (66)	2.64 (1.60)	−2.50 (2.00)	0.92 (5)	0.78
Taskinen [79]	235	255 (78)	2.01 (1.76)	−3.06 (1.50)	0.90 (23)	0.56
Jurs [76]	320	179 (87)	2.81 (2.47)	−3.30 (2.35)	0.97 (9)	0.73
AstraZeneca[h]	152	394 (98)	3.57 (1.44)	−5.50 (1.37)		0.34

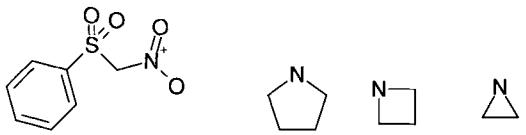
[a] Database published in the corresponding paper,
[b] number of compounds calculated in the present work (may differ from the original work because of difficulties in retrieving some structures),
[c] mean molecular weight (MW) and standard deviation (sd),
[d] mean $ClogP_{oct}$ and standard deviation (sd),
[e] mean $\log S_w$ and standard deviation (sd),
[f] published squared coefficient of correlation of the calculated $\log S_w$ with the measured $\log S_w$ and number of descriptors involved in the model (nb descr.),
[g] squared coefficient of correlation of $ClogP_{oct}$ with measured $\log S_w$,
[h] data set of non-ionizable compounds with solubility measured in-house at pH 7.4 in 0.01M phosphate buffer after 3 days equilibrium.

It would be unfair to be critical of any of these methods because they do not explain data that as a pharmaceutical industry we have refused to let people see. Nevertheless for compounds which are typical of our medicinal programs, it is still the case that none of the methods we have seen produce reasonable estimates of solubility *de novo* from structure. Perhaps more worrying is that, in our experience, few of the methods produce reasonable estimates consistently, even within related chemical families, as the structural complexity increases.

3.4 Prediction of pK_a

In our opinion the charge state of a compound is one of the most fundamental properties because it influences all the other physical properties. For example through the pH-partition

hypothesis it influences the partition coefficient (see Section 3.2.5), and in an analogous way the solubility, assuming the solubility of the salt is not limiting. Consequently in the Rule -of five [43] – which are guidelines to help decide whether a compound is likely to be bioavailable – the absence of the charge state is a clear limitation. We do not believe the pK_a was left out through choice, but because there are simply no good methods around to estimate the pK_a. Over the years there have been a number of attempts to come up with predictive models for the pK_a of a compound, and these are reviewed here. In most cases this is a tough challenge because ionization involves a proton transfer step and so effects which are often second order in ground state processes such as partitioning or solubility can become very important in ionization. For example, the effect of charge delocalization in the carbon acid **1** in Figure 3.1 means that the pK_a is 4.8 [112], in the same range as carboxylic acids. There are also examples where protonation of the basic center causes changes in pK_a for **2–4** as the ring size changes.



	1	2	3	4
pKa	4.6	11.3	11.0	8.0

Figure 3.1. The effects of charge delocalization or protonation on the pK_a of structures 1–4 (see text).

The methods available for the prediction of pK_a fall into two main categories, which will be discussed in the following part of this Chapter: Fragment-Based Approaches, and methods based on Molecular Properties.

3.4.1 Fragment-Based Methods

Klopman and co-workers used their MULTICASE [81] methodology initially on a set of 121 acids and non acids and then extended it to assess the acidity of nearly 2500 organic acids [82]. MULTICASE linearly fragments each structure, and then assigns a relative weight to the contribution of each fragment to the acidity. Acidity is treated as a class variable, and this instance there were 704 acids (pK_a < 6.5), 55 marginals (pK_a 6.5–7.8) and 1705 non acids (pK_a > 7.8). The results look chemically sensible in that the program picked out the main acid families, although it did seem to have problems with phenols where the pK_a can vary from non-acid to acid depending on the substitution, so the phenol biophore appeared several times with different substitution patterns. The standard deviations varied depending on the acid family, but for 601 carboxylic acids it was 0.5 log units. This is not as good as can be expected from the best LFER approaches. The resulting model was then used to predict the acidity of

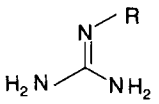
192 drug-like molecules, after 22 structures were removed because they were in the original training set. The standard deviation was 1.5 log units; however it should be remembered that these kinds of fragment methods can only be expected to work well when there are examples of the acid already in the database.

3.4.2 Methods Based on Molecular Properties

In 1935, Hammett [83] suggested that the difference in pK_a between a substituted benzoic acid and benzoic acid itself could be used to model the electronic effects of the substituents as in Eq. (3.19)

$$\log K_x = \rho\sigma + \log K_H \tag{3.19}$$

where K_x = Dissociation constant of substituted benzoic acid
K_H = Dissociation constant of unsubstituted benzoic acid
ρ = Sensitivity to substituent effect
σ = effect of substituent



5

This basic scheme has been extended and modified over the years, and different kinds of s value have been proposed for various situations [84,85]. This formalism too, has been very successful in forming the basis for the estimation of pK_a values within chemical families, and there are many examples [86]. So for example, in a series of substituted guanidines (**5**) measured some years ago [87], the effect of R on the basic pK_a can be described simply by Eq. (3.20) where σ_F is the substituent constant describing the field effect.

$$pK_a = 22.5\,\sigma_F + 14.2 \tag{3.20}$$
$$(\text{with } n = 16, r^2 = 0.92, s = 0.5)$$

For this set of guanidines, Eq. (3.20) shows that we can estimate the pK_a to within 0.5 pK units providing we know the substituent constants. Linear Free Energy Relationships (LFER) are still the simplest and most successful way to predict pK_a from chemical structure, and there are compilations of measured pK_a values [88–90] and substituent constants [91]. However the technique has a number of disadvantages:

1. Measurements are needed on closely related model compounds to set up the regression equation.

2. Hypothetical compounds need to have a tabulated substituent constant.
3. The approach is limited to closely related analogues. Therefore separate relationships are needed for each chemical family.

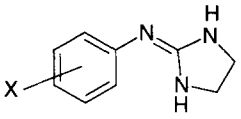
Recently these kinds of approaches have been automated, and software is offered by two companies, ACD (ACD/pK_a) and Compudrug (PKCalc). As far as we are aware there is no refereed publication detailing the methodology used in ACDpK_a, or the source of their 9000 experimental pK_a values. To our knowledge, the only references are on the ACD web-site, where there is a comparison of the estimated pK_a of 22 known drugs with experimental measurements. This is a pity because the methodology looks to be sensible and well put together. In particular they have provided a way to estimate substituent constants of different kinds, when measured values are not available. Nevertheless in our opinion it is quite impossible to assess the likely error in pK_a estimation from the information provided. pKCalc is an expert system which is based on the approach taken by Perrin [92]. The approach and the validation have been published [49]. Both of the approaches described above should in theory provide reasonable estimates for known ionizable systems within closely congeneric series. Where the substituent constant is unknown, then the pK_a will be subject to variable error, depending how good the guess is. The error depends on the sensitivity of the pK_a of the system under study to the substituent effect. In the guanidine example shown above, an error of say 0.05 in the substituent constant will produce an error of nearly 1 unit in the pK_a because of the extreme sensitivity of this system. Whereas the error in a set of carboxylic acids, for example, should be much lower because the sensitvity to substituent effects is much lower, too.

SPARC [93] is another promising approach to the prediction of pK_a. It has been tested on 4300 compounds with apparently astounding results [94]. The approach is complex and relies on breaking the structure into reacting centers (C), say an acid group, and perturbers which are non-acidic structural features affecting the acidity of the reacting center. The pK_a is then given by Eq. (3.21)

$$pK_a = (pK_a)_c + \delta(pK_a)_c \tag{3.21}$$

where $(pK_a)_c$ = pK of reactive centre
$\delta(pK_a)_c$ = effect of substructural feature on reacting group

The perturbers are further broken into substituents, that is groups which exert an influence (S), and conductors, those which transmit an influence (R), to go along with the reaction centers (C). For each of these an effect is calculated, using simple models to account for field effects, electrostatic effects, solvation effects etc. To go with the size of the calculated effects there is sensitivity parameter for each of the reactive centers. These are used to adjust the value of the reacting center pK_a to give the final estimate. It is assumed that the size effects and sensitivity effects are transferable between models. The results on the 4300 compounds look very impressive, and for a test set of 200 structures held out of the analysis the standard deviation is 0.37 units. The approach was further tested on a set of β-blockers and benzodiazepines [95], again with good results. However as the authors emphasize, for the approach to work it is necessary to have good estimate of the pK_a of the reacting center, in other words the approach is only viable for known systems.



6

CoMFA [96] has been used to model the pK_a of some clonidine analogues (**6**), and for 23 compounds the standard error for prediction was reported as 0.27 units. The technique can only be applied to structurally related families because the structures have to be aligned, and there is still the problem of which conformer to choose (see Chapter 10). Nevertheless the approach has the advantage that predictions can be made relatively easily for all related structures; there is no "missing fragment" problem.

For a set of 135 simple alcohols, phenols and carboxylic acids, Jurs [97] developed a model relating the pK_a to an empirical description of the charge. The method is based upon a comparison of the charges on the heavy atoms for the neutral and charged species. The contribution of these charges to the p*K*a is then weighted by assigning each atom to one of four atom types. These categories are halogen, N, O, F and *sp* hybridized C, sp^2 C, and sp^3 C. The weights were determined by regression. The rms error is 0.45 units, and prediction was tested on a further set of 14 compounds. The substituent constant approach outlined above will give comparable or perhaps better results for each family separately, but Jurs method at least offers the prospect of calculating pK_a from structure alone, across the families for these simple systems.

3.4.3 Conclusions

If the aim is to use these methods to scan databases of assorted structures and come up with reasonable estimates of pK_a, then none of these methods are sufficiently reliable. However if the ionizable group is part of a well-studied chemical family, then there are methods which can give reasonable results if carefully used.

3.5 Prediction of Protein Binding

The ability of drugs to bind to serum albumin, effectively reduces the free concentration available to bind at the site of action. Although the affinities for albumin are often much less than those measured for binding to the receptor or enzyme target. However the high concentration of albumin in plasma (approx. 5×10^{-4}M) means that by the law of mass action there may be very little free drug available. Indeed, it has been shown that it is the free drug concentration *in vivo* which correlates with the concentration required for *in vivo* efficacy [3]. Binding to plasma proteins can also have consequences *in vitro* where almost all cell as-

says contain serum albumin of one form or another in order to keep the cells viable. Consequently, we have found that knowledge of serum protein binding helps to interpret this kind of data.

Hervé *et al.* [98] reviewed the role and the characteristics of drug binding to plasma proteins. Plasma binding of drugs not only involves several different proteins, mainly albumin and α-1-acid-glycoprotein (AGP), but at least in the case of albumin up to six different binding sites have been identified for the human protein. Presumably unless there is some knowledge about where the compounds are binding, this will make predictive models quite difficult outside congeneric series. It has been recognized that albumin tends to bind neutral and negatively charged compounds, while AGP prefers positively charged compounds [99] (see also [100] for a recent review).

In addition the crystal structure of albumin is now available [101–103], and there have been some attempts to look at the mode of binding of different drugs. We anticipate there will be more studies of this type in the future.

Most of the published studies find a strong relationship with some form of lipophilicity, and none of them have looked much beyond that (some typical examples follow below). Early work [104] showed that for 25 diverse neutral compounds Eq. (3.22) was obtained.

$$\begin{gathered}\text{Log}1/C = 0.67\ \log P + 2.6 \\ (n = 25, r^2 = 0.89, s = 0.24)\end{gathered} \tag{3.22}$$

However the compounds studied here were not as well functionalized as many medicinally relevant structures. Sholtan [105] looked at the binding of 57 drugs to bovine serum albumin drawn from six families: sulphonamides, tetracylines, penicillins, steroids, carbenolides, and acridines. He found excellent linear relationships within each family between the binding constant to albumin and log$P$ measured in the isobutanol–water system. In each family the slope was 0.9, but the intercept varied depending on the chemical type. When these relationships were converted onto the n-octanol scale [106], identical slopes were obtained to that shown in Eq. (3.22) derived from much simpler structures.

Other published studies concern congeneric series such as cephalosporins [107] (Eq. 3.23) where K_a is the association constant of the drug with human serum albumin, and there is parabolic relationship between the binding constant and the chromatographic column capacity factor k'_w in a RP-HPLC system.

$$\log K_a = 3.91\ \log k'_w - 13.01\ \log k'_w + 13.68 \tag{3.23}$$

Eqs. (3.24) and (3.25) show the relationship for arylpropionic acids between the affinity, and lipophilicity again measured as a HPLC capacity factor [108].

$$\log n_1 K_1 = 0.83\ \log k'_w + 0.18 \tag{3.24}$$

$$\log n_2 K_2 = 1.05\ \log k'_w - 1.68 \tag{3.25}$$

Here n is the number of binding sites per mole of albumin for site I (n_1) or site II (n_2), K is the association constant of the drug with site I (K_1) or site II (K_2).

Other examples show the relationship (Eq. 3.26) between lipophilicity and binding to AGP, for β-adrenolytic and antihistamine drugs [109], and in Eq. (3.27) for 52 assorted basic drugs [110].

$$\log k'_{AGP} = 0.942 + 1.045 \log k'_{IAM} + 1.625 N_{ch} - 0.013 S_T \quad (3.26)$$

$$\log k'_{AGP} = 1.69 + 0.66 \log k'_{IAM} + 3.34 N_{ch} - 0.01 S_T \quad (3.27)$$

In this case, k'_{AGP} is the association constant of the drug with AGP, k'_{IAM} is the chromatographic capacity factor by chromatography using immobilized artificial membrane column, N_{ch} is the electron excess charge on aliphatic nitrogen and S_T is the width of the solute. Although no correlation was found between the percentage of protein bound drug and the partition coefficient measured at pH7.4 ($\log D_{7.4}$) for a series of 18 gyrase inhibitors [111], we have developed (and currently use) a simple model relating whole serum protein binding properties to either measured $\log P$ for neutral and acidic compounds, or $\log D_{7.4}$ for basic compounds [112]. Like Yalkowsky's model for solubility, this model is useful but needs measured values of $\log P$ and pK_a since it is no longer satisfactory when using calculated $\log P$ and pK_a values.

3.6 Conclusions

Considering the importance of the physical properties reviewed here, perhaps it is a surprise to some that predictive ability is not better. Of course it depends on the question that is being asked, and the method that is chosen will reflect the precision needed in the answer. For example, when selecting constituents for a library, or analyzing the results of a HTS campaign, all that may be required is an indication that the compounds are likely to have better than 1μM solubility, or $\log P$ less than say 3. However this may not be good enough in lead optimization where measurements are likely to be needed. We do best with $\log P$ octanol–water where there is a large database of measured values with which to calibrate the calculations. However even here, in our test set of 1300 or so neutral compounds which none of the methods had ever seen before, the standard error of prediction for the best methods was around 0.5 log units, much higher than is often claimed by method developers. To put this error into context, around 20% of the estimates in our database will be wrong by more than 1 log unit. For pK_a estimation, the only practical solution is the use of substituent constants, and estimation by analogy. These methods have been automated, but the validation leaves much to be desired, and we believe that they can be relied upon only when there are very close analogues in the database.

Methods for prediction of solubility are the poorest because the database of measured values is so small and so limited in chemical functionality. In our opinion this makes it difficult to have confidence that any of the methods developed on these databases can be applied to medicinally relevant compounds. Neither does it seem likely that they could be re-cast to predict say soluble or not at some arbitrary concentration say 1μM. Unless and until the public databases contain the structural diversity found for $\log P$, it is difficult to see how this situation can be improved.

Very little systematic work has been reported on albumin binding. What has been done is within chemical series, and is mainly limited to relationships with log*P*.

However, in no way should this review be seen as a criticism of the method developers. In many cases methods were developed with the environmental sciences in mind, where by and large the structures are simpler than in the pharmaceutical industry. To be honest, it is seems harsh to complain that methods do not predict structures that we, as a pharmaceutical industry, are not prepared to release. Until this situation changes it will be difficult to see how predictive methods are going to improve.

References

[1] D. E. Leahy, J. Lynch, D.C. Taylor, in *Novel Drug Delivery and its Therapeutic Implications*, Prestcott, Nimmo (Eds.), John Wiley, Chichester **1989**, pp. 33–44.
[2] W. M. Pardridge, D. Triguero, J. Yang, P. Cancilla, *J. Pharmacol. Exp. Ther.* **1990**, *253*, 884–891.
[3] D. A. Smith, in *Computer Assisted Lead Finding and Optimisation*, Wiley-VCH, **1997**, pp. 265–277.
[4] D. E. Leahy, J. J. Morris, P.J. Taylor, A.R. Wait, *J. Chem. Soc. Perk. Trans. 2* **1992**, 723–731.
[5] C. Hansch, A. Leo, in *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*, A.C.S. Washington, D.C. **1995**.
[6] Pomona or MedChem log*P* Database; available from Daylight CIS or BioByte.
[7] A. J. Leo, *Chem. Rev.* **1993**, *93*, 1281–1306.
[8] G. L. Biagi, M. Recanatini, A. M. Barbaro, P. A. Borea, *Process Control and Quality* **1997**, *10*, 129–149.
[9] A. J. Leo, *Methods and Principles in Medicinal Chemistry Vol 4. (Lipophilicity in Drug Action and Toxicology)*, V. Pliska, B. Testa, H. Van de Waterbeemd (Eds.), VCH Weinheim, Cambridge **1996**, pp157–172.
[10] R. Mannhold, R. Rekker, *The Calculation of Drug Lipophilicity*, VCH, Weinheim **1992**.
[11] P.-A. Carrupt, B. Testa, *Rev. Comput. Chem.* **1997**, 241–315.
[12] CLOGP v 4.61. DAYLIGHT CIS, 27401 Los Altos Suite 360 Mission Viego, CA 92691.
[13] A. J. Leo, *217th ACS Meeting Anaheim Ca Book of Abstracts* **1999**, 105.
[14] A. K. Ghose, A. Pritchett, G. M. Crippen, *J. Comp. Chem.* **1988**, *9*, 80–90.
[15] P. Broto, G. Moreau, C. Vandyke, *Eur. J. Med. Chem.* **1984**, *19*, 71–78.
[16] PrologP Compudrug Chemistry Ltd, 1362 Budapest, POB 405, Hungary.
[17] Advanced Chemical Development Inc., 133 Richmond Street West, Suite 605, Toronto, Ontario, Canada M5H 2L3.
[18] G. O Spessard, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1250–1253.
[19] A. K. Ghose, V. Viswananadhan, J. Wendoloski, *J. Phys. Chem. B* **1998**, *102*, 3762–3772.
[20] S. A. Wildman, G. M. Crippen, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.
[21] SMARTS Toolkit v. 4.6 DAYLIGHT CIS, 27401 Los Altos Suite 360 Mission Viego, CA 92691.
[22] R. Wang, Y. Fu, L. Lai, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615–621.
[23] J. Devillers, D. Domine, C. Guillon, W. Karcher, *J. Pharm. Sci.* **1998**, *87*, 1086–1090.
[24] J. Devillers, *Analusis* **1999**, *27*, 23–29.
[25] L. B. Kier, L. W Hall, *Molecular Structure Description. The Electrotopological State*, Academic Press, San Diego **1999**.
[26] J. J. Huuskonen, A. E. P Villa, I. V. Tetko, *J. Pharm. Sci.* **1999**, *88*, 229–233.
[27] K.-J. Schaper, M. Samitier, *Quant. Struct. Act. Relat.* **1997**, *16*, 224–230.
[28] M. H. Abraham, *Chem. Soc. Rev.* **1993**, *22*, 73–83.
[29] M. H. Abraham, H. Chadha, in *Lipophilicity in Drug Action and Toxicology*, V. Pliska, B.Testa (Eds.),VCH, Weinheim **1996**.
[30] P. Politzer, J. S. Murray, *Theoretical and Computational Chemistry* **1994**, *1(Quantitative Treatment of Solute/Solvent Interactions)*, 243–289.
[31] J. S. Murray, P. Politzer, G. R. Famini, *J. Mol. Struct. (Theochem.)* **1998**, *454*, 299–306.
[32] J. Platts, D. Butina, M. H. Abraham, A. Hersey, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835–845.
[33] G. Klopman, J.-Y. Li, M.Dimayuga, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752–781.
[34] Breindle, B. Beck, T. Clark, R.C. Glen, *J. Mol. Model.* **1997**, *3*, 142–145.

[35] G. D. Hawkins, D. A. Liotard, C. J. Cramer, D. G. Truhlar, *J. Org. Chem.*, **1998**, *63*, 4305–4313.
[36] C. J. Cramer, D. G. Truhlar, *Chem. Rev.* **1999**, *99*, 2161–2200.
[37] H. van de Waterbeemd, R. Mannhold, *Methods and Principles in Medicinal Chemistry Vol4, (Lipophilicity in Drug Action and Toxicology)*, V. Pliska, B. Testa, H. Van de Waterbeemd (Eds), VCH Weinheim, Cambridge **1996**, pp 401–408.
[38] R. Mannhold, R. Rekker, C. Sonntag, A. M. TerLaak, K. Dross, E. Polymeropoulis, *J. Pharm. Sci.* **1995**, *84*, 1410–1419.
[39] J. Devillers, D. Domine, *SAR QSAR Env. Res.* **1997**, *7*, 195–232.
[40] J. J. Morris, A. R. Wait, unpublished observations.
[41] I. Moriguchi, S. Hirono, I. Nakagome, H. Hirano, *Chem. Pharm. Bull.* **1994**, *42*, 976–978.
[42] A. Leo, *Chem. Pharm. Bull.* **1995**, *43*, 512–513.
[43] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
[44] P. Gaillard, P. A. Carrupt, B. Testa, A. J. Boudon, *J. Comput.-Aided Mol. Design* **1994**, *8*, 83–96.
[45] R. C. Young, R. C. Mitchell, T. H Brown, C. R. Ganellin, R. Griffiths, M. Jones, K. R. Rana, D. Saunders, I. R. Smith, N. R. Sore, T. J. Wilks, *J. Med. Chem.* **1988**, *31*, 656–671.
[46] R. H. Davies, R. D. Bagnall, W. G. M. Jones, *Int. J. Quantum Chem. Quantum Biol. Symp.* **1974**, *1*, 201–212.
[47] M. H. Abraham, J. Platts, A. Hersey, A. J. Leo, R. W. Taft, *J. Pharm. Sci.*, in press.
[48] F. Csizmadia, A. Tsantili-Kakoulidou, I. Panderi, F. Darvas, *J. Pharm. Sci.* **1997**, *86*, 865–870.
[49] A. Tsantili-Kakoulidou, I. Panderi, F. Csizmadia, F. Darvas, *J. Pharm Sci.* **1997**, *86*, 1173–1179.
[50] S. H. Yalkowsky, S. Bannerjee, *Aqueous Solubility, Methods of Estimation for Organic Compounds*, Marcel Dekker, New York **1992**.
[51] M. H. Abraham, J. Le, *J. Pharm. Sci.* **1999**, *88*, 868–880.
[52] F. Irmann, *Chem. Ing. Tech.* **1965**, *37*, 789–798.
[53] K. Wakita, M. Yoshimoto, S. Miyamoto, H. Watanabe, *Chem. Pharm. Bull.* **1986**, *34*, 4663–4681.
[54] T. Suzuky, *J. Comput.-Aided Mol. Design* **1991**, *5*, 149–166.
[55] D. Tiegs, J. Gmehling, P. Ramussen, A. Fredenslund, *Ind. Eng. Chem. Res.* **1987**, *26*, 159–161.
[56] A. Kan, M. B. Tomson, *Environ. Sci. Technol.* **1996**, *30*, 1369–1376.
[57] P. B. Myrdal, A. M. Manka, S. H. Yalkowsky, *Chemosphere* **1995**, *30*, 1619–1637.
[58] Y. C. Lee, P. B. Myrdal, S. H. Yalkowsky, *Chemosphere* **1996**, *33*, 2129–2144.
[59] Y. C. Lee, S. Pinssuwan, S. H. Yalkowsky, *Chemosphere* **1997**, *35*, 775–782.
[60] G. Klopman, S. Wang, D. M. Balthasar, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474–482.
[61] C. Hansch, J. E. Quinlan, G. L. Lawrence, *J. Org. Chem.* **1968**, *33*, 347–350.
[62] J. Hilderbrand, R. Scott, *Regular Solutions*, Prentice Hall, Englewood Cliffs, N.J. **1962**.
[63] W. M. Meylan, P. H. Howard, R. S. Boethling, *Env. Toxicol. Chem.* **1996**, *15*, 100–106.
[64] J. Nouwen, B.Hansen, *Quant. Struct. Act. Relat.* **1996**, *15*, 17–30.
[65] P. Bruneau, unpublished observations.
[66] J. J. Morris, A. R. Wait, unpublished observations.
[67] A. R. Katritzky, U. Maran, M. Karelson, V. S. Lobanov, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 913–919.
[68] R-M. Dannenfelser, S. H. Yalkowsky, *J. Pharm. Sci.* **1999**, *88*, 722–724.
[69] P. Ruelle, U. W. Kesselring, *J. Pharm. Sci.* **1998**, *87*, 987–997.
[70] P. Ruelle, U. W. Kesselring, *J. Pharm. Sci.* **1998**, *87*, 998–1014.
[71] E. Silla, I. Tunon, F. Villar, J. L Pascual-Ahuir, *J. Mol. Struct. (Theochem.)* **1992**, *254*, 369–377.
[72] N. Bodor, A. Harget, M.-J. Huang, *J. Am. Chem. Soc.* **1991**, *113*, 9480–9483.
[73] N. Bodor, M.-J. Huang, A. Harget, *Int. J. Quantum Chem. Quantum Chem. Symp.* **1992**, *26*, 853–867.
[74] N. Bodor, M.-J., Huang, *J. Pharm. Sci.* **1992**, *81*, 954–960.
[75] J. M. Sutter, P. C. Jurs, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100–107.
[76] B. E. Mitchell, P.C. Jurs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489–496.
[77] A. R. Katritzky, Y. Wang, S. Sild, T. Tamm, M. Karelson, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720–725.
[78] J. Huuskonen, M. Salo, J. Taskinen, *J. Pharm. Sci.* **1997**, *86*, 450–454.
[79] J. Huuskonnen, M. Salo, J. Taskinen, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450–456.
[80] B. Lucic, N. Trinajstic, S. Sild, M. Karlson, A. R. Katritzky, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610–621.
[81] G. Klopman, *Quant. Struct. Act. Relat.* **1992**, *11*, 176–184.
[82] G. Klopman, D. Fercu, *J. Comp. Chem.* **1994**, *15*, 1041–1050.
[83] L. P. Hammett, *Physical Organic Chemistry*, McGraw-Hill, N.Y., **1970**.

[84] M. Charton, *Prog. Phys. Org. Chem.* **1980**, *13*, 119–251.
[85] C. Hansch, A. Leo, R. W. Taft, *Chem. Rev.* **1991**, *91*, 165–195.
[86] D. D. Perrin, B. Dempsey, E. P. Serjeant, p*Ka Prediction for Organic Acids and Bases*, Chapman and Hall, London **1981**.
[87] P. J. Taylor, A. R. Wait, *J. Chem. Soc. Perk Trans 2*, **1986**, 1765–1770.
[88] E. P. Serjeant, B. Dempsey, *Ionisation constants of Organic Acids in Aqueous Solution*, Pergamon, London **1979**.
[89] D. D. Perrin, *Dissociation Constants of Organic Bases in Aqueous Solution*, Butterworths, London **1965**.
[90] D. D. Perrin, *Dissociation Constants of Organic Bases in Aqueous Solution Supplement*, Butterworths, London **1972**.
[91] C. Hansch, A. Leo, D. Hoekman, *Exploring QSAR Hydrophobic Electronic and Steric Constants*, ACS Washington, **1995**.
[92] F. Csizmadia, J. Szegezdi, F. Darvas, *Proc. 9th European Symposium on Structure-Activity Relationships: QSAR and Molecular Modelling*, C. G Wermuth (Ed.), ESCOM, Leiden **1993**, pp 507–510.
[93] S. H. Hilal, L. A. Carreira, S. W. Karrickhoff, in *Theoretical and Computational Chemistry*, P. Politzer, J. S. Murray (Eds.), Elsevier Science, **1994**, pp. 291–353.
[94] S. H. Hilal, S. W. Karickhoff, *Quant. Struct. Act. Relat.* **1995**, *14*, 348–355.
[95] S. H. Hilal, Y. El-Shabrawy, L. A. Carreira, S. W. Karickhoff, S. S. Toubar, M. Rizk, *Talanta* **1996**, *43*, 607–619.
[96] K. H. Kim, Y. C. Martin, *J. Med. Chem.* **1991**, *34*, 2056–2060.
[97] S. L. Dixon, P. Jurs, *J. Comput. Chem.* **1993**, *14*, 1460–1467.
[98] F. Hervé, S. Urien, E. Albengres, J.-C. Duché, J.-P. Tillement, *Clin. Pharmacokinet.* **1994**, *26*, 44–58.
[99] R. Kaliszan, A. Nasal, M. Turowski, *Biomedical Chromatography* **1995**, *9*, 211–215.
[100] R. Olsom, D. Christ, *Ann. Rep. Med. Chem.* **1996**, *31*, 327–336.
[101] S. Sugio, A. Kashima, S. Mochizuki, M. Noda, K. Kobayashi, *Protein Engineering* **1999**, *12*, 439–446.
[102] S. Curry, H. Mandelkow, P. Brick, N. Franks, *Nature Struct. Biol.* **1998**, *5*, 827–834.
[103] D. C. Carter, J. X. Ho, *Adv. Prot. Chem.* **1994**, *45*, 153–203.
[104] J. M. Vandenbelt, C. Hansh, C. Church, *J. Med. Chem.* **1972**, *15*, 787–789.
[105] W. Scholtan, *Arzneimittelforschung* **1968**, *18*, 505–517.
[106] C. Hansch, A. Leo, in Exploring *QSAR Fundamentals and Applications in Chemistry and Biology*, A.C.S. Washington, D.C. **1995**, pp. 226–227.
[107] F. Desmotes-Mainard, F. Péhourcq, A. Radouane, L. Labat, B. Bannwarth, *Pharm. Res.* **1995**, *12*, 1535–1538.
[108] L. Deschamps-Labat, F. Péhourcq, M. Jagou, B. Bannwarth, *J. Pharm. Biomed. Anal.* **1997**, *16*, 223–229.
[109] A. Nasal, A. Radwanska, K. Osmialowski, A. Bucinski, R. Kaliszan, G. Barker, P. Sun, R. Hartwick, *Biomedical Chromatography* **1994**, *8*, 125–129.
[110] R. Kaliszan, A. Nasal, M. Turowski, *Biomedical Chromatography* **1995**, *9*, 211–215.
[111] G. Zlotos, A. Bucker, M. Kinzig-Schippers, F. Sorgel, U. Holzgrabe, *J. Pharm. Sci.* **1998**, *87*, 215–220.
[112] A. Wait, unpublished results.
[113] S. C. Valvani, S. H. Yalkowsky, T. J. Roseman, *J. Pharm. Sci.* **1981**, *70*, 502–507.

Virtual Screening for Bioactive Molecules
Edited by Hans-Joachim Böhm and Gisbert Schneider

4 Descriptor-Based Similarity Measures for Screening Chemical Databases

John M. Barnard, Geoffrey M. Downs, Peter Willett

Excerpted with permission from *Journal of Chemical Information and Computer Sciences* 1998, *38*, 983–996.

4.1 Introduction

As this book demonstrates, there are many ways in which databases of chemical structures can be scanned to find those molecules that meet some structural criterion (or criteria). In this chapter we discuss perhaps the simplest and longest-established of the available approaches, *viz* the use of similarity searching [1–6]. Similarity searching was first developed in the mid-1980s as a complement to the facilities that were already available for chemical substructure searching. The latter involves the retrieval of all those molecules in a database that contain a user-defined query substructure, irrespective of the environment in which the query substructure occurs [7]. This has proved to be a valuable tool for accessing databases of chemical structures, especially now that three-dimensional (3-D) pharmacophore searching is available to complement the long-established facilities for two-dimensional (2-D) substructure searching [8, 9]. It does, however, have several limitations that arise from the requirement that a database structure must contain the entire query substructure if it is to be retrieved, which implies that the user who is posing a database query must, prior to the commencement of the search, already have formed a fairly clear view of the types of structure that will be retrieved. These limitations, which are discussed in detail by Willett [1], led to the development of the alternative, and complementary, similarity searching.

Similarity searching generally involves the specification of an entire query molecule, the target structure, rather than the partial structure required for substructure searching (although the target can be a substructure of another, larger molecule if desired). The target is characterized by one or more structural descriptors that are compared with the corresponding sets of descriptors for each of the molecules in the database. These comparisons enable the calculation of a measure of similarity between the target structure and each of the database structures; and the latter are then sorted into order of decreasing similarity with the target. The output from the search is a ranked list in which the structures calculated to be most similar to the target structure, the nearest neighbours, are located at the top of the list.

The principal rationale for similarity searching is given by the similar property principle, which states that structurally similar molecules are expected to exhibit similar properties or activities [2]. Thus, given an appropriate measure of inter-molecular structural similarity, the nearest neighbours of a bioactive target structure, such as one that yielded a positive result in a high-throughput screen, are also expected to exhibit the same activity. It must be empha-

sized at this point that similarity searching provides a rather crude way of screening a database, since it is appropriate when just a single bioactive molecule is available, without any evidence as to which parts of that molecule are responsible for the observed activity. Progressively more sophisticated approaches are appropriate as more structural data become available: substructure or pharmacophore searching when sufficient bioactive molecules are available to generate a query specification, and a docking search when the 3-D structure of the biological target is known. Even so, it makes obvious sense to exploit whatever information is available, and much effort has thus gone into the development of similarity searching as a precursor and complement to these more detailed types of screening mechanism. As we shall see, there are many ways in which the similarity between a pair of molecules can be calculated. Here we focus on descriptor-based methods, where each molecule is characterized by a list of attribute-values, such as a set of physicochemical property values or a set of binary (present/absent) fragment substructures, as discussed in the next Section.

4.2 Fragment-Based Similarity Searching

The first two reports of similarity searching appeared in the mid-1980s, based on work carried out at Lederle Laboratories [10] and at Pfizer [11]. The starting points for these two, near-contemporaneous studies were very different, but both groups of workers realized that counts of the numbers of fragment substructures common to a pair of molecules provided a computationally efficient, and surprisingly effective, basis for quantifying the degree of structural resemblance between the two molecules under consideration.

The Lederle study was carried out as part of a project to develop simple, robust techniques for the prediction of biological activity that would not suffer from the sample-to-feature problems which affect many types of high-dimensionality descriptors [12]. Molecules were represented by their constituent atom pairs, where an atom pair is a substructural fragment comprising two non-hydrogen atoms together with the number of intervening bonds (see Section 4.4 below). These characterizations were used for two applications: for similarity searching, with the set of atom-pairs describing a user-defined target structure being matched against the corresponding sets for each of the database structures; and substructural analysis, where weights were calculated that relate the presence of a specific substructural moiety in a molecule to the probability that the molecule is active in some biological test system [13]. The similarity search allowed users to request either some number of the top-ranked molecules or all those that had a similarity with the target structure greater than a minimal value. The latter search option does require the user to have at least some feeling for the magnitude of the values resulting from the chosen similarity measure, but it serves to restrict the output to those molecules that do have a significant level of resemblance to the target structure.

The work at Pfizer started out as a way of prioritizing the outputs of 2-D substructure searches from their in-house chemical information system. A user of the Pfizer system would submit not only a conventional substructural query, but also a target molecule typical of the sorts of structure that were required. A conventional screen search and atom-by-atom search [7] were used to identify the matches to the query substructure, and then a similarity measure based on the screens common to the target and to each of these matches was used to rank the

substructure search output in order of decreasing similarity with the query. Specifically, the similarities were calculated using the Tanimoto coefficient discussed in the next Section of this Chapter. At least in part, the initial substructural query was used to minimize the elapsed time required for the calculation of the similarities, by restricting the similarity calculation to just that small fraction of the database not eliminated by the substructure search. The subsequent development of a much faster nearest neighbour search algorithm, based on an inverted file, allowed the ranking of an entire database against the target structure in real time, without the need for the specification of the initial substructural query.

Interactive, fragment-based similarity searching has proved to be extremely popular, both for property prediction purposes (as in the work at Lederle) and for allowing end-users to pose "give me ten more like this" queries (as in the work at Pfizer), and it is now a standard retrieval mechanism in nearly all operational systems for chemical information management. Developments since the Lederle and Pfizer systems were first reported have involved both enhancements of fragment-based searching and the use of different types of similarity measure.

An example of an enhanced fragment-based system is provide by Hagadone's work on substructure similarity (or subsimilarity) searching [14]. Conventional similarity searching is appropriate when the need is to identify complete structures that are similar to the target structure. Such a global similarity search [4], *i.e.* one in which the entire matching structures are involved in the similarity calculation, is far less effective when the need is to identify molecules containing a substructure similar to a target structure or target substructure. This is an example of a local similarity search [4], *i.e.* one in which account must be taken of parts of the molecules being compared and in which a more detailed similarity calculation is required. In subsimilarity searching, a simple fragment-based similarity search is used to calculate an upperbound to the size (in terms of the numbers of constituent atoms or bonds) of the maximal common substructure (or MCS) between the target (sub)structure and each database structure. These upperbounds are then used to prioritize database structures for an MCS search that uses a rapid, but approximate, maximal common subgraph isomorphism algorithm.

Another example of a similarity search system using fragment occurrence information in combination with a second-level search is described by Fisanick *et al.* [15–17] as part of a project to develop facilities for similarity searching in the Chemical Abstracts Service (CAS) Registry File, using 2-D, 3-D and molecular property data. The 2-D studies involved subsets of the substructural fragments that comprise the CAS Online screen dictionary [18], focusing on the different types of similarity relationships that can be identified between a target structure and a database structure when different classes of substructural fragment are employed. For example, the selection of *augmented atoms* (an atom and its pendant atoms and bonds) and *atom sequences* (unbranched chains of atoms) gives a very different view of the structural resemblances between a pair of molecules from that provided by the selection of ring composition fragments (the atoms within a ring and the bonds between them). This suggests that further analysis into mixed descriptor types could give users an even more flexible approach to similarity searching, perhaps using the data fusion techniques discussed in Section 4.5. Another part of the CAS work includes a second-stage search based on *reduced-graphs* [19–21], which, unlike substructural fragments, retain some of the topological relationships between areas of a molecule and which are thus capable of providing a local measure of similarity.

Both of the studies above thus involve the combination of a global, fragment-based, similarity algorithm with a more sophisticated, local, graph algorithm that allows some degree of substructural matching. Ways of combining substructural constraints in a global similarity measure are discussed by Willett [22] and by Grethe and Hounshell [23].

Further developments have focused on the similarity measure used to quantify the degree of structural resemblance between the target structure and each of the structures in the database to be searched. Many different types of similarity measure have been discussed in the literature [1–6] but they generally involve three principal components: the *representation* that is used to characterize the molecules being compared, the *weighting scheme* that is used to assign differing degrees of importance to the various components of these representations, and the *similarity coefficient* that is used to provide a quantitative measure of the degree of structural relatedness between a pair of structural representations. While there has been some interest in the extent to which the weighting scheme affects the utility of a similarity measure [14, 24, 25], there is a much more extended literature relating to the other two components. These topics are reviewed in the next two Sections of this Chapter which, while continuing to focus on fragment-based measures of similarity, also cover other types of descriptor-based representation.

4.3 Association and Distance Coefficients for Similarity Searching

The idea of determining a numerical measure of the similarity (or conversely, the distance) between two objects, each characterized by a common set of attributes, is common to a wide range of disciplines, including biology, psychology and bibliographic information retrieval. Because of the diversity of these application areas, and the lack of communication between them, there has been a great deal of duplication of effort, and commonly used similarity coefficients have been reinvented a number of times. This partly accounts for the variety of different names applied to some of these coefficients. This section reviews those coefficients that have found widespread use in chemical information systems; more comprehensive surveys of the very many coefficients available are provided by Hubálek [26], Gower [27] and Ellis *et al.* [28], *inter alia*.

An object A can be described by means of a vector X_A of n attributes such that:

$$X_A = \{ x_{1A}, x_{2A}, x_{3A}, \ldots x_{jA}, \ldots x_{nA} \}, \tag{4.1}$$

where x_{jA} is the value of the of the j^{th} attribute of object A, as detailed in Table 4.1 (which provides a complete list of the symbols used in this chapter). The values of the attributes may be real numbers over any range (and may involve some weighting factor applied to the basic property value involved), or they may be confined to dichotomous (*i.e.* binary) values, indicating the absence (0) or presence (1) of some particular feature of the object. In the case of a molecular object, the attributes might be a set of n topological indexes or calculated physico-chemical properties, or the on/off state of each of the n bits in the fingerprint representing the molecule.

Table 4.1. Symbols used in Table 4.2 and equations in Chapter 4.

i, j	Attributes
A, B	Objects (or molecules)
n	Total number of attributes of an object (e.g. bits in a fingerprint)
X_A	Attribute vector describing object A
x_{jA}	Value of j^{th} attribute in object A
χ_A	Set of "on" bits in binary vector X_A
a	Number of bits "on" in molecule A
b	Number of bits "on" in molecule B
c	Number of bits "on" in both molecules A and B
d	Number of bits "off" in both molecules A and B
$S_{A,B}$	Similarity between objects A and B
$D_{A,B}$	Distance between objects A and B

Some coefficients are measures of the distance, or dissimilarity between objects (and have a value of 0 for identical objects), while others measure similarity directly (and have their maximum value for identical objects). In most cases the values that can be taken by a coefficient lie in the range 0–1, or can be normalized to that range. This is typically effected by means of a function based on the values of the attributes for the two objects that are being compared, with the resulting coefficients being referred to as association coefficients. The zero-to-unity range provides a simple means for converting between a similarity coefficient and a complementary distance coefficient, namely subtraction from unity. In some cases a similarity coefficient and its complement have been developed independently and are known by different names, *e.g.* the Soergel distance coefficient is the complement of the Tanimoto (or Jaccard) association coefficient.

Distance coefficients are analogous to distances in multi-dimensional geometric space, though they are not necessarily precisely equivalent to such distances. For a distance coefficient to be described as metric, it must have the following properties:

1. Distance values must be zero or positive, and the distance from an object to itself must be zero.

$$D_{A,B} \geq 0, \quad D_{A,A} = D_{B,B} = 0 \tag{4.2}$$

2. Distance values must be symmetric.

$$D_{A,B} = D_{B,A} \tag{4.3}$$

3. Distance values must obey the triangular inequality.

$$D_{A,B} \leq D_{A,C} + D_{C,B} \tag{4.4}$$

4. The distance between non-identical objects must be greater than zero.

$$A \neq B \quad \Leftrightarrow \quad D_{A,B} > 0 \tag{4.5}$$

A distance coefficient which has only the first three of these properties is called pseudo-metric, and one which does not have the third property is non-metric. Though a particular distance coefficient may have all four properties, this is not sufficient to imply that the distances involved can be embedded in a Euclidean space of any given dimensionality (*e.g.* n, the number of properties). Certain other properties are necessary, and even then the dimensionality of the space required may be much larger than n. The requirements for Euclidean embedding are discussed by Gower [27].

Though a large number of similarity and distance coefficients have been defined (and often redefined by different authors), many of them are closely related to each other. In some cases, the same coefficient can be obtained by different routes; in other cases coefficients which are different when calculated for continuous attributes become equivalent when applied to binary attributes. Certain coefficients are described as being monotonic with each other, which means that it can be shown analytically that they will always produce identical similarity rankings of objects against a specified target, even though the actual coefficient values are different. Even though two coefficients may not be completely monotonic, the values resulting from their use may well exhibit a high degree of correlation, as demonstrated by Holliday *et al.* in a comparison of the Cosine and Tanimoto Coefficients [29]. Some pairs of coefficients, conversely, exhibit very low correlations, suggesting that they are reflecting very different characteristics of the objects that are being compared [28]. An extended empirical study of the monotonicity relationships existing between no less than 43 different coefficients is reported by Hubálek [26].

Where the attribute values are restricted to 0 and 1, the expressions used for the various similarity and distance measures can often be substantially simplified. In this context a number of useful symbols can be defined. For objects A and B characterized by vectors X_A and X_B containing n binary values (such as fingerprints) we can write:

$$a = \sum_{j=1}^{j=n} x_{jA} \qquad \text{(number of bits "on" in } A\text{)} \tag{4.6}$$

$$b = \sum_{j=1}^{j=n} x_{jB} \qquad \text{(number of bits "on" in } B\text{)} \tag{4.7}$$

$$c = \sum_{j=1}^{j=n} x_{jA} x_{jB} \qquad \text{(number of bits "on" in both } A \text{ and } B\text{)} \tag{4.8}$$

$$d = \sum_{j=1}^{j=n} (1 - x_{jA} - x_{jB} + x_{jA} x_{jB} \qquad \text{(number of bits "off" in both } A \text{ and } B\text{)} \tag{4.9}$$

and hence:

$$n = a + b - c + d \tag{4.10}$$

Note that the definitions of a and b shown here differ from those given by Gower [27] and by Ellis *et al.* [28]; they are, however, the definitions that have been more commonly used in the chemical information literature. The various quantities above can also be expressed in set-theoretic notation, if we define χ_A as the set of all elements x_{jA} in vector X_A whose value is 1 (the "on" bits), and χ_B as the set of all elements x_{jB} in vector X_B whose value is 1. Then:

$$a = |\chi_A| \tag{4.11}$$

$$b = |\chi_B| \tag{4.12}$$

$$c = |\chi_A \cap \chi_B| \tag{4.13}$$

$$d = n - |\chi_A \cup \chi_B| \tag{4.14}$$

and, as a corollary to the above, the number of bits "on" in at least one of the molecules is given by:

$$a + b - c = |\chi_A \cup \chi_B| \tag{4.15}$$

Given the above definitions, Table 4.2 describes a number of similarity and distance coefficients commonly used in chemical information, with expressions for calculating them for continuous-variable or dichotomous attributes, or using set notation.

Table 4.2. Descriptions of some distance metrics and similarity coefficients commonly used in chemical information. Definitions of the symbols used are shown in Table 4.1. Note that the negative lowerbound values for the three association coefficients apply only if negative attribute values are possible.

Hamming Distance	
Other names:	• Manhattan Distance • City–Block Distance • Normalized complement for dichotomous data called Simple Matching Coefficient
Formula for continuous variables:	$D_{A,B} = \sum_{j=1}^{j=n} \lvert x_{jA} - x_{jB} \rvert$
Formula for dichotomous variables:	$D_{A,B} = a + b - 2c$
Set-theoretic definition:	$D_{A,B} = \lvert \chi_A \cup \chi_B \rvert - \lvert \chi_A \cap \chi_B \rvert$
Range:	• ∞–0 (continuous), n–0 (dichotomous)
Metric properties:	• Obeys all four metric properties
Notes:	• Equivalent to the squared Euclidean Distance for dichotomous variables • Can be normalized to the range 1–0 if the values of all attributes are normalized to this range and the result divided by n

Table 4.2. (continue)

Euclidean Distance	
Other names:	
Formula for continuous variables:	$D_{A,B} = \sqrt{\sum_{j=1}^{j=n} \left(x_{jA} - x_{jB}\right)^2}$
Formula for dichotomous variables:	$D_{A,B} = \sqrt{a + b - 2c}$
Set-theoretic definition:	$D_{A,B} = \sqrt{\lvert \chi_A \cup \chi_B \rvert - \lvert \chi_A \cap \chi_B \rvert}$
Range:	• ∞–0 (continuous), n–0 (dichotomous)
Metric properties:	• Obeys all four metric properties
Notes:	• Frequently used as its square (with which it is of course, monotonic) which avoids the need to take the square root in the calculation • Monotonic with the Hamming Distance in all cases (and its square is equivalent to the Hamming Distance for dichotomous variables) • Can be normalized to the range 1–0 if the values of all attributes are normalized to this range and the result divided by n

Soergel Distance	
Other names:	
Formula for continuous variables:	$D_{A,B} = \dfrac{\sum_{j=1}^{j=n} \lvert x_{jA} - x_{jB} \rvert}{\sum_{j=1}^{j=n} \max\left(x_{jA}, x_{jB}\right)}$
Formula for dichotomous variables:	$D_{A,B} = 1 - \dfrac{c}{a + b - c} = \dfrac{a + b - 2c}{a + b - c}$
Set-theoretic definition:	$D_{A,B} = \dfrac{\lvert \chi_A \cup \chi_B \rvert - \lvert \chi_A \cap \chi_B \rvert}{\lvert \chi_A \cup \chi_B \rvert}$
Range:	• 1–0
Metric properties:	• Obeys all four metric properties provided all attributes have non-negative values
Notes:	• For dichotomous variables only, the Soergel distance is identical to the complement of the Tanimoto Coefficient

Table 4.2. (continue)

Tanimoto Coefficient	
Other names:	• Jaccard Coefficient
Formula for continuous variables:	$S_{A,B} = \frac{\sum_{j=1}^{j=n} x_{jA} x_{jB}}{\sum_{j=1}^{j=n} \left(x_{jA}\right)^2 + \sum_{j=1}^{j=n} \left(x_{jB}\right)^2 - \sum_{j=1}^{j=n} x_{jA} x_{jB}}$
Formula for dichotomous variables:	$S_{A,B} = \frac{c}{a+b-c}$
Set-theoretic definition:	$S_{A,B} = \frac{\lvert \chi_A \cap \chi_B \rvert}{\lvert \chi_A \cup \chi_B \rvert}$
Range:	• –0.333–1 (continuous), 0–1 (dichotomous)
Metric properties:	• Complement does not obey the triangular inequality in general, though does obey it if dichotomous variables are used
Notes:	• Monotonic with the Dice Coefficient • Complement of the dichotomous version is identical to the Soergel Distance

Dice Coefficient	
Other names:	• Czekanowski Coefficient • Sørenson Coefficient • Essentially equivalent to the Hodgkin Index for overlap of electron density functions
Formula for continuous variables:	$S_{A,B} = \frac{2\sum_{j=1}^{j=n} x_{jA} x_{jB}}{\sum_{j=1}^{j=n} \left(x_{jA}\right)^2 + \sum_{j=1}^{j=n} \left(x_{jB}\right)^2}$
Formula for dichotomous variables:	$S_{A,B} = \frac{2c}{a+b}$
Set-theoretic definition:	$S_{A,B} = \frac{2\lvert \chi_A \cap \chi_B \rvert}{\lvert \chi_A \rvert + \lvert \chi_B \rvert}$
Range:	• –1–1 (continuous), 0–1 (dichotomous)
Metric properties:	• Complement does not obey the triangular inequality
Notes:	• Monotonic with the Tanimoto coefficient

Table 4.2. (continue)

Cosine Coefficient	
Other names:	• Ochiai coefficient • Essentially equivalent to the Carbo Index for overlap of electron density functions
Formula for continuous variables:	$S_{A,B} = \frac{\sum_{j=1}^{j=n} x_{jA} x_{jB}}{\sqrt{\sum_{j=1}^{j=n} (x_{jA})^2 \cdot \sum_{j=1}^{j=n} (x_{jB})^2}}$
Formula for dichotomous variables:	$S_{A,B} = \frac{c}{\sqrt{a.b}}$
Set-theoretic definition:	$S_{A,B} = \frac{\lvert \chi_A \cap \chi_B \rvert}{\sqrt{\lvert \chi_A \rvert \cdot \lvert \chi_B \rvert}}$
Range:	• –1–1 (continuous), 0–1 (dichotomous)
Metric properties:	• Complement does not obey the triangular inequality
Notes:	• Highly correlated with the Tanimoto coefficient, though not strictly monotonic with it

Both the Hamming Distance and the Euclidean Distance are examples of a more general class of distance metrics called Minkowski Distances which are given by the general formula:

$$D_{A,B} = \sqrt[t]{\sum_{j=1}^{j=n} \left(\left| x_{jA} - x_{jB} \right| \right)^t} \tag{4.16}$$

where $t = 1$ for the Hamming Distance and $t = 2$ for the Euclidean Distance.

A fundamental difference between the Hamming and Euclidean Distances on the one hand, and the Tanimoto, Dice and Cosine Coefficients on the other, is that the former effectively consider a common absence of attributes (or common low values in the case of continuous variables) as evidence of similarity, whereas the latter do not. This is a basic philosophical argument, which has been much discussed in the literature. In the context of numerical taxonomy, Sokal and Sneath [30] have commented: "*The absence of wings ... among a group of distantly related organisms (such as a camel, a horse and a nematode) would surely be an absurd indication of affinity. Yet a positive character such as the presence of wings ... could mislead equally ... for a similar heterogeneous assemblage (for example, bat, heron and dragonfly).*"

In the chemical context, James *et al.* [31] have suggested that Hamming and Euclidean distances are useful only for "relative" distance comparisons (*i.e.* the distance of two molecules to the same target) but not for "absolute" comparisons (between two independent pairs of molecules), for which they prefer the Tanimoto coefficient. Nevertheless, Euclidean Distance comparisons form the basis of Ward's hierarchical agglomerative clustering method [32],

which has been shown to be particularly effective for the clustering of databases of chemical structures [1, 33, 34]. It is also worth noting that a number of familiar chemical concepts are essentially negatively defined; for example, the common feature of carbocycles is the lack of heteroatoms and the common feature of aliphatic systems is the lack of aromaticity.

A further fundamental difference is that the association coefficients involve a normalization factor that helps to lessen molecular size effects in some cases. Thus, in a similarity search using fragment bit-strings or fingerprints, a large molecule in the database is *a priori* much more likely to have bits in common with the target structure than is a small molecule, and it is thus appropriate to include some degree of size normalization in the coefficient to avoid a bias in the nearest neighbours towards the largest database molecules. The converse of this problem can arise in diversity applications, for example in dissimilarity-based compound selection procedures where one seeks to identify database subsets for which the constituent molecules are as dissimilar as possible [35]. Small molecules are likely to have few bits set in a fingerprint: since the Tanimoto Coefficient, for example, does not take account of common absence of features and since $c \leq \min(a,b)$ (in the coefficient's expression in Table 4.2), low similarity (and thus high dissimilarity) values will be obtained with small molecules, thus possibly biasing the size distribution in the final subset that is selected. A solution adopted at Pharmacia and Upjohn [36] is to use a composite coefficient essentially involving both the Tanimoto Coefficient and the Simple Matching Coefficient (the complement of the normalized Hamming Distance). Further discussions of the relationship between molecular size and the magnitude of fragment-based similarity values are provided by Flower [37] and by Schuffenhauer *et al.* [38].

Following earlier work by Adamson and Bush [39], Willett and Winterman [40] compared the performance of a range of association and distance coefficients by the extent to which they obeyed the similar property principle [2]. Specifically, they assessed the effectiveness of a coefficient by the extent to which it was able to predict correctly a compound's measured property or activity value as the value of the most similar compound in the same dataset. In this study, the Tanimoto and Cosine Coefficients performed rather better than the Hamming and Euclidean Distance measures, and in an operational system implemented subsequently [11], the Tanimoto Coefficient was preferred, partly on the basis of a subjective evaluation of the similarity search rankings it produced, and partly because its calculation does not involve a square root, making it faster. Since then, the Tanimoto Coefficient has been the measure of choice for fragment-based chemical similarity work, though the Hamming Distance (equivalent to the squared Euclidean Distance for binary data) retains its adherents and the Euclidean Distance remains the most popular measure for continuous data, such as molecules characterized by sets of physicochemical properties.

Other criteria can be used to evaluate similarity coefficients. For example, Cheng *et al.* [41] have described four association coefficients for assessing the degree of relatedness between pairs of different similarity coefficients. Their study, which again draws upon the similar property principle, was used to compare different coefficients based on different descriptor sets (Euclidean Distances with topological indexes, and Tanimoto Coefficient with 2-D bit-strings) but the same principles could also be applied to coefficients based on the same descriptors. Computational efficiency can also merit consideration as a basis for comparison. For example, the Cosine Coefficient allows the calculation of the average similarity between all pairs of compounds in two disjointed datasets extremely rapidly, something that is not

possible with the Tanimoto Coefficient [29], but may be necessary for some similarity applications. Finally, the behavior of a coefficient over its range of possible values may give guidance as to its suitability for use in a particular application domain, as evidenced by the continuing discussion as to which similarity coefficient is most appropriate for the calculation of field-based similarities [25, 42–45].

Bradshaw [46] has drawn attention to the use of asymmetric similarity coefficients (in which $S_{A,B} \neq S_{B,A}$) based on the ideas of Tversky [47]. The general form for Tversky similarity is defined for binary data as follows:

$$S_{A,B} = \frac{c}{\alpha(a-c) + \beta(b-c) + c} \tag{4.17}$$

where α and β are user-defined constants. If α and β are equal, the resulting similarity coefficient is symmetric, and in the case of certain values, the expression reduces to one of the commonly known coefficients: the Tanimoto Coefficient when $\alpha = \beta = 1$, and the Dice Coefficient when $\alpha = \beta = {}^1/_2$. If α and β are different, the resulting coefficient is asymmetric, and when $\alpha = 1$ and $\beta = 0$ this yields $S_{A,B} = {}^c/_a$, which can be interpreted as the "fraction of A" which it has in common with B; the coefficient will become equal to 1 when all the features of A are also in B – i.e. when A is (within the constraints of a fragment-based representation) a substructure of B, and features of B which do not occur in A are irrelevant to the similarity value. This type of subsimilarity expression has also been derived by Maggiora *et al.* [48] using an approach based in fuzzy-set theory, and provides an alternative subsimilarity measure to the MCS procedure described by Hagadone [14] and discussed in Section 4.2 above.

In conclusion, we reiterate the fact that the discussion here has concentrated on those coefficients, and their close relations, that have been most extensively used, thus far, for chemical applications. There are many others that have been discussed in the literature of, *e.g.* multivariate statistics [27], information retrieval [28], and numerical taxonomy [30], and it must not be assumed that there is any single "best" coefficient even if we restrict attention to the domain of chemical structure handling. Indeed, as noted by Jones and Curtice [49] in a discussion of the association between indexing terms in information retrieval systems: "*What is annoying is that no clear-cut criterion for choice among the alternatives has emerged. As a result, few candidate measures have been permanently dismissed from consideration, and a rather large set of formulae remains available.*" There is hence a continuing need for both empirical and analytical comparisons of the available coefficients to ensure that the most appropriate one(s) are employed in any specific similarity-based system.

4.4 Structural Representations for Similarity Searching

Similarity-based screening of large chemical databases needs representations of the molecules that are both *effective, i.e.* can differentiate between molecules that are different, and *efficient, i.e.* quick to calculate in operation. There is a general conflict between these two requirements in that the most effective methods of representation tend to be the least efficient to calculate, and *vice versa*, and so a suitable compromize needs to be made. Here we focus on the descriptors based on 2-D and 3-D substructural fragments or properties that provide

an appropriate balance between these conflicting requirements, and are thus currently most commonly used for similarity searching. This overview is necessarily brief and further details are given in the recent review by Brown [50].

The representation of molecules by descriptors involves the generation of suitable descriptors, and if desired, the selection of a subset of them, and then the encoding of the chosen descriptors in a form that will enable similarity calculation between pairs of representations. Many of the descriptors are described in the literature along with a particular encoding method; however, the two are largely independent, and it is usually possible to encode a given descriptor in a variety of ways. We have thus deliberately separated descriptor selection and descriptor encoding in this Section to highlight the two stages.

4.4.1 Descriptor Selection

There is an infinite variety of potential descriptors, and so descriptor selection is necessary as an exercise in data reduction to select those most appropriate to a given application. The following subsections will examine examples of counts, 2-D fragment, 3-D fragment, physicochemical property descriptors, topological indices, whole molecule comparisons, and the issue of descriptor choice.

The simplest descriptors are counts of individual atoms, bonds, degrees of connectivity *etc.* These can be extended to counts of rings, pharmacophore points and any other feature that can be represented as a single node or arc in the graph or reduced-graph representation of the molecule.

2-D fragment descriptors were first studied in detail by Lynch and his co-workers (see, *e.g.* [51]), who investigated the use of various types of atom-centred, bond-centred, and ring-centred fragments for substructure searching. This work led to the widespread adoption of augmented atom, atom sequence and ring fragments in substructure search systems, *e.g.* the fragments in CAS Online [18]. Some typical 2-D fragment definitions are shown in Figure 4.1. Augmented atoms comprise a central atom with the neighbouring attached atoms and intervening bonds. Atom sequences are linear sequences of a given number of connected atoms, with their intervening bonds. Ring fragments can be of several different types, for instance the ring sequence (atom sequence round a ring), and ring fusion sequence (ring-connectivity counts round a ring) fragments. Other fragment definitions, originally developed at Lederle Laboratories, that have become popular for similarity searching are the atom pair [10] and topological torsion [52] fragments. Atom pairs comprise a pair of atom types and the intervening distance between them, in terms of the shortest bond-by-bond path between them. The atom type describes the elemental type, the number of non-hydrogen attachments, and the number of π-bonds. The topological torsion fragment comprises a linear sequence of four connected atoms, with each atom type described in the same way as for atom pairs.

Workers at CAS found that the use of specific elemental and bond types for atoms and bonds can be too specific for substructure searching, and generalized forms of these fragments are thus often used. For instance, atoms can be generalized to groups such as their group in the periodic table, and bonds to ring or chain, allowing the specification of many combinations of generalized atoms and bonds. Such generalized forms can also be used for

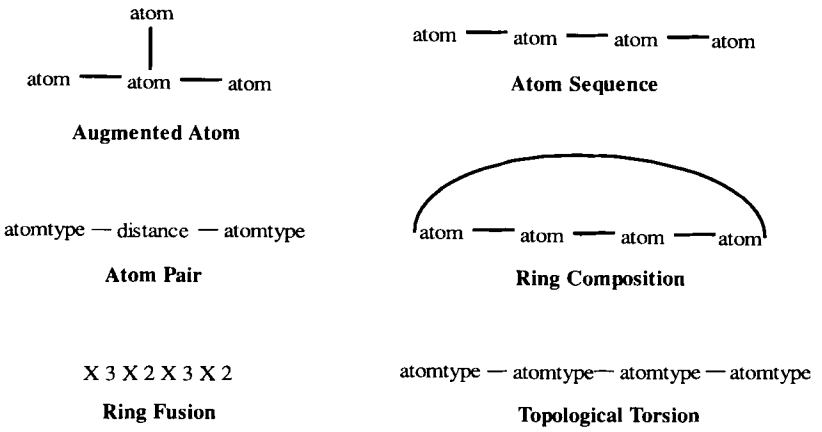


Figure 4.1. Examples of 2-D fragment descriptors.

similarity searching. The generalization of atom pairs and topological torsions by use of physicochemical atom types is mentioned later.

An alternative to the algorithmically generated descriptors described above is to define and search for particular functional groups, which may be expressed in specific or general terms. Once defined, the functional groups can be detected by scanning the connection tables for instances of them. A more efficient way is to use string-searching of a linear representation of the molecule, for instance by defining the functional groups in terms of SMARTS strings and using them to search SMILES representations [53].

As noted in Section 4.2, 2-D fragment descriptors rapidly established themselves as the basis for operational similarity searching, and it was some years before attempts were made to develop fragments for 3-D similarity searching. Some examples of 3-D fragment descriptors are shown in Figure 4.2. Many of the 2-D fragments that can be generated from a 2-D connection table have equivalents in the 3-D fragments that can be generated from a 3-D connection table. However, there is much less consensus in the 3-D area as to which are the best descriptors to use. Moreover, the variety of available descriptors is greater and new fragment types continue to be developed. Due to the fully-connected nature of a 3-D connection table, and the flexibility of 3-D structures, the number of 3-D fragments generated for a given class can be much larger than for the 2-D equivalent, and the generation process can be more time-consuming.

Willett and his co-workers have described both distance-based [54] and angle-based [55] descriptors for the calculation of 3-D similarity. The simplest of the distance-based descriptors is the distance distribution in which each distance in a molecule increments a count in an associated distance-range bin. The resultant frequency distribution of distances is used to represent the molecule. To include elemental types, individual-distance descriptors comprise a pair of atoms and the inter-atomic distance between them. The angle-based descriptors were based on generalized valence angles and torsion angles, in which the atoms comprising the angle do not need to be directly bonded to each other. The distance-based descriptor described by Bemis and Kuntz [56] is an extension of the distance-distribution descriptor, and

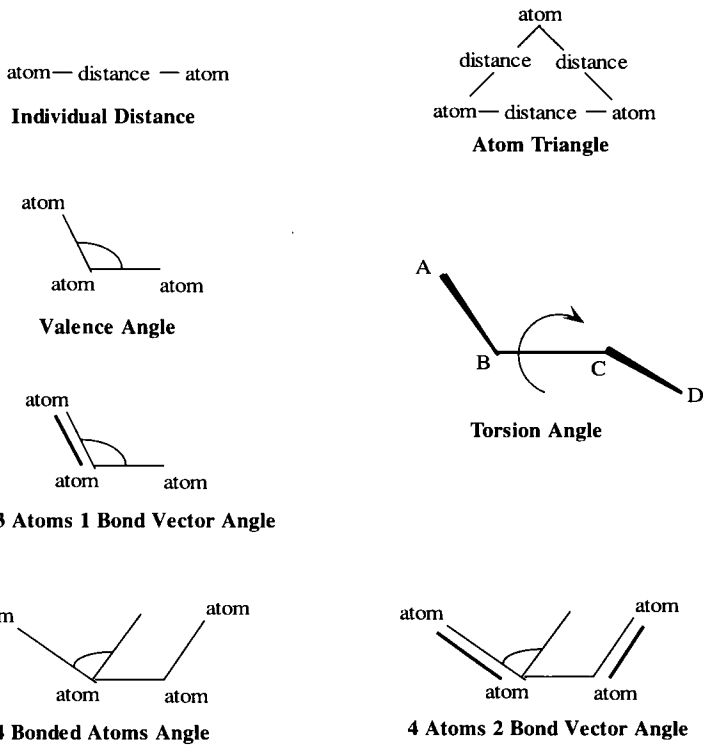


Figure 4.2. Examples of 3-D fragment descriptors.

uses the distances between triplets of atoms. For each triplet in a molecule, the three interatomic distances are squared and summed to give a single value, and the distribution of these values is used to describe the molecule. Closely related to this is the atom triplet descriptor of Nilakantan *et al.* [57]. Here, the three distances between the three atoms comprising the triplet are sorted into increasing length; the first is left alone, the second is multiplied by 10^3, the third by 10^6, and then they are summed to produce a single integer value. A further representation of a molecule that is based on a distance distribution is described by Bauknecht *et al.* [58], drawing on a previous study of the use of autocorrelation vectors for similarity calculations [59]. Here, the vector encodes information about the separations of pairs of atoms with particular electronic characteristics (although the method as described would appear to allow the use of any sort of atom-centred property for the calculation of a vector).

The group at Merck have described 3-D variants of the atom pair, referred to as geometric atom pairs and geometric binding property pairs [60]. In geometric atom pairs, the atom types are as defined for standard atom pairs, but the distance between them is the through-space distance rather than the through-bond distance. In geometric binding property pairs, the distance is through-space distance and the atom type is generalized to one of seven binding classes (cation, anion, H-bond donor, H-bond acceptor, polar, hydrophobic, and other). Similarly, the group at Abbott Laboratories has compared a wide variety of descriptors [34]

including two in-house descriptors based on Potential Pharmacophore Points (PPPs). Five points are defined: H-bond donor, H-bond acceptor, positively-charged, negatively-charged, and hydrophobic. All atoms of the molecule are analyzed to see whether they can be classed potentially as one of the point types. The descriptors are PPP-pairs and PPP-triangles. PPP-pairs are similar to geometric atom pairs, with the atom types represented by PPP types. PPP-triangles are triplets of PPPs and their associated distances (categorized into bin ranges).

Eight classes of 3-D descriptor have been investigated at CAS [15]. The atom pair distance and 3-bonded atoms angle are generalized distance-distribution and valence angle descriptors, respectively (with atom types Carbon, Hetero and Any). The 3 atoms and 1 bond vector angle descriptor is a hybrid atom triangle, and the 4-bonded atoms and 4 atoms and 2 bond vectors angle are hybrid topological torsion descriptors, each with selected angle information added to the generalized atom type information. The "atom triangles" are atom triangles with generalized atom types, and the atom triangle 3-slot, and atom triangle 5-slot are reduced and generalized atom triangle descriptors using two atoms, and one or three of the interatomic distances, respectively. The detailed search results provided by Fisanick *et al.* [15] demonstrate that these triangle-based features provide a simple and effective mechanism for similarity searching based on size and shape.

The CAS workers have also investigated the use of calculated molecular properties [15]. Twenty whole-molecule properties were tested, such as *C*log*P*, molar refractivity, ionization potential, HOMO and LUMO. The resultant values can be used directly as descriptors. In addition, several localized properties were included, such as atomic electron densities and eigenvalues for molecular orbitals, with the resultant values being binned into ranges to provide sets of descriptors for the whole molecule. A subset of the global properties subsequently formed the basis for the similarity measures used in a comparison of various clustering methods [33]. Similar work has been reported by Kearsley *et al.* [61], who generalized the atom pair and topological torsion descriptors by replacing the atom types by physicochemical properties [61]. Binding property pairs and binding property torsions have the atom types replaced by one of seven binding property groups (cations, anions, H-donor, H-acceptor, polar, hydrophobic, and other). Hydrophobic pairs and torsions, and charge pairs and torsions have the continuous values split into seven overlapping bins. Unlike most descriptors, the charge pairs and torsions consider hydrogen atoms as atoms.

Topological indices are similar to physicochemical properties in that they characterize some aspect of molecular data by a single value. Very many different topological indices have been, and continue to be, described in the Quantitative Structure–Activity Relationship (QSAR) literature, but most are highly correlated. An early example of the use of topological indices for similarity calculation in large databases is provided by Basak *et al.* [62], who generated 90 topological indices, encoding shape, size, bonding pattern and branching pattern. Principal components analysis of these indices identified ten principal components that were then used as the descriptors for the similarity analysis. Topological indices are often used in conjunction with other descriptors, as exemplified by work at Rhone-Poulenc Rorer [63]. This study of 49 molecular properties (of which just over half were topological indices, and a quarter were counts) identified six descriptors that were relatively uncorrelated and covered steric, electronic and hydrophobic aspects. The selected descriptors comprised three topological indices (flexibility, normalized electrotopological, and aromatic density), two fragment-based properties (H-donor and H-acceptor), and one physicochemical property (*C*log*P*).

Given the wide variety of descriptor types available, it is necessary to select the most appropriate structural representation for a given application. A recent, detailed comparative study is reported by Brown and Martin [64], who have analyzed various descriptors (and encoding methods) to find those most relevant to ligand-receptor binding. 2-D structural descriptors contain a lot of information about the physical properties and reactivity of a molecule, and are quick to calculate. Augmented atoms are very localized, atom sequences are less so, and atom pairs span the whole of a 2-D structure. Ring descriptors are essential for cyclic structures, and topological torsions correlate well with 3-D torsions except in highly folded molecules. Physical properties and topological indices can be useful representations of hydrophobic and electrostatic interactions. 3-D shape descriptors can give useful information about dispersion and steric interactions. The larger 3-D descriptors tend to be the most effective, but the flexibility of many molecules can increase the time required to generate 3-D descriptors, and/or decrease their effectiveness. Generalization of atom types from specific elements to groups or properties can help to reveal broader similarities. Overall, the trend is to use combined descriptors or descriptor sets which contain many different descriptor types, at many different levels of generalization.

4.4.2 Descriptor Encoding

Having discussed some of the types of descriptor that are available, we now describe how they can be encoded to enable similarity calculations to be carried out.

The representation that is overwhelmingly used as a basis for similarity calculations in large databases is the fixed-length bit-string. This contains a fixed number of bits in which each bit can represent the absence (0) or presence (1) of some feature, either on its own or in conjunction with other bits in the bit-string. The binary bit-string is usually used for 2-D and 3-D fragment descriptors. Discrete variables, with more than two values, can be represented in the binary bit-string by using a bit for each possible value, or for given ranges of values. Continuous variables can be represented by defining ranges of values and then assigning a bit to each range, a process known as binning. The ranges covered by each bin can be separate or overlapping, as is done, *e.g.* in the geometric atom pair descriptors of Sheridan *et al.* [60]. The ranges can also be equidistant, equifrequent, or user/application defined. Equidistant ranges, as the name implies, have the same interval. Equifrequent ranges have different intervals, each interval being derived from examination of the frequency distribution of the descriptor being represented, or an equation of the distribution. For specialized applications, where distinct peaks in the distribution are known, the user may define the bin ranges manually.

Bit-strings can be directly, dictionary-, or hash-assigned, as described below. Examples of alternatives (dataprints and distribution-comparisons) applicable to certain descriptor types are also mentioned.

Descriptors with fixed limits (*e.g.* number, range of sizes, elemental composition) can be directly assigned to positions in a bit-string, with offsets being calculated to assign different groups or different descriptors to separate areas of the bit-string. For example, the ring descriptors devized by Downs *et al.* [65], were directly assigned to the end of a bit-string, offset to avoid the beginning (which was reserved for dictionary-assigned augmented atom and

atom sequence descriptors). The Diverse-Property Derived (DPD) code developed at Rhone-Poulenc Rorer [63] could also could be directly assigned. The DPD contains the six descriptors described earlier, each split into a number of classes (2–4) giving a total of 17 bit positions (432 combinations).

Systems based on dictionary-assigned bit-strings employ a dictionary that specifies correspondences between particular functional groups or fragments and bit positions in a bit-string, with each entry (structural key) in the dictionary being assigned a bit position (screen number). Dictionaries of functional groups tend to be fairly small and so all groups can be listed in the dictionary and assigned to a short bit-string. However, analysis of a database to generate several different fragment types typically produces many tens or hundreds of thousands of distinct fragments, with a highly skewed distribution (*i.e.* a few fragments occur very frequently and many occur very infrequently).

At least three methods are available to reduce the number of potential fragments to fit into a fixed-length bit-string of a few thousand bits, whilst retaining those fragments that act as the best descriptors (as was carried out, for example, in the development of the CAS ONLINE Screen Dictionary [18]). Statistical analysis of the fragment frequencies can be carried out to to remove very frequent/infrequent fragments (for substructure searching, equifrequent occurrence of fragments gives better screenout; for similarity calculations frequency is less important). Specific fragments can be generalized to less specific forms which cover many different, but related, specific fragments. Finally, the same screen number can be assigned to several different, but related (*e.g.* by co-occurrence or composition) fragments; co-occurrence is particularly relevant for similarity calculations since it biases the measure towards that feature. Careful selection of very generalized fragments can give good representations for similarity calculations using relatively few bits; for example, Brown and Martin found that effective searches could be achieved using a small subset of MDL Information Systems' MACCS keys [34, 64]. However, selection of appropriate fragments to include in a dictionary is tedious, features not represented in the dictionary can never be reflected in the similarity measure, and the resulting structural resemblances may be strongly database-dependent.

Rather than selecting a subset of fragments for inclusion in a dictionary, so that the number of screens is reduced to the same as the length of the bit-string, hash-assigned bit-strings are created by fitting all of the fragments into the bit-string. This can be achieved by hashing the fragment to generate one or more integers that fall within the length, or a given subrange of the length, of the bit-string (fingerprint). The more integers generated by the hash function, the more unique patterns can be superimposed on the bit-string, and so the more fragments can be included. This fingerprint approach is used, for example, by Daylight Chemical Information Systems Inc. and Tripos Inc. for both substructure searching and similarity searching. However, overlaps between patterns can lead to many patterns being overlaid by other patterns, with a consequent loss of information. For similarity calculations this can give rise to false similarities since common bits in two bit-strings may have been set by completely unrelated fragments. Adding all fragments can also give problems by including many co-occurring fragments, by including large numbers of fragments that are unrelated to the similarity relationships that the measure is seeking to quantify, and by swamping the effects of those fewer fragments that are so related. These problems are exacerbated by the technique of folding the bit-strings to condense the information further.

Physicochemical property and topological index descriptors are usually represented using a fixed-length string of real numbers (sometimes referred to as a *dataprint*). Dataprints typically have far fewer elements than bit-strings, and each element has a value. Dataprints thus describe molecular space by a full matrix rather than the sparse matrix description given by bit-strings. To avoid biases caused by differences in magnitude of the descriptors (particularly physicochemical properties), it is usual to normalize each element of a dataprint by the range or standard deviation of that element throughout the dataset [24]. As noted previously when discussing distance-based 3-D descriptors, it is also possible to use the frequency distribution of many descriptors directly to encode the descriptors for similarity calculation. If the distributions have the same number of elements, then a similarity coefficient or distance can be calculated in much the same way as for dataprints.

4.5 Conclusions

The previous Sections of this Chapter have reviewed the origins and current status of similarity searching in databases of 2-D and 3-D chemical structures. Although we have discussed a large number of ways in which the similarity between a pair of molecules can be quantified, it must be emphasized that we have focused our attention on simple, descriptor-based measures that can be computed sufficiently rapidly to enable them to be used for searching databases of non-trivial size. There are, of course, many other measures that can be used to quantify the degree of resemblance between pairs of structures, but they are currently too demanding of computational resources to permit their use for database searching. Thus, Downs and Willett [4] divide the available measures into three broad classes, depending on their computational requirements. The first class contains measures that are likely to remain too time-consuming for use in a database context for many years, as with many of the quantum-mechanical similarity measures that have been described (see, *e.g.* [60, 67]). The second class contains the descriptor-based approaches that are already used for similarity searching and formed the focus of this Chapter. The remaining class thus includes measures that are intermediate in their computational requirements and might thus be applicable to database searching, given an appropriately fast algorithm. The development of such algorithms is one of the principal challenges facing workers in the field of molecular similarity, but the potential benefits are large, given the very different types of structural relationship that might be uncovered by new types of similarity measure.

Many of these intermediate measures involve the generation of a molecular superposition, so that the output from the matching algorithm is not just some quantitative measure of the degree of similarity between two molecules, but also an alignment that superimposes the matching features in the two structures under consideration. An example of such an approach is the *feature tree* described by Rarey and Dixon [68], which provides a graphic representation of the hydrophobic fragments and functional groups in a molecule, and which matches pairs of such representations using an MCS-like procedure. Two other examples of such whole-molecule approaches are the atom-mapping method developed by Pepperrell *et al.* [54, 69], and the Superpositioning by PERMutations (SPERM) method developed at Organon [70] following work by Dean *et al.* [71]. Atom mapping compares the 3-D environment of each atom in one molecule with the 3-D environment of each atom in a second mol-

ecule to give a list of interatomic similarities. These provide the input to a construction procedure that identifies large, but not necessarily maximal, common substructures. In SPERM, a molecule is placed at the centre of a tessellated icosahedron, each vertex of which encodes the distance from that vertex to the molecule's surface, and pairs of molecules are aligned by maximizing the degree of fit between these sets of distances. Dean's group has described a range of methods for calculating 3-D similarities [72]; thus far, only the tessellated icosahedron approach has been applied to database searching, but it is likely that improvements in computing speeds will enable other of his methods to be used in this context in the future. Whole-molecule methods for generating alignments increasingly make use of the distribution of physicochemical properties in 3-D space, and involve scoring functions that are much more complex than conventional similarity calculations (see, *e.g.* [38, 73–78]). Examples of work in this area are discussed elsewhere in this book, and there seems little doubt that such methods will play an increasingly important role in virtual screening systems in the future, especially when robust techniques are identified for the inclusion of conformational flexibility in the matching function [74–76].

Validating the statistical significance of structure–property relationships is an important component of any QSAR study. It is thus surprising that there have been only two studies to date that considered the significance of the similarities calculated in molecular similarity studies, with Bradshaw and Sayle [79] and Sheridan and Miller [80] using randomization experiments to assess the significance of fragment-based and MCS-based similarities, respectively. It is to be hoped that more use will be made of such techniques in the future to validate both the measures that have been discussed thus far and the many new measures that continue to be described in the literature (see, *e.g.* [81–84]). Once an individual measure has been validated in this way, the next challenge is to compare it with other measures to determine its effectiveness. As noted previously, this is normally done by means of simulated property prediction experiments based on the similar property principle, with the aim of identifying the most effective method(s) from amongst those being compared. An alternative approach recognizes that different similarity measures reflect different types of molecular characteristic, and the multifaceted nature of biological activity would thus suggest that no single measure will be optimal for all sorts of similarity search that one might wish to carry out. Instead, one can combine the rankings of a database produced by different measures to give a single resultant ranking using data fusion methods [60, 61, 85, 86]. We expect that such approaches will become increasingly attractive as improvements in computer power mean that it is feasible to carry out several different searches of a database for a given target structure.

In conclusion, similarity searching provides a simple, but effective way of screening chemical databases and thus already plays an important role in lead-discovery programmes in the pharmaceutical and agrochemical industries. Its importance can only increase further as 3-D similarity measures become established, complementing the 2-D similarity measures that form the basis for most current chemical information systems. However, we wish to reiterate the point made at the start of this Chapter about the inherent simplicity of similarity searching. Its strength is that it enables screening to be carried out when just a single bioactive molecule is available; if additional structure and/or activity information (such as a 2-D substructural query, a 3-D pharmacophoric pattern, or the 3-D structure of a protein binding site) is available, then other screening methods are to be preferred, as discussed elsewhere in this book.

Acknowledgements

We thank the American Chemical Society for permission to base much of the material in this chapter on a paper previously published in the *Journal of Chemical Information and Computer Sciences* [6]. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

References

[1] P. Willett, *Similarity and Clustering in Chemical Information Systems*, Research Studies Press, Letchworth **1987**.
[2] M. A. Johnson, G. M. Maggiora (Eds.), *Concepts and Applications of Molecular Similarity*, John Wiley, New York **1990**.
[3] P. M. Dean (Ed.), *Molecular Similarity in Drug Design*, Chapman and Hall, Glasgow **1994**.
[4] G. M. Downs, P. Willett, *Rev. Comput. Chem.* **1995**, *7*, 1–66.
[5] A. C. Good, J. S. Mason, *Rev. Comput. Chem.* **1995**, *7*, 67–117.
[6] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
[7] J. M. Barnard, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532–538.
[8] P. Willett, *J. Mol. Recognit.* **1995**, *8*, 290–303.
[9] Y. C. Martin, P. Willett (Eds.), *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*, American Chemical Society, Washington **1998**.
[10] R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
[11] P. Willett, V. Winterman, D. Bawden, *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36–41.
[12] J. G. Topliss, R. P. Edwards, *Am. Chem. Soc. Symp. Ser.* **1979**, *112*, 131–145.
[13] R. D. Cramer, G. Redl, C. E. Berkoff, *J. Med. Chem.* **1973**, *17*, 533–535.
[14] T. R. Hagadone, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 515–521.
[15] W. Fisanick, K. P. Cross, A. Rusinko, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664–674.
[16] W. Fisanick, K. P. Cross, J. C. Forman, A. Rusinko, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 548–559.
[17] W. Fisanick, A. H. Lipkus, A. Rusinko, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 130–140.
[18] P. G. Dittmar, N. A. Farmer, W. Fisanick, R. C. Haines, J. Mockus, *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.
[19] V. J. Gillet, G. M. Downs, A. Ling, M. F. Lynch, P. Venkataram, J. V. Wood, W. Dethlefsen, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126–137.
[20] Y. Takahashi, M. Sukekawa, S. Sasaki, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 639–644.
[21] W. Fisanick, *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145–154.
[22] P. Willett, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 114–116.
[23] G. Grethe, W. D. Hounshell, in *Chemical Structures 2*, W. A. Warr (Ed.), Springer-Verlag, Berlin **1993**, pp. 399–407.
[24] P. A. Bath, C. A. Morris, P. Willett, *J. Chemomet.* **1993**, *7*, 543–550.
[25] D. B. Turner, P. Willett, A. M. Ferguson, T. W. Heritage, *SAR QSAR Environ. Res.* **1995**, *3*, 101–130.
[26] Z. Hubálek, *Biol. Rev. Cambridge Phil. Soc.* **1982**, *57*, 669–689.
[27] J. C. Gower, in *Encyclopaedia of Statistical Sciences*, S. Kotz, N. L. Johnson, C. B. Read (Eds.), John Wiley, Chichester **1982**, pp. 397–405.
[28] D. Ellis, J. Furner-Hines, P. Willett, *Perspect. Inform. Manag.* **1994**, *3*, 128–149.
[29] J. D. Holliday, S. S. Ranade, P. Willett, *Quant. Struct.-Act. Relat.* **1995**, *14*, 501–506.
[30] R. R. Sokal, P. H. Sneath, *Principles of Numerical Taxonomy*, W. H. Freeman, San Francisco **1963**.
[31] C. A. James, D. Weininger, J. Delaney, *Daylight Theory Manual, Fingerprints – Screening and Similarity*, at URL http://www.daylight.com/dayhtml/doc/theory/theory.toc.html
[32] J. H. Ward, *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.
[33] G. M. Downs, P. Willett, W. Fisanick, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094–1102.
[34] R. D. Brown, Y. C. Martin, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
[35] M. S. Lajiness, *Perspect. Drug. Disc. Design* **1997**, *7/8*, 65–84.
[36] M. S. Lajiness, personal communication, February **1998**.
[37] D. Flower, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379–386.
[38] A. Schuffenhauer, V. J. Gillet, P. Willett, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295–307.
[39] G. W. Adamson, J. A. Bush, *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 55–58.

[40] P. Willett, V. Winterman, *Quant. Struct.-Act. Relat.* **1986**, *5*, 18–25.
[41] C. Cheng, G. M. Maggiora, M. S. Lajiness, M. A. Johnson, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 909–915.
[42] R. Carbo, M. Arnau, L. Leyda, *Int. J. Quantum Chem.* **1980**, *17*, 1185–1189.
[43] C. A. Reynolds, C. Burt, W. G. Richards, *Quant. Struct.-Activ. Relat.* **1992**, *11*, 34–35.
[44] A. C. Good, *J. Mol. Graph.* **1992**, *10*, 144–151.
[45] J. D. Petke, *J. Comput. Chem.* **1993**, *14*, 928–933.
[46] J. Bradshaw, *Introduction to the Tversky Similarity Measure*, at URL http://www.daylight.com/meetings/mug97/Bradshaw/MUG97/tv_tversky.html
[47] A. Tversky, *Psycholog. Rev.* **1977**, *84*, 327–352.
[48] G. M. Maggiora, J. Mestres, T. R. Hagadone, M. S. Lajiness, *Asymmetric Similarity and Molecular Diversity*, Presented at the 213th National Meeting of the American Chemical Society, San Francisco, California, April 13–17, **1997**.
[49] P. E. Jones, R. M. Curtice, *Am. Docum.* **1967**, *18*, 153–161.
[50] R. D. Brown, *Perspect. Drug Disc. Design* **1997**, *7/8*, 31–49.
[51] G. W. Adamson, J. Cowell, M. F. Lynch, A. H. W. McLure, W. G. Town, A. M. Yapp, *J. Chem. Docum.* **1973**, *13*, 153–157.
[52] R. Nilakantan, N. Bauman, J. S. Dixon, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82–85.
[53] C. A. James, D. Weininger, J. Delaney, *Daylight Theory Manual, SMARTS*, at URL http://www.daylight.com/dayhtml/doc/theory/theory.toc.html
[54] C. A. Pepperrell, P. Willett, *J. Comput.-Aided Mol. Design* **1991**, *5*, 455–474.
[55] P. A. Bath, A. R. Poirrette, P. Willett, F. H. Allen, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141–147.
[56] G. W. Bemis, I. D. Kuntz, *J. Comput.-Aided Mol. Design* **1992**, *6*, 607–628.
[57] R. Nilakantan, N. Bauman, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79–85.
[58] H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205–1231.
[59] G. Moreau, P. Broto, *Nouv. J. Chim.* **1980**, *4*, 757–764.
[60] R. P. Sheridan, M. D. Miller, D. J. Underwood, S. K. Kearsley, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
[61] S. K. Kearsley, S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley, R. P. Sheridan, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118–127.
[62] S. C. Basak, V. R. Magnuson, G. J. Niemi, R. R. Regal, *Discrete App. Math.* **1988**, *19*, 17–44.
[63] R. A. Lewis, J. S. Mason, I. M. McLay, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 559–614.
[64] R. D. Brown, Y. C. Martin, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
[65] G. M. Downs, V. J. Gillet, J. D. Holliday, M. F. Lynch, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 215–224.
[66] R. Carbo, B. Calabuig, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 600–606.
[67] R. B. Hermann, D. K. Herronn, *J. Comput.-Aided Mol. Design* **1991**, *5*, 511–524.
[68] M. Rarey, J. S. Dixon, *J. Comput.-Aided Mol. Design* **1998**, *12*, 471–490.
[69] C. A. Pepperrell, R. Taylor, P. Willett, *Tetrahed. Comput. Methodol.* **1990**, *3*, 575–593.
[70] N. C. Perry, V. J. van Geerestein, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 607–616.
[71] P. M. Dean, P. Callow, P. L. Chau, *J. Mol. Graph.* **1988**, *6*, 28–34.
[72] P. M. Dean, T. D. J. Perkins, in *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*, Y. C. Martin, P. Willett (Eds.), American Chemical Society, Washington **1998**, pp. 199–218.
[73] S. K. Kearsley, G. M. Smith, *Tetrahed. Comput. Methodol.* **1990**, *3*, 615–633.
[74] T. D. J. Perkins, J. E. J. Mills, P. M. Dean, *J. Comput.-Aided Mol. Design* **1995**, *9*, 479–490.
[75] D. A. Thorner, D. J. Wild, P. Willett, P. M. Wright, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 900–908.
[76] G. Klebe, T. Mietzner, F. Weber, *J. Comput.-Aided Mol. Design* **1999**, *13*, 35–49.
[77] C. Lemmen, C. Hiller, T. Lengauer, *J. Comput.-Aided Mol. Design* **1998**, *12*, 491–502.
[78] J. Mestres, D. C. Rorher, G. M. Maggiora, *J. Comput.-Aided Mol. Design* **1999**, *13*, 79–93.
[79] J. Bradshaw, R. Sayle, *Some Thoughts on Significant Similarity and Sufficient Diversity*, at URL http://www.daylight.com/meetings/emug97/Bradshaw/Significant_Similarity
[80] R. P. Sheridan, M. D. Miller, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 915–924.
[81] D. D. Robinson, T. W. Barlow, W. G. Richards, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 943–950.
[82] H. Briem, I. D. Kuntz, *J. Med. Chem.* **1996**, *39*, 3401–3408.
[83] A. H. Lipkus, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 582–586.
[84] B. D. Silverman, M. C. Pitman, D. E. Platt, I. Rigoutsos, *J. Comput.-Aided Mol. Design* **1998**, *12*, 525–532.
[85] C. M. R. Ginn, D. B. Turner, P. Willett, A. M. Ferguson, T. W. Heritage, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23–37.
[86] C. M. R. Ginn, P. Willett, J. Bradshaw, *Perspect. Drug Discov. Design*, in press.

Virtual Screening for Bioactive Molecules
Edited by Hans-Joachim Böhm and Gisbert Schneider

5 Modelling Structure–Activity Relationships

Gianpaolo Bravi, Emanuela Gancia, Darren V. S. Green, Mike M. Hann

5.1 Introduction

It is generally recognized that the structure and the physico-chemical properties of a molecule are important in determining its biological activity. QSAR (Quantitative Structure–Activity Relationship) and pattern recognition are techniques which enable this relationship to be expressed mathematically. Historically, the primary objective of these methods was the understanding of which factors are important for the specific activity of a series of compounds, with a secondary goal of predicting the activity of unknown compounds. Modelling structure–activity relationships typically involves two main steps:

- computing or measuring, for each compound in a collection, a number of molecular descriptors which reflect their physical and chemical properties, and
- deriving a quantitative correlation between the molecular descriptors (predictor variables) and the activity (response variable).

In the past few decades, a large number of different molecular descriptors, designed to code new physical and chemical information or to be easily and quickly computable, have been proposed. The number of statistical tools applied to QSAR and pattern recognition, either developed by chemometricians or borrowed from mathematicians, has also increased to allow the use of binary as well as continuous variables, and to enable the extraction of relevant information from large matrices of highly inter-correlated molecular descriptors.

The explosion of so many different techniques has opened new perspectives: QSAR and pattern recognition methods, that historically have been dedicated to solve lead optimization problems, have been more recently applied to prioritize compound screening in the very early phases of a project.

Our aim is to give an overview of the most commonly used methods, focusing on those which have proven to be consistently successful in a number of applications, and attempting to cover the different phases of the drug discovery process, from lead generation to lead optimization. Promising new approaches from the recent literature are also reviewed.

The review covers classical Hansch analysis, 3-D QSAR methods, 3-D alignment-free and topological descriptors, pharmacophores, and pharmacophoric keys applied to QSAR. In the text we mainly focus on the molecular description aspects and on the applications, whereas details on the statistical techniques, mentioned during the course of the Chapter, are given in Section 5.8.

5.2 Hansch Analysis

The classical QSAR approach (also known as Hansch analysis) correlates biological activity values with physical and chemical properties by linear, multiple linear or non-linear regression analysis (see Section 5.8.2.1). The first equation explaining a biological activity through the linear combination of parameters describing hydrophobic (log$P$) and electronic (σ) effects was proposed by Hansch and Fujita in 1964 [1]:

$$\text{Log}1/C = k_1 \times \log P + k_2 \times \sigma + k_3 \tag{5.1}$$

where C is the concentration that produces a certain biological response and k_1, k_2, k_3 are the regression coefficients. Eq. (5.1) was extended to model *in vivo* data, by adding a quadratic term ($k_4 \times \log P^2$), in an attempt to include non-linear effects of hydrophobicity.

Thousands of QSAR analyses derived using physical and chemical parameters, have been reported since this first formulation. The recently published book "Exploring QSAR" [2] contains a collection of successful classical QSAR applications in modelling non-specific toxicity, protein and enzyme activity, metabolism, mutagenesis, carcinogenesis, microbial, and pesticide activity. The C-QSAR database [3], assembled at Pomona College, is probably the most comprehensive electronic collection of QSAR equations. It has been developed over 20 years and contains more than 10000 datasets, covering both biological and physical organic chemistry. It reflects the philosophy that the best validation of a QSAR equation is not in the value of the correlation coefficient, the F-test or any other statistical parameter (see Section 5.8.2.1), but rather via a lateral correlation with many other analogous QSARs.

The number and variety of parameters used in QSAR has kept increasing as well. They can be roughly classified in three classes, according to the effects they model [4]:

- electronic effects: Hammett constants (σ), pK_a, field and resonance parameters (F and R), dipole moments, charge-transfer constants, parameters from spectroscopic data, and quantum-mechanical parameters,
- hydrophobic effects: partition coefficients (mainly Log$P$ and π) and chromatographic parameters (α and β, that is hydrogen bond basicity and acidity), and
- steric and polarizability effects: molecular weight (MW), molar refractivity (MR), molar volume (MV), Taft steric effect (E_s), Verloop sterimol (L, B_1, B_2, B_3, B_4, and B_5) and other geometrically derived parameters.

Many indicator variables (for instance the number of hydrogen bond acceptors/donors, the presence/absence of a fragment and so forth) have been also introduced.

Many published QSAR studies use molecular descriptors derived experimentally (the second volume of "Exploring QSAR" [5] contains approximately 17000 partition coefficients from octanol-water and a comprehensive listing of electronic and steric parameters). So long as QSAR studies concentrated on relatively small datasets (typically 20–50 molecules) of structurally related molecules, this approach was perfectly valid. However, despite the compilation of experimental tables, values for uncommon structures are still missing, not to mention the inconsistencies among values coming from different sources. Moreover, the robotic approaches to synthesis and screening (combinatorial chemistry and high-throughput

screening) are likely to dramatically increase the size of datasets available for QSAR. It is becoming unfeasible to manually insert property values from experimental tables, and in the future, the access to accurate methods for calculating physical and chemical properties will be more and more crucial.

Lipophilicity [6] and steric-polarizability [7] parameters have been calculated for many years, using fragment-based methods, which are now routinely used by pharmaceutical companies. More recently, methods for predicting electronic parameters have been published: Hammett constants can be calculated using additive contributions of fragments [8], partial atomic charges [9] and Frontier Molecular Orbital (FMO) energies [10]. Some of the experimental hydrogen bonding scales derived from solvatochromic parameters have been reproduced by Molecular Electrostatic Potential (MEP) [11], semi-empirical atomic charges, and FMO energies [12]. One aspect to be taken into account when developing a method for predicting physico-chemical properties for QSAR is the ease of calculation and automation of the whole process. The Hammett prediction through partial atomic charges is a good example: the procedure is very easy to automate and the results can be exported in a straightforward way (the method has been implemented in a Web-based property calculator [9]).

Unfortunately, the great expansion in the number of descriptors for QSAR over the years does not necessarily mean that new, meaningful chemical information has been encoded.

Many studies have been reported, investigating inter-correlations among properties, mainly by means of PCA (see Section 5.8.1.1). For example, in one study a total of 74 parameters for a set of 59 aromatic substituents [13,14] were reduced to five principal components, still retaining about 84% of the information (variance) contained in the original 74 variables. A similar study [15,16], performed on amino acid sidechain properties, allowed the derivation of three principal properties (z-scores), that can be used for QSAR studies instead of the initial set of highly inter-correlated descriptors.

Although it seems that many equivalent parameters are available, some interactions (for example the desolvation contribution to binding, the membrane partitioning, and the strength of hydrogen bonds) still lack good characterizations [17]. The search for new descriptors better able to codify the information contained in a chemical structure continues to be a very active field in QSAR, and has produced many different techniques that will be discussed in the rest of this Chapter.

As an example of the use of Hansch analysis in a medicinal chemistry project, the work of Hamilton and co-workers at Parke-Davies [18], who studied xanthine derivatives as adenosine antagonists, is instructive. A predictive QSAR model was derived using classical substituent parameters augmented with indicator variables for hydrogen bond donors and acceptors. The correlation among the parameters was removed by the use of a data reduction technique and a QSAR model subsequently derived using multiple regression. This model was used not only to predict potency, but also to conclude that the most active compounds in a series had already been made. In addition, the model allowed the identification of a substitution point on the molecules, which could be used to attach solubilizing groups without affecting the potency.

Protein crystallography, homology modelling, docking techniques, pharmacophore modelling, and the recently developed 3-D QSAR methods are powerful means to better understand specific biological interactions, but they are of limited value to optimize ADME (Ab-

sorption, Distribution, Metabolism, and Excretion) profile. When modelling pharmacokinetic properties, Hansch analysis is still amongst the most successful approaches [19].

A large number of drugs are developed to be orally administered and therefore they must cross the intestinal epithelium to be delivered into the systemic circulation and reach their target. A number of QSAR models have been proposed to predict the passive intestinal absorption, based either on experimental *in vitro* data or on physicochemical parameters. The Caco-2 cell monolayer has been proposed as *in vitro* model of human intestinal absorption [20]: a strong correlation ($R = 0.95$) was observed between *in vivo* human absorption and the *in vitro* Caco-2 permeability coefficient for 35 compounds.

The passive intestinal absorption of 18 structurally diverse drugs has been correlated with the number of hydrogen bond donors, the Polar Surface Area (PSA) and either log*D*5.5 or log*D*6.5 (octanol/water distribution coefficient at pH 5.5. and 6.5 respectively) [21]. Alternatively, the "rule of five" [22] (see Chapter 3) or the value of the PSA [23] can offer a crude, but fast estimate of the chance of absorption.

Another property, whose prediction is extremely useful in Drug Discovery, is the capability to penetrate the Blood–Brain barrier (BBB).

Abraham *et al.* [24] in 1994 proposed the following general equation for the interpretation and prediction of log*BB* (where *BB* is defined as C_{brain}/C_{blood}):

$$\begin{aligned} &\text{Log}BB = 0.198 \times R_2 - 0.687 \times \pi_2{}^H - 0.715 \times \Sigma\alpha_2{}^H \\ &- 0.698 \times \Sigma\beta_2{}^H + 0.995 \times Vx - 0.038 \\ &(\text{with } n = 57, r = 0.95, s = 0.197, F = 99) \end{aligned} \tag{5.2}$$

R_2 is an excess molar refraction, $\pi_2{}^H$ is the dipolarity/polarizability, $\Sigma\alpha_2{}^H$ and $\Sigma\beta_2{}^H$ are summation hydrogen bond acidity and basicity, and Vx is the characteristic volume of McGowan. The authors also describe methods to estimate the above descriptors by fragment-based approaches, so that log*BB* can be straightforwardly calculated from the chemical structure.

A more recent, simpler QSAR model [25] predicts Log*BB* using the *PSA* and the calculated log*P* (*C*log*P*):

$$\begin{aligned} &\text{Log}BB = -0.148 \times PSA + 0.152 \times C\text{log}P + 0.139 \\ &(\text{with } n = 55, r = 0.89, s = 0.35, F = 95) \end{aligned} \tag{5.3}$$

With respect to previously reported studies, Eq. (5.3) has the advantage of being readily interpretable, fast to calculate and easy to automate. Indeed, it is often the case that a less statistically robust, but interpretable model may be used in preference to a superior model, based on mathematical parameters such as those derived from a PCA.

5.3 3-D QSAR

For many years, 3-D QSAR has been synonymous with CoMFA (Comparative Molecular Field Analysis) [26]. CoMFA was the first technique to implement in a QSAR approach the concept that a specific biological activity of a molecule is an inherent property of its three-di-

mensional structure and that any binding between a receptor and a ligand is mostly the product of non-covalent weak interactions. The idea behind CoMFA is that the 3-D steric and electrostatic properties of a molecule can be described by embedding the molecule in a grid and calculating its interaction energies with a probe atom (typically Csp^3 atom with charge +1) at each node of the grid. For QSAR purposes, all the training set molecules are aligned and the steric and the electrostatic interaction energies are calculated at each grid point (typically 1–2 Å separation). The relative spatial orientation of each molecule within the grid (alignment) clearly plays a crucial role in the analysis. A description matrix, whose rows represent the molecules and whose columns contain the interaction energies, is filled. Such a matrix, where columns are highly inter-correlated and their number largely overwhelms the number of rows, cannot be analyzed by multiple linear regression. The PLS technique (see Section 5.8.2.2) is a statistical tool able to derive a QSAR model from this type of data. The results of a PLS analysis are still in the form of a linear equation. Each predictor variable is characterized by a coefficient, which reflects its importance. The equation can then be used as in traditional Hansch analysis to predict the activities of unknown molecules.

In spite of using PLS, spurious results can still occur due the high level of noise hidden in the description matrix. The program GOLPE (Generating Optimal Linear PLS Estimations) [27] was developed to identify which variables are meaningfully related to biological activity and to remove those detrimental to predictivity. Within this approach, Fractional Factorial Design (FFD) is initially applied to select multiple combinations of variables. For each combination, a PLS model is then derived and only variables able to significantly increase the predictivity are considered. To minimize the risk of chance correlations, the effect of each variable on the predictivity is compared with the average effect of dummy variables. Variables are finally classified as helpful, uncertain or detrimental. A further advance in GOLPE is the implementation of the SRD (Smart Region Definition) approach [28]. This is aimed at selecting the cluster of variables, rather than the single variable, mainly responsible for activity. This selection technique seems less prone to chance correlation than any single variable selection, and improves the interpretability of the results.

Several deficiencies of CoMFA in its original implementation have been reported over the years. The Lennard–Jones 6–12 potential appears inadequate at giving a reliable steric description [29] and also the electrostatic potential at giving a good-quality H-bonding description [30]. It was observed that CoMFA steric and electrostatic fields, in their original form, are able to properly represent only the enthalpic contribution to the free energy of interaction [31]. Several efforts have been made to either improve current fields or introduce new ones. For a detailed discussion on this matter see the reviews by Norinder [32] and Kroemer [33]. An alternative is to compute the molecular fields with the GRID program [34]. GRID includes a wider variety of probes so that more different types of interactions can be modelled. It has been successfully applied in combination with GOLPE in a number of cases [35,36]. Klebe replaced CoMFA potentials with Gaussian-type functions in his CoMSIA (Comparative Molecular Similarity Indices Analysis) approach [37]. Statistical results from these functions seem to be more stable and reproducible than those from CoMFA [38]. For more details on this technique refer to Chapter 10.

Since its launch by TRIPOS in 1988, more than 500 applications of CoMFA have been reported. The possibility of using heterogeneous sets of molecules (and not being restricted to the analysis of molecules with different substituents around a common skeleton), the accu-

rate local description in terms of molecular fields, and the straightforward interpretability of the statistical results are undoubtedly the major reasons for its popularity. The latter point is one of the biggest breakthroughs achieved by 3-D QSAR over classical QSAR. Equation coefficients can be plotted in the 3-D space and contoured. Upon visual analysis, regions of space contributing most to the activity (in either a positive or negative way) can be easily identified. This provides a simple and immediate check of the reliability of the statistical results, in that a reliable model should be able to pick up trivial SAR embedded in the training set. Additionally, the contours might suggest modifications to existing molecules which could either improve or retain the activity. A typical problem in Drug Design is not only optimizing the affinity, but also the selectivity towards a specific enzyme, in order to retain the required therapeutic activity, avoiding as much as possible undesirable side effects. In this respect 3-D QSAR methods like CoMFA, GRID/GOLPE and CoMSIA have proven to be very useful. To cite one such application, Matter *et al.* [39] recently described the combined use of 3-D QSAR and crystallographic information to understand the selectivity of the inhibitors of two Matrix Metalloproteinases (MMP-8 and MMP-3). The first step of the study consisted in deriving two separate models, one for MMP-8 and one for MMP-3, and in comparing the respective contour plots. As they were not sufficient to highlight macroscopic differences, a new QSAR model was derived by using the affinity ratio IC_{50}(MMP-8)/IC_{50}(MMP-3). The latter model allowed the identification of key determinants for selectivity towards one of the two enzymes.

Another promising CoMFA-related technique is the recently described SOMFA (Self-Organizing Molecular Field Analysis) [40]. This technique is still based on aligning molecules, embedding them in a grid and computing molecular fields. In its current implementation, indicator variables are used rather than steric field (that is a value of one is assigned to any grid node within the van der Waals envelope, a null value to all the others) while the Coulomb potential is used to derive the electrostatic field. What is crucial to SOMFA and makes this approach different from CoMFA is the notion of mean centred activity (y_{mca}). The response variable is centred around the mean, so that a scale is obtained where active molecules have positive values and inactive molecules negative values (Figure 5.1). For each property, a SOMFA master grid is then calculated. It has been defined as:

$$\mathrm{SOMFA}(x, y, z) = \sum_{i=1}^{n} \mathrm{property}_i\,(x, y, z) \times y_{i,mca} \tag{5.4}$$

The SOMFA master grid provides a straightforward visual representation of the regions positively contributing to activity as well as of the detrimental regions. For each molecule and each property a SOMFA value is then computed from:

$$\mathrm{SOMFA}_i = \sum_x \sum_y \sum_z \mathrm{property}_i\,(x, y, z) \times \mathrm{SOMFA}\,(x, y, z) \tag{5.5}$$

These values can be used as descriptors to derive regression models with a biological activity. PLS, which is still a barrier to many potential CoMFA users, is not required in the process. SOMFA looks very promising due to its inherent simplicity. However the method is still in its early days and requires further validation.

Structure	Activity	Mean Centered Act.
	1	-1
	2	0
	3	1

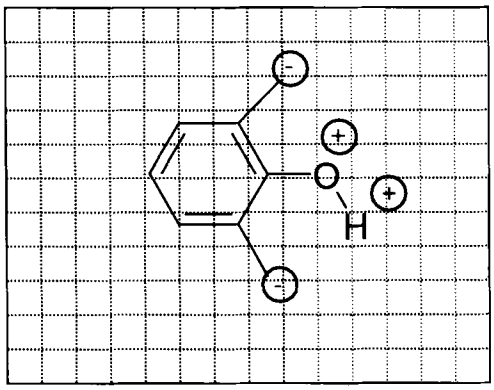


Figure 5.1. This simplified example illustrates the concept of mean centred activity in SOMFA. The example is set up such that a hydroxyl group enhances the activity while steric bulk adjacent to the hydroxyl reduces binding. Activities are mean centred so that high active compounds have positive values and low active compounds negative ones. Molecules are then superimposed using the common core and embedded in a grid. The mean centred activity of each molecule is finally added to every grid point crossed by the molecule. The hydroxyl group, present in the two most potent molecules, is left with a net positive value on it. The benzenoid core, common to all the datasets, has no net value associated with it, and the methyl groups have negative values. Adapted from reference [40].

An approach which attempts to maintain the advantages of field-based analysis, whilst introducing the concept of receptor size and shape, is the Receptor Surface Models (RSM) of MSI [41,42]. As in the CoMFA and SOMFA approaches, a set of pre-aligned molecules is required. A super-molecular surface is then generated to represent the shape of the molecular aggregate. This shape is assumed to be complementary to the shape of the receptor site itself. The option is given to generate a closed or open surface model. A closed model completely encloses the volume common to the molecules. An open model has holes on its surface, which may represent solvent-accessible regions or regions of uncertain information. Properties like partial charge, electrostatic potential, H-bonding propensity and hydrophobicity are calculated and stored with each surface point. The surface can be color-coded according to the property values, so that the intensity of the color corresponds to the magnitude of the property. The visual analysis of a receptor model surface is *per se* very informative. It gives a direct view on regions of space where aligned molecules share common features and where molecules differ. The surface property values are used to compute interaction energies between a molecule and the receptor surface model. Molecules can also be minimized within the surface. This process is analogous to minimizing a structure within a receptor site while keeping the site atoms fixed. The minimization process is a valuable tool that can help in the superimposition of molecules not used to derive the initial alignment, for instance molecules belonging to a test set. The following energy values can be calculated:

- E_{interact}: the non-bonded interaction energy
- E_{inside}: internal energy of the "bound" conformation
- E_{relax}: internal energy of the structure allowed to relax without feeling the surface

The difference between E_{inside} and E_{relax} gives an estimate of the strain energy in the molecule upon minimization. For each energy term, the total value plus the van der Waals and electrostatic components are computed. A simple QSAR equation can be directly derived using the energy values as descriptors and multiple linear regression analysis. The surface points and their individual interaction energy values (van der Waals, electrostatic and both energies combined) can also be used to derive a QSAR equation using PLS. This approach, called RSA [43], resembles CoMFA, the difference being that the points are located on a surface rather than on a grid. The authors claim that improved models can be obtained applying a non-linear, "genetic PLS". A genetic algorithm is used to select relevant points and a spline term is used to introduce non-linearity in the model. The selected points can be highlighted on the 3-D screen and visually analyzed. Though the spline term might have a beneficial effect, the use of GA with PLS in presence of hundreds of highly correlated descriptors is prone to produce chance correlation [44].

A comparison between RSM and CoMFA has been recently reported [45] on a set of 30 glycogen phosphorylase (GP) inhibitors. This comparison appears particularly useful since the receptor-bound conformation of the inhibitors is known. RSM and CoMFA provided comparable statistics, the former yielding a slightly higher q^2 value (0.70 against 0.60). The visual analysis of the receptor surface and of the contour maps confirmed that overall the information extracted by the two methods is in fact very similar and in qualitative agreement with the GP binding site. It is worth noting that the q^2 values derived from this experiment (even when molecular alignment, the greatest problem associated with 3-D QSAR, is solved) show that the statistical methods do not give perfect answers. Thus, in many practical applications, a QSAR with q^2 value of 0.6–0.7 is considered a good model.

All these 3-D QSAR methodologies share common advantages in that an accurate 3-D representation of molecular properties and an easy visual interpretation of statistical results can be carried out. However, they also share a crucial drawback, which is represented by the alignment step. The results depend not only on the 3-D conformation of each molecule but also on how different molecules are orientated relative to each other in 3-D space. Hence, if the alignment supplied is incorrect the 3-D QSAR analysis will fail or yield misleading results.

Different strategies are applied to align structures, depending on the amount of experimental information available. When a crystallographic complex is known for all the structures in the training set, the alignment can be swiftly accomplished by simply superimposing the structures of the complexes. If at least one X-ray complex (or the structure of the free receptor) is available, docking the training set molecules might provide a plausible alignment. Unfortunately the most common scenario is that no structural information is available, making the selection of the appropriate conformation and alignment an even more complex matter. In general, the more diverse and flexible the molecules are, the larger is the complexity of the problem. Certainly the presence of rigid molecules helps the alignment generation, but even so considerable work will be needed. For instance, the presence of a hydroxyl is not generally considered to increase the flexibility of a molecule. However, its orientation does have a profound effect on the electrostatic field and will influence the 3-D QSAR analysis. Rigid molecules often incorporate symmetry elements (due to the presence of ring systems), so that their alignment is a multiple solution problem even in the case of apparently trivial datasets. Hence, much attention must be paid to every detail and multiple alignments need to be explored, making any 3-D QSAR very time consuming.

Usually, when analyzing a congeneric series, a low-energy conformation of an active and possibly rigid molecule is chosen and used as a template (of course any insights on its possible bioactive conformation are used). The basic assumption of this approach is that the most active molecule in a series fits the steric and electronic requirements of the receptor site most effectively. This strategy has yielded good results in a number of cases [46,47]. However, it cannot claim general validity. In principle, each molecule will bind to a receptor so as to maximize the number of favorable interactions and minimize the unfavorable ones. Therefore, unexpected binding modes might well occur [48]. A significant example in this sense is the one involving antiviral (WIN) compounds complexed to rhinovirus 14 (HRV14) where very closely related structures bind in a reverse orientation [49,50].

For flexible and diverse structures, pharmacophore model generation methods [51,52] are used to derive an alignment. The crucial assumption of these methods is that a common binding mode must exist for all the molecules, or at least a considerable overlap. However this is not always the case, as highlighted by some structurally heterogeneous inhibitors of acetylcholinesterase (AChE). The recent X-ray determination [53] of the complexes of the enzyme with tacrine, edrophonium and decamethonium provided evidence that each of the three inhibitors has a unique binding mode with very little overlap. Their alignment would probably never have been predicted by any of the available tools.

In summary, 3-D QSAR methods provide a lot of useful information which can help in the understanding of structure-activity relationships, suggest chemical modifications to improve or retain a specific activity, and predict the activity of unknown structures. However, their application is quite time-consuming and requires a lot of *a priori* knowledge or at least a synergistic interaction with molecular modelling techniques to generate a plausible structure alignment. 3-D QSAR methods are specific for lead optimization tasks and, because the alignment is often not amenable to automation, they are usually restricted to the analysis of small datasets (less than 100 compounds).

5.4 Alignment-Free 3-D Descriptors

Considerable efforts continue to be dedicated to the development of descriptors which, whilst still retaining 3-D information, are not dependent on how molecules are superimposed to each other. In this Section we focus our attention mainly on molecular moments (WHIM and CoMMA), descriptors derived from molecular surfaces (MS-WHIM and MS-ACOR) and descriptors derived from spectra (EVA and CoSA).

The moments of atomic properties are used as descriptors by the WHIM [54] and CoMMA (Comparative Molecular Moments Analysis) [55] approaches. The two approaches mainly differ in the order of moments and the type of atomic properties used.

WHIM describes a chemical structure in terms of size, shape, symmetry and atom distribution. This is achieved by applying a weighted PCA on atomic coordinates (any atomic property can be used as weight) and extracting for each principal axis m (m = 1,2,3) moments from 2nd order to 4th order:

- PCA eigenvalues: $$\lambda_m = \frac{\sum_{i=1}^{n} w_i \times t_{im}^2}{\sum_{i=1}^{n} w_i} \qquad (5.6)$$

and their proportions: $\vartheta_m = \frac{\lambda_m}{\sum_{m=1}^{3} \lambda_m}$ (5.7)

- Skewness: $\gamma_m = \frac{\sum_{i=1}^{n} w_i \times t_{im}{}^3}{\sum_{i=1}^{n} w_i} \times \frac{1}{\lambda_m{}^{3/2}}$ (5.8)

- Kurtosis: $K_m = \frac{\sum_{i=1}^{n} w_i \times t_{im}{}^4}{\sum_{i=1}^{n} w_i} \times \frac{1}{\lambda_m{}^2}$ (5.9)

where t_{im} are elements of the score matrix after PCA, w_i is the weight vector, and n is the number of atoms in the molecule. The above parameters are defined as directional descriptors (in a more recent version the third order moment, skewness, was replaced by an information content index [56]). The use of PCA ensures they are invariant to the reference system. A set of non-directional descriptors, which are derived from the directional ones, are also included [57]. In non-directional descriptors the information about principal axes disappears and a global view of the molecule is gained. Depending on the atomic property used as the weight, different types of information can be extracted. For instance, when using atomic mass, the PCA eigenvalues correspond to the moments of inertia. Six atomic properties are used by WHIM (unweighted value, mass, van der Waals radius, electronegativity, polarizability, and electrotopological index) for a total of 99 descriptors.

CoMMA instead uses moments from the 0^{th} to 2^{nd} order [55]. Atomic mass and atomic charge are used as properties. In other words, the molecular weight, the moments of inertia, the total charge, the dipole, and the quadrupole are calculated. Additional parameters, describing the relationship between moments of the mass and between charge distributions, are extracted for a total of 14 descriptors, invariant to rotation and translation.

It is generally accepted that receptors and substrates recognize each other at their molecular surface. Therefore the binding of a ligand depends on the shape of its surface as well as the distribution of certain properties (for instance electrostatic potential) on its surface. Two methods have been described recently which attempt to capture this information and condense it into a brief numerical vector, which can be used for QSAR purposes (Figure 5.2).

The first method applies the WHIM approach to molecular surfaces to derive a set of descriptors called MS-WHIM [47,58]. Molecular surface point coordinates replace the atomic frame. Weighted PCA provides a unique reference system, independent of the original orientation of the surface. Weights are given by the values of six properties computed at the molecular surface level: 1. unweighted value, 2. positive, and 3. negative electrostatic potential, 4. hydrogen bond acceptor, and 5. donor ability, and 6. hydrophobicity. The directional descriptors above-mentioned (Eqs. 5.6-5.9) are then extracted.

MS-WHIM descriptors have been applied to several QSAR problems obtaining promising results [47,58–62]. Furthermore, they have been compared to the original WHIM indices on

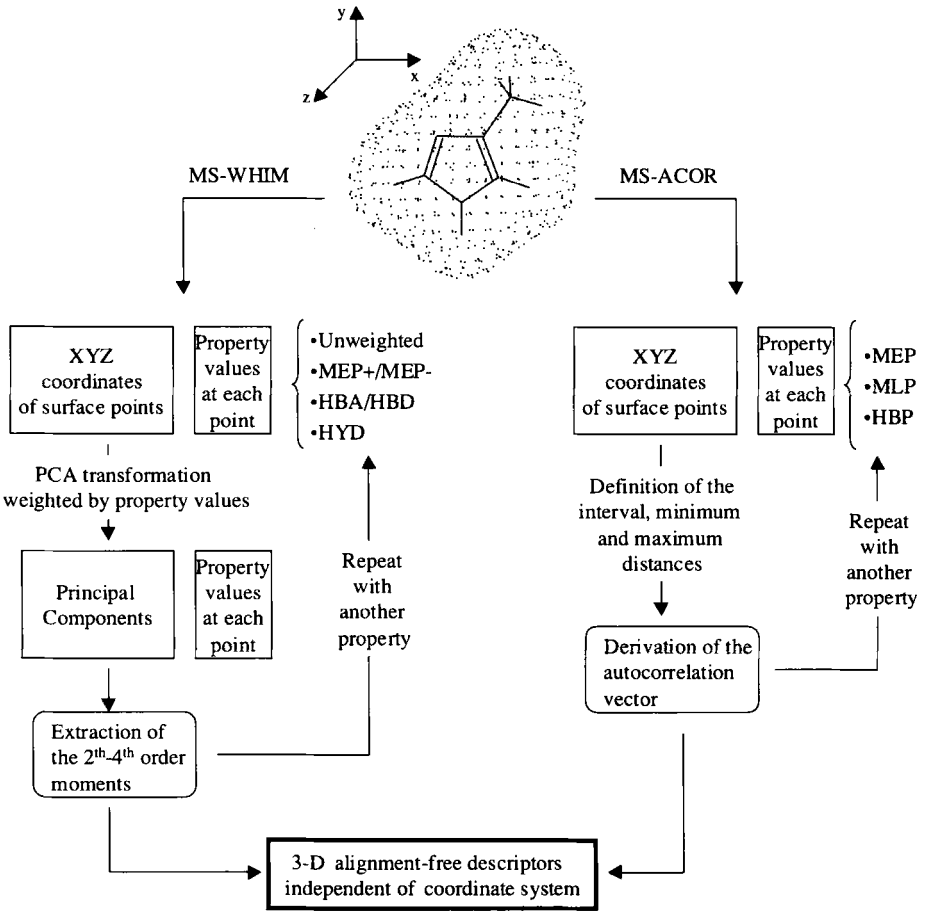
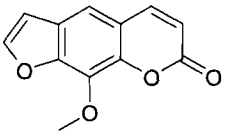


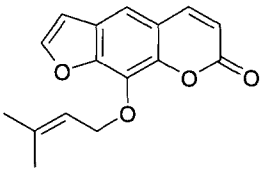
Figure 5.2. Flow chart of MS-WHIM and MS-ACOR approaches aimed at deriving 3-D alignment-free descriptors from molecular surface properties. For a given molecule, the molecular surface is computed as well as a number of molecular surface properties (details are given in [47, 58, 65]). A matrix containing as many rows as the number of molecular surface points and their x,y,z coordinates is then filled up. In the case of MS-WHIM, property values are used, in turn, to perform a weighted PCA and to extract moments from 2nd to 4th order. In the case of MS-ACOR, property values are used, in turn, in the derivation of the autocorrelation vector, upon previous selection of minimum, maximum and interval distances. For more details and mathematical formula see text (Section 5.4).

a number of datasets [47,62]: the use of molecular surface properties consistently resulted in more informative descriptors and yielded significantly better models than atomic properties. Very recently, a set of inhibitors and substrates of cytochrome P450 2A5 (CYP2A5) was analyzed [52] with MS-WHIM. This is a particularly appropriate example because the molecules, despite being structurally quite simple, contain symmetry elements which might render a CoMFA-like analysis ambiguous. Poso *et al.* [63] derived a CoMFA model on this dataset, obtaining a q^2 of 0.56 (with Leave-One-Out, see Section 5.8.2.2) and a poor prediction of im-

peratorin (Figure 5.3). This result is not surprising, given the high similarity of this molecule to the most potent in the series, methoxsalen, and the alignment adopted. MS-WHIM descriptors yielded a q^2 of 0.70 (with five random groups repeated 100 times, see Section 5.8.2.2), with size, positive MEP, H-bond acceptor ability and hydrophobicity the dominant properties. Not only was there a general improvement on the q^2 value, but also the prediction for imperatorin was better.



0.07 μM



2.80 μM

Figure 5.3. Chemical structures of methoxsalen and imperatorin along with CYP2A5 binding affinity values (K_i).

The second method based on molecular surface properties is the autocorrelation approach and we will refer to it as MS-ACOR. The autocorrelation technique was first applied to molecular description by Broto *et al.* [64], as a way to describe the distribution of a given atomic property over a molecular structure. Each term of the autocorrelation vector of a property p distributed over n atoms is defined as follows:

$$A_d = \sum_{i,j} p_i \times p_j \tag{5.10}$$

Each A_d coefficient depends on the way i and j are selected: A_0 is simply the sum of p squared over all the atoms, A_1 is the sum of the products of p over all the atom pairs separated by one bond, A_2 is the sum of products of p over all the atom pairs separated by two bonds, and so on. The same procedure can be applied by replacing the topological distance (number of bonds) with the interatomic distance in the 3-D space. Recently, Wagener *et al.* [65] extended the autocorrelation concept to the molecular surface, by calculating the autocorrelogram of the molecular electrostatic potential (MEP), molecular lipophilic potential (MLP), and hydrogen bonding potential (HBP). Each term of the autocorrelation vector for a potential P is calculated as follows:

$$A(d_{\text{lower}}, d_{\text{upper}}) = \frac{1}{N} \times \sum_{i,j} P_i \times P_j \tag{5.11}$$

where N is the number of terms in the sum and the sum is run over all the i and j surface points, whose distance is within the spatial interval defined by d_{lower} and d_{upper}. Considering an interval of 1 Å the following autocorrelation terms are calculated: MS-ACOR(0,1) where all the pairs of surface points whose distance is lower than 1 Å are included in the sum, MS-ACOR(1,2), where all pairs of points, whose distance is between 1 and 2 Å are taken into account, and so forth until the user-defined threshold distance is reached. Usually this corresponds to the size of the smallest molecular surface in the set. Results are dependent on the chosen interval. Values of 0.5 and 1 Å are recommended [62,65].

Analogous to the WHIM approach, autocorrelation has proven to yield more informative descriptors if applied to the molecular surface, rather than to the molecular frame [62]. This can be exemplified by looking at the values of representative descriptors of three HIV-RT inhibitors that, in spite of showing a high structural similarity, are characterized by different activities. Figure 5.4 highlights that MS-ACOR descriptors are able to discriminate among the three isomers, while autocorrelation based on atomic properties gives very similar descriptors.

MS-ACOR was recently applied in combination with Kohonen networks (see Section 5.8.1.3) to the analysis of high-throughput screening data of combinatorial libraries [66]. A three-reagent library based on the hydantoin scaffold was synthesized. From the screening, 185 hits were identified out of 5248 compounds tested, a hit rate of 3.5%. Compounds were divided into a training set (3567) and a test set (1761). Compounds occupying neurons containing at least one hit, and compounds in neurons next to them were classified as hits. The remaining compounds were classified as non-hits. Using MS-ACOR with HBP, 96% of the test set hits were predicted correctly with only 8% of false positives.

The final class of descriptors covered in this section are those derived from spectra. The EVA (EigenVAlue) descriptors [67,68] are derived from infrared (IR) frequencies, which can be calculated or extracted from experimental spectra. As the number of vibrational frequencies of a molecule is equal to $3N$–6, N being the number of atoms, they can not be directly used as QSAR descriptors, unless all the molecules in the dataset have the same number of atoms. The dimension of the EVA vector must be unified across the entire set of molecules through a standardization process. First, the original frequency values are projected on a bounded frequency scale (1–4000 cm^{-1}), where each vibration is represented as a point on the scale. Then, a Gaussian function of fixed standard deviation (σ) is placed on each frequency value, f_i ($i = 1 - 3N$–6). The value of EVA at each point of the scale (x) is calculated by summing up the contribution of all the $3N$–6 Gaussians at that point according to:

$$\text{EVA}(x) = \sum_{i=1}^{3N-6} \frac{1}{\sigma\sqrt{2\pi}} \times \exp\left(-\frac{(x - f_i)^2}{2\sigma^2}\right) \tag{5.12}$$

In this way, the same number of descriptors is derived for all the structures, according to the chosen sampling interval. The proper selection of the interval and of the standard deviation of the Gaussian functions is important for the success of the method. Typically the interval range is 2–5 cm^{-1} resulting in 2000–800 descriptors, while standard deviations of 10–40 cm^{-1} are considered reasonable [67]. Different ways to compute normal mode frequencies

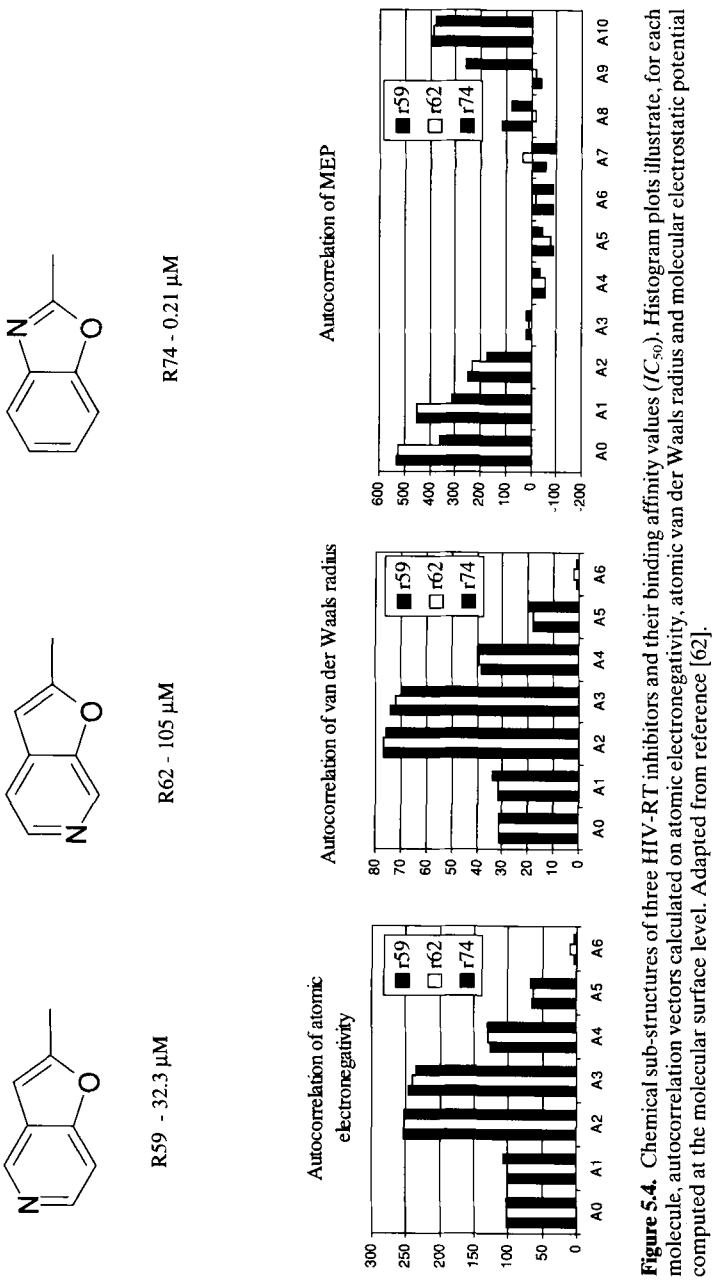


Figure 5.4. Chemical sub-structures of three HIV-RT inhibitors and their binding affinity values (IC_{50}). Histogram plots illustrate, for each molecule, autocorrelation vectors calculated on atomic electronegativity, atomic van der Waals radius and molecular electrostatic potential computed at the molecular surface level. Adapted from reference [62].

have been also compared [68]. In general, semiempirical methods (AM1 and PM3) performed equally well and outperformed molecular mechanics (AMBER). EVA has been tested on several different datasets, providing good results, often comparable to CoMFA [68].

The extraction of QSAR descriptors from experimental or simulated spectra was further explored by Bursi *et al.* [69] in their CoSA (Comparative Spectra Analysis) approach. They used the information extracted from experimentally determined mass, IR and ^{1}H NMR spectra, and simulated IR and ^{13}C NMR spectra to analyze 45 progestagens with PLS. Individually used, all the spectra provided good correlations, ^{1}H NMR giving the best q^2 value (0.548). Since in many ways spectra contain complementary information, combinations of different spectra were explored as well. Combinations of two spectra gave better results than individual ones (^{1}H NMR with simulated IR gave q^2 of 0.602), but combinations of more spectra could not further improve the results. A CoMFA analysis was also performed on the same dataset yielding a q^2 of 0.395. Interestingly this result was significantly improved by adding ^{1}H NMR spectra ($q^2 = 0.640$).

Certainly the most appealing property of the methods reviewed in this Section is their invariance to molecular orientation, meaning that the molecular structures do not need to be aligned prior to their calculation. For this reason they can be preferred to CoMFA-like methods for analyzing datasets, where the binding mode is suspected to be not unique, or there is no *a priori* information about alignment.

However, even though it is not necessary to perform molecular superimpositions, one must still address the issue of conformer choice in the case of flexible molecules. The dependence on molecular flexibility varies from one class of descriptors to another. Descriptors from simulated spectra are not very sensitive to the conformation, whereas moments and autocorrelation indices are more conformation-dependent. Obviously this problem will affect different types of analyses to a different extent: when modelling molecular properties, for instance log$P$, the choice of the appropriate conformation will have less influence than when modelling a specific interaction like the binding to a receptor.

Another important issue concerning all these alignment-free descriptors is that the interpretation of statistical results is not straightforward. It is in general difficult to visualize the chemical and physical content of these descriptors, mainly because of their global nature and the mathematical approach used in deriving them. This means that the corresponding QSAR models cannot be directly used to suggest which local chemical modifications are likely to improve or retain the activity, and which ones are detrimental. Hence they can be used only for predictive purposes, for instance for the virtual screening of databases. Many groups are currently working on this problem, as a key bottleneck to wider application of these methods is the ability to present the results in a way which can be interpreted by non-experts.

5.5 Topological Descriptors

In the past few years, there has been a consistent return to the use and development of 2-D approaches. This tendency was mainly driven by technologies like combinatorial chemistry and high-throughput screening, which typically make available large amounts of data, and require, more than anything else, easily computable descriptors. This Section will cover some of the most applied 2-D descriptors, in particular fragment-based descriptors, atom pairs, topological torsions, Kier and Hall connectivity, shape, and electrotopological indices.

Molecular holograms (HQSAR) were developed at TRIPOS for QSAR analysis [70]. A molecular hologram is an array containing counts of fragments and is strictly related to tra-

ditional 2-D fingerprints employed in database searching and molecular diversity techniques [71]. The molecular structure is broken down into all possible linear and branched fragments of connected atoms of size between M and N atoms, where M and N are integers defined by the user (typically M assumes a value of 4 and N 7). Different levels of information can be included in the hologram, by specifying parameters like atomic number, bond type, connectivity, hydrogen atoms, and chirality. At the lowest level of information only atom types are distinguished, so that furan and thiophene are distinguishable, but not ethane and ethylene. At the highest level, two enantiomers would have different holograms. An integer array of length L is created with all of the array elements set initially to zero. Each of the fragments is then hashed into this array incrementing by one the appropriate array element. Identical fragments will always hash to the same element. However, because the number of fragments is typically greater than L, some identifiers will collide. In other words, the same element will contain fragments of different types. The L-arrays are used as descriptors and PLS is used to build regression analyses.

In a recent study the implications of using a hashing algorithm in HQSAR were investigated [72]. In particular, a strong dependence of the PLS models on the length of the hologram (L) was found. The shorter the length, the stronger the dependence and the greater the variation in q^2. The comparison was extended to unhashed fragment bit-strings, which yielded the best results. The authors suggest that unhashed descriptors should be used, when computationally feasible, because hashed fragments tend to obfuscate PLS modelling. In principle this would help also in terms of model interpretation. Thus results from HQSAR analysis can be graphically displayed as a color-coded structure diagram in which the color of each atom reflects the contribution of that atom to the molecule's overall activity. The contribution of each atom in a fragment is obtained by dividing the PLS coefficient associated with the fragment by the number of atoms in the fragment. The overall contribution of each atom is obtained by summing up the individual atomic contributions from each of the fragments containing that atom.

The use of fragment-based descriptors for QSAR dates back to the Free–Wilson and Fujita–Ban (FW/FB) analysis methods [73,74]. With these approaches, molecules in a dataset are broken down in user-selected fragments, typically the substituents around a common core. There are limitations in the choice of the fragments: fragments must occur at least twice, and linear dependencies, that is fragments A and B always occurring together, must be avoided. Multiple linear regression is then used to derive a correlation between indicator variables, reflecting the presence/absence of fragments, and biological activity. The underlying assumption of this method is that fragments are considered to be independent, each giving an additive contribution to the activity. In this regard, the coefficients of the regression equation are directly proportional to the free-energy change when replacing one fragment with another [75]. This approach is fairly simple and has worked in several studies [76,77]. In other cases, it failed to produce satisfying correlations and this was ascribed to the presence of non-additive effects [78].

An inherent problem of the FW/FB approach is that it is inadequate at predicting compounds containing new unexplored fragments. Molecular holograms might suffer less from this limitation because they employ fragment patterns; that is, each atom in a molecule belongs to several different fragments. However, it is still a reasonable assumption that, in prediction, the confidence limits of HQSAR are narrower with respect to 2-D descriptors

encoding physical properties and 3-D methods based on molecular fields or molecular surfaces.

Despite their limited predictive ability, fragment-based methods like HQSAR may be valuable for data mining purposes, for a rapid understanding of the SAR embedded in the data. The STIGMATA algorithm [79] of Blankley and co-workers is another example of the utility of 2-D fingerprint descriptors, and is less statistically rigorous, but more suited to data mining than HQSAR. The method makes use of the concept of a modal fingerprint, which is a fingerprint of fixed length. A bit is set to 1 in the modal fingerprint if it is set in at least the threshold percentage of molecules in the input dataset. As an example, a set of dopamine D_2 ligands was analyzed at both the 50% and 100% thresholds. The modal fingerprint can be used in a similarity search to find novel molecules, or as in HQSAR the individual bits can be mapped back to the functional groups that set them and used to color code a displayed structure diagram. In addition, the algorithm can be used for compound acquisition or library design.

In a further example of data mining, the following involves the use of atom pairs and topological torsions in combination with SIMCA (see Section 5.8.3.2). Atom pairs [80] and topological torsions [81] were developed at Lederle laboratories in the 1980s. An atom pair is defined as a substructure composed of two non-hydrogen atoms and an interatomic separation measured in bonds along the shortest path connecting the two atoms. A topological torsion is a linear sequence of four consecutively bonded non-hydrogen atoms. The description uses the atom types and includes the number of heavy-atom connections and the number of π electron pairs on each atom. Atom pairs are long-range descriptors, in that they can capture long-range correlations between atoms in active molecules. Topological torsions, instead, are short-range descriptors and are aimed to complement atom pairs. These descriptors have been further developed by replacing atom types with numbers, which reflect their hydrophobic, donor/acceptor and partial charge properties, according to a particular binning scheme [82].

Hunt [83] recently applied these descriptors in combination with SIMCA to the analysis of a dataset of 405 NMDA glycine site antagonists. SIMCA is a tool for discriminant analysis based on PCA. Compounds are not represented by their activity values, but they are assigned to a category of activity (for instance high, medium, or low). Such a tool is valuable when activity values are affected by considerable uncertainty. In this case, 384 training set compounds were assigned to five activity classes and SIMCA was able to correctly classify 90% of them. Of 21 test compounds, 9 were well predicted, 11 were one category off, and only 1 was two categories off. The author suggests that running multiple experiments with staggered activity categories is a useful way to improve the results and at the same time increment prediction reliability. For instance, in this case, the combination of two SIMCA models predicted well 13 test compounds and 8 were one category off. The interpretation of SIMCA results was made more effective by coloring probe molecules according to the PCA loading values. Particularly useful was the identification of atoms in inactive molecules that could be targeted for change, as they contribute to descriptors detrimental to activity.

The topological indices (χ and κ), developed by Kier and Hall [84], are based on molecular graph theory. A molecular graph is the representation of a chemical structure as an ensemble of vertices (all non-hydrogen atoms) and edges (all bonds connecting non-hydrogen atoms). To calculate the χ connectivity indices, each non-hydrogen atom is characterized by two values, δ_i and δ_i^v, defined as follows:

$$\delta_i = \sigma_i - h_i \tag{5.13}$$

$$\delta_i^v = Z_i^v - h_i \tag{5.14}$$

where σ_i and Z_i^v are the number of σ and valence electrons of the i^{th} atom, respectively, and h_i is the number of hydrogen atoms bonded to the i^{th} atom. The χ indices are extracted from the molecular graph, by summing the reciprocal square root of δ and δ_i^v values in various ways. The 0^{th} order index ($^0\chi$) is the sum over all the atoms in the graph:

$$^0\chi = \sum_i (\delta_i)^{-1/2} \tag{5.15}$$

$$^0\chi^v = \sum_i (\delta_i^v)^{-1/2} \tag{5.16}$$

The first order index is the sum over all the edges:

$$^1\chi = \sum_{i,j} (\delta_i \times \delta_j)^{-1/2} \tag{5.17}$$

$$^1\chi^v = \sum_{i,j} (\delta_i^v \times \delta_j^v)^{-1/2} \tag{5.18}$$

Higher order indices are summations over sequences of two, three, etc. edges. The κ shape indices do not include information on the identity of the atoms, but simply involve counts over fragments of bonds. Similar to χ, indices of different order are defined depending on the length of the fragment. The first order index $^1\kappa$, is a count over single-bond fragments (paths of length 1):

$$^1K = \frac{2 \times ^1P_{max} \times ^1P_{min}}{^1P^2} \tag{5.19}$$

where ^1P is the number of paths of length 1 in the graph, and P_{max} and P_{min} are the maximum and minimum number of paths of length 1 possible for the same number of atoms. The second order index is determined by the count of two-bond fragments and so on.

A further development of the molecular connectivity approach led to the electrotopological (E-state) indices [85], where the amount of electronic information is significantly increased. The E-state index is an atomic rather than a molecular attribute, and is based on the intrinsic state of an atom (defined as $\delta^v + 1/\delta$). The influence of all the other atoms is introduced as a perturbation of this intrinsic state.

Kier and Hall descriptors have been recently applied by Zheng *et al.* [86] to the design of a combinatorial library in an approach called inverse QSAR. Within this approach 28 bradykinin (BK) potentiating pentapeptides were selected as a training set and modelled by PLS. The resultant equation was used to predict a virtual library of pentapeptides. The theoretical size of a pentapetide library is too large (3.2 million) to run a full enumeration. Hence a genetic algorithm was applied according to the following scheme. Initially, a random population of 100 peptides is generated and encoded using Kier and Hall descriptors. The fitness of each peptide is evaluated by its biological activity predicted by the PLS model previously

derived. The population is then evolved through crossover and mutation experiments. The most frequent amino acids (building blocks) found in the final population were those selected for the experiment.

An inherent drawback of Kier and Hall topological indices is their poor interpretability. This example highlights that low interpretable descriptors can still be useful when the purpose of the analysis is the development of a predictive model and prioritization of building blocks rather than understanding SAR. However, the prediction of virtual libraries is not an easy task and requires a high degree of extrapolation, as emphasized by the authors, since new libraries typically contain fragments previously unexplored. The newer E-state descriptors are much easier to interpret, being based on atom types and familiar concepts such as electronegativity. Also, as they are atom-based rather that fragment-based, they are more generally applicable.

Another approach recently developed that makes use of Kier and Hall descriptors is the so-called *binary QSAR* [87]. This approach was specifically developed to handle data from high-throughput screening (HTS). Typically molecular sets from HTS are much larger then the datasets used in traditional QSAR analyses and cover far more chemical variations. The response variable is characterized by an extremely low precision and a significant error rate. The method derives its name from the nature of the biological variable, which can assume only discrete values (active/inactive). The molecular structures are coded by using connectivity and shape indices and a Bayesian inference technique is used to estimate the probability that a new molecule is active.

This strategy was applied to a set of 463 diverse estrogen receptor (ER) ligands [88], randomly divided into training set (410) and test set (53). Molecules were assigned to inactive and active classes according to an arbitrary threshold. Different activity thresholds were used to test the robustness of the method. The overall accuracy in classifying active and inactive compounds remained stable, close to 90%, the fluctuation being approximately 10%. When used in prediction, the model correctly classified 78% (7 out of 9) of the active compounds and 98% of the inactive ones (43 out of 44).

The goal of this approach is not the accurate prediction of activity but rather the virtual screening of databases, the analysis of HTS data and the prioritization of future screening campaigns. As highlighted by the authors, in the above example the hit rate obtained by using binary QSAR was five-fold higher than random selection.

5.6 Pharmacophores and Pharmacophoric Keys

Methods reviewed in the previous Section employ 2-D descriptors, which are easy and fast to compute; hence they are well suited for analyzing large datasets. However, when the purpose of the study is the discovery of new chemical series, structurally different from known available leads, 3-D descriptors outperform topological indices [89,90]. In this regard pharmacophores are widely used. The pharmacophore concept is based on the kinds of interaction observed to be important in ligand–protein interactions: hydrogen bonding, charge, and hydrophobic interactions. A pharmacophore is commonly derived using a set of active molecules and looking for a common spatial arrangement of common functionalities (pharma-

cophoric groups). Information from inactive molecules can be added as well. However, active/inactive represent class behaviors and usually no quantitative data are used. More details on pharmacophore generation and database searching can be found in Chapter 7.

In this regard, Catalyst/HypoGen [91] developed at BioCad represents a unique method since it does require individual activity values, and it attempts to derive a pharmacophore, which is able to explain the activity differences among the training set. Normally, multiple pharmacophore hypotheses are retrieved which might differ for the type or the number of the features, or just for their spatial disposition. Hypotheses are ranked according to their "cost", which reflects their ability to reproduce the observed activity values and their complexity. The larger the number of pharmacophoric features needed to explain the activity differences, the greater the complexity and the cost of the hypothesis. The cost of each pharmacophore model is also compared with the cost of the null hypothesis. The latter is obtained by randomization of the activity values. The larger the difference between the costs, the greater is the significance of the model.

The main advantages of Catalyst/HypoGen are in the ease of visual interpretation of the results, its potential use for either molecule design or 3-D database searches, and for providing quantitative estimations of activities of a test set. However, Catalyst/HypoGen aims to derive a quantitative structure–activity correlation considering only the geometrical disposition of pharmacophoric features. This is rather an oversimplification of the problem. A description in terms of pharmacophoric features is rather crude (that is groups like *n*-butyl and *t*-butyl are both defined as hydrophobic although they are evidently characterized by different shapes), and can only explain large variations of activity. Also, the underlying assumption of HypoGen is that inactive molecules lack potency because they are unable to provide a good match for the pharmacophoric features. However, multiple reasons for inactivity may exist. The simplest example involves a molecule that, in spite of matching the pharmacophore, has an additional bulky group, which clashes with the receptor and reduces its activity. Although the latest version of the software allows the inclusion of exclusion spheres in an attempt to address this problem, a careful selection of weakly active and inactive molecules is needed. Negative information can be very useful but it must be used with caution. Particularly informative in this sense is the inclusion of enantiomers where a change in the configuration leads to a drastic change in activity. Clearly the inactive enantiomer must be unable to properly orient its pharmacophoric groups.

HypoGen is a sophisticated tool, which involves molecular superimposition of all the molecules analyzed, and typically can be applied only to small datasets. However, techniques have been developed to encode 3-D pharmacophoric information while avoiding the need to align molecules, resulting in a more manageable descriptor called the pharmacophoric key. Keys involving 2-, 3-, and even 4-point pharmacophores have been derived [90,92,93]. This operation usually involves the definition of a number of atom types along with a set of distance ranges. A molecule is subjected to conformational analysis and a fingerprint is then generated with each bit representing presence/absence in any molecular conformation of a certain doublet or triplet or quartet (for more details refer to Chapter 7). Pharmacophoric keys may be considered 4-D descriptors, since they also encode the conformational behavior of a molecule. They have been widely applied to diversity analysis for library design and compound acquisition [94,95]. More recently they have been applied to QSAR and pattern recognition studies.

PharmPrint is a method developed at Affymax [96] for the rapid fingerprinting of 3-point pharmacophores. Atom types are identified as hydrogen bond acceptor and donor, with formal positive and negative charge, hydrophobic and aromatic. PharmPrint has been applied to the analysis of sets of human and rat estrogen receptor (ER) ligands. PLS was used to derive linear regressions between fingerprints and binding affinity. A considerable improvement in the results, at least in terms of q^2 (obtained with Leave-One-Out protocol), was achieved by introducing a seventh type, called X, which represents any atom not labelled with any of the first six types. The authors ascribed this improvement to probable extra information about molecular volume. However, given that the inclusion of X increases the number of descriptors from 6726 to over 10459, it would be necessary to apply a more robust cross-validation procedure and to use repeated scrambling of the response variable to check for the presence of chance correlations (see Section 5.8.2.2).

In an attempt to reproduce datasets where activity data are characterized by low accuracy (which typically come out from primary screening), the authors merged 15 ER ligands taken from previously analyzed sets with 750 non-ER compounds randomly selected from a clean version of MDDR (MDL Drug Data Report) database. Active compounds were assigned an activity value of 1.0, while inactive ones were given a value of 0.0. PLS was used to discriminate between the two classes (active compounds were duplicated 50 times to equalize the weights between the two classes). The resulting model was used to predict 8290 non-ER and 250 ER compounds from the MDDR database, which had been previously excluded. A further test set was established considering 85 compounds from the Affymax corporate database, active at the sub-micromolar level, and structurally different from the training set compounds. A satisfactory discrimination between the active and inactive classes was achieved. Given an arbitrary activity cut-off of 0.2, the model was able to correctly classify 89.7% of 8 290 inactive MDDR compounds, 87.4% of 250 active MDDR compounds, and 87.2% of 85 active Affymax compounds.

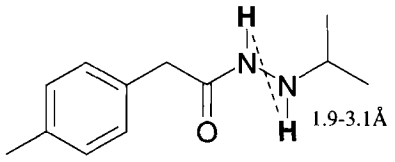
One of the major advantages of using pharmacophoric keys lies in their simple interpretability. PLS weights for the first latent variable were used to assess the relative importance of the pharmacophores. The most positively contributing pharmacophores are illustrated in Figure 7.15 (see Chapter 7) for two structurally different ER compounds. Interestingly, the top ten negatively contributing pharmacophores all contain an X functionality, and their interpretation is therefore difficult.

The strategy outlined in this application represents a nice example of how QSAR can be applied to prioritize screening in the very early phases of a project.

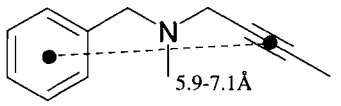
Pharmacophoric keys appear to be particularly valuable when multiple binding modes are expected to occur (and, in this regard, outperform methods requiring molecular superimposition, whose underlying assumption is that a common binding mode must exist). Such a situation is typically present with the results of a primary screening. The application of these descriptors in combination with decision trees (also known as recursive partitioning, see Section 5.8.3.3) appears a promising technique to analyze massive datasets and detect different activity trends.

Two-point pharmacophoric keys have been used with SCAM to analyze a set of 1 650 monoamine oxidase (MAO) inhibitors [97]. SCAM is a recursive partitioning method developed to deal with thousands of binary descriptors. In this application MAO inhibitors were assigned to four different classes according to their potency. SCAM was used to recursively

partition the whole dataset into smaller and more homogenous subsets, until each subset could be no longer split. At each split, molecules were divided into two subsets according to whether or not a particular 2-point pharmacophore was in the structure. The 2-point pharmacophore causing the best splitting was identified according to a Student's *t*-test. The final model was in a form of a tree with each node in the tree assigned to an activity class. In this application, two active nodes were identified which were characterized by considerably different 2-point pharmacophore patterns. Representative compounds for each node are illustrated in Figure 5.5 along with the most significant key features. Compound AL16432 is a hydrazide, which can be hydrolyzed to acetylhydrazine, and act as an irreversible inhibitor. Propargylamine compound AL19120 is a suicide inhibitor itself, which irreversibly inhibits MAO through covalent attachment to its flavin cofactor. This is a demonstration of how SCAM can detect multiple mechanisms of action.



AL16432



AL19120

Figure 5.5. Chemical structures of MAO inhibitors AL19120 and AL16432 along with significant pharmacophoric features found by SCAM. Adapted from reference [97].

In spite of obtaining a good partition of the dataset, it was observed that for one of the active nodes the simultaneous presence in a single conformation of all the 2-point pharmacophores leading to that node was actually impossible. This was a direct consequence of how the pharmacophore key was assembled; *i.e.* a bit is turned on if the corresponding pharmacophore is present in any of the molecular conformations. This is equivalent to performing an OR operation over all the conformations of a given molecule. To overcome such a problem, an improved version of SCAM named SCAMPI [98] has been recently described.

As in SCAM a conformational analysis is performed on each molecule to encode 2-point pharmacophores in a bit string. The most significant 2-point pharmacophore is identified and used to split the dataset. The compounds containing this feature are then subjected to a new conformational analysis using the distance between the two pharmacophoric points as a con-

straint. Now, the presence of 3-point pharmacophores is recorded in a bit-string. These 3-point descriptors retain the 2-points already identified, so that they are only composed of a new chemical feature related to the previous two. The most significant 3-point descriptor is used again to split the data. The process continues finding 4-point, 5-point, etc. until no new pharmacophoric points can be found. Repeating the conformational analysis at each split of the tree dramatically increases the computational requirements of the program. However, two benefits are gained. Firstly, a better exploration of the conformational space is achieved and, secondly, it assures that all the molecules present in a certain subnode do have at least one conformation containing all the pharmacophoric points that lead to that subnode.

The use of decision trees like SCAM has been criticized, as results depend strictly on how the split is performed. Only the single most significant descriptor is chosen. There might be equivalent or slightly less significant splits, which are not explored. As a consequence the optimal tree may be missed. This is analogous to the multiple minimum problem. An obvious extension of the method is to consider multiple splitting or multiple tree generation, and work in this direction is ongoing [98].

Decision trees can be very useful tools for data mining, rapid visualization of SAR in a large dataset and detection of possible different binding modes. They can also be applied to prioritize screening as they can be used for class prediction. For instance, a screening set of compounds can be assayed, the results analyzed with SCAM, and a corporate collection or a virtual library dropped down the tree to suggest additional compounds for biological screening or synthesis. The loop can be repeated for a number of iterations. Sequential screening may well prove to be an efficient replacement of the ultra-high throughput screening of a corporate sample collection [99].

5.7 Conclusions

It is not possible to give an exhaustive description of all the QSAR and pattern recognition methods developed in the past 40 years in a single Chapter. We hope to have given at least an idea of the variety of statistical techniques and molecular descriptors available to modellers and medicinal chemists to analyze Structure–Activity Relationships. In conclusion, it is important to stress that the large number of methods does not make most of them redundant, as each method has its own strengths and limitations (see Chapter 4). The proper choice of technique to analyze a particular problem is the first condition for success. The different stages of the Drug Discovery process are characterized by different time scales, by different levels of information, by the production of chemical and biological data of different natures, and by different objectives. All these considerations should be taken into account in the selection of the method for the analysis of a dataset.

At the beginning of a project, when looking for lead compounds, the level of noise contained in the biological data is high, active molecules may only be at micromolar level and the active/inactive ratio is typically very low. The number of molecules being analyzed is large and molecules can be very heterogeneous. The aim at this stage is to prioritize the screening of new molecules and possibly deliver new chemical series as soon as possible. Alignment-free descriptors encoding 3-D information (as described in Sections 5.4 and 5.6) appear par-

ticularly attractive in this regard. They are certainly computationally more expensive than topological descriptors, however they are more likely to yield structurally different leads. Classification and/or data reduction techniques are typically preferred to regression techniques at this stage, due to inaccuracies within the activity data and to the presence of multiple mechanisms of action.

Once lead molecules have been identified and the explosion/optimization phase begins, different computational methods are needed because the objectives are different. Activity data are, in general, more accurate; and congeneric series of molecules are typically explored. Explosion and optimization are often achieved by sequential synthesis of arrays and small libraries. This might lead to the generation of hundreds of molecules, which makes SAR analysis problematic. Methods employing fragment-based descriptors (as described in Section 5.5), in combination with regression or classification methods, can be usefully applied to highlight fragments contributing to activity in a positive or negative fashion.

More chemical series may be optimized in parallel so that a certain degree of structural diversity is still present. 3-D QSAR techniques like Catalyst/Hypogen (Section 5.6) may be usefully applied to compare different series and detect similarities. It is common practice in medicinal chemistry to transfer a successful chemical group from one series to another. Hypogen could facilitate such an operation, given that a common binding mode exists.

During the lead optimization phase, the primary objective of the analysis is the understanding of those features which determine activity and/or selectivity in order to help in the design of new molecules. Hence molecular descriptors must be as interpretable as possible. In this regard, CoMFA-like methods (as described in Section 5.3) are very valuable (see also Chapter 10). However, quite stringent conditions are needed for their application. All these techniques require the preliminary alignment of the molecules, a task that can be more or less feasible depending on the nature of the molecules (degree of structural diversity and flexibility within the series) and on *a priori* information (availability of X-ray or NMR data). They are also time-consuming and it might be the case that less informative but faster analyses are preferred instead. These might be accomplished using alignment-free 3-D descriptors, 2-D topological and electronic descriptors.

The transition of data from *in vitro* to *in vivo* models imposes new conditions. CoMFA was appositely developed to model specific ligand–receptor interactions and its application when dealing with *in vivo* models, which involve transport and absorption phenomena, may be not appropriate [100]. Potency and selectivity are not enough for a drug candidate; it must posses an appropriate ADME profile to reach the site of action and remain there for a sufficient length of time. The actual aim of the lead optimization process is indeed the discovery of the right balance between potency, selectivity and physicochemical properties. In this regard, global 3-D descriptors, which provide a holistic view of the molecule, as well as classical Hansch analysis, may be usefully applied.

In conclusion, QSAR techniques can be employed at every step in the drug discovery process, from lead generation to candidate selection. In order for QSAR to be effectively applied, the right technique must be selected for the problem in hand. An appreciation of the available methods, their strengths and weaknesses, is essential if QSAR is to aid a medicinal chemistry program. The authors hope that this Chapter will provide the reader with some assistance.

Acknowledgements

Andrew R. Leach (Glaxo Wellcome) and David T. Manallack (Celltech Chiroscience) made useful comments on the manuscript and their contributions are much appreciated.

5.8 Appendix – Statistical Techniques in QSAR and Pattern Recognition

The purpose of this Appendix is to offer a brief description of some of the statistical techniques mentioned throughout the Chapter. This Appendix covers three topics, namely data reduction (unsupervized learning), regression, and classification (supervized learning) methods. It is certainly not a comprehensive list but rather an overview of assessed methods widely applied to model SAR. Mathematical details are avoided where possible; however, for the sake of clarity, some formulas are necessarily shown. In this Appendix the number of objects (molecules) and variables (molecular descriptors) are denoted by n and p, respectively.

5.8.1 Data Reduction and Display

5.8.1.1 Principal Component Analysis

Principal Component Analysis (PCA) [101] is a technique for reducing the number of variables while retaining most of the original information. PCA transforms a data matrix $X(n,p)$ of n objects and p variables into a new matrix $T(n,m)$ where m ranges from 1 to p. The new variables, called principal components (PCs), are linear combinations of the original variables. They are orthogonal to each other, and so they carry independent information. They are generated according to the variance explained, so that the first PC explains the largest amount of variance in the original data, and the last PC the lowest. The elements of T are called *scores* and the coefficients of the linear combinations *loadings*. The new matrix T still retains the whole original information if m equals p. However, if the original variables are highly inter-correlated, the first few components retain most of the information, although the dimensionality of the matrix is greatly reduced ($m<<p$).

Several algorithms have been implemented, which calculate all the PCs at once (matrix diagonalization), or in a step-wise manner (the NIPALS algorithm [102]). The latter is commonly used since it is less time-consuming, especially for very large data matrices. The number of significant PCs can be estimated by one of the following criteria [102]:

- Percentage of cumulative explained variance.
- Eigenvalue-one: only components with eigenvalue (explained variance) greater than 1.0 are considered important.
- Score-plot: the eigenvalues are plotted against the number of components and the point where the slope drops is used as the last significant component.

- Cross-validation: this is probably the most reliable method and takes advantage of the fact that the NIPALS algorithm can tolerate missing values. The elements of the X matrix are distributed into 3–5 groups so that all the rows and all the columns of the matrix are represented within each group. One group is then excluded and a PCA carried out on the remaining elements. Obtained parameters are used to predict the missing elements, and the predictions are compared with the real values. This operation is repeated for each of the groups. The optimal number of components is reached when no significant improvement in the predictions can be achieved.

The interpretation of the PCA results is usually carried out by plotting the scores and loadings for the significant components. The score-plot allows the visualization of complex spatial relationships among the objects. The position of new objects on the plot can be predicted. The loading-plot can help in highlighting the contribution of the original variables in each PC, and detecting existing correlations.

PCA scores are often called Principal Properties (PPs) and can replace the original variables. In this regard they are widely applied in combination with experimental design techniques [103] prior to PLS modelling (Section 5.8.2.2). Also, they can be used to build a regression model with PCR (Section 5.8.2.2).

5.8.1.2 Non-Linear Mapping

Non-Linear Mapping (NLM) [104] allows the display of highly multi-dimensional datasets in a 2-D, or 3-D space. The first step in NLM is the calculation of all the distances d_{ik} between any pair of objects (i,k) in the original n-D space. Object separation is typically evaluated using the Euclidean distance. Then the objects are randomly placed in a 2-D, or 3-D space, and the distances (d'_{ik}) calculated again. The difference between the inter-objects distances in the p-D space and those in the 2-D, or 3-D space is expressed as an error function (EF):

$$EF = \sum_{i>k} \frac{(d'_{ik} - d_{ik})^2}{d'^{\rho}_{ik}} \tag{5.20}$$

The minimization of this error results in a 2-D, or 3-D plot, where the inter-objects distances are as similar as possible to the original distances. The power term ρ serves to alter the emphasis on the relative importance of large *versus* small distances. A good compromise is a value of 2 [105].

An NLM plot is very useful to give an idea of the data distribution. Unlike PCA, NLM preserves the spatial relationships among the objects. However, NLM cannot be used to predict the position of new objects, as each axis of the plot is an unknown, non-linear combination of the original variables. Furthermore, the projection to a 2-D, or 3-D map only makes sense if a major percentage of the variance is contained in so few dimensions.

5.8.1.3 Neural Networks

Two types of Neural Networks (NN) techniques are mostly applied to produce 2-D, or 3-D displays from high-dimensional data: Kohonen and ReNDeR networks. As in the NLM technique, the idea is to compress the information contained in a dataset in order to display it, while trying to preserve the original relationships among the data (topology).

A Kohonen network [106] is based on one layer of neurons arranged on a 2-D plane. Each neuron is connected to each of the inputs (molecular descriptors) via a connection weight. The objective is to map in that plane a set of multi-dimensional objects, so that similar objects are mapped in close positions (in NN terminology: similar objects excite close neurons).

The essential steps in the training of a Kohonen network are as follows:

1. random m weights (where m is the number of original variables) are assigned to each neuron,
2. the input signal (first object) is passed through the network,
3. a response is calculated for all the neurons to find which neuron has the weights most similar to the input signal (this is called winner neuron),
4. the weights of the winner neuron are corrected to better approximate the signal in the next cycle,
5. the weights of its neighbor neurons are corrected as well, scaled depending on their distance from the winner neuron, and
6. steps 2–5 are repeated for all the objects.

The training procedure is repeated for a number of cycles sufficient to stabilize the weights. At the end of the training, each neuron is sensitized to a particular region of the original space. Therefore, when new objects are presented to the trained network, they will be mapped onto the neurons corresponding to the most similar objects in the training set.

A ReNDeR (Reversible Nonlinear Dimension Reduction) Network [107] is a feed-forward back-propagation network, containing an input layer, an output layer, and a few (4–5) hidden layers of neurons. Whilst the input and output layers contain as many neurons as the input parameters, the hidden layers contain only a few (2–3) neurons (for example 7:3:2:3:7, for seven parameters in a network with three hidden layers). Once the network has been trained, the output from the central layer (2) can be used to plot the objects in a 2-D space.

A detailed description of Neural Networks applied to QSAR and other chemistry problems can be found in references [108] and [109].

5.8.2 Regression Techniques

5.8.2.1 Multiple Linear Regression

Multiple Linear Regression (MLR) expresses a single dependent variable (y) as a linear combination of multiple independent variables (x):

$$y = ax_1 + bx_2 \ldots + k \tag{5.21}$$

where a, b are the coefficients of the regression, and k is a constant. The regression model can be built in a stepwise manner: starting with all the x_j and, at each step, removing the least contributing variable (backward), or starting with the variable which is most correlated to y and adding up variables until no more improvement is achieved (forward).

A number of statistical parameters are used to evaluate regression models. The overall fit of the model is given by r^2:

$$r^2 = \frac{\sum_{i=1}^{n}(y_{i,calc} - y_{i,mean})^2}{\sum_{i=1}^{n}(y_{i,obs} - y_{i,mean})^2} = 1 - \frac{\sum_{i=1}^{n}(y_{i,calc} - y_{i,obs})^2}{\sum_{i=1}^{n}(y_{i,obs} - y_{i,mean})^2} \tag{5.22}$$

The r^2 coefficient can vary from 0 (none of the variance associated with y is explained by the model) to 1 (all the experimental variance is explained by the model). The statistical significance of the model is measured by the F value:

$$F = \frac{n-p-1}{p} \frac{\sum_{i=1}^{n}(y_{i,calc} - y_{i,mean})^2}{\sum_{i=1}^{n}(y_{i,obs} - y_{i,calc})^2} \tag{5.23}$$

The larger the F value, the greater the significance of the model. In particular F must be larger than tabulated F values with p and $(n–p–1)$ degrees of freedom at a chosen confidence level (for instance 95%). The significance of each variable can be assessed by the t value:

$$t = \frac{c}{SE_c} \tag{5.24}$$

where c is the regression coefficient, and SE_c its corresponding standard error. The t parameter must be larger than tabulated t values with $(n–p–1)$ degrees of freedom.

Good statistics is a necessary condition but not sufficient for a meaningful regression model. Especially when increasing the number of variables, the number of possible models becomes larger and the risk of a chance correlation increases as well. Chance effects have been investigated on sets of random numbers [110] and it has been shown that the higher the ratio of variables to the number of objects, the greater the risk of chance correlation. For example, given a dataset of ten objects, the combination of five variables can correlate with random "activities" producing r^2 superior to 0.5. For medium-size datasets (n less or equal to 30), having at least 5–6 objects for each variable has been suggested to avoid chance correlation [111]. Finally, MLR is based on a number of assumptions about the dependent variable y (the errors on y are randomly distributed and roughly the same size) as well as on x (predictor variables are independent and error-free). In particular, the above conditions are generally not satisfied for datasets where the number of variables largely exceeds the number of objects, making MLR inappropriate.

5.8.2.2 Principal Component Regression and Partial Least Squares

Unlike MLR, Principal Component Regression (PCR) and Partial Least Squares (PLS) can be applied to datasets characterized by large numbers of descriptors and low numbers of objects. Both rely on the assumption that all the descriptors can be seen as a combination of a small number of intrinsic variables (called principal components in PCR and latent variables in PLS) plus some errors, and both are aimed at extracting this relevant information from the original descriptor matrix X and correlating it to the biological activity Y [112].

The PCR method accomplishes this task step-wise by:

1. executing a PCA on the X matrix and saving the scores,
2. selecting the optimal number p of components (based on explained variance), and
3. using the first p PCA scores of X to build a regression model with Y.

Because the PCA scores and the regression coefficients are calculated independently, variables important for explaining the biological response may have already been removed at the regression stage.

The two steps (PCA and regression analysis) can be effectively combined by using the PLS method. PLS is aimed at finding linear combinations of the descriptors (latent variables) that not only approximate the original matrix X, but simultaneously correlate with the biological activity Y. Latent variables (LVs) retain the same properties of PCs in the sense that they are linear combinations of the original variables and they are an orthogonal set, but they differ because LVs are built maximizing the covariance between X and Y.

As with PCA, plots of the scores and coefficients of the linear combinations can be generated, and they help the interpretation of the model and the identification of outliers, as well as non-linear relationships.

Regression coefficients in terms of original variables can also be computed, so that the PLS solution can be still reported in the traditional form (Eq. 5.21). Unlike MLR, PLS can simultaneously handle more activities. In this case, a different set of regression coefficients is produced for each activity. More details on PLS mathematical treatment can be found in the appendix to reference [112].

The optimal number of LVs of a PLS model is usually estimated by *cross-validation* (CV). CV means that the objects are divided in n groups, a model is derived with n–1 groups and used to predict the excluded group of objects. This is repeated until all groups have been excluded one at a time. The q^2 value is calculated from the predictions as follows:

$$q^2 = 1 - \frac{\sum_{i=1}^{n} (y_{i,pred} - y_{i,obs})^2}{\sum_{i=1}^{n} (y_{i,obs} - y_{i,mean})^2} \tag{5.25}$$

The above formula is analogous to r^2 (Eq. 5.22), where calculated y values are replaced by predicted values. Other statistical parameters commonly calculated are the *SDEP* and the s_{PRESS}:

$$SDEP = \sqrt{\frac{\sum_{i=1}^{n} (y_{i,pred} - y_{i,obs})^2}{n}} \tag{5.26}$$

$$s_{PRESS} = \sqrt{\frac{\sum_{i=1}^{n} (y_{i,pred} - y_{i,obs})^2}{n - a - 1}} \tag{5.27}$$

where a is the number of LVs. The optimal number of LVs corresponds to the highest q^2 or to the lowest *SDEP*. However, as the number of LVs increases, PLS suffers the same limitations as MLR; hence the number of latent variables should be kept as small as possible. As a rule of thumb, a new LV is added only if it leads to an increment of at least 5% in the q^2. Alternatively the s_{PRESS} can be used, because it does take into account the number of LVs and penalizes high-dimensional models.

The simplest CV protocol is the leave-one-out (LOO), where one object at time is removed and predicted. A more robust and reliable protocol consists of removing 20–25% of the objects at each time. This means dividing the objects into five or four groups. Usually this operation is performed randomly and the final q^2 is highly dependent on the initial assignation. Hence it is strongly recommended to repeat the whole protocol a large number of times (100), and compute mean as well as standard deviation values associated with q^2, *SDEP* and s_{PRESS}. This procedure can be rather time-consuming. A good compromise between speed and robustness is to divide the object in 4–5 groups according to their activity, so that each group contains active, medium, and low-activity compounds.

CV is used to estimate the actual predictive ability of a model. In this regard, it can be used not only with PLS but with any regression or classification methods. Ideally, cross-validated and not-cross-validated parameters (for instance q^2 and r^2) should be very close. A large difference may indicate overfitting problems or the presence of outliers.

PLS is typically applied to datasets where the ratio of variables to the number of objects is high (for instance CoMFA). Hence chance correlations may well occur, in particular when variable selection methods, aimed at the mere optimization of r^2 or q^2, are applied. An easy way to detect chance correlations is to perform the scrambling of Y. Activity values are randomly permuted so that each value is no longer assigned to the correct molecule (the correlation between the original Y vector and the new scrambled Y vector should be kept low). A cross-validated model is then run and the scrambled q^2 calculated. The whole procedure is repeated for a number of times. Ideally all the scrambled q^2 values should be significantly lower than the value obtained with the original Y. As for CV, Y-scrambling can be applied with any regression or classification methods. Its use is particularly recommended when dealing with poorly interpretable descriptors.

Finally PLS, when used in prediction, provides a rapid evaluation of how similar test set molecules are to those in the training set, and hence of how reliable the predictions may be. This method is based on a comparison between the unexplained variance (s_{new}^2) of the molecule to be predicted and the unexplained variance of the training set (s_0^2) by means of an *F*-test [112]:

$$F = \frac{{s_{new}}^2}{{s_0}^2} \tag{5.28}$$

If F is larger than a critical value with approximately $(n-a)$ and $(n-a)^2$ degrees of freedom (being typically $p>>n$), the molecule to be predicted is judged to be dissimilar to the training set. Hence, the prediction by the PLS model cannot be considered reliable.

5.8.3 Classification Techniques

5.8.3.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [113] is a technique for separating categories of objects, based on a combination of the independent variables called discriminant function. The simplest case is when all the objects belong to two categories: active and inactive. The discriminating function W is obtained as a linear combination of the independent variables, as in regression analysis:

$$W = \sum_{j=1}^{p} a_j \times x_j \tag{5.28}$$

where x_j are the independent variables and a_j the coefficients of the linear combination. The discriminant parameters a_i are defined so that active compounds have $W>W_0$ and inactive ones have $W<W_0$, where W_0 is called critical value.

The coefficients can be determined in a number of different ways. Typically they are derived so as to maximize the ratio of the inter-category sum of squared distances to the intra-category sum of squared distances (that is, maximizing variance between the categories and minimizing variance within the individual categories). Problems may arise when one category is embedded within the other, so that no linear separation is possible.

Results of LDA are usually expressed as percentages of compounds correctly and incorrectly classified. Cross-validation can be applied, as in the case of regression analysis, to test the predictive ability of the model. The critical value is used to classify new objects. If the calculated value of the discriminant function for an object is lower than the critical value, the object is assigned to the inactive category, if higher to the active one.

LDA is also able to handle more than two categories; up to k–1 discriminant functions can be necessary to describe k categories.

As in regression analysis, the probability of achieving a separation simply by chance increases with the number of variables. The recommendation is to keep the ratio of data points to descriptor variables equal to at least three. Also a skewed distribution of objects may be a problem. Ideally the categories should be equally populated. Otherwise, trivial separations may be achieved, despite the fact that the above-specified ratio is greater than three [114].

5.8.3.2 Soft Independent Modelling of Class Analogy

Soft Independent Modelling of Class Analogy (SIMCA) is aimed at discriminating between objects belonging to different categories by means of separate mathematical models. It can

be usefully applied when the response variable is categorical rather then continuous (for instance high, medium, low activity) and when the relationship between structure and activity is expected to be non-linear. SIMCA was originally developed by Wold [115], and has been extensively applied in process applications and analytical chemistry. SIMCA classifies the objects by using PCA. A separate PCA is performed for each category of activity, using cross-validation to determine the number of components necessary to distinguish among the categories. The principal components for each category define a hyper-volume in which the category members lie.

As in PCA, the analysis of SIMCA loadings helps the interpretation of the models. The matrix of residuals (unexplained variance) within each category is also used to derive important parameters for each descriptor, namely the Modelling Power and the Discriminating Power. The former describes the fraction of variance of the descriptor, explained by the PCA components for each category. A value near zero indicates that that variable is irrelevant and may be removed from the description matrix. Discriminating Power is a measure of the ability of a descriptor to separate categories from each other. It is computed as the ratio of the sum of squared residuals for the descriptor when the objects are fitted to all categories apart from their true category, to the sum of squared residuals when they are fitted to the true category. The larger this ratio, the better the descriptor at differentiating categories.

SIMCA can also be used in prediction. Each new object is projected into each category component space. If the obtained point falls within one of the defined volumes, it belongs to the corresponding category. The object can belong to more than one category (if their volumes overlap) as well as to an unknown category (if it does not fall in any of the volumes). In this regard considerable differences exist between different implementations of SIMCA, and have been highlighted by Hunt in reference [83].

5.8.3.3 Recursive Partitioning

Recursive Partitioning (RP), also known as decision tree analysis, is a rather intuitive technique aimed at discovering logical patterns within datasets. Well known implementations of RP include, among the others, CART [116], C4.5 [117], FIRM [118] and the above mentioned SCAM [97]. Given a set of data characterized by a number of descriptors and belonging to different categories, the goal of RP is the derivation of a set of rules based on the descriptors (for instance $x_j > k$, where k is a constant), which correctly categorizes as many observations as possible. The rules are derived by successive partitioning of the data into subsets, each partition being accomplished on the basis of the values of a single descriptor only. In other words, the initial dataset is split into two subsets, called nodes, which in turn are further split into two subnodes. The process continues until there is no meaningful way to further split the data, and terminal nodes are reached. The terminal nodes may contain objects belonging to one category only, or may be characterized by a level of purity (that is the ratio of objects in the node belonging to a category to the objects in the node) higher than a user-defined threshold. Such nodes would be called leaves and assigned to a category. Also, terminal nodes may contain a number of objects, which is considered too small for further splitting, or there may be no evidence for a further statistically valid split. These nodes would be left unassigned.

The number of possible splitting rules is extremely high even for small datasets with few descriptors. The exhaustive exploration of all possible combinations is therefore not feasible, and a number of algorithms have been developed to deal with this problem. Most methods employ, at each step, the rule that maximizes some local measure of the progress in the partitioning, and avoids exploring the consequences of alternative choices. The method used to identify the best splitting rule at each step is therefore of crucial importance, and different scoring systems have been developed. For instance, the so-called *Gini Impurity* score is aimed at minimizing the impurity of the resultant nodes. As a consequence, a node is often split into a pure node with very few examples and a larger impure node. In contrast, the *Twoing* score tends to generate nodes of similar size and the model looks more balanced [116,119].

The relative population of each category is a key issue with RP and will influence the results. Appropriate scaling can be applied to equalize the populations, making the cost of misclassifying samples from categories with few examples higher than the cost of misclassifying samples from larger populated categories [119].

The output of a RP analysis is in the form of a tree diagram or dendrogram. Tree diagrams are in general easy to interpret, if meaningful descriptors are chosen. For instance, rules which lead to leaves containing mainly active molecules can provide a clue to the key features associated with activity. However, depending on the data and the input parameters, complex, deep trees may well be generated. Increasing model complexity, as in the case of regression analysis, provides a better fit for the data, but it does not necessarily help predictivity and interpretability. A crude approach to avoid overfitting is thus to restrict the number of nodes which are allowed in the tree. However, more robust techniques have been developed to obtain less complex and more interpretable models (*pruning* operation). Among the others, cost-complexity [116] consists of selecting a test set, which can be either an external set or a random selection of the dataset. Trees of decreasing size from the original are then generated and used to predict the test set. The tree providing the highest accuracy is finally selected. Cross-validation may also be usefully applied to define the optimal depth of the tree [119].

References

[1] C. Hansch, P. P. Maloney, T. Fujita, *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.
[2] C. Hansch, A. Leo, *Exploring QSAR Fundamentals and Applications in Chemistry and Biology, Vol. 1*, ACS, Washington, DC **1995**.
[3] C. Hansch, A. Leo, L. Zhang, *Comprehensive Quantitative Structure–Activity relationships*: C-QSAR, Pomona College, Claremont, CA, USA **1992**.
[4] H. Kubinyi, *QSAR: Hansch Analysis and Related Approaches*, VCH Publishers, New York **1993**, pp. 21–55.
[5] C. Hansch, A. Leo, Exploring QSAR Fundamentals and Applications in Chemistry and Biology, Vol. 2, ACS, Washington, DC 1995.
[6] A. Leo, CLOGP, Daylight Chemical Information Systems, Irvine, USA.
[7] A. Leo, CMR, Daylight Chemical Information Systems, Irvine, USA.
[8] ACD/Labs 4.0, Advanced Chemistry Development Inc., Toronto, Canada.
[9] P. Ertl, *Quant. Struct-Act. Relat.* **1997**, *16*, 377–382.
[10] R. D. Gilliom, J.-P. Beck, W. P. Purcell, *J. Comput. Chem.* **1985**, *6*, 437–440.
[11] J. S. Murray, P. Politzer, *J. Org. Chem.* **1991**, *56*, 6715–6717.
[12] J. C. Dearden, T. Ghaufurian, in *QSAR and Molecular Modelling: Concepts, Computational tools,*

and Biological Applications, F. Sanz, J. Giraldo, F. Manaut (Eds.), Prous Science Publishers, Barcelona **1995**, pp. 117–119.
[13] H. Van de Waterbeemd, B. Testa, P.-A. Carrupt, N. El Tayar, *Prog. Clin. Biol. Res.* **1989**, *291*, 123–126.
[14] H. Van de Waterbeemd, N. El Tayar, P. -A. Carrupt, B. Testa, *J. Comput.-Aided Mol. Design* **1989**, *3*, 111–132.
[15] M. Sjoestroem, S. Wold, *J. Mol. Evol.* **1985**, *22*, 272–277.
[16] J. Jonsson, L. Eriksson, S. Hellberg, M. Sjoestroem, S. Wold, *Quant. Struct.-Act. Relat.* **1989**, *8*, 204–209.
[17] H. Kubinyi, *QSAR: Hansch Analysis and Related Approaches*, VCH Publishers, New York **1993**, p. 21.
[18] B. H. W. Hamilton, D. F. Ortwine, D. F. Worth, E. W. Badger, J. A. Bristol, R. F. Bruns, S. J. Haleen, R. P. Steffen, *J. Med. Chem.* **1985**, *28*, 1071–1079.
[19] H. Kubinyi, *QSAR: Hansch Analysis and Related Approaches*, VCH Publishers, New York **1993**, p. 3.
[20] S. Yee, *Pharm. Res.* **1997**, *14*, 763-766.
[21] A. Karlen, S. Winiwarter, M. Bohnam, H. Lennernas, A. Hallberg, Correlation of Intestinal Drug Permeability in Humans (in vivo) with Experimentally and Theoretically Derived Parameters, Molecular Modelling and Prediction of Bioactivity, *Abstract book 12th European Symposium on QSAR*, Copenhagen 1998.
[22] C. A. Lipinsky, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug. Deliv. Rev.* **1997**, *23*, 3–25.
[23] D. E. Clark, *J. Pharm. Sci.* **1999**, *88*, 807–814.
[24] M. H. Abraham, H. S. Chadha, R. C. Mitchell, *J. Pharm. Sci.* **1994**, *83*, 1257–1268.
[25] D. E. Clark, *J. Pharm. Sci.* **1999**, *88*, 815–821.
[26] R. D. Cramer, D. E. Patterson, J. D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
[27] M. Baroni, G. Costantini, G. Cruciani, D. Riganelli, R. Valigi, S. Clementi, *Quant. Struct-Act. Relat.* **1993**, *12*, 9–20.
[28] M. Pastor, G. Cruciani, S. Clementi, *J. Med. Chem.* **1997**, *40*, 1455–1464.
[29] R. T. Kroemer, P. Hecht, *J. Comput.-Aided Mol. Design* **1995**, *9*, 205–212.
[30] P. W. Kenny, *J. Chem. Soc. Perkin Trans.* **1994**, *2*, 199–202.
[31] G. Klebe, U. Abraham, *J. Med. Chem.* **1993**, *36*, 70–80.
[32] U. Norinder, in *3D QSAR in Drug Design: Recent Advances*, H. Kubinyi, G. Folkers, Y. C. Martin (Eds.), Kluwer Academic Publishers, The Netherlands **1998**, pp. 23–39.
[33] R. T. Kroemer, P. Hecht, S. Guessregen, K. R. Liedl, in *3D QSAR in Drug Design: Recent Advances*, H. Kubinyi, G. Folkers, Y. C. Martin (Eds.), Kluwer Academic Publishers, The Netherlands **1998**, pp. 41–56.
[34] P. J. Goodford, *J. Med. Chem.* **1985**, *28*, 849–857.
[35] G. Cruciani, K. A. Watson, *J. Med. Chem.* **1994**, *37*, 2589–2601.
[36] J. Nilsson, H. Wikstroem, A. Smilde, S. Glase, T. Pugsley, G. Cruciani, M. Pastor, S. Clementi, *J. Med. Chem.* **1997**, *40*, 833–840.
[37] G. Klebe, in *3D QSAR in Drug Design: Recent Advances*, H. Kubinyi, G. Folkers, Y. C. Martin (Eds.), Kluwer Academic Publishers, The Netherlands **1998**, pp. 87–104.
[38] M. Bohm, J. Sturzebecher, G. Klebe, *J. Med. Chem.* **1999**, *42*, 458–477.
[39] H. Matter, W. Schwab, *J. Med. Chem.* **1999**, *42*, 4506–4523.
[40] D. D. Robinson, P. J. Winn, P. D. Lyne, W. G. Richards, *J. Med. Chem.* **1999**, *42*, 573–583.
[41] M. Hahn, *J. Med. Chem.* **1995**, *38*, 2080–2090.
[42] M. Hahn, D. Rogers, *J. Med. Chem.* **1995**, *38*, 2091–2112.
[43] M. Hahn, D. Rogers, in *3D QSAR in Drug Design: Recent Advances*, H. Kubinyi, G. Folkers, Y. C. Martin (Eds.), Kluwer Academic Publishers, The Netherlands **1998**, 117–133.
[44] S. J. Cho, D. Cummins, J. Bentley, C. W. Andrews, A. Tropsha, *J. Comput.-Aided Mol. Design*, submitted for publication.
[45] S.-S. So, M. Karplus, *J. Comput.-Aided Mol. Design* **1999**, *13*, 243–258.
[46] R. E. Wilcox, T. Tseng, M.-Y. K. Brusniak, B. Ginsburg, R. S. Pearlman, M. Teeter, C. DuRand, S. Starr, K. A. Neve, *J. Med. Chem.* **1998**, *41*, 4385–4399.
[47] G. Bravi, E. Gancia, P. Mascagni, M. Pegna, R. Todeschini, A. Zaliani, *J. Comput.-Aided Mol. Design* **1997**, *11*, 79–92.
[48] C. Mattos, D. Ringe, in *3D QSAR in Drug Design: Theory, Methods and Applications*, H. Kubinyi (Ed.), ESCOM Science Publishers B. V., The Netherlands **1993**, pp. 226–254.
[49] D. J. Smith, M. H. Kremer, M. Luo, G. Vriend, E. Arnold, G. Kamer, M. G. Rossmann, M. A. McKinlay, G. D. Diana, M. J. Otto, *Science* **1986**, *233*, 1286–1293.

[50] J. Badger, I. Minor, M. A. Oliveira, T. J. Smith, M. G. Rossmann, *Proteins Struct. Funct. Genet.* **1989**, *6*, 1–19.
[51] Y. C. Martin, M. G. Bures, E. A. Danaher, J. DeLazer, I. Lico, P.A. Pavlik, *J. Comput.-Aided Mol. Design* **1993**, *7*, 83–102.
[52] D. Barnum, J. Greene, A. Smellie, P. Sprague, *J. Chem. Inf. Comput. Sci.* **1999**, *36*, 525–533.
[53] M. Harel, I. Schalk, L. Ehret-Sabatier, F. Bouet, M. Goeldener, C. Hirth, P. H. Asselsen, I. Silman, J. L. Sussman, *Proc. Natl. Acad. Sci.* **USA 1993**, *90*, 9031–9035.
[54] R. Todeschini, M. Lasagni, E. Marengo, *J. Chemometrics* **1994**, *8*, 263–272.
[55] B.D. Silverman, D. E. Platt, *J. Med. Chem.* **1996**, *39*, 2129–2140.
[56] R. Todeschini, P. Gramatica, R. Provenzani, E. Marengo, *Chemom. Intell. Lab. Systems* **1995**, *27*, 221–229.
[57] R. Todeschini, P. Gramatica, *Quant. Struct-Act. Relat.* **1997**, *16*,113–119.
[58] G. Bravi, J. H. Wikel, *Quant. Struct-Act. Relat.* **2000**, *19*, in press.
[59] A. Zaliani, E. Gancia, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 525–533.
[60] G. Bravi, J. H. Wikel, *Quant. Struct-Act. Relat.* **2000**, *19*, in press.
[61] S. Ekins, G. Bravi, B. J. Ring, T. A. Gillespie, J. S. Gillespie, M. Vandenbranden, S. A. Wrighton, J. H. Wikel, *J. Pharmacol. Exp. Ther.* **1999**, *288*, 21–29.
[62] E. Gancia, G. Bravi, P. Mascagni, A. Zaliani, *J. Comput.-Aided Mol. Design* **2000**, *14*, 293-306.
[63] A. Poso, R. Juvonen, J. Gynther, *Quant. Struct.-Act. Relat.* **1995**, *14*, 507–511.
[64] G. Moreau, P. Broto, *Nouv. J. Chim.* **1980**, *4*, 757–764.
[65] M. Wagener, J. Sadowski, J. Gasteiger, *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.
[66] A. Teckentrup, H. Briem, J. Gasteiger, Analysing HTS Data: an Approach Using Neural Networks, *Abstract Book 5th International Conference on Chemical Structure*, Noordwijkerhout, The Nertherlands, **1999**.
[67] A. M. Ferguson, T. W. Heritage, P. Jonathan, S. E. Pack, L. Phillips, J. Rogan, P. J. Snaith, *J. Comput.-Aided Mol. Design* **1997**, *11*, 143–152.
[68] D. B. Turner, P. Willett, A. M. Ferguson, T. W. Heritage, *J. Comput.-Aided Mol. Design* **1997**, *11*, 409–422.
[69] R. Bursi, T. Dao, T. van Wijk, M. de Gooyer, E. Kellenbach, P. Verwer, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 861–867.
[70] HQSAR, Tripos Inc., St. Louis, USA.
[71] C. A. James, D. Weininger, *Daylight Theory Manual*, Daylight Chemical Information Systems Inc., **1995**.
[72] M. Seel, D. B. Turner, P. Willett, *Quant. Struct-Act. Relat.* **1999**, *18*, 245–252.
[73] S. M. Free, J. W. Wilson, *J. Med. Chem.* **1964**, *7*, 359–399.
[74] T. Fujita, T. Ban, *J. Med. Chem.* **1971**, *14*, 148–152.
[75] H. Kubinyi, in *Comprehensive Medicinal Chemistry, Vol. 4, Quantitative Drug Design*, C. Hansch, P. G. Sammes, J. B. Taylor, C. A. Ramsden (Eds.), Pergamon Press, USA **1990**, pp. 589–643.
[76] M. Higginbottom, C. Kneen, G. S. Ratcliffe, *J. Med. Chem.* **1992**, *35*, 1572–1577.
[77] I. C. Muszynsky, L. Scapozza, K.-A. Kovar, G. Folkers, *Quant. Struct-Act. Relat.* **1999**, *18*, 342–353.
[78] K.-J. Schaper, *Quant. Struct-Act. Relat.* **1999**, *4*, 354–360.
[79] N. E. Shemetulskis, D. Weininger, C. J. Blankley, J. J. Yang and C. J. Humblet, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862–871.
[80] R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
[81] R. Nilakantan, N. Bauman, J. S. Dixon, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82–85.
[82] S. K. Kearsley, S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley, R. P. Sheridan, *J. Chem. Inf. Comput. Sci.* 1996, 36, 118–127.
[83] P. Hunt, *J. Comput.-Aided Mol. Design*, **1999**, *13*, 453–467.
[84] L. H. Hall, L. B. Kier, in *Reviews in Computational Chemistry Vol. 2*, K. B. Lipkowitz, D. D. Boyd (Eds.), VCH Publishers, USA **1991**, pp. 367–422.
[85] L. H. Hall, L. B. Kier, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.
[86] S. J. Cho, W. Zheng, A. Tropsha, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 259–268.
[87] P. Labute, *Pac. Symp. Biocomput. '99* **1999**, 444–455.
[88] H. Gao, C. Williams, P. Labute, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164–168.
[89] O. F. Guner, M. Hahn, H. Li, M. Hassan, 2D versus 3D similarity: Use of molecular shape-based 3D searching techniques for identifying novel compounds, *Book of Abstracts 213th ACS National Meeting*, San Francisco, USA **1997**.

[90] R. P. Sheridan, M. D. Miller, D. J. Underwood, S. K. Kearsley, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
[91] Catalyst, Molecular Simulations Inc., San Diego, USA.
[92] E. K. Davies, in *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery*, American Chemical Society USA **1996**, pp. 309–316.
[93] J. S. Mason, I. Morize, P. R. Menard, D. L. Cheney, C. Hulme, R. F. Labaudiniere, *J. Med. Chem.* **1999**, *42*, 3251–3264.
[94] A. C. Good, R. A. Lewis, *J. Med. Chem.* **1997**, *40*, 3926–3936.
[95] J. S. Mason, S. D. Pickett, in *Computational Methods for the Analysis of Molecular Diversity*, P. Willet (Ed.), Kluwer Academic Publishers, The Netherlands **1997**, pp. 85–114.
[96] M. J. McGregor, S. M. Muskal, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569–574.
[97] X. Chen, A. Rusinko III, S. S. Young, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1054–1062.
[98] X. Chen, A. Rusinko III, A. Tropsha, S. S. Young, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 887–896.
[99] S. S. Young, Analysis of Large, High-Throughput Screening Data Using Recursive Partitioning, *Abstract book 12th European Symposium on QSAR*, Copenhagen **1998**.
[100] H. Kubinyi, *QSAR: Hansch Analysis and Related Approaches*, VCH Publishers, New York **1993**, p. 171.
[101] S. Wold, K. Esbensen, P. Geladi, *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.
[102] M. Otto, *Chemometrics*, Wiley-VCH, Weinheim **1999**, p. 125.
[103] S. Clementi, G. Cruciani, M. Baroni, G. Costantino, in *3D QSAR in Drug Design: Theory, Methods and Applications*, H. Kubinyi (Ed.), ESCOM Science Publishers B. V., The Netherlands **1993**, pp. 567–582.
[104] B. R. Kowalski, C. F. Bender, *J. Am. Chem. Soc.* **1973**, *95*, 686–693.
[105] D. Livingstone, *Data Analysis for Chemist*, Oxford University Press, Oxford **1995**, p. 85.
[106] T. Kohonen, *Proc. IEEE* **1990**, *78*, 1464–1480.
[107] D. J. Livingstone, in *Neural Networks in QSAR and Drug Design*, J. Devillers (Ed.), Academic Press, London **1996**, pp. 157–176.
[108] J. Zupan, J. Gasteiger, *Neural Networks for Chemists*, VCH Publishers, Weinheim **1993**.
[109] J. Devillers (Ed.) *Neural Networks in QSAR and Drug Design*, Academic Press, London **1996**.
[110] J. G. Topliss, R. P. Edwards, *J. Med. Chem.* **1979**, *22*, 1238–1244.
[111] H. Kubinyi, *QSAR: Hansch Analysis and Related Approaches*, VCH Publishers, New York **1993**, p. 59.
[112] S. Wold, E. Johansson, M. Cocchi, Marina, in *3D QSAR in Drug Design: Theory, Methods and Applications*, H. Kubinyi (Ed.) ESCOM Science Publishers B. V., The Netherlands **1993**, pp. 523–550.
[113] M. Otto, *Chemometrics*, Wiley-VCH, Weinheim **1999**, p. 162.
[114] D. Livingstone, *Data Analysis for Chemist*, Oxford University Press, Oxford **1995**, p. 149.
[115] S. Wold, M. Sjostrom, in *Chemometrics: Teory and Application*, B. R. Kowalsky (Ed.), ACD Symposium Series n. 52, American Chemical Society, USA **1977**, pp. 243–282.
[116] L. Breiman, J. H. Freidman, R. A. Olshen, C. J. Stone, *Classification and Regression Trees*, Wadsworth, New York **1984**.
[117] J. R. Quinlan, *C4.5 Programs for Machine Learning*, Morgan Kaufman Publishers, San Mate **1992**.
[118] FIRM, Formal Inference-Based Recursive Modelling, University of Minnesota, St. Paul, USA, **1995**.
[119] CSAR, Molecular Simulations Inc., San Diego, USA.

Virtual Screening for Bioactive Molecules
Edited by Hans-Joachim Böhm and Gisbert Schneider

6 Database Profiling by Neural Networks

Jens Sadowski

6.1 "Drug-Likeness": a General Compound Property?

When combinatorial chemistry and high-throughput screening started to produce an avalanche of chemical compounds and biological results, computational chemistry and molecular modelling were faced with the problem of handling much larger sets of chemical structures than before. With the problem of reducing large databases to reasonably small subsets (e.g. for compound purchase, or for selection from HTS hit lists) without an explicit target but with the intention to test the compounds in different biological assays, computational chemistry first came up with answers to questions like:

- Which subset fills most effectively "holes" in an existing in-house inventory?
- Which subset of compounds spans the most diverse chemical space?

These questions are closely related to similarity and diversity as inter-compound relationships. In recent years, many methods and computer programs became available for such applications (for a review, see [1]). There are extensive studies on the value of several descriptor types and clustering methods available [2]. Evidence was found that a diversity-driven selection procedure can indeed enrich compound sets with compounds that interact with diverse biological targets [3].

The inherent problem in diversity selection is that there is no direct link to biological activities. At the very lower end of a diversity scale where compounds are rather similar this link is obvious: similar compounds should act similarly. The more similar structures are the more similar activities are found. This is the well-known rational in all structure–activity relationships. Far away at the high-diversity end of the scale, the situation becomes quite different. There is nothing like "some compounds are more diverse than others". They are simply different. On the other hand, there is of course a number of other criteria that should also be covered when selecting compounds. It is obvious, that "drug-likeness" is related to biological, chemical, and physical properties like activity, reactivity, synthesizability, bioavailability, and toxicity. In a recent study [4] Lipinski demonstrated the importance of solubility and showed approaches for assessing this problem by computational filters.

As mentioned above, there is a bundle of additional compound properties that cannot be assessed easily by computational procedures. Experienced medicinal chemists have a complex feeling for suitable structures. Sometimes there are certain reactive or otherwise explicitly unwanted substructures that can be used to exclude compounds [5] (see also Section 2.3). But often the chemists cannot even name substructures or rules that discriminate between drugs and non-drugs. It seems to be more like a gut feeling for toxicity, mutagenicity, or preferable properties. This is a strong hint that there is such a general compound property as

"drug-likeness" which cannot be correlated with certain physicochemical properties or substructures. The knowledge behind this – the experience of the medicinal chemist – should nowadays be implicitly contained in databases of drugs and basic chemicals. Such databases are very large collections of the same examples of drugs and non-drugs that a medicinal chemist saw over many years. Very recently, the idea to use such databases for the construction of computational filters that recognize and rank the drug-likeness of chemical compounds was almost simultaneously realized by several groups [6–8]. In the following Section, the principles of such methods will be explained and their validity will be demonstrated by a number of examples.

6.2 Methods and Programs

6.2.1 Databases

Knowledge about drugs and non-drugs in the form of many examples is implicitly contained in large public databases. It is obvious that databases like World Drug Index (WDI) [9], Comprehensive Medicinal Chemistry (CMC) [10], or MACCS Drug Data Report (MDDR) [11] are good sources of drugs and drug-like molecules. Of course, one could discuss that many of these compounds never got to the market, or that some of the drug-classes like cytostatics seem not to be typical drugs. But at least for the vast majority of these compounds, one can assume that they were designed and synthesized by medicinal chemists who intended to make drugs.

It is a bit more complicated to find an analogy for the non-drugs. A practical approach is to use collections of available chemicals or general collections of organic compounds like ACD [12] or SPRESI [13], and to assume that the rate of drug-like molecules contained in them is sufficiently small and can in fact be ignored. In Section 6.3, the pre-processing of these databases will be discussed in detail.

6.2.2 Descriptors

For a computational assessment, chemical structures often are translated into so-called descriptors, i.e. numbers that characterize certain properties or substructures. There are a number of standard descriptors available in various molecular modelling packages from, e.g. Daylight, Tripos, or Molecular Design. To a large extent, these descriptors have been evaluated recently [2]. The choice for a certain kind of descriptor should be made mainly by considering its information content. Sometimes too detailed descriptors make it complicated to extract general rules. A simple atom type descriptor with about 100 entries can work better than a 2000-bit fingerprint [14]. Section 6.3 will give an example of such a descriptor that works.

6.2.3 Classification Tools

In order to derive knowledge from the above-mentioned databases which are encoded by a suited set of descriptors, a statistical classification method is needed. In principle, a clustering approach or any other unsupervised learning algorithm could be used to derive clusters of similar compounds. The properties of the drug-like clusters could then be used for the classification of unknown compounds. A much more straight-forward approach is to use a supervised learning method like linear regression, decision trees [15], or neural networks. Due to their non-linear behavior, neural networks seem to be superior. Good descriptions of neural network techniques can be found in textbooks [16,17]. Professional neural networks software can also be found in abundance (see, e.g. [18]). In principle, a neural network can be treated as a black box, which resembles a very simple brain with neurons and axons. It can similarly be trained by confronting it with suitable training data that contain sufficiently general examples of the problem under consideration – in this case drugs and non-drugs. The properly trained network should then be able to predict the drug-likeness of compounds it never saw during training.

6.2.4 Complete Algorithm

The overall procedure is based on the data and methods mentioned above. The first step is the proper selection of databases, descriptors, and classification approach. The second step is the learning phase. Here, the approach tries to extract knowledge from the databases and to translate it into a classification algorithm. In step three, this classification scheme can then be applied to arbitrary compounds. The resulting classification scheme is illustrated in Figure 6.1. A given chemical compound is translated into a descriptor which in turn is forwarded to the classification tool (here a trained neural network) which comes up with a decision whether the compound is drug-like or not.

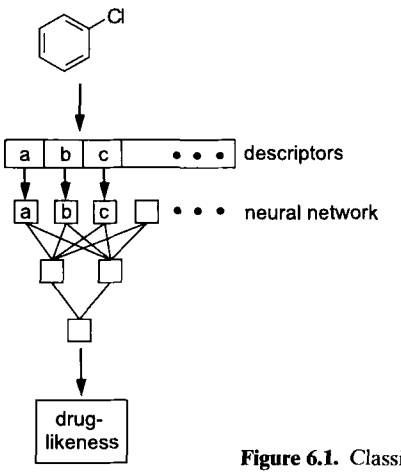


Figure 6.1. Classification scheme for drug-like compounds.

6.3 Applications

6.3.1 Drug-Likeness and a Recipe for a Computational Filter

Almost simultaneously, three different groups came up with different solutions for the problem of drug-likeness. Gillet *et al.* [6] used global structural features like the number of hydrogen-bond donors and acceptors, the numbers of rotatable bonds and aromatic rings, the molecular weights, and a shape descriptor as descriptors for the compounds in the World Drug Index and the SPRESI database as collections of typical drugs and non-drugs. A genetic algorithm was used to derive a weighting scheme from the databases, which calculates the drug-likeness of a compound from these features. A significant discrimination between drugs and non-drugs was achieved by this scheme.

Ajay *et al.* at Vertex [7] investigated the use of two types of descriptors (seven global molecular descriptors as in the Gillet paper and 166 MDL keys), and two classification methods (decision trees and feed-forward neural networks). The best approach was the combination of 78 descriptors with a neural network which classified 80% of the drugs in the MDDR database correctly as drugs while classifying about 90% of the ACD correctly as non-drugs.

Sadowski *et al.* at BASF [8] used the WDI and the ACD as drug/non-drug representations, an atom type based descriptor and a feed-forward neural network for the classification. In the following, a recipe of how to construct a computational filter for drug-like molecules is presented. The following steps were performed:

1. The ACD and WDI databases were preprocessed by removing reactive compounds and duplicates from both databases and by removing the exact matches of WDI compounds from the ACD. This left 169331 non-drugs and 38416 drugs.
2. All ACD and WDI compounds were assigned drug-likeness scores of 0 and 1, respectively.
3. An atom type based descriptor was calculated for each molecule. This descriptor is simply the counts of the 120 Ghose/Crippen atom types [14] in a given molecule. These atom types represent very simple functional groups as, e.g. "aliphatic CH_2", "aromatic carbon", or "amide nitrogen". This descriptor – a very simple type of a fingerprint – was found to have the ideal information content. The originally 120 Ghose/Crippen types were reduced to those 92 which were populated at least 20 times in the training dataset.
4. 5000 randomly chosen compounds from each of the two databases were forwarded to the training of a feedforward neural network based on the public domain program SNNS [18]. SNNS was used to construct and train a neural network with 92 input neurons (the atom type counts), 5 hidden neurons, and 1 output neuron (the drug-likeness score). The training following the "back-propagation with momentum" algorithm was performed over 2000 cycles, with a learning rate of 0.2 and a momentum term of 0.1. The training pursued the aim of minimizing the classification error for the learning set of 5000 drugs and 5000 non-drugs.

In order to assess the predictive power of the trained neural network, it was used to predict the drug-likeness of the 10000 training compounds as well as the drug-likeness of the remaining 165331 non-drugs from the ACD and the remaining 33416 drugs from the WDI. Fig-

ure 6.2 shows separately the distribution of the predicted drug-likeness scores in the ACD and WDI subsets in the training (closed lines) and test (dashed lines) datasets. Two conclusions can be drawn from this. First, the non-drugs are to about 80% on the left-hand side (i.e. the non-drug side) and the drugs are to about 80% on the right-hand side (the drug side) of the plot. Secondly, the differences between training and test sets are insignificant. The good classification is by no means a result of mere overtraining the neural net. Thus, it is indeed possible to create a computational filter that can distinguish between drugs and non-drugs.

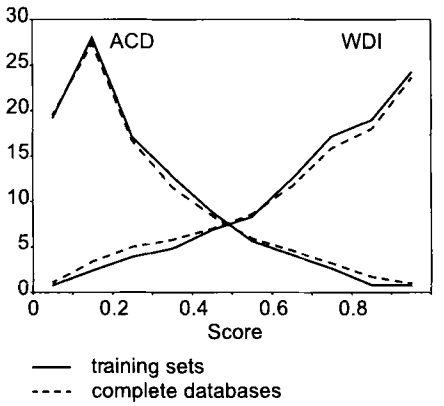


Figure 6.2. Distribution of calculated drug-likeness over training and test sets from ACD and WDI.

A simple additional verification is shown in Table 6.1 – the drug-likeness scores for a number of best-selling drugs ordered with respect to decreasing market volume [19]. Again, with one exception (diclofenac) the compounds were clearly classified as drugs with a score greater than 0.5.

A more comprehensive evaluation is based on a retrospective analysis of five HTS runs at BASF. This study demonstrates the similar behavior of the trained network compared to medicinal chemists. Figure 6.3 shows the percentage of predicted drug-like molecules (score > 0.3) over four stages in the HTS cycle: the totally screened compounds (some 100000 compounds), the compounds above a certain level of % inhibition (several thousand compounds), the compounds above a certain IC_{50} level (several hundreds), and finally the compounds chosen from this last list by medicinal chemists as leads for further development (about ten). Obviously, for the five HTS runs the level of predicted drug-likeness remains more or less the same for the first three stages – about 50–60%. It jumps significantly up to 70–100% drug-likeness after the chemists selected promising compounds by hand. This means, the trained network makes on average the same decisions as the chemists. This and nothing more must be expected from such an approach.

A number of additional statistical tests in all three papers [6–8] hardened the finding that the results of such an approach are absolutely valid. It is indeed possible to construct a filter, which estimates the drug-likeness of molecules like a large collective of medicinal chemists over many years.

Table 6.1. Calculated drug-likeness score for a number of best selling drugs.

Name	Score
Ranitidine	0.78
Enalapril	0.82
Fluoxetin	0.53
Aciclovir	0.64
Simvastatin	0.80
Co-amoxiclav	
amoxicillin	0.80
clavulanic acid	0.68
Diclofenac	0.40
Ciprofloxacin	0.93
Nifedipin	0.76
Captopril	0.82
Diltiazem	0.80
Lovastatin	0.89
Cimetidine	0.72
Omeprazol	0.85
Cefaclor	0.67
Ceftriaxon	0.97
Estrogene	
estrone	0.62
equilin	0.73
Cyclosporin	0.84
Beclometason	0.82
Famotidin	0.65
Salbutamol	0.93
Sertralin	0.65

6.3.2 Crop Protection Compounds

An interesting extension of the above describe approach would be a similar filter for crop protection compounds. Interestingly, there is no such database with the chemical structures in a computer-readable format available. Therefore, two databases that were assembled in-house at BASF were used: 986 crop protection compounds that are on the market or under development (CP) and 1203 compounds from new crop protection patents (PAT). The structural overlap of these two databases is less than 1%. The two databases were pre-processed with the same procedure as the drugs in reference [8]. The CP database (986 compounds) was

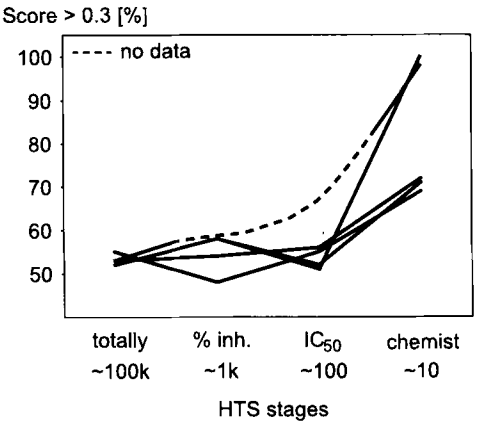


Figure 6.3. Retrospective analysis of HTS data.

used along with 1000 "non-crop protection" compounds from the ACD [12] as training set. The CP database can be considered as the currently available world of crop protection compounds. The PAT database consisting of 1203 mostly newly patented compounds in this area can be considered as the future of crop protection chemistry. This database was used as a test set.

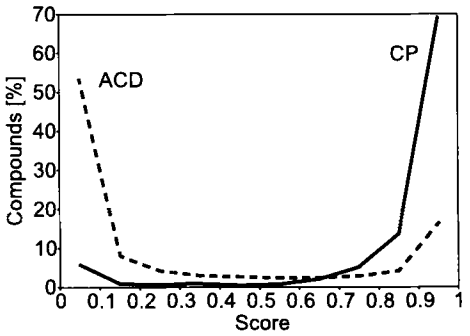


Figure 6.4. Distribution of crop protection scores in the CP (closed line) and ACD (dashed line) datasets.

The neural network training was performed like for the drug filter described above [8]. The results are shown in Figure 6.4. The vast majority of the CP data (91%) are on the right hand side of the diagram. These are the correctly classified crop protection compounds. The majority of the non-crop protection compounds (71%) are on the left hand side. Thus, the new filter is able to distinguish between compounds that are suited for crop protection purposes and those which are not. In order to assess the predictivity of the approach, the second

crop protection database PAT (being the future of crop protection chemistry) was sent through the trained neural network. Figure 6.5 shows the distribution of the score for this dataset. Clearly, these 1203 compounds were also classified mostly correct (69%).

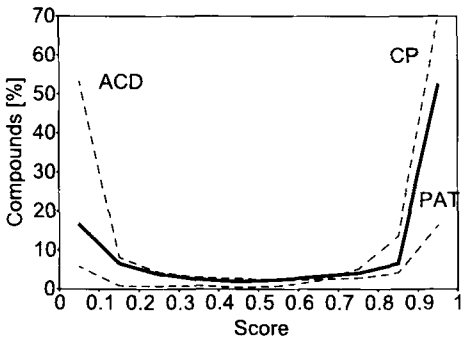


Figure 6.5. Crop protection score distribution in the PAT dataset (solid line). For comparison, the distributions for the training data (Figure 6.4) are shown with thin dashed lines.

In addition, the two filters for drugs and crop protection compounds were cross-validated by applying the crop protection network to the World Drug Index and the drug-likeness network to the crop protection compounds. The crop protection filter found 67% compounds not suited for crop protection in the WDI. On the other hand, the drug filter found 77% non-drugs in the two crop protection databases (CP and PAT, 2189 compounds). Therefore, the filters behave as expected since a general assumption is that most drugs do not act in crop protection and vice versa.

6.3.3 Virtual High-Throughput Screens

In principle, the technique demonstrated in detail above for drug and crop protection filters can also be applied to arbitrary classification tasks. The only prerequisite is that there are sufficiently validated data available; at least 1000 examples for each class. Such data naturally comes out of primary HTS runs with [%] inhibition as the only rough criterion for activity. This would give access to "virtual high-throughput screens" when the networks succeed to behave similar to the real HTS assays. In order to demonstrate this enhancement, two explicitly named screening sets (AIDS antiviral activity and cytotoxicity) and data from six in-house HTS runs were used.

From the public part of the NCI database [20], 1502 "confirmed active" and "confirmed moderately active" compounds, and 41017 "confirmed inactive" compounds (tested for AIDS antiviral activity) were extracted. From this total dataset, 1000 active and 1000 inactive compounds were used to train a neural network. The remaining compounds were used as test set. Figure 6.6 shows the distribution of the resulting AIDS score in these datasets. The result is not as clear as for the drug-likeness score – the reason might be the relatively small

number of actives (1502). But the trained network is still able to classify about 65% of the active and inactive compounds in the test sets correctly. This is a significant improvement over a merely random decision (50% error probability).

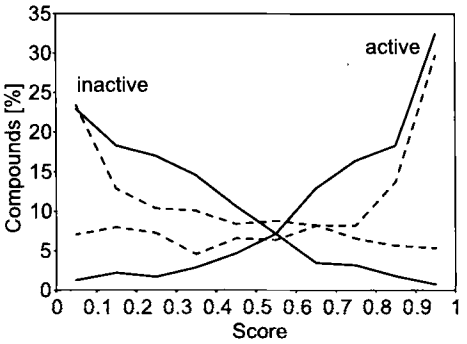


Figure 6.6. Distribution of the AIDS antiviral score in the inactive and active datasets (training data: closed lines, test data: dashed lines).

The next example was drawn from an in-house cytotoxicity screen. 2350 compounds with a confirmed cytotoxic activity and 16400 inactive compounds were used. From these, 1500 randomly drawn toxic and 1500 randomly drawn non-toxic compounds served as training set and the remaining compounds as test set. Figure 6.7 shows the resulting score distributions. Clearly, in most cases the trained network correctly predicts whether compounds are toxic for both the training (closed lines) and test data (dashed lines). For the test sets, it correctly classified about 75% of both classes. This filter can now be used as an additional early warning system for pointing out potentially problematic compounds and for excluding them from purchase.

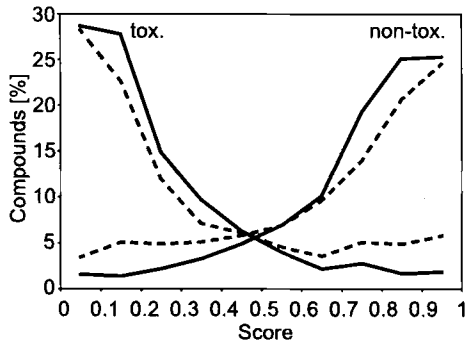


Figure 6.7. Distribution of the cytotoxicity score in the toxic and non-toxic datasets (training data: closed lines, test data: dashed lines).

As mentioned above, this procedure was applied also to six in-house HTS runs. Properly drawn training sets of HTS hits and non-hits were used to create virtual HTS screens. Table 6.2 shows together with the results of the AIDS and toxicity screens the percentages of correctly classified compounds in the training and test sets. It is obvious that in some cases the classification is good enough to trust such virtual screens, and in some cases not. Based on the data in Table 6.2, one could use the screens 1–4 as virtual HTS for the further selection of compounds from various sources for these projects. In the cases of screens 5 and 6, such a procedure would not be recommended since the prediction of the test data is poor.

Table 6.2. Percentages of correctly classified compounds from several screening runs.

Test name	Training result		Test result	
	Inactive	Active	Inactive	Active
AIDS antiviral	83	87	65	66
Cytotoxic	87	87	76	76
Screen 1	83	77	79	75
Screen 2	91	90	85	86
Screen 3	86	74	74	64
Screen 4	87	70	73	58
Screen 5	78	82	53	62
Screen 6	88	86	60	61

6.3.4 Optimization of Combinatorial Libraries

In order to extend the compound profiling approach described above to combinatorial libraries, some additional principles have to be considered. An example might be the following. From 7262 carboxylic acids and 1761 aldehydes taken from the ACD [12], a virtual library of 13×10^6 products can be derived (proprietary scaffold). The aim is to synthesize a much smaller number of some hundreds of diverse compounds with the focus on crop protection (see above) and a moderate price for the starting materials. These are three criteria which are relatively easy to be computed. Of course one could follow a "cherry-picking" scheme and simply select the top-ranked compounds from the virtual library. But in order to achieve synthetic effectivity, the sub-library should fulfill the additional combinatorial restriction that it is built up by complete enumeration of all combinations of a small number of building blocks. This is in clear contrast to a "cherry-picking" approach which would simply select the best compounds from the whole virtual library. Given 15 acids and 15 aldehydes, which leads to 225 products, there are 10^{82} possible selections of sub-libraries out of the 13×10^6 compounds in the virtual library. A systematic exploration is currently impossible for such a large number of compounds.

Gillett *et al.* [21] proposed solving such problems by a genetic algorithm [22]. Genetic algorithms optimize a population of individuals (possible solutions) by improving their "fit-

ness" (i.e. the adaption to the problem) by applying principles of the natural evolution like "mutation" and "cross-over". The implementation used here is based on the Genesis program [23]. The individuals in a population are different 15 × 15 sub-libraries out of the virtual library described above. Their fitness is the weighted sum of:

1. the percentage of compounds with a crop protection score greater than 0.3,
2. a diversity index, and
3. the reciprocal prices of the starting materials.

The GA was run with a population size of 50, a maximum number of generations of 200, a mutation rate of 0.1%, and a cross-over rate of 60%. These are more or less the recommended default values [23]. Sufficient convergence could be reached with these parameters.

The diversity index is the normalized sum of the absolute differences of the Ghose/Crippen fingerprints [14] of all pairs of compounds within a given 15 × 15 sub-library. The Ghose/Crippen fingerprints of the 225 products in a given sub-library are calculated from the fingerprints of the individual building blocks. Since these fingerprints also provide the basis for the crop protection score, the computation of the fitness function is very effective in terms of computer resources. It takes less than 30 minutes to do the 10000 fitness function calculations needed for one optimization run on an SGI R10000 processor.

Ten optimal 15 × 15 sub-libraries where generated by the genetic algorithm described above. In order to assess the quality of the results, they were compared to:

1. 10000 randomly generated libraries,
2. ten libraries optimal with respect to maximal diversity regardless of the other criteria, and
3. ten libraries with minimal diversity.

The ten optimal libraries had about 78% compounds with a crop protection score greater than 0.3 and a diversity index of 5.2 at costs of about 3000 USD (20 grams per building block). The 10000 randomly drawn compounds behaved much less favorably. The best few libraries contained up to 30% suited compounds at a minimal cost of 30000 USD. The diversity index of the 10000 random libraries was between 3.9 and 5.4. Thus, the GA optimization is by far more advantageous.

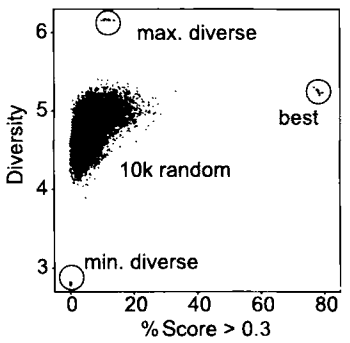


Figure 6.8. Diversity index *vs.* percentage of suitable compounds (score > 0.3): 10000 randomly drawn 15 x 15 libraries, the best library after optimization, the maximal and the minimal diverse libraries.

Figure 6.8 illustrates this superiority. The diversity index of the sub-libraries is plotted against the percentage of suitable compounds (crop protection score > 0.3) for the 10000 random libraries, ten optimal libraries with respect to score, diversity and cost, ten maximal diverse libraries regardless of the other criteria, and ten minimal diverse libraries regardless of the other criteria. The last two groups of libraries where obtained by GA runs for maximizing and minimizing diversity alone without the other criteria (score and cost) in order to find the lower and upper ends of the diversity scale for this type of combinatorial library (there is no absolute diversity scale).

The diagram shows the ten best libraries on the right hand side of the score axis and in the upper third of the diversity scale with the minimal diverse and maximal diverse libraries as end points. This is much better with respect to both criteria than the 10000 random libraries and much better with respect to the score than the maximal diverse libraries. Thus, the GA found a sufficient compromise between several independent criteria for combinatorial library optimization.

6.4 Conclusions

Drug-likeness is a general compound property beyond physicochemical properties, certain explicit structural features, or diversity. It is possible to derive knowledge about drug-likeness from public databases of drugs and non-drugs. This approach is able to classify about 80% of the drugs and non-drugs correctly.

Here it is stressed again that drug-likeness in this context simply means: Medicinal chemists would have voted similarly. There is no guarantee that a compound termed "drug-like" will indeed become a drug. This is something the chemists themselves cannot guarantee. On the other hand, a drug-likeness filter can enrich compound sets with compounds having better chances of becoming drugs. The neural network approach is one possible recipe for constructing such a computational filter, by using public databases and suited descriptors. The approach can be extended to a number of other related properties based on data about crop protection suitability, toxicity, HTS hits, and combinatorial library optimization.

References

[1] W. A. Warr, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134–140.
[2] R. D. Brown, Y. C. Martin, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
[3] H. Matter, *J. Med. Chem.* **1997**, *40*, 1219–1229.
[4] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
[5] G. M. Rishton, *Drug Discovery Today* **1997**, *2*, 382–384
[6] V. J. Gillet, J. Bradshaw, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165–179.
[7] Ajay, W. P. Walters, M. A. Murcko, *J. Med. Chem.* **1998**, *41*, 3314–3324.
[8] J. Sadowski, H. Kubinyi, *J. Med. Chem.* **1998**, *41*, 3325–3329.
[9] WDI: World Drug Index; Version 2/96, Derwent Information **1996**.
[10] CMC: Comprehensive Medicinal Chemistry; MDL Information Systems.
[11] MDDR: MACCS Drug Data Report; MDL Information Systems.
[12] ACD: Available Chemicals Directory; Version 2/96, MDL Information Systems **1996**.
[13] SPRESI: Daylight Information Systems.
[14] A. K. Ghose, V. N. Viswanadhan, J. J. Wendoloski, *J. Phys. Chem. A* **1998**, *102*, 3762–3772.

[15] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, Los Angeles **1993**.
[16] J. Zupan, J. Gasteiger, *Neural Networks for Chemists: An Introduction*. VCH Publishers, Weinheim **1993**.
[17] A. Zell, *Simulation neuronaler Netze*, Addison-Wesley, München **1994**.
[18] SNNS: Stuttgart Neural Network Simulator; Version 4.0, University of Stuttgart **1995**.
[19] SCRIP No. 2040, July 7, **1995**, p. 23.
[20] NCI: National Cancer Institute database **1999**.
Online download: http://cactus.cit.nih.gov/ncidb/download.html.
[21] V. J. Gillet, P. Willett, J. Bradshaw, D. V. S. Green, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169–177.
[22] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA **1989**.
[23] Genesis program by J. J. Grefenstette, Naval Research Laboratory, Washington DC **1987**.

Virtual Screening for Bioactive Molecules
Edited by Hans-Joachim Böhm and Gisbert Schneider

7 Pharmacophore Pattern Application in Virtual Screening, Library Design and QSAR

Andrew C. Good, Jonathan S. Mason, Stephen D. Pickett

7.1 Introduction

A pharmacophore is commonly defined as a critical three-dimensional (3-D) geometric arrangement of molecular features or fragments forming a necessary but not sufficient condition for biological activity [1–2]. The application of pharmacophore descriptors has formed a mainstay of ligand based virtual screening for much of the last decade. In traditional 3-D database searches, pharmacophores are applied singly after determination from structure activity studies of multiple active ligands. More recently they have been exploited as constraints for structure-based virtual screens. The new technologies of combinatorial chemistry, high throughput screening and genomics have required further expansion of their application, with pharmacophore fingerprints being developed as screening tools. 3-D pharmacophore fingerprints can be generated for both ligands and protein binding sites, providing powerful techniques for incorporating active site constraints in similarity/diversity analyses.

All the above techniques are reviewed within this Chapter, with particular reference to the recent advances, advantages and challenges of each methodology. We begin (Section 7.2) with a brief description of the underlying technological background for pharmacophore generation. We then detail their use in "traditional" applications, including single pharmacophore searching and model derivation (Section 7.3). We also discuss their application as constraints in protein structure-based virtual screening (Section 7.4). Finally, we review the application of pharmacophore fingerprints as full molecular descriptors (Section 7.5), including their use as descriptors in screening, library design, and QSAR (and combinations thereof).

7.2 Preparations for Pharmacophore Screening

Before pharmacophore screening can take place, certain key preparations are necessary. The first requirement for all such searches is that the structural data to be screened is available in 3-D form, and that pharmacophoric atom-type definitions have been created to allow pharmacophore formalization. Further, it is generally essential that a method for undertaking conformational searching of the structures is available.

7.2.1 3-D Structure Generation

The creation of a 3-D structural dataset first requires the generation of appropriate "2-D" connectivity files with atom and bonding descriptions for each molecule. This is used as input for an automated 3-D structure-generation program. Such connectivity files generally come in the form of SMILES strings [3], SLN strings [4,5], and 2-D SD connection tables [6].

Once the connectivity file is available, a number of different programs can be used to build the 3-D structures. These techniques undertake the conversion using a mixture of rules [7,8] (linking atom hybridization bond lengths, bond angles and non-bonded forces), pre-calculated 3-D fragment databases [9], and distance geometry techniques [10,11] to the connectivity data.

Historically, the Concord [7] program has been the most popular 3-D conversion software. It combines a knowledge base of rules with energy minimization to generate a low energy 3-D conformation for each structure. Cyclic and acyclic portions are constructed separately, with the resulting substructures fused to form the complete molecule. Optimum acyclic bond lengths, angles and torsions are extracted from a table of published values. Similarly, bond lengths and torsion angles of single cyclic portions are built using precalculated topological rules. Ring systems are constructed through the assignment of gross conformations of each ring. Each constituent ring is then fused into the system in order, with a strain minimization function employed to create an acceptable geometry. Once constructed, the structure can be further optimized through energy. Concord is able to produce reasonable low energy conformations for most small organic structures, though large rings and peptides are not handled well. The program is rapid, provides structures with consistent geometries (bond lengths etc.), gives useful structure quality information (e.g. a close-contact ratio), and allows the output of many file formats.

Another popular 3-D conversion program, Corina [8], was created more recently in an attempt to address some of Concord's shortcomings. Improvements include more extensive bond length and angle tables, Hückel molecular orbital refinement of conjugated systems, and a backtracking procedure to rebuild highly strained ring systems using alternative conformational templates. Overall, many of the techniques have been found to produce reasonable 3-D structures for most molecules (although there are notable exceptions), with Corina showing a slight edge in terms of conversion rate. For a more detailed analysis of 3-D structure building techniques see the excellent review by Green [12].

7.2.2 Pharmacophore Atom-Typing

Pharmacophore searches require that molecules be broken down by their functional properties before screening can take place. The most common properties used to describe a structure's potential pharmacophoric centers are:

1. Base, for example sp^3 N aliphatic amines, hydrazines, sp^2 N amidines, guanidines and 2/4 amino pyridines (positively charged centers at physiological pH 7),
2. Acid, such as carboxylic acid, acyl sulphonamide, unsubstituted tetrazole and on occasion phenols (negatively charged centers at physiological pH 7),

3. Acceptor, for example carbonyl, aliphatic ether and hydroxyl,
4. Donor, such as primary/secondary amide/aniline nitrogens and hydroxyl,
5. Aromatic, generally (but not always) in the form of ring centroids, and
6. Hydrophobes, for example certain 5/6 member aromatic rings, isopropyl, butyl and cyclohexyl.

Figure 7.1 illustrates how a molecule can be broken down into pharmacophoric elements.

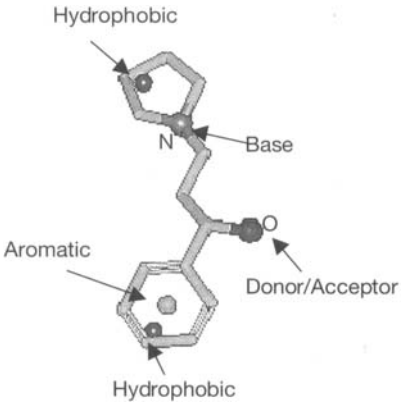


Figure 7.1. Example of how a 3-D molecular structure can be broken down into its constituent pharmacophoric elements.

This atom-typing either occurs on compound registration (this is the method used in database registration in Chem-DBS3D/Chem-X [9]), or at search time (DOCK [13–14] and Unity [5] apply this technique). Table 7.1 illustrates possible DOCK search time definitions. The advantage of defining pharmacophoric elements at search time is that they can be easily tailored for particular screens. For example, if a user wished to include tyrosine phenol oxygens as acid moieties for a particular search, it would simply be a matter of editing the acid definitions file before the run was undertaken. The atom types generated by Chem-X/Chem-Diverse are automatically assigned when reading a molecule into Chem-X, through a customizable parameterization file and fragment database [15]. The dummy atoms that are used to represent the hydrophobic regions are added by an automatic method within Chem-X that uses bond polarities (hydrophobic regions defined for groups of three or more atoms that are not bonded to atoms with a large electronegativity difference).

While we will deal almost exclusively with these atom definitions, it is worth noting that other pharmacophore atom types have also been developed [16,17].

7.2.3 Conformational Flexibility

Initial pharmacophore search paradigms relied on a single low energy conformation to represent each molecule within a given database [18–22]. This was quickly recognized as a major

Table 7.1. Sample DOCK pharmacophore atom-type definitions. The format given is the same as that used by DOCK in its *chem.defn* file. For more information on the syntax of these assignments, take a look at the online DOCK manual (URL: *http://www.cmpharm.ucsf.edu/kuntz/dock4/html/Manual.45.html#pgfId=4869*).

Name: acceptor

```
#generic O acceptor - note that ether excluded here
definition O [ H ] [ N ( 2 O ) ] [ 2 C ] ( * )

# ester
definition O.3 [ H ] ( C.2 ( O.2 ) )

#generic N acceptor
definition N [ H ] [ 2 O ] [ 3 C ] ( * )
```

Name: acid

```
# optional tyrosine
#definition O.3 ( H ) ( C.ar )

# protonated and deprotonated carboxyls
definition O.co2 ( C )
definition O.3 ( H ) ( C.2 ( O.2 ) )
definition O.2 ( C ( O.3 ( H ) ) )

# tetrazole
definition N.pl3 ( H ) ( N.2 ( N.2 ( N.2 ( C.2 ) ) ) )
definition N.pl3 ( H ) ( N.2 ( N.2 ( C.2 ( N.2 ) ) ) )
definition N.2 ( N.2 ( N.2 ( C.2 ( N.pl3 ( H ) ) ) ) )
definition C.2 ( N.pl3 ( H ) ( N.2 ( N.2 ) ) )
definition N.2 ( N.2 ( C.2 ( N.pl3 ( H ) ( N.2 ) ) ) )
definition N.2 ( C.2 ( N.2 ( N.pl3 ( H ) ( N.2 ) ) ) )
definition N.2 ( N.2 ( C.2 ( N.2 ( N.pl3 ( H ) ) ) ) )
definition N.2 ( N.pl3 ( H ) ( N.2 ( N.2 ( C.2 ) ) ) )

# acyl sulphonamide
definition N.am ( S ( 2 O.2 ) ) ( C.2 ( O.2 ) )
definition O.2 ( C.2 ( N.am ( H ) ( S ( 2 O.2 ) ) ) )
definition O.2 ( S ( O.2 ) ( N.am ( H ) ( C.2 ( O.2 ) ) )
```

Name: aromatic

```
definition C.ar

#thiophene
definition S.3 ( C.2 ( C.2 ( C.2 ( C.2 ) ) ) )
definition C.2 ( S.3 ( C.2 ( C.2 ( C.2 ) ) ) )
definition C.2 ( C.2 ( S.3 ( C.2 ( C.2 ) ) ) )
```

deficiency and new methodology was developed to obviate the issue. The most popular techniques basically fall into two categories. The first employs explicit conformational information regarding database structures, storing the flexibility information within the database, and/or determining conformational possibilities at search time [23–26]. The second uses pharmacophore constrained torsion minimization at search time [27–28].

7.2.3.1 Conformation Search Techniques

A popular method for dealing with conformational flexibility has been through the explicit storage of multiple molecular conformations [29]. The problem with such an approach is that, if one wishes to keep the storage and search times down to a reasonable level, only a small number of conformers may be stored per molecule. Thus, if this technique is to be successful, careful attention must be paid to which molecular conformations should be retained in the database. A general approach has been to apply conformational sampling methods to search conformational space, and then use a difference metric (e.g. root mean squared [RMS] interatomic distance difference) to cluster conformations into families. This form of post-processing methodology has traditionally suffered from a number of drawbacks, chief among them being to ensure a good coverage of conformational space. Sampling methods often miss regions of conformational space unless they are run for prohibitively long times. This leads to the second drawback of conformational redundancy. All these techniques tend to produce many thousands of conformations that are subsequently pared down to a just a few families. The problem is further exacerbated if one requires each conformation be a local minimum, since all conformations lying in the same energy well will collapse down into a single family. As a consequence, conformational sampling can be inefficient both in terms of the keying time required as well as the amount of conformational space covered.

Explicit Conformational Storage. To mitigate these problems, a number of techniques have been used. One method for improved conformational sampling is known as Poling [24,30–31]. This procedure is designed to promote conformational variation through the addition of a "poling function" to a standard molecular mechanics force. The function has the effect of changing the energy surface being minimized, to penalize conformational space around any previously accepted conformers. As a result the technique both increases conformational variation and eliminates redundancy within the limits imposed by the function. Poling has been applied within the Catalyst search software[32].

New paradigms for conformational analysis continue to be researched [25,26]. An example of this is the Confort software [26]. Confort is designed to rapidly identify the local energy minim bounds of each rotatable bond. For each combination of these ranges, the bounded region hyperspace centroid is used as an initial "raw conformation". This is then relaxed applying analytic gradients to the internal coordinate sub-space. The resultant population of "relaxed conformations" can be optimized and filtered by energy and diversity to rapidly determine a set of diverse low energy conformers for the system.

Conformational Analysis at Search Time. An alternative approach to explicit conformation storage is to calculate conformations at search time. Speed is maintained through the application of techniques such as allowed torsion look up tables [33–36], van der Waals (vdW) clash checks with look ahead to quickly remove sterically disallowed conformers [37], and random analysis with fixed maximum CPU time for the more flexible database structures. Both Chem-DBS3D [9] and DOCK [14] employ this methodology.

7.2.3.2 Torsion Fitting

This technique was designed specifically for screening versus a single pharmacophore model. It has the advantage that no explicit conformational search need be undertaken, with each candidate molecule undergoing pharmacophore-constrained torsion optimization instead. Specific molecular torsions are tweaked to determine whether all the pharmacophore constraints can be met simultaneously, with an optional check for internal vdW clashes. Variants of this approach have been implemented within Chem-DBS3D [9], Unity [5,27], Isis-3D [28] and Catalyst [32].

7.3 Screening by Single Pharmacophore: Elucidation and Execution

Single pharmacophore searching has been a successful method of virtual screening for many years now. The method provides an excellent paradigm for exploiting the potential binding modes of known ligands to discover novel active chemotypes. This can be crucial when attempting to avoid potential patent problems with competitor compounds, or if the current in-house lead chemotype contains intrinsic problems (e.g. toxicity). The technique's importance is such that it has been reviewed many times [38–40]. Nevertheless, the technique continues to be an active area of research [41–49] and as such is still deserving of some attention.

The utility of such screens is illustrated below by the work of Marriott et al. [41] (see Figure 7.2). An additional prerequisite for these searches, over those described in the previous section is the need for a technique to elucidate the crucial pharmacophore from within a set of active molecules. In this study, 3-D conformer models of known muscarinic m3 antagonists were used in conjunction with the DISCO program [50] to determine such pharmacophore models. DISCO takes as its input a series of low energy conformations for each active molecule, and potential pharmacophore site points are generated automatically for each conformer. A useful feature of DISCO is its ability to define the locations of potential protein hydrogen bond donors and acceptors. This can be important if ligands are found to approach the same polar site point from different directions, something not easily accounted for when purely atomic superimposition is used. The program then utilizes clique detection [51] to determine matching groups of site points distances between the conformers of a specified number of molecules within the study set, using the most rigid molecule as the reference. The resulting site point matches are used to overlay the ligands. It should be noted that the science of pharmacophore model determination has received much attention, with many alternative techniques devised for their elucidation [44–49,52–53].

The resulting DISCO models were analyzed and two were selected to screen in-house 3-D databases using the Unity program. The resultant search produced datasets from which 172 compounds were sent for screening. Three hits were determined, containing two structurally distinct chemotypes from the original pharmacophore template structures. It is this ability to find novel chemotypes that makes pharmacophore searches so attractive.

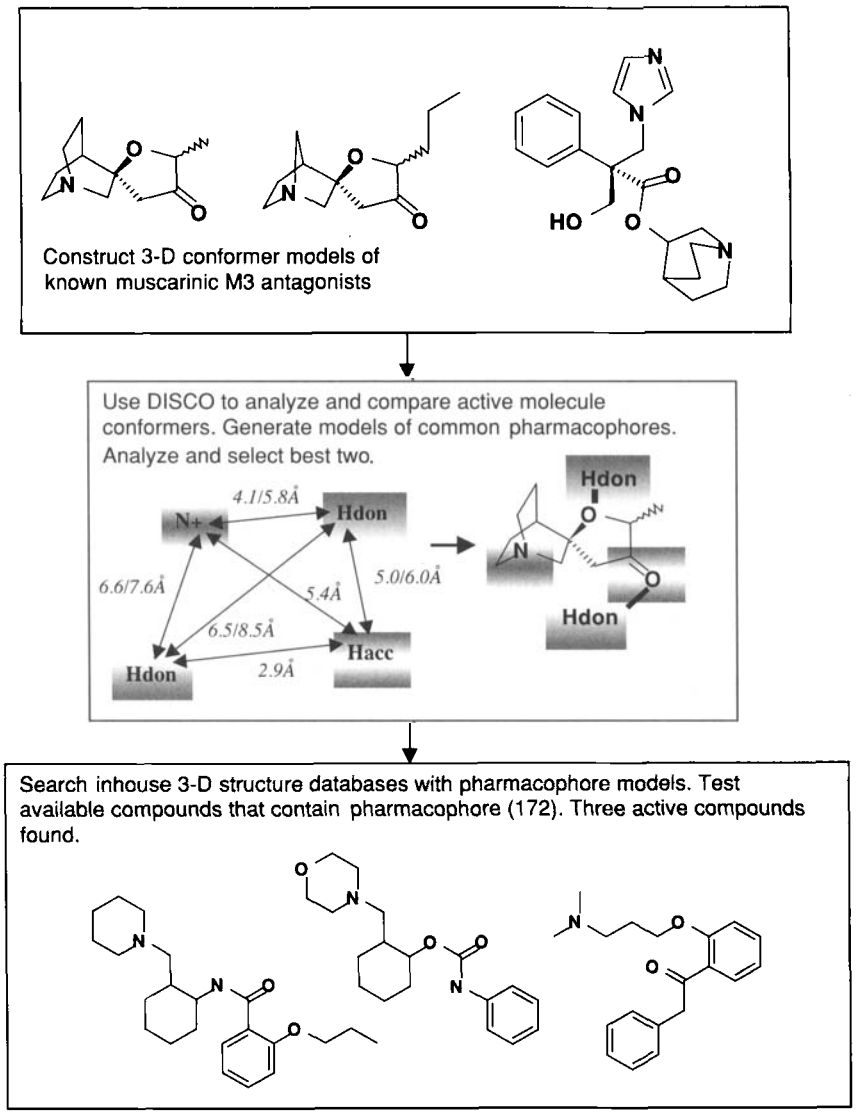


Figure 7.2. Outline of Muscarinic M3 antagonist study undertaken by Marriott *et al.* [41]. Both distances for the two selected pharmacophore models are given where difference exists between them. **Hdon** corresponds to potential active site donor atom positions, **N+** refers to ligand base and ligand acceptor is designated by **Hacc**. The shading highlights matching functionalities.

7.4 Pharmacophore Constrained Structure-Based Virtual Screening

The rate of X-ray and NMR protein target structure elucidation is continually accelerating. Such data can prove invaluable, since by exploiting structural information taken directly from the target active site, it is possible to discover ligands with both diverse chemotypes and binding modes. As a result, structure-based screening is potentially even more powerful than ligand-based pharmacophore screens. Nevertheless, the application of pharmacophore information can dramatically improve the performance of such calculations.

A number of different structure-based screening programs make use of pharmacophore-like constraints [14,54–58]. Perhaps the most widely used form of this 3-D search paradigm is the DOCK program developed by Kuntz and co-workers [13–14]. As originally designed [13], DOCK describes the receptor active site using sets of overlapping spheres derived from the Connolly molecular surface [59] of the site (the SPHGEN program [60]). The sphere centers are used to define potential ligand atom positions and as such are potential pharmacophore points. In the original implementation ligand atoms and sphere centers are matched using a form of clique detection [13,61] (see Figure 7.3). For each molecule, every time the prerequisite number of nodes (ligand atom–receptor sphere center pairs) are found to match, the ligand atoms are mapped onto the sphere centers by a least squares fit [62]. When the resulting fit leaves the ligand positioned within the active site so as to be under the bump limit (the bump limit equals the number of times a ligand and receptor atom clash), the com-

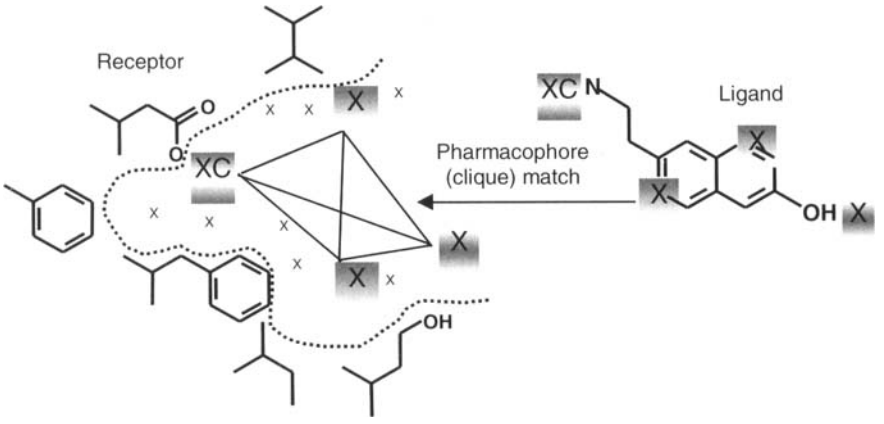


Figure 7.3. Schematic of clique detection as applied to the receptor site point/ligand atom paradigm used in DOCK. As originally applied only distances needed to correspond to produce a match [13]. More recently, the feature to force chemical matching of the atoms and hence allow the use of pharmacophore constraints was added [14,68–69] (the shading highlights matching functionalities). An additional change was also made to permit the definition of critical regions. The resulting constraints allow tight control over allowed ligand-docking modes. For example, here the XC site point could be defined as needing to be matched by a base and also as being a critical region. All resulting docked ligands would thus have to contain a base in this region of the binding site. It is worth noting that, while this figure relates specifically to receptor–ligand site points, it nevertheless provides a reasonable schematic representation of the clique detection [51] mentioned in Section 7.3. Adapted from [38].

plementarity of the resulting superposition is determined. In this way, DOCK explores the six degrees of freedom of the ligand with respect to the active site, allowing the examination of a large number of putative ligand–receptor orientations. More recent versions of DOCK also allow the consideration of ligand torsion flexibility [14]. As we have mentioned, an outstanding feature of this approach is that, rather than searching against only a single postulated binding mode (i.e. pharmacophore-based searches), whole receptor sites can be explored, allowing a large diversity of potential receptor–ligand interactions to be considered. New classes of ligands can thus be discovered which may bind with the receptor in hitherto unforeseen ways. Docking has been used successfully in this way in a number of projects [63–67]. This advantage is also the source of a major problem for structure-based searching tools such as DOCK, however, since the thousands of putative ligand–receptor orientations tested make the searching process very CPU-intensive. In the past, it was not uncommon to spend one or two weeks searching a database of 100000 structures (single conformers only).

To combat this, methods were developed to introduce pharmacophore constraints into the DOCK search paradigm [14,68–69]. Since DOCK already relies on site points to orient putative ligands, the program lends itself naturally to pharmacophore constraints. This is accomplished by simply defining each site point to include pharmacophore atom-type information complementary to the local active site region. A variety of techniques can be used to introduce this data, including site analysis using the GRID program [70], and abstraction of pharmacophoric group positions from bound ligand complexes. Subsequent DOCK searches only permit ligand atoms with the same pharmacophoric type to match with each point. This can be also be combined with the definition of critical regions [54,69,71] in the active site (for example pockets considered critical for binding). Site point(s) within these regions must always be included in a match for it to be valid. The resulting DOCK node search essentially becomes a series of n point pharmacophore screens (n being the number of matching nodes requested). Using such an approach dramatically decreases the number of non-productive (high energy and low chemical complementarity) ligand–receptor orientations tested by DOCK, which is where the program spends much of its time.

A nice example of the advantages of such an approach comes from work at Bristol-Myers Squibb using DOCK on adipocyte lipid-binding protein (AP2). Two runs were set up. In the first, the site points for the AP2 active site were derived using the ligand acid oxygens from two known AP2 pdb [72] complexes, 2ANS [73] and 1LIC [74]. Additional site points were derived from analysis of the active site using an in-house program, Makepoint [75]. The software uses the normals (plus associated atoms) of a Connolly active site molecular surface calculation to derive pharmacophoric site point positions within the site. Visual analysis is then used to modify and delete points as required. The acid oxygen points were assigned to be critical, so that all node matches had to contain at least one of these points. Amber force field scoring [76] was used to rank the hits (the electrostatic term was disabled since the chemical information in the site points was deemed sufficient). In the second DOCK run, the standard SPHGEN program from DOCK was run and the search was undertaken without constraints using Amber force field scoring (with electrostatics to allow DOCK to "see" the acid-binding region). Both datasets were executed against a database of ~10000 compounds containing 28 molecules known to be active against AP2. Table 7.2 shows the general DOCK parameters used in both searches. Figure 7.4 shows the resulting hit rate for the searches. It is clear that the pharmacophore constraint search performs the better of two. Of equal impor-

tance is the fact that the pharmacophore-constrained search run is ~20 times faster (~2 CPU days on a Silicon Graphics 250MHz R10000). This is particularly impressive when one considers that the constrained search contained more than twice the number of site points (107) in the vanilla run. In-house tests show that, when the number of site points used is equivalent, the speed-up can be up to 100 times. This is crucial to be able to run multi-conformer DOCK searches over large databases (>> 100000 compounds).

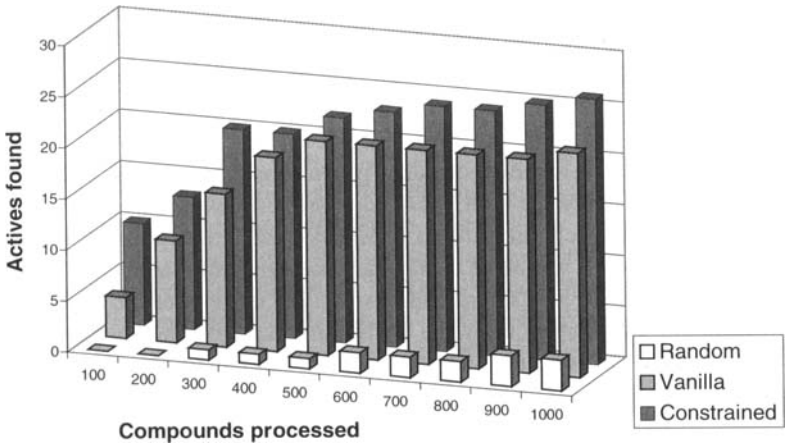


Figure 7.4. Comparison of vanilla (unconstrained) and pharmacophore-constrained run hit rates using DOCK on AP2. The number of hits that would be expected from a random selection is presented for reference (dataset size ~10000 including 28 known active "seed" molecules). The pharmacophore-constrained run was completed in ~1/20th of the time required for the vanilla runs and produced superior results.

The DiR (Design in Receptor) module of Chem-X [58,77] provides an alternative flexible search paradigm similar to DOCK in nature. DiR uses potential complementary three- or four-point pharmacophores within active sites as 3-D molecular database search queries. Molecules that hit are docked into the site by a least squares fit of the matching pharmacophore points, and a bump check is run as a quick shape complementarity test. An interesting feature of DiR is the application of "adaptive" queries to enable more rapid searching. This is accomplished by eliminating pharmacophores from consideration for new conformations of a particular molecule once they have been matched once.

7.5 Pharmacophores as Full Molecular Descriptors

Explicit 3-D molecular similarity calculations encompass the quantitative comparison of selected spatial molecular descriptors through the application of a similarity equation (index). Such evaluations have played an important role in many CADD calculations. Numerous reviews have been written detailing many aspects of explicit molecular similarity [78–84]. In

Table 7.2. DOCK calculation settings for AP2 runs. These are presented to highlight the large number of variables (of which these are the most influential) which can affect a DOCK calculation. Note that conformational flexibility and rigid minimization were included in the search paradigm. An analysis of the resulting DOCK run showed that a total of ~240000 ligand conformers were docked and minimized during the search. For more information on these settings see the online DOCK manual (URL: *http://www.cmpharm.ucsf.edu/kuntz/dock4/html/Manual.19.html#pgfId=8491*).

DOCK parameter	Value assigned	Effect on DOCK run
nodes min/max	4	Min/max atom count for matching (pharmacphore) clique
distance_tolerance	0.5	Maximum permissible distance error between clique atom pairs
distance_minimum	3.0	Minimum distance for any given clique atom pair
heavy_atoms_minimum	10	Minimum molecule size
heavy_atoms_maximum	50	Maximum molecule size
bump_filter	4	Maximum permissible number of clashing atoms between ligand and protein before minimization after matching cliques have undergone least squares fit
energy_size_penalty	0.25	Energy penalty assigned to ligand per heavy atom
conformation_cutoff_factor	5	Number of conformers to generate per rotatable bond
flexible_bond_maximum	15	Maximum number of rotatable bonds allowed in ligand
clash_overlap	0.7	Scaling factor for vdW radius assigned for clash calculation
minimize_ligand	yes	Flag to turn on rigid body minimization after least squares fit of the clique match if bump check passed
maximum_orientations	5000	Maximum number of scoring orientations to test per conformation

classic CADD investigations such calculations have typically been applied to the determination of optimum molecular superposition and as QSAR descriptors. Even the fastest 3-D database-searching techniques have traditionally required that molecules are superimposed in co-ordinate space before a similarity analysis can begin [85]. At the same time newer technologies are beginning to force a different set of constraints for many potential CADD studies. For example, genomics is lining up an ever-increasing queue of potential targets for the pharmaceutical industry. In general, the amount of initial structural information for these targets will be limited (possibly a target homologue and substrate). As we have already intimated, traditional ligand-based pharmacophore searches usually exploit a significant body of information (for example competitor lead series, as in the example described in Section 7.4) to derive the required pharmacophore. Such a data store is unlikely to be available (at least at the beginning of the project) for most genomics targets, so new methods are required to ad-

dress this situation. A different problem exists with respect to high throughput screening and combinatorial chemistry technology. Here the constraints are the huge sizes of the molecular datasets available. HTS hit lists are often large, heterogeneous, and of variable accuracy. Fast general SAR elucidation techniques are thus required for dataset analysis. Similarly, the potential virtual dataset sizes of many combinatorial libraries is such that the ability to analyze molecular similarity / diversity rapidly is crucial. While many research groups have often turned to less demanding 2-D descriptors in an effort to deal with this data overload [86–89], the obvious attractions of 3-D descriptors (ligand–protein binding is a 3-D spatial property only partially described using 2-D descriptors) remain [90].

The overall requirement was thus to provide molecular descriptors with 3-D content, whilst obviating the need for molecular superposition or pharmacophore hypothesis generation. A number of different techniques have been developed to accomplish this. Most rely on binary signatures of atom pair / triplet / quartet distance combinations found within a molecule. In their original guises, such descriptors were primarily designed as shape similarity measures based on matching inter-atomic distance distributions [91–94]. It was soon realized, however, that the make up of a molecule's atomic chemistry could be incorporated to produce full pharmacophoric fingerprints [16,95–98].

7.5.1 Screening by Molecular Similarity

One application of pharmacophore fingerprints is as a ranking descriptor in database searches. Pharmacophore fingerprints can be pre-calculated for database compounds, with conformational sampling, and stored in an efficient format. An example of this are pre-stored four-center pharmacophore fingerprints, where one line of encoded information uses ~11 kilobytes of space for 1000 pharmacophores, (S. J. Cho and J. S. Mason, unpublished results). Comparisons of such databases against probe fingerprints can be accomplished at speeds of more than 700000 compounds per hour on a single Silicon Graphics R10000, 250Mhz CPU. The results of these comparisons provide a powerful 3-D virtual screening method, both directly and as an adjunct to other methods. Similarity is measured using potential pharmacophore counts (per compound and in common) and similarity indices such as the modified Tanimoto index [15,99] shown in Eq. (7.1):

$$T^{M}{}_{AB} = \frac{\sum_{i=1}^{n} P_{Ai} P_{Bi}}{0.25 \sum_{i=1}^{n} P_{Ai} \left(1 - \frac{\sum_{i=1}^{n} P_{Ai} P_{Bi}}{\sum_{i=1}^{n} P_{Bi}}\right) + 0.05 \sum_{i=1}^{n} P_{Bi} + \sum_{i=1}^{n} P_{Ai} P_{Bi}} \tag{7.1}$$

where T^{M} is the modified Tanimoto similarity, P is a property (in this case 0/1 from binary pharmacophore fingerprint), A is a database molecule, B is the probe molecule, n is the number of theoretically accessible pharmacophores, and i is the i^{th} pharmacophore. Detailed examples highlighting variants of this approach are discussed below.

7.5.1.1 Geometric Atom Pair Descriptors

Sheridan and co-workers have produced a number of seminal papers in the area of chemical similarity screening [16–17,19–20,49,54,100] (see also Chapter 4). Among these was a study of database screening using an extension of topological atom pair descriptors [101] to geometric atom pairs [16]. The method works by using a series of pre-calculated (10–25) conformations for each molecule within a flexibase [29]. For each structure, the atoms were broken down into two pharmacophoric centers types. The first are termed geometric binding property pairs and use basically the same atom-typing defined in Section 7.2.2 (donor, acceptor, acid, base, hydrophobic, polar, and other). The second are termed geometric atom pairs with atoms broken down by a combination of element type, number of neighbors, and electron count [101]. For each conformation, all combinations of atom pairs are analyzed, with the atom pair and distance combination corresponding to a particular bin in the fingerprint. Inter-atom distance is partitioned into a series of 30 continuous bins over 1–75.3Å. The contribution of each atom pair to the descriptor is divided across these bins according to the position of the distance relative to the bin centers (Figure 7.5). The resultant histograms of each probe and database molecule conformation are then compared according to Eq. (7.2), giving the similarity of molecules A and B determined from the overlap of each descriptor bin fk:

$$Sim_{AB} = \frac{\sum_k \min(f_{Ak} f_{Bk})}{0.5\left[\sum_k f_{Ak} \sum_k f_{Bk}\right]} \tag{7.2}$$

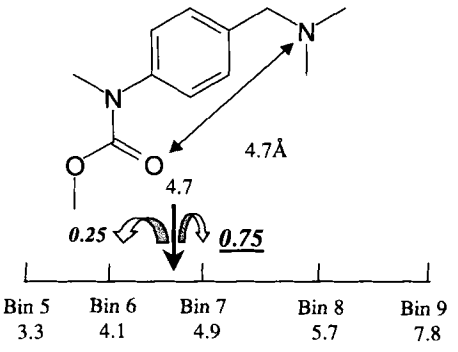


Figure 7.5. Bin assignment for geometric atom pair descriptors. Contribution of each atom pair (shown here underlined) is inversely dependent to the distance of the neighboring bins (here bins 6 and 7). The resultant descriptor is thus a histogram distribution of atom pairs with a sum total bin contribution of $n(n-1)/2$ per molecule. Adapted from [16].

The technique was compared with its topological equivalent (counting bond connections between atoms to determine inter-atomic "distance"). Comparisons were undertaken on ~30000 structures from the Derwent Standard drug file (SDF) [102], using probe molecules with known activity against a particular target to rank the database using combinations of

geometric and topological descriptors. Both methods were able to significantly enrich the top of their respective hit lists with other actives molecules for the same target (~20–30-fold enhancement over random in the top 300 compounds). Nevertheless, as was illustrated previous sections, the 3-D geometric descriptors were again able to pick out active chemotypes with greater structural variation relative to the 2-D screens (Figure 7.6).

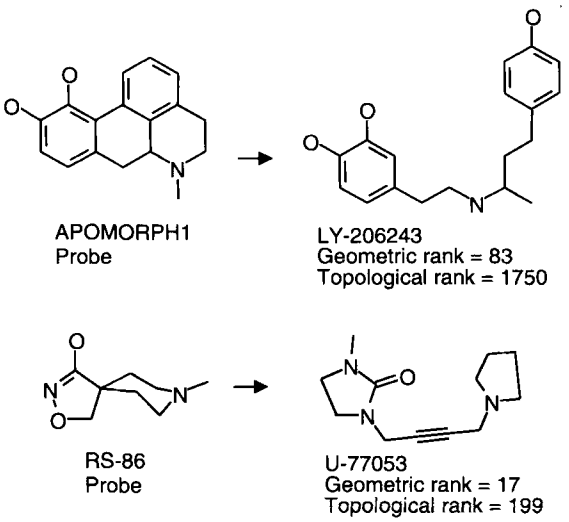


Figure 7.6. Sample molecules illustrating the difference in 2-D topological and 3-D descriptor rankings from geometric atom pair studies. Both examples illustrate how geometric descriptors rank active chemotypes higher when their structures begin to diverge from the parent probe. Adapted from [16].

7.5.1.2 Fingerprints Based on Pharmacophore Triplets and Quartets

The atom pair descriptors described above have been extended by a number of research groups to cover pharmacophore triplets and quartets [77,95–98,103]. With such descriptors, triangles (tetrahedra) are formed from all combinations of 3 (4) pharmacophoric points for all conformations of a given molecule. As before, each descriptor bin represents a particular combination of pharmacophore points (Donor–Acceptor–Aromatic, Donor–Basic–Aromatic, etc.) and distances. Unlike the atom pair measures, such descriptors are generally applied in a binary manner, with the resultant bit string containing the pharmacophores from the whole conformational ensemble of the molecule (Figure 7.7). A perceived advantage of these descriptors over atom pairs is their potentially superior descriptive power due to the increased "shape" information (inter-pharmacophore distance relationships) content of the individual descriptors [15]. Further, since three-point pharmacophores represent planes or slices through 3-D shapes, four-point descriptors offer further potential 3-D content including information on volume and chirality [15,77]. Determination of the relative merits of two, three and four-point pharmacophore descriptors is an area of ongoing study [15,77,104].

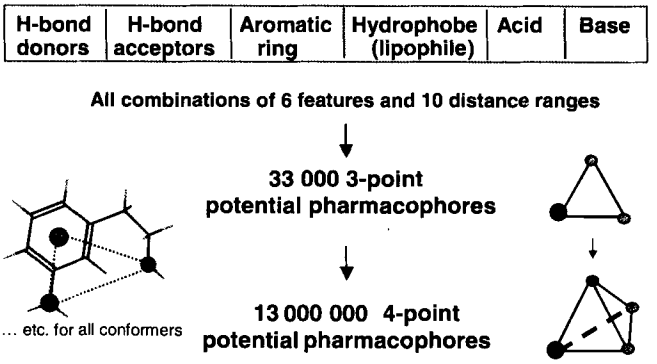


Figure 7.7. Pharmacophore triplet/quartet fingerprint creation. Unlike with geometric atom pairs (Figure 7.5), assignment is purely binary (on or off), with no „bleeding" between bins. Note the large difference in bin count between three- and four-center pharmacophores. These extra bins provide additional shape information, thus increasing molecular separation in similarity and diversity studies. Separation plays a central role in determining the final result of such calculations. Too little separation results in a noisy descriptor and hence too many molecules being defined as similar. When too large a separation exists, trivial differences can have a disproportionately negative effect on the similarity value. Achieving the optimum balance for molecular separation is an essential but tricky task. Adapted from [77].

As we have already shown, pharmacophore-based geometric descriptors have a major advantage over topological descriptors in being able to pick up chemotypes with major variations in structural makeup. This is of particular importance when exploiting a peptide lead, where screen success is dependent on the ability remove the peptidic nature of the structure. A nice example of this comes from the work of Pickett *et al.* [98]. A customized [97,15,105] version of Chem-Diverse [103] was used to calculate four-point pharmacophore fingerprints for a 100000 compound database. For the four-point pharmacophores, ten distance ranges were used in conjunction with the pharmacophore atom types mentioned in Section 7.2.2, resulting in a binary key with 24.4 x 106 theoretically accessible pharmacophores. The known RGD binding motif found in fibrinogen [106] was then used to screen the database, which had been seeded with fibrinogen receptor antagonists covering a wide range of structural classes [107] (Figure 7.8). Solution NMR and crystal structures of the fibronectin Type III domain locate the RGD motif [108–110] in a disordered loop region, suggesting a degree of flexibility. Given this flexibility, it would be difficult to identify a reasonable single pharmacophore from knowledge of the crystal structure alone. This highlights an advantage of the conformational ensemble descriptor. By merging the fingerprints of all the conformations for a particular molecule, the resultant descriptor provides an overall measure of the pharmacophores available to that structure. This is particularly valuable when a hypothesized bioactive conformation is not known, since no such model is required to undertake the search. As a result screens can be undertaken even when only limited structural information is available, as is the case here.

The four-point pharmacophore key of the RGD tripeptide probe, N- and C-capped with amide groups, were generated and compared against the combined (RPR + Fibrinogen) set. When the small molecule leads are compared with the RGD motif (Figure 7.8), it is easy to

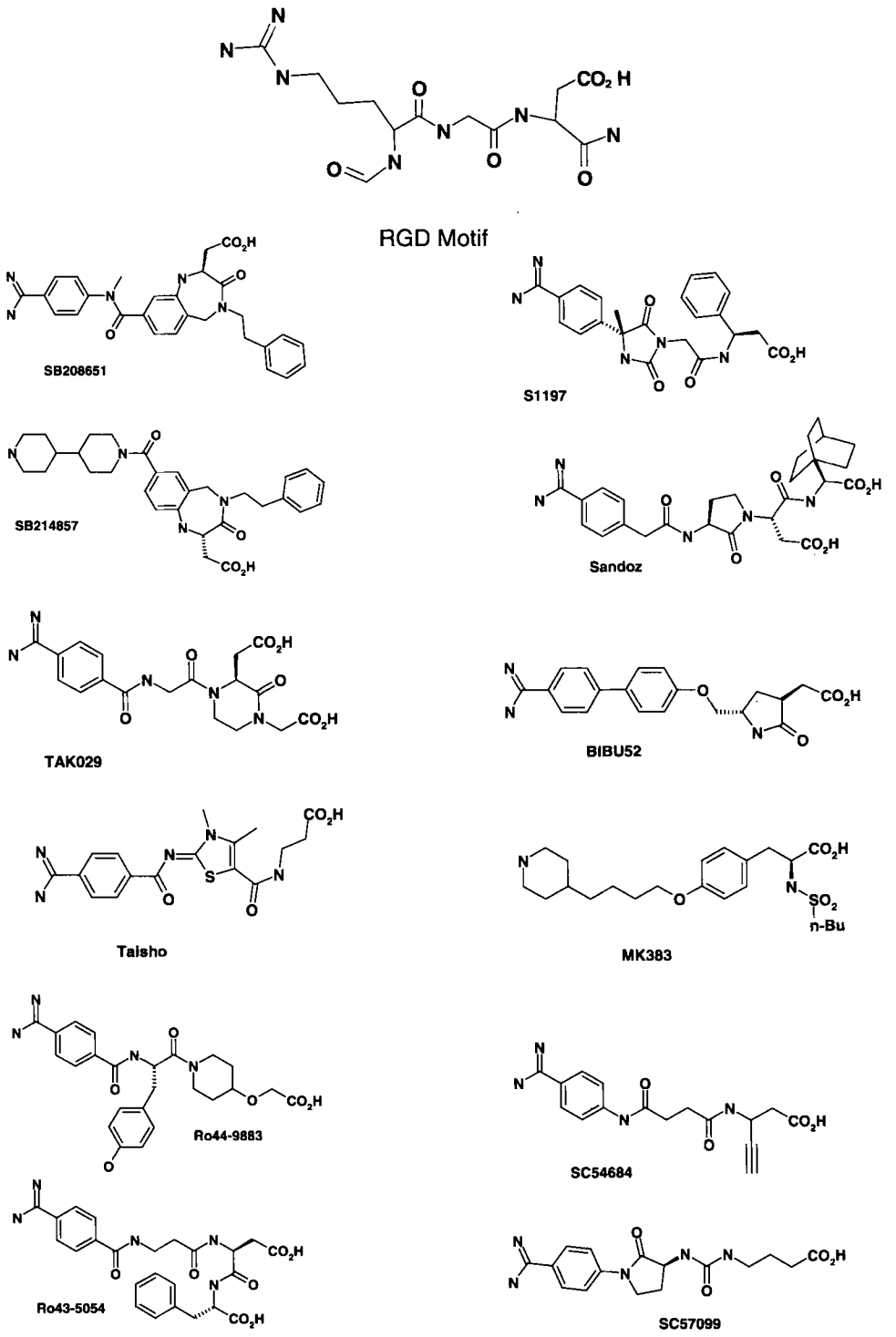


Figure 7.8. RGD motif used in fibronectin search, together with 12 structurally diverse fibronectin

see that there is very limited 2-D resemblance. A 2-D similarity search will thus clearly not do well in this context. Nevertheless one might argue that a simple search for molecules containing both an acid and a base would yield as good a result as a more complicated pharmacophore fingerprint analysis. To test this, all such molecules were abstracted from the database – 1645 in all – with 881 exhibiting at least one pharmacophore in common with the RGD motif. These 881 compounds were subjected to ranking by pharmacophore overlap to the RGD probe. Results are presented in Figure 7.9. Even in this severely constrained case a large enrichment is observed, with over two-thirds of the actives appearing in the first 100 molecules. When no such constraint is applied, all the actives appear in the top 3% of the dataset of 100000 compounds. Similarly impressive results were obtained when each of the active chemotypes was used in place of RGD as the search probe. These investigations provide yet another excellent example of the ability of pharmacophore descriptors to extract novel active chemotypes, in this case using only a minimum of structural information.

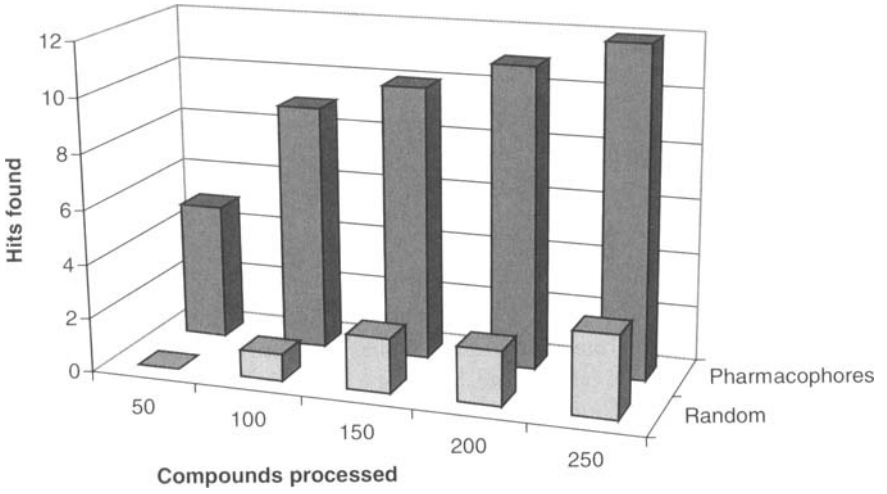


Figure 7.9. Hit rate enrichment results of RGD pharmacophore fingerprint search [98]. Expected random hit rate is given for reference.

7.5.1.3 Relative Diversity/Similarity Using Pharmacophores

The ability to separate chemical structure from ligand-binding properties has allowed researchers to make an extremely useful modification to the standard descriptor evaluation. By forcing one of the points in the pharmacophoric description to be a group, substructure or site-point of interest, a fingerprint can be generated that describes the possible pharmacophoric shapes from the viewpoint of that special point/substructure (Figure 7.10). This creates a "relative" or "internally referenced" measure of diversity, enabling new design and analysis methods. The technique has been extensively used to design combinatorial libraries that contain "privileged" substructures focused on seven-transmembrane G-protein coupled receptors [15].

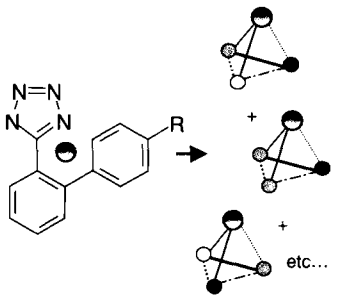


Figure 7.10. Example of privileged pharmacophore center-typing. Here biphenyl tetrazole, a substructure seen in a number of GPCR inhibitors, is specifically defined as a pharmacophore feature (center-type), using a centroid dummy atom. Only pharmacophores which include this type are included in the fingerprint, thus providing a relative measure of diversity / similarity with respect to the privileged center. Adapted from [15].

7.5.1.4 Pharmacophore Fingerprints from Protein-Binding Sites

Pharmacophore fingerprints calculated from a ligand can also be compared to the pharmacophore fingerprints of complementary site-points in a target binding site. The pharmacophore fingerprints of different target binding sites can also be compared, providing a measure of structure-based diversity (the binding sites defining the relevant pharmacophore diversity). The multiple potential pharmacophore method thus provides a novel method to measure similarity when comparing ligands to their binding site targets, with applications such as virtual screening and structure-based combinatorial library design, and to compare binding sites themselves. These protein binding site pharmacophore fingerprints can also be used to allow for flexibility of the binding site, by using a union fingerprint generated from several different binding site conformations.

An example of the method has been reported [15,77] from studies on three closely related serine proteases thrombin, factor Xa and trypsin. Using four-point pharmacophores, fingerprints were generated from site-points positioned in the active sites using the results of GRID [70] analyses (Figure 7.11). Fingerprints were also generated using full conformational flexibility for some highly selective and potent thrombin and factor Xa inhibitors and receptor-based similarity was investigated as a function of common potential four-point pharmacophores for each ligand–receptor pair to resolve enzyme selectivity. Identical studies were also undertaken using three-point pharmacophores. The goal of the studies was to see if common potential pharmacophores could give information pertaining to relative enzyme selectivity. The results indicated that the use of just four-point potential pharmacophores gave correct indications as to relative selectivity for this set of related enzymes. The thrombin and factor Xa inhibitors exhibit greater similarity with the complementary four-point pharmacophore fingerprints of the thrombin and factor Xa active sites, respectively, than with the potential pharmacophore keys generated from the other enzymes. When three-point pharmacophores were used, however, poor resolution of enzyme selectivity was observed. These results suggest that, for this study, four-point pharmacophores are better suited for comparisons based on pharmacophore overlap alone.

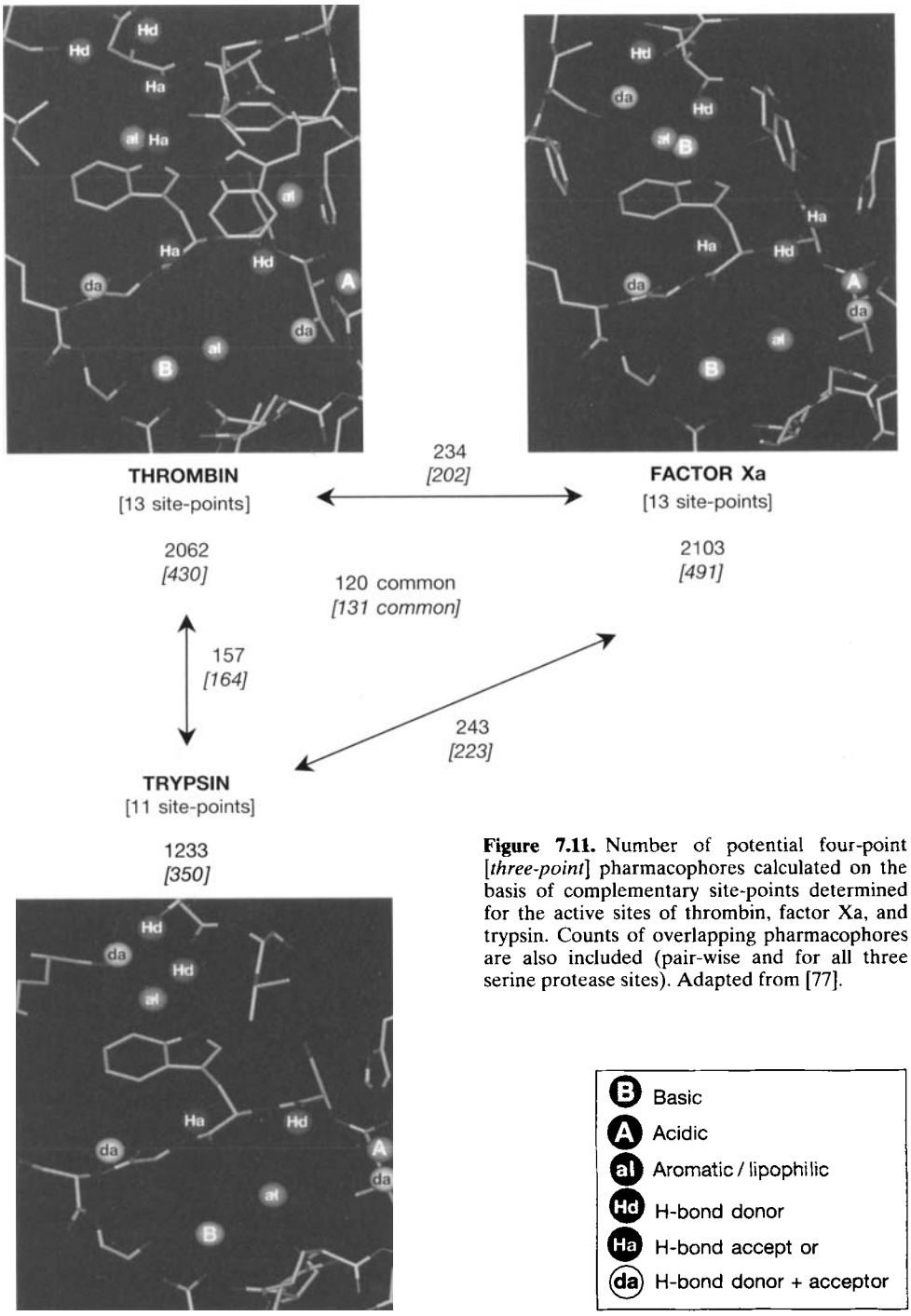


Figure 7.11. Number of potential four-point [*three-point*] pharmacophores calculated on the basis of complementary site-points determined for the active sites of thrombin, factor Xa, and trypsin. Counts of overlapping pharmacophores are also included (pair-wise and for all three serine protease sites). Adapted from [77].

7.5.2 Combinatorial Library Design Using Pharmacophore Fingerprint Ensemble

The above examples illustrate that multi-pharmacophore descriptors contain important information relating to biological activity. It is also possible to apply such descriptors to combinatorial library design through their application as a constraint in set selection, where the ability to determine the similarities and differences between structurally diverse molecules is particularly attractive. This is generally accomplished through the creation of an ensemble pharmacophore dataset measure. Such a descriptor attempts to condense the individual dataset molecule pharmacophore fingerprints into a single measure that describes the important features of the dataset as a whole [15,103,111–116].

7.5.2.1 Binary Pharmacophore Ensemble Descriptors and Beyond

A number of commercial programs make use of this approach [103,116]. The first implementation was the Chem-Diverse program [103], which was developed to exploit three- and four-point pharmacophore information in molecular diversity profiling. The Chem-Diverse protocol for molecular diversity is based on trying to obtain the maximum coverage of binary pharmacophore space by potential combinatorial chemistry products. The general protocol for the program is shown in Figure 7.12. A central part of the Chem-Diverse compound selection procedure requires on-the-fly conformational analysis of all potential library products. Any pharmacophores found are added to a single pharmacophore key, which describes the ensemble of selected molecules. Compounds are only selected if the set of pharmacophores they express overlaps with the ensemble key by less than a user-defined amount, that is, if the molecule contains a significant number of previously unseen pharmacophores. As a consequence, the results of such searches are dependent on the order in which the molecules are extracted from the database. Chem-Diverse chooses molecules based on what it considers to be the most diverse set of products ("cherry-picking"), with no explicit reference to the constituent reagents. While this is theoretically the most efficient method for ensuring a maximum diversity library, it is a combinatorially inefficient selection for synthesis, where reagent array selection is preferred. A reagent array as defined here means that all reagents from one component of a combinatorial library are reacted with all reagents in the other components. For example, a possible array for a 100 compound two-component library would be a set of 10 x 10 reagents). Within Chem-Diverse, direct modification of the search criteria to include additional molecular properties such as shape is not generally feasible. This is because the diversity function employed by Chem-Diverse uses only pharmacophoric properties. Such constraints are instead accomplished implicitly by assigning upper and lower bounds for given properties, or controlling the order in which molecules are processed. The binary form of the ensemble fingerprints also limits their utility somewhat, since such a key only registers whether or not a particular pharmacophore exists in the selected molecular ensemble, not how many times it is found. This makes the key prone to saturation, even when artificially small distance bins are applied, as is the case with the default bin settings in Chem-Diverse (32 to be exact).

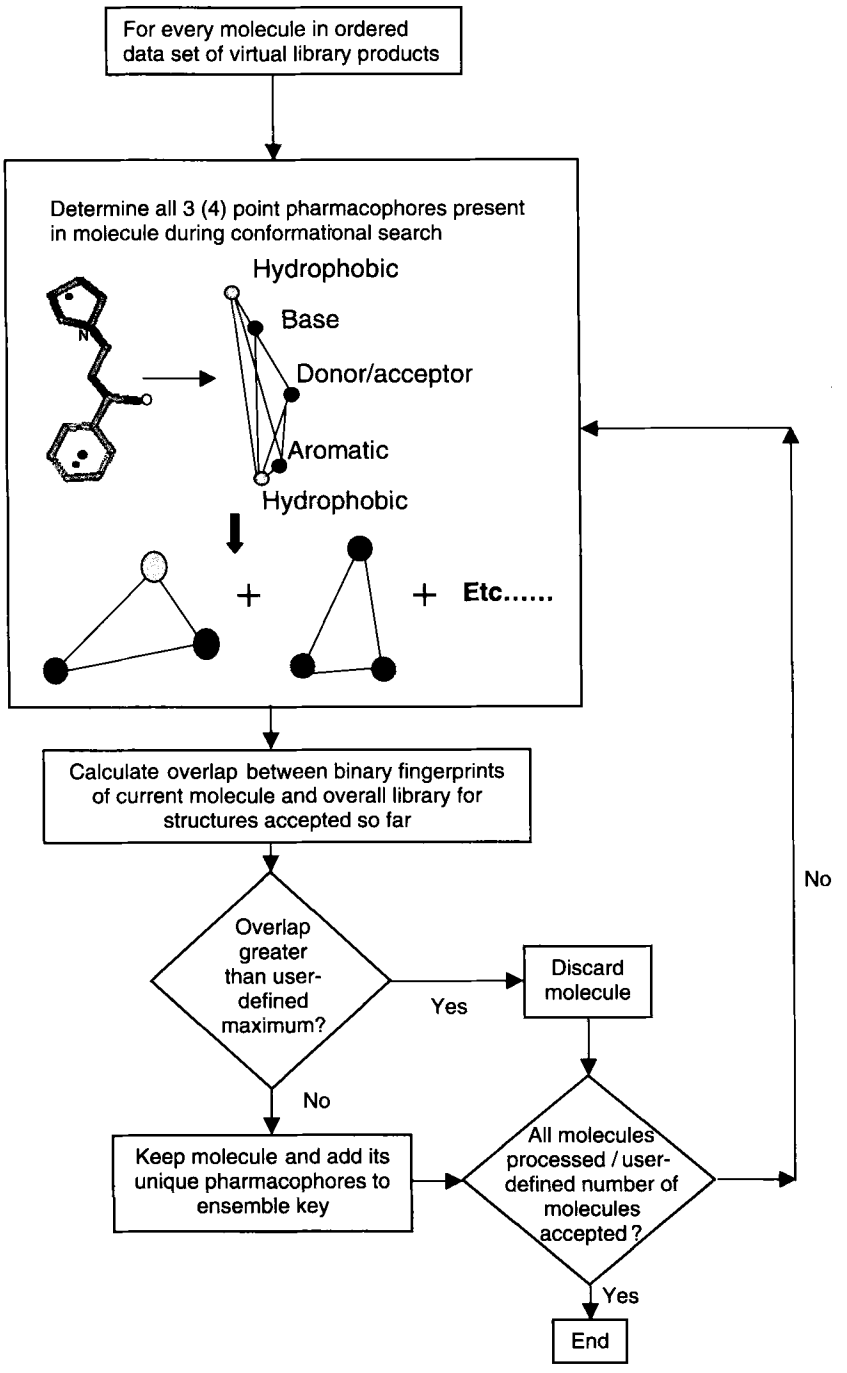


Figure 7.12. Outline of Chem-Diverse compound selection strategy. Adapted from [113].

Good and co-workers [111–113] created the HARPick program to address many of these Chem-Diverse features. The basic outline of HARPick is illustrated in Figure 7.13. The main structure of the program revolves around a Monte Carlo Simulated Annealing [117] protocol. By applying such a stochastic optimization technique, reagent selection is divorced from the diversity evaluation. Selections can then be made in reagent space, while diversity is calculated in product space. This allows the user direct control over the number of reagents selected, allowing the method to easily permit selections according to reagent array. Further, it is a simple task to introduce flexible scoring functions incorporating many diverse properties. It is thus possible to add additional secondary descriptors, including those that help prevent the selection of non drug-like (e.g. promiscuous and ultra-flexible), as well as alternative diversity descriptors.

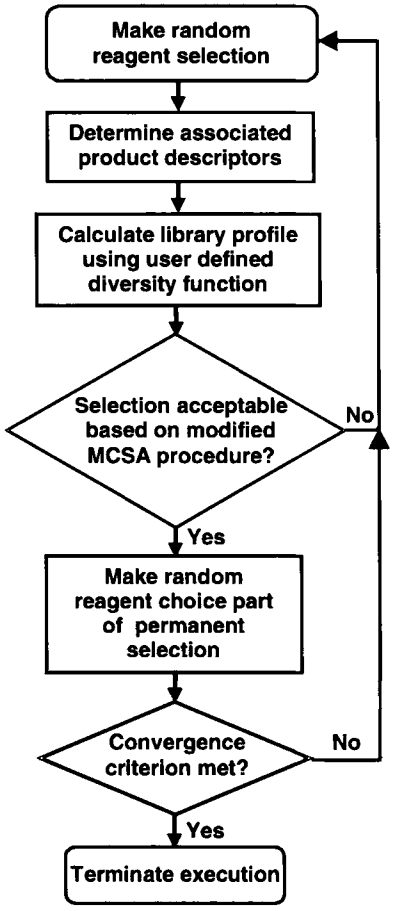


Figure 7.13. Outline of HARPick compound selection strategy. Advantages over the Chem-Diverse paradigm include reagent-based compound through the use of stochastic optimization (in this case Monte Carlo Simulated Annealing), and a multiple component product-based diversity function (including non-binary pharmacophore fingerprint, compound shape, and flexibility terms). Adapted from [113].

Major alterations were also made to the pharmacophore descriptor methodology. Pharmacophore center types, distance bins, and conformational search parameters were all al-

tered to improve performance. Compressed pharmacophore fingerprints were calculated ahead of time rather than on-the-fly and stored in memory at search time to permit rapid access. Finally, and perhaps most importantly, a non-binary description of pharmacophore space was employed, permitting not only knowledge of which pharmacophores are hit, but also how many times. The importance of this approach is highlighted by an analysis of just 20168 structures from the SDF [102]. Over 68% of all HARPick theoretically accessible pharmacophores were found to be present in the dataset. Considering that the average size of a company compound collection will be far in excess of 20000, this illustrates the problem of binary fingerprint saturation. Further, the distribution of pharmacophores is far from even, with ~56000 hit 1–10 times and ~27000 hit more than 50 times. The full results of this analysis are shown in Figure 7.14. HARPick diversity was thus tuned to include a term (*Conscore*) forcing molecules to occupy relative rather than absolute voids in pharmacophore space:

$$Conscore = \sum_{i=1}^{a} O_i S_i \tag{7.3}$$

where *Conscore* is the pharmacophore constraint score, O_i is the number of times pharmacophore i has been hit for molecules selected from current dataset, S_i is the score associated with pharmacophore i for the constraining library, and a is the number of accessible pharmacophores.

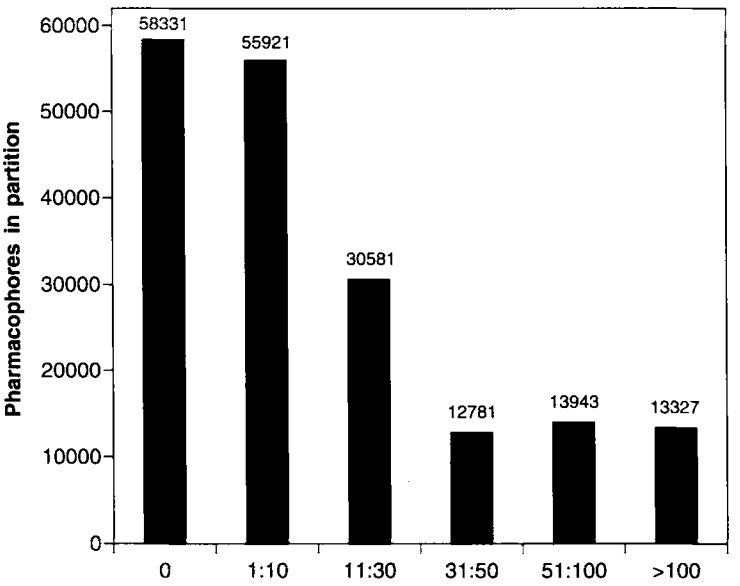


Figure 7.14. Triplet pharmacophore frequency distribution histogram for 20169 SDF molecules. Total pharmacophore count = 4797745. Number of theoretically accessible pharmacophores = 184884. Number of unique pharmacophore triplets found in library = 126553, representing more than 68% coverage (126553/184884) of theoretically accessible pharmacophores. Adapted from [113].

$$S_i = \left[\max(0, (avcov - Oc_i))\right]^v \tag{7.4}$$

where max(0,(*avcov*–Oc_i)) is the maximum of the values 0 and *avcov*–Oc_i, *avcov* is the average pharmacophore count across all occupied pharmacophores in constraining library, Oc_i is the number of molecules containing pharmacophore *i* in the constraining library, and *v* is the user defined weight.

$$avcov = \frac{\sum_{i=1}^{a} \min(Oc_i, \beta)}{Unique_c} \tag{7.5}$$

where min(Oc_i,β) is the minimum of the values Oc_i and β (the user-defined maximum contribution to *avcov* by any single pharmacophore), and $Unique_c$ is the number of pharmacophore bins occupied in constraining library.

Conscore forms part of the full HARPick diversity function:

$$Energy = \frac{Unique^w \times Conscore \times Partscore_{pp}{}^x \times Partscore_{pa}{}^x \times Partscore_{ha}{}^x \times S}{Totpharm^y \times Flex^z \times n} \tag{7.6}$$

where *w,x,y,z* are user-defined weights, *Unique* is the number of unique pharmacophores found (equivalent to the Chem-Diverse binary ensemble key), $Partscore_{xx}$ are shape descriptor distribution functions, *totpharm* is the total number of pharmacophores present in selected molecule set, and *flex* is the library flexibility measure.

The utility of this function was tested on a selection of 1000 compounds (20 x 50 reagent array) from a two-component reaction involving a total pool of 33835 potential product molecules (67 x 505 reagent set). The selections were compared against the SDF dataset. Two runs were undertaken. In the first *Conscore* was deactivated (*v* weight set to 0), allowing selections based purely on internal diversity. For the second run the *avcov* value was set to 10 (which for the SDF dataset meant that all pharmacophore bins with an occupation level of 7 or less scored), and *v* was set to 1. The resulting selections were thus constrained to find pharmacophores present seven times or less in the SDF dataset. In addition three random 20 x 50 reagent selections were run to provide baseline results. The results highlight the utility of the non-binary term (Table 7.3). For ~20% additional total pharmacophores in the library, a greater than 100% increase in pharmacophores occupying low occupancy SDF pharma-

Table 7.3. Results of pharmacophore distribution (*Conscore* – see Eqs. 7.3–7.6) constrained HARPick run. Adapted from [113].

Calculation	Unique pharmacophore count	Total pharmacophore count	Pharmacophores found in scoring bins
Unconstrained	78567	806539	99696 (12%)
Conscore constrained	79125	1079998	203061 (19%)
Random	51791	1033772	51837 (5%)

cophores is achieved. By inverting the *Conscore* term, the procedure can be turned into a method for selecting for high occupancy bins, which is the requirement when undertaking focused library designs. The great flexibility of this kind of stochastic optimization methodology has led to its use by many other researchers for library design [90,118–119].

7.5.3 Pharmacophore Fingerprint Ensembles as QSAR Descriptors

A recent development in the analysis of pharmacophore fingerprint ensemble analysis has been their application in QSAR studies. McGregor and Muskal [115] developed their own variant of three-point pharmacophore fingerprint to undertake these studies. The six pharmacophoric atom types already described in Section 7.2.2 were applied together with an additional definition of *other* for all remaining unassigned atoms. Interestingly, the addition of this extra type improved the resulting QSAR model statistics, the reason for which was ascribed to indirect volume description. Distances were divided into six bins as described by Pickett *et al.* [97], giving a total of 10549 accessible pharmacophores after removal of pharmacophores failing the triangle rule (the length of one side cannot exceed the length of the other two), and/or redundant by symmetry. It should be noted that these pharmacophore elimination techniques are generally applied by all pharmacophore fingerprint generation methods. The pharmacophore descriptors were used as PLS [120] QSAR descriptors for three estrogen receptor (ER) datasets previously analyzed using other QSAR descriptors [121]. The resultant models were found to be more predictive than those developed for the same sets using CoMFA, CODESSA and Holographic QSAR approaches [121]. In the fourth test a model was derived from 15 actives (activity set to 1.0) from one dataset of the initial QSAR studies, plus 750 "inactives" taken from the non-ER active structures in the MDDR [122] (activity set to 0.0). This test was designed to mimic the kind of "noisy" data one would see from HTS screens. Analysis of the resultant model was undertaken by scoring 250 MDDR ER active molecules, 86 ER actives from a combinatorial library, and 8290 "inactives" from the remainder of the MDDR not used in model training. Setting a cutoff value of 0.2 to define the boundary between active and inactive, in all three cases >87% of each test set were assigned correctly. The results of these studies provide further evidence of the utility of these descriptors, with an obvious application of the fourth test model as the primary

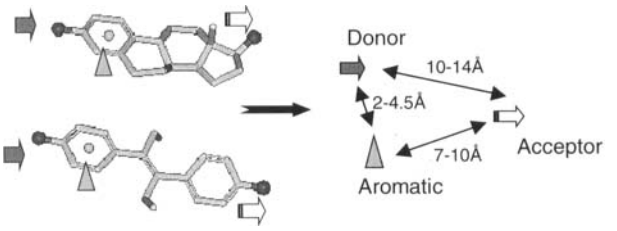


Figure 7.15. Pharmacophore with greatest positive weighting in study 4 of McGregor and Muskal [115] mapped to the natural ligand estradiol (top) and the most potent dataset compound diethylstilbestrol (bottom). Again, the diverse nature of the chemotypes being mapped is worthy of note. Adapted from [115].

constraint in a focused library design. Of added attraction is the ability to turn the QSAR equation back into a graphic based on the highly weighted pharmacophores in the model. This is illustrated in Figure 7.15, which shows the highest positive weighted pharmacophore for the test 4 QSAR equation in two diverse ER-active chemotypes.

7.6 Conclusions

This Chapter highlights the ability of pharmacophores to divorce the 3-D structural requirements for biological activity from the 2-D chemical makeup of a ligand. The resultant descriptors are thus able to exploit even limited data regarding a target to discover novel active chemotypes. The methodology can be alignment-independent if necessary, with calculation speeds rapid enough to permit their use on datasets deemed too large for most other 3-D descriptors. It is this combination of properties that makes pharmacophores such a powerful weapon in the computational chemist's arsenal.

References

[1] P. Gund, *Progress in Molecular and Subcellular Biology*. Vol. 5, Springer-Verlag, Berlin **1977**, pp. 117–143.
[2] G.R. Marshall, *3D QSAR in Drug Design*, ESCOM, Leiden **1993**, pp. 80–116.
[3] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
[4] S. Ash, M. A. Cline, W.R. Homer, T. Hurst, G. B. Smith, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71–79.
[5] Unity/SLN manual available from Tripos Inc., 1699 South Hanley Road, Suite 303, St Louis, MO 63144, USA. URL: *http://www.tripos.com*
[6] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, J. Laufer, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244–255.
[7] R. S. Pearlman, in *3D QSAR in Drug Design: Theory, Methods and Applications*, ESCOM, Leiden, Netherlands **1993**, pp. 41–80. Concord is distributed by Tripos Inc. (see [5]).
[8] J. Sadowski, C. Rudolph, J. Gasteiger, *J. Anal. Chim. Acta.* **1992**, *265*, 233-241. Corina is distributed by Oxford Molecular PLC, the Medewar Centre, Oxford Science Park, Oxford OX4 4GA, UK. URL: *http://www.oxmol.com*
[9] Chem-DBS3D/Chem-X, developed and distributed by Oxford Molecular PLC (see [8]).
[10] Convertor, available from MSI Inc, 9685 North Scanton Road, San Diego, CA 92121, USA. URL: *http://www.msi.com*
[11] E. M. Ricketts, J. Bradshaw, M. Hann, F. Hayes, N. Tanna, D. M. Ricketts, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 905–925.
[12] D. V. S. Green, in *Des. Bioact. Mol.*, American Chemical Society, Washington D.C. **1998**, pp. 47–71.
[13] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, T. E. Ferrin, *J. Mol. Biol.* **1982**, *161*, 269–288. DOCK is developed and distributed by the Kuntz group, Dept. of Pharmaceutical Chemistry, 512 Parnassus, University of California, San Francisco, CA 94143-0446, USA. URL: *http://www.cmpharm.ucsf.edu/kuntz*
[14] T. J. A Ewing, I. D. Kuntz, *J. Comput. Chem.* **1997**, *18*, 1175–1189.
[15] J. S. Mason, I. Morize, P. R. Menard, D. L. Cheney, C. Hulme, R. F. Labaudiniere, *J. Med. Chem.* **1999**, *42*, 3251–3264.
[16] R. P. Sheridan, M. D. Miller, D. J. Underwood, S. K. Kearsley, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
[17] S. K. Kearsley, S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley, R. P. Sheridan, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118–27.
[18] J. H. Van Drie, D. Weininger, Y. C. Martin, *J. Comput.-Aided Mol. Des.* **1989**, *3*, 225–251.

[19] R. P. Sheridan, A. Rusinko III, R. Nilakantan, R. Venkataraghavan, *Proc. Natl. Acad. Sci. USA* **1989**, *86*, 8165–8169.
[20] R. P. Sheridan, R. Nilakantan, A. Rusinko III, N. Bauman, K. S. Haraki, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 251–255.
[21] J. B. Moon, W. J. Howe, *Tetrahedron Comput. Methodol.* **1990**, *3*, 697–711.
[22] O. F. Guner, D. W. Hughes, L. M. Dumont, *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 408–414.
[23] N. W. Murrall, E. K. Davies, *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312–316.
[24] A. Smellie, S. L. Teig, P. Towbin, *J. Comput. Chem.* **1995**, *16*, 171–187.
[25] M. T. Stahl, A. Nicholls, R. A Anthony, A. J. Grant, *Book of Abstracts*, 217th ACS National Meeting, Anaheim, Calif., March 21–25 **1999**, American Chemical Society, Washington, D.C., COMP-026.
[26] R. Balducci, R. S. Pearlman, *Book of Abstracts*, 217th ACS National Meeting, Anaheim, Calif., March 21–25 **1999**, American Chemical Society, Washington, D.C., COMP-011.
[27] T. Hurst, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190–196.
[28] T. E. Moock, D. R. Henry, A. G. Ozkabak, M. Alamgir, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 184–189. ISIS-3D is developed and distributed by MDL Information Systems Inc., San Leandro, CA, USA. URL: *http://www.mdl.com*
[29] S. K. Kearsley, D. J. Underwood, R. P. Sheridan, R. T. Mosley, *J. Comput.-Aided Mol. Design* **1994**, *8*, 565–582.
[30] A. Smellie, S. D. Kahn, S. L. Teig, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 285–294.
[31] A. Smellie, S.D. Kahn, S.L. Teig, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 295–304.
[32] Catalyst 3D database searching and Pharmacophore hypothesis software, developed and distributed by MSI (see [10]).
[33] D. P. Dolata, A. R. Leach, K. Prout, *J. Comput.-Aided Mol. Design* **1987**, *1*, 73–85.
[34] A. R. Leach, K. Prout, D. P. Dolata, *J. Comput. Chem.* **1990**, *11*, 680–693.
[35] P. W. Walter, D. P. Dolata, *J. Mol. Graphics* **1994**, *12*, 130–138.
[36] D. P. Dolata, S. Karabatsorides, *Book of Abstracts*, 213th ACS National Meeting, San Francisco, April 13–17 **1997**, American Chemical Society, Washington DC, COMP-389.
[37] G. R. Marshall, C. D. Barry, H. E. Bosshard, R. A. Dahmkoeller, D. A. Dunn, *ACS Symposium Series* **1979**, *112*, 205–226.
[38] A. C. Good, J. S. Mason, in *Reviews in Computational Chemistry Vol. 7*, VCH, New York **1995**, pp. 67–127.
[39] G. W. A. Milne, M. C. Nicklaus, S. Wang, *SAR QSAR Environ. Res.* **1998**, *9*, 23–38.
[40] W. A. Warr, P. Willett, in *Des. Bioact. Mol.*, American Chemical Society, Washington DC **1998**, pp. 73–95.
[41] D. P. Marriott, I. G. Dougall, P. Meghani, Y. J. Liu, D. R. Flower, *J. Med. Chem.* **1999**, *42*, 3210–3216.
[42] P. Traxler, P. Furet, *Pharmacol. Ther.* **1999**, *82*, 195–206.
[43] R. C. Glen, P. Willett, G. Jones, *Book of Abstracts*, 213th ACS National Meeting, San Francisco, April 13–17 **1997**, American Chemical Society, Washington DC, COMP-007.
[44] J. D. Holliday, P. Willett, *J. Mol. Graphics Model.* **1997**, *15*, 221–232.
[45] D. Barnum, J. Greene, A. Smellie, P. Sprague, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563–571.
[46] X. Chen, A. Rusinko III, A. Tropsha, S. S. Young, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 887–896.
[47] A. K. T. Ting, P. Johnson, S. Green, R. McGuire, *Book of Abstracts*, 218th ACS National Meeting, New Orleans, Aug. 22–26 **1999**, American Chemical Society, Washington DC, COMP-141.
[48] I. B. Bersuker, S. Bahceci, J. E. Boggs, R. S. Pearlman, *SAR QSAR Environ. Res.* **1999**, *10*, 157–173.
[49] M. D. Miller, R. P. Sheridan, S. K. Kearsley, *J. Med. Chem.* **1999**, *42*, 1505–1514.
[50] Y. C. Martin, M. G. Bures, E. A. Danaher, J. DeLazzer, I. Lico, P. A. Pavlik, *J. Comput.-Aided Mol. Design* **1993**, *7*, 83–102. DISCO is Distributed by Tripos Inc. (see [5]).
[51] A. T. Brint, P. J. Willett, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 152–158.
[52] G. Jones, P. Willett, R. C. Glen, *J. Comput.-Aided Mol. Design* **1995**, *9*, 532–549.
[53] C. Lemmen, T. Lengauer, G. Klebe, *J. Med. Chem.* **1998**, *41*, 4502–4520.
[54] M. D. Miller, S. K. Kearsley, D. J. Underwood, R. P. Sheridan, *J. Comput.-Aided Mol. Design* **1994**, *8*, 153–174.
[55] M. C. Lawrence, P. C. Davis, *Proteins* **1992**, *12*, 31–41.
[56] P. Burkhard, P. Taylor, M. D. Walkinshaw, *J. Mol. Biol.* **1998**, *277*, 449–466.
[57] Cerius 2 / Structure-Based Focussing; developed and distributed by MSI Inc. (see [10]).
[58] Design in Receptor (DiR); developed and distributed by Oxford Molecular (see [8]).

[59] M. L. Connolly, *J. Mol. Graph.* **1993**, *11*, 139.
[60] For more information on SPHGEN, consult the following URL: *http://www.cmpharm.ucsf.edu/kuntz/dock4/html/Manual.20.html#pgfId=6353*
[61] B. K. Shoichet, D. L. Bodian, I. D. Kuntz, *J. Comput. Chem.* **1992**, *13*, 380–397.
[62] D. R. Ferro, J. A. Herrmans, *Acta Crystallogr.* **1977**, *A33*, 345–352.
[63] I. D. Kuntz, *Science* **1992**, *257*, 1078–1082.
[64] B. K. Shoichet, R. M. Stroud, D. V. Santi, I. D. Kuntz, K. M. Perry, *Science* **1993**, *259*, 1445–1450.
[65] C. S. Ring, E. Sun, J. H. McKerrow, G. K. Lee, P. J. Rosenthal, I. D. Kuntz, F. E. Cohen, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 3853–3857.
[66] I. D. Kuntz, E. C. Meng, B. K. Shoichet, *Acc. Chem. Res.* **1994**, *27*, 117–123.
[67] P. Burkhard, U. Hommel, M. Sanner, M. D. Walkinshaw, *J. Mol. Biol.* **1999**, *287*, 853–858.
[68] B. K. Shoichet, I. D. Kuntz, *Protein. Engineering* **1993**, *6*, 723–732.
[69] R. L. Desjarlais, J. S. Dixon, *J. Comput.-Aided Mol. Design* **1994**, *8*, 231–242.
[70] P. J. Goodford, *J. Med. Chem.* **1985**, *28*, 849-857. The GRID program is developed and distributed by Molecular Discovery Ltd., Oxford, UK.
[71] For more information, consult the following URL: *http://www.cmpharm.ucsf.edu/kuntz/dock4/html/Manual.e.html#75586*
[72] Rutgers University Research Collaboratory for Bioinformatics. URL: *http://www.rcsb.org/pdb/*
[73] 2ANS, J. J. Ory, L. J.Banaszak; to be published.
[74] 1LIC, J. M. Lalonde, D. A. Bernlohr, L. J. Banaszak; to be published.
[75] The MAKEPOINT program has some similarities with the methodology applied to the creation of site points in the SANDOCK program (see [56]).
[76] E. C. Meng, B. K. Shoichet, I. D. Kuntz, *J. Comput. Chem.* **1992**, *13*, 505–524.
[77] J. S. Mason, D. L. Cheney, *Pac. Symp. Biocomput. '99* **1999**, 456–467.
[78] A. C. Good, in *Molecular Similarity in Drug Design*, Blackie Academic and Professional, Glasgow, UK **1995**, pp. 24–56.
[79] A. C. Good, W. G. Richards, *Perspect. Drug Discovery Des.* **1998**, *9–11*, pp. 321–338.
[80] W. G. Richards, D. D. Robinson, Molecular similarity, *Vol. Math. Its Appl.* **1999**, *108*, 39–49.
[81] R. Carbo-Dorca, E. Besalu, *Theochem.* **1998**, *451*, 11–23.
[82] D. A. Thorner, D. J. Wild, P Willett, P. M. Wright, *Perspect. Drug Discovery Des.* **1998**, *9–11*, 301–320.
[83] T. Langer, *Perspect. Drug Discovery Des.* **1998**, *12–14*, 215–231.
[84] H. Kubinyi, *Comput.-Assisted Lead Find. Optim.*, 11th Eur. Symp. Quant. Struct.-Act. Relat., Verlag Helvetica Chimica Acta, Basel **1997**, pp. 9–28.
[85] N. C. Perry, V. J. Van Geerestein, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 607–616.
[86] R. D. Brown, Y. C. Martin, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
[87] R. D. Brown, *Perspect. Drug Discovery Des.* **1997**, *7/8*, 31–49.
[88] R. D. Brown, Y. C. Martin, *SAR QSAR Environ. Res.* **1998**, *8*, 23–39.
[89] Y. C.Martin, M. G. Bures, R. D. Brown, *Pharm. Pharmacol. Commun.* **1998**, *4*, 147–152.
[90] M. G. Bures, Y. C. Martin, *Curr. Opin. Chem. Biol.* **1998**, *2*, 376–380.
[91] G. W. Bemis, I. D. Kuntz, *J. Comput-Aided Mol. Design* **1992**, *6*, 607–628.
[92] R. Nilakantan, N. Bauman, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79–85.
[93] A. C. Good, T. J. A. Ewing, D. A. Gschwend, I. D. Kuntz, *J. Comput-Aided Mol. Design* **1995**, *9*, 1–12.
[94] W. Fisanick, K. P. Cross, A. Rusinko III, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664–674.
[95] A. C. Good, I. D. Kuntz, *J. Comput-Aided Mol. Des.* **1995**, *9*, 373–379.
[96] M. J. Ashton, M. Jaye, J. S. Mason, *Drug Discovery Today* **1996**, *1*, 71–78.
[97] S. D. Pickett, J. S. Mason, I. M. Mclay, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214–1233.
[98] S. D. Pickett, I. M. McLay, D. E. Clark, *J. Chem. Inf. Comput. Sci.* **2000**, in press.
[99] P. Willett, in *Molecular Similarity in Drug Design*, Blackie Academic and Professional, Glasgow **1995**, pp. 110–137.
[100] R. P. Sheridan, R. B. Nachbar, B. L. Bush, *J. Comput.-Aided Mol. Design* **1994**, *8*, 323–340.
[101] R. E. Cahart, D. H. Smith, R. Ventkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
[102] Standard Drug File version 6, developed and distributed by Derwent Publications Ltd., London, England **1991**. Now known as the World Drug Index (WDI).
[103] E. K. Davies, in *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery, American Chemical Society*, Washington DC **1996**, pp. 309–316.
[104] A. C. Good, *Internet J. Chem.*, in press.

[105] S. D. Pickett, C. Luttmann, V. Guerin, A. Laoui, E. James, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 144–150.
[106] C. D. Eldred, B. D. Judkins, *Prog. Med. Chem.* **1999**, *36*, 29–90.
[107] S. A. Mousa, D. A. Cheresh, *Drug Discovery Today* **1997**, *2*, 187–199.
[108] A. L. Main, T. S. Harvey, M. Baron, J. Boyd, I. D. Campbell, *Cell* **1992**, *71*, 671–678.
[109] C. D. Dickinson, B. Veerapandian, X. P. Dai, R. C. Hamlin, H. Nguyen, E. Ruoslahti, K. R. Ely, *J. Mol. Biol.* **1994**, 236, 1079–1092.
[110] D. J. Leahy, I. Aukhil, H. P. Erickson, *Cell* **1996**, *84*, 155–164.
[111] R. A. Lewis, A. C. Good, S. D. Pickett, in *Computer-Assisted Lead Finding and Optimization*, 11th Eur. Symp. Quant. Struct.-Act. Relat., Verlag Helvetica Chimica Acta, Basel **1997**, pp. 135–156.
[112] A. C. Good, S. D. Pickett, R. A. Lewis, *Book of Abstracts*, 213th ACS National Meeting, San Francisco, April 13–17 **1997**, American Chemical Society, Washington DC, COMP-339.
[113] A. C. Good, R. A. Lewis, *J. Med. Chem.* **1997**, *40*, 3926–3936.
[114] S. D. Pickett, D. E. Clark, R. A. Lewis, *Book of Abstracts*, 215th ACS National Meeting, 29 March – 2 April **1998**, Dallas, TX., American Chemical Society, Washington DC, COMP-009.
[115] M. J. McGregor, S. M. Muskal, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569–574.
[116] MSI's Cerius/Catalyst (see [10]) Tripos's Pharmacophore Triplets (see [5]) both have differing implementations of pharmacophore triplet and quartet fingerprints descriptors.
[117] S. Kirkpatrick, C. D. Gelatt, M. P.Vecchi, *Science* **1983**, *220*, 671–680.
[118] V. J. Gillet, P. Willett, J. Bradshaw, D. V. S. Green, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169–177.
[119] L. Weber, *Drug Discovery Today* **1998**, *3*, 379–385.
[120] S. Wold, M. Sjöström, L. Eriksson, in *Encyclopedia of Computational Chemistry*, John Wiley & Sons, New York **1998**, pp. 2006–2021.
[121] W. Tong, D. R. Lowis, R. Perkins, Y. Chen, W. J. Welsh, D. W. Goddette, T. W. Heritage, D. M. Sheehan, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 669–677.
[122] MDDR, developed and distributed MDL Information Systems Inc. (see [28]).

Virtual Screening for Bioactive Molecules
Edited by Hans-Joachim Böhm and Gisbert Schneider

8 Evolutionary Molecular Design in Virtual Fitness Landscapes

Gisbert Schneider

8.1 Introduction

One of the goals of virtual screening is to generate molecules with desired properties and bioactivity "from scratch". This idea of computer-assisted *de novo* design is not new. Several, for the most part structure-based techniques were developed during the past years, which led to a number of very successful applications [1–5]. Among the prominent representatives are the software packages LUDI [6], BUILDER [7], CAVEAT [8], PRO-LIGAND [9], PRO-SELECT [10], GENSTAR [11], and CONCERTS [12]. These algorithms generally require a three-dimensional (3-D) structural model of a binding or receptor site, identify potential ligand–receptor interaction or attachment points, and construct novel molecular entities by combinatorial or piecewise amalgamation of atoms and molecular fragments. In most applications, the compatibility of novel structures or individual substituents in a given position is estimated by empirical scoring functions correlated to binding free energy (see Chapter 11) [13]. Although combinatorial docking procedures clearly proved their applicability to *de novo* design [14], one of the major problems still to be solved is the accurate prediction of binding energies [4,15].

A complementary approach to starting from a receptor structure is to build on a pharmacophore hypothesis that was derived from a known bioactive molecule or ligand [16,17]. Based on a pharmacophore model, alternative molecular architectures can be virtually assembled mimicking the pharmacophore pattern present in the original template structure. This methodology and related tactics represent practicable approaches to ligand *de novo* design when a high-resolution receptor structure is not available, which is particularly the case for many neuroreceptors in CNS research, including the large group of different G-protein-coupled receptors (GPCR) (see Chapter 1) [18]. The NEWLEAD program was an early attempt following the idea of pharmacophore-based design [19], more recent examples are MOLMAKER [20] and CATS [21]. The latter two do not explicitly generate novel structures. Instead, a pharmacophore similarity search is performed in large virtual compound collections and "historical" databases (see Chapter 7). Moreover, a small number of peptide design algorithms were developed that either build peptide-ligands based on the receptor structure, e.g. PRO-LIGAND [22], and the early Moon and Howe approach [23], or generate novel bioactive amino acid sequences based on a ligand-derived scoring function, like the PROSA [24] and SME software [17,25].

This brief compilation of computer programs for *de novo* design could easily be extended, as over 20 algorithms have been reported in the literature (for review of structure-based *de novo* design software, see [4]), and many advanced methods will enter the scene in the future.

There are, however, two principal problems to be solved by any *de novo* design method, irrespective of their range of potential applications, and independent of their two-dimensional or 3-D structural character:

1. **The problem of large numbers:** Usually it is impossible to perform an exhaustive search for a desired compound within a reasonable period of time. There are far too many virtual structures that can be generated by atom and fragment linking, and hence rapidly lead to a combinatorial explosion. Therefore, a systematic and efficient search strategy must be available for navigation through chemical space, i.e. here the virtual space spanned by all algorithmically tractable molecules.
2. **The scoring problem:** This topic has a twofold meaning. On the one hand, the scoring function must be appropriate for the given design project, accurately predicting realistic activity estimates (e.g., binding energy, K_i, IC_{50}). On the other, the structure of the virtual chemical space must allow for a systematic search to be performed at all. This means that an adequate, function-related ordering of molecules must exist in chemical space; otherwise a random search will take place. The scoring problem is especially hard when a receptor structure is not available, and a context-specific empirical scoring function must be developed on basis of only a small number of known active molecules.

For the purpose of this Chapter it is convenient to focus on a discussion of these two dominant issues. Additional emphasis is put on the description of a special evolutionary technique for template-based *de novo* design (TOPAS, TOPology-Assigning System) that may be employed when more conventional structure-based techniques cannot be applied, e.g. due to a lack of high-resolution receptor models. Several additional objectives, which also motivated the development of TOPAS, are of general interest for medicinal chemistry and will be briefly addressed here:

- Identification of "fast-followers" taking a known lead or drug as the template structure,
- Generation of focused libraries that are biased towards a given activity for secondary screening,
- Preferred generation of novel "drug-like" molecules for combinatorial library design,
- Development of peptide-derived molecules with a non-peptide backbone architecture, and
- Exploration of the full diversity of molecules that could potentially bind to an active site [4].

8.2 *De Novo* Design is an Optimization Process

If one wants to start a drug design project, most often the initial knowledge is very limited about relevant structural features, pharmacophore patterns, or structure–activity relationships (SAR). In the course of ongoing discovery this knowledge grows, and towards the end of the project much information will have been collected about a promising lead structure candidate. Ideally, the growth of knowledge is accompanied by decreasing library diversity

(Figure 8.1a). Here the term "diversity" is thought to describe the degree of structural variety inherent to a particular compound collection. Initially there are no, or only very soft, constraints put on a compound collection for screening, e.g. empirical heuristics like cutoff values for molecular weight and predicted lipophilicity. As soon as crucial parts of a structure are revealed, the picture of a lead structure candidate becomes clearer, and library diversity decreases in subsequent screening rounds. In other words, screening becomes more and more focused or biased towards a given set of molecular architectures. A drawing by M. C. Escher adequately illustrates this pattern-elucidation process (Figure 8.1b). One aim of virtual screening is definitely to assist in dragging interesting structures out of the "twilight zone", where a main advantage of *de novo* design is the possibility to generate virtually millions of molecules that would escape our notion in more conventional database-searching techniques.

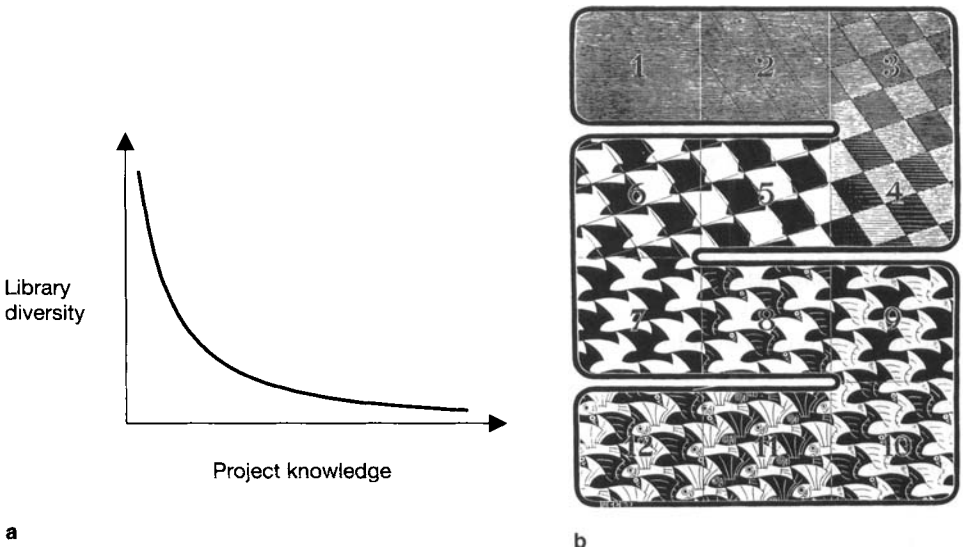


Figure 8.1. a Idealized course of library diversity with increasing knowledge in a drug discovery project. **b** Schematic illustration of molecular feature extraction during lead identification and optimization. M.C. Escher's "Regular Division of the Plane I" © 2000 Cordon Art B.V., Baarn, The Netherlands. All rights reserved. With kind permission of the copyright holder.

The molecular optimization process can be regarded as an "evolutionary" interplay between variant generation and selection-of-the-fittest. The design cycle shown in Figure 8.2 can be entered at any position, and it may consist of real bench-experiments and computer-based parts. Its most important aspect is that the diversity of the screening library is tailored in each cycle to adapt the search process to the local shape of the "fitness landscape". This turns the blind random search for desired molecular structures into a systematic exploration of chemical space ("adaptive walk"), thereby enabling to start with a broad distribution of molecules ("universal" library), and focus on more specific features as soon as some "bioac-

tivity" has been found ("focusing" library). The concept of adaptive chemical diversity permits the making of large steps in search space, e.g. on a plateau of the associated fitness landscape (wide distribution of variants), or the fine-tuning of a structural framework towards the end of a design project, i.e. climbing the optimum (narrow distribution of variants). An in-depth treatment of adaptive walks in fitness landscapes can be found elsewhere [26,27].

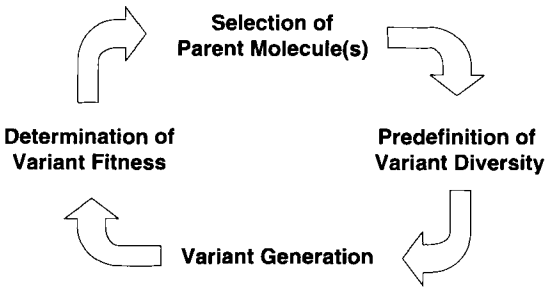


Figure 8.2. An evolutionary drug design cycle. Starting from a "parent" molecule, offspring ("variants") are generated in such a way that the distribution of variant structures exhibits a desired degree of diversity. Fitness is determined for each of the novel structures, and the best is/are selected as the parent(s) of the next cycle. Note that this scheme can be applied to both virtual screening and real bench experiments.

A simplifying picture of a fitness landscape is shown in Figure 8.3. The co-ordinates X and Y represent arbitrary dimensions of a chemical space, e.g. "lipophilicity" and "polar surface". The fitness axis is thought to reflect bioactivity, which may be some predicted value or an experimental measure. Any systematic search will follow a path along increasing fitness (dashed line), where it must be avoided getting trapped in a local optimum. It is important to note that the shape of the fitness landscape varies with the different axes spanning a chemical space, and depends on the fitness function employed. Some extremes are depicted in Figure 8.4.

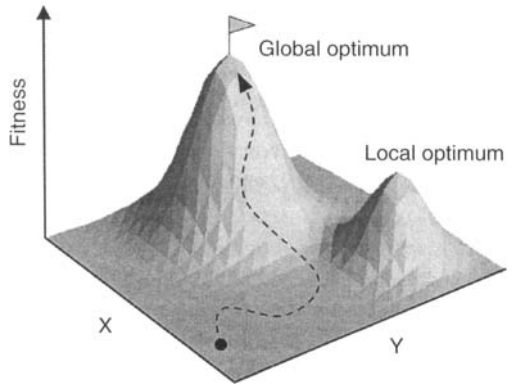


Figure 8.3. Model of a two-dimensional fitness landscape with two optima. X and Y represent appropriate molecular descriptors, e.g. lipophilicity and polar surface area. The dotted line indicates a possible optimization path.

Smooth, unimodal landscapes are highly desirable, as finding the optimum is a straightforward task here (Figure 8.4a). It will still be possible to converge in the global optimum whenever there is an underlying ordering of a multimodal landscape (Figure 8.4b). Special search strategies are required in such a case, because simplistic gradient techniques will easily get stuck in the local optimum located next to the starting point. Fortunately however, very complex multimodal fitness landscapes are expected to be rare in realistic experimental systems [28,29]. The landscape in Figure 8.4c is a nightmare for virtual screening. Although a distinct global optimum is present, only random searching can eventually find it. The conclusion of this short discussion of fitness landscapes is that we cannot expect molecular optimization to succeed if either the fitness measure is inadequate (malfunctioning bioassay, error-prone prediction method), or the axes spanning the search space represent a poor choice.

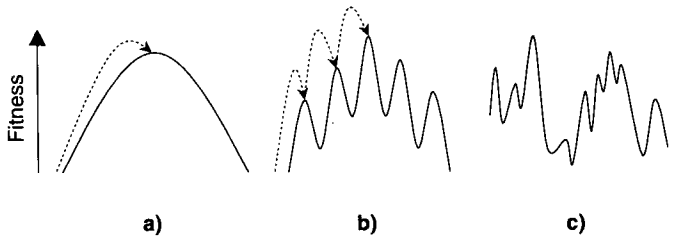


Figure 8.4. Three types of fitness landscapes: **a** unimodal smooth landscape ("the dream"), **b** multimodal landscape with an underlying ordering of local optima, and **c** multimodal landscape with a random distribution of local optima ("the nightmare").

8.3 An Evolution Strategy for Systematic Search in Chemical Space

There is no general best optimization or search technique. In a smooth fitness landscape, advanced steepest descent or gradient search represent a reasonable choice to find the global optimum, and several variations of this technique have been described to facilitate navigation in a landscape with some few local optima [30]. Typically, virtual screening is confronted with a high-dimensional search space and convoluted fitness functions that easily lead to a rugged fitness landscape. Robust approximation algorithms must be employed to find practicable solutions here [26,31]. Stochastic search methods sometimes provide answers that are provably close to optimal in these cases. In fact, several systematic investigations demonstrated that a simple stochastic search technique may be comparable or sometimes superior to sophisticated gradient techniques, depending on the structure of fitness landscape or the associated error surface [32,33]. Special types of stochastic search methods, namely evolutionary strategies (ES) [34] and genetic algorithms (GA) [35], are very appealing for molecular design, as they provide an intuitively easy implementation of the design cycle shown in Figure 8.2 [17,36,37]. In the following, the discussion of these techniques is restricted to their application to molecular design. Other important algorithmic approaches have been devel-

oped for approximation tasks, and the reader is referred to the literature for more comprehensive surveys [38,39].

The common feature of both ES and GA is a cyclic variation–selection process. "Parents" breed "offspring", and the best of each "generation" according to a fitness measure becomes the parent of the subsequent optimization cycle. Evolution strategies have often been applied to real-valued function optimization problems. These algorithms generally operate directly on the real values to be optimized, in contrast to genetic algorithms, which usually operate on a separately coded transformation of the objective variables (the so-called "chromosome"). Furthermore, ES include a second-level optimization of so-called "strategy parameters", i.e. tunable variables that in part determine how each parent will generate offspring. This special algorithmic feature is perfectly suited to account for compound library diversity (Figure 8.1a; *vide infra*) [17,40]. Comparisons of GA and ES appear to favor the original evolution strategy approach [26,41].

A useful plain ES employs a single parent and a number of λ variants per generation, "(1,λ) ES" in the notation according to Rechenberg [34]. Selection-of-the-best is performed among the offspring only, i.e. the parent dies out. This characteristic helps navigate in a multimodal fitness landscape because it facilitates to escape from local optima a parent may reside on [34]. Within a generation the offspring are approximately Gaussian-distributed around the parent. This enables a local stochastic search to be performed, and guarantees that there is a finite probability of breeding very dissimilar structures ("snoopers" in search space). The width of the variant distribution is determined by the variance or standard deviation, σ, of the bell-shaped curve reflecting the distance-to-parent probability (Figure 8.5). The σ value can be regarded as a measure of library diversity. Small values lead to narrow distributions of variants, large values result in the generation of novel structures that are very dissimilar to the parent molecule as judged by a chemical similarity or distance measure (Figure 8.5). In the beginning of a virtual screening experiment using ES, σ should be large to explore chemical space for potential optima, later on during the design experiment σ will have to be small to facilitate local hill-climbing.

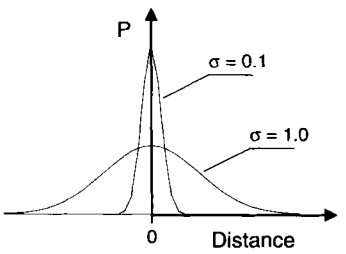


Figure 8.5. Idealized distributions of offspring around a parent structure. The value of the standard deviation, σ, can be used to determine library "diversity".

This scheme exactly follows the idea of adaptive chemical diversity in molecular design (Figure 8.1). σ represents an ES strategy-parameter. This means that σ itself is subjected to optimization and will be propagated from generation to generation, thereby enabling an adaptive stochastic search to be carried out. In contrast to related techniques like simulated annealing, there exists no predefined cooling schedule or a fixed decaying function for σ. Its

value may freely change during optimization to adapt to the actual requirements of the search. Large σ values will automatically evolve if large steps are favorable, small values will result when only small moves lead to success in the fitness landscape. If there is an underlying ordering of local optima as depicted in Figure 8.4b, this strategy will provide a simple means to perform peak-hopping towards the global optimum. A straightforward (1,λ) ES that proved to be successful in many molecular design applications can easily be formulated in pseudo-code:

initialize parent: (S_P, σ_P, F_P)
for each generation:
  generate λ variants, (S_V, σ_V, F_V)
  select best variant structure, $(S_{best}, \sigma_{best}, F_{best})$
  set $(S_P, \sigma_P, F_P) = (S_{best}, \sigma_{best}, F_{best})$

For our purpose of molecular optimization, S represents a chemical structure, and F is the fitness value associated with S. The indices P and V refer to parent or variant (offspring) attributes respectively. In the first generation σ_P is commonly set to unity, and S_P is a randomly assembled or selected molecular structure. A known drug or active compound may also serve as the initiatory parent. Novel variant structures, S_V, and the associated step-sizes, σ_V, can be selected using the Box-Muller formula (Eq. 8.1), where g is a Gaussian-distributed random number and i and j are random numbers in]0,1]:

$$g = \sigma_P \sqrt{-2 \ln(i)} \sin(2\pi j) \tag{8.1}$$

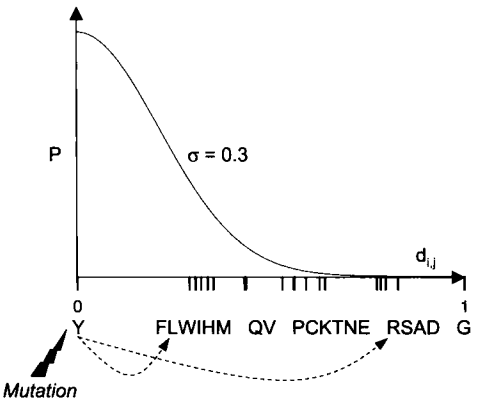


Figure 8.6. Mutation probability of the amino acid residue tyrosine (Y). The ordering of the natural amino acids is based on a distance measure (*cf.* Figure 8.5). In the example there is a high probability of the Y→L transition, and a marginal probability of the Y→R mutation.

An example of a mutation event based on this scheme is shown in Figure 8.6. For reasons of clarity, the mutation of an amino acid residue is illustrated. However, a similar scheme is appropriate for arbitrarily defined combinatorial building blocks or whole compounds. In the example shown, the amino acid tyrosine (Y) represents the parent structure, and the remaining 19 genetically encoded amino acids provide the stock of structures. With decreasing prob-

ability the parent is substituted by more distant residues. In the example, distance between two residues is defined by their Euclidian distance, $d_{i,j}$, in a primitive two-dimensional space spanned by a hydrophobicity [42] and a volume axis [43] (Figure 8.7). This amino acid similarity measure proved to be useful in several applications [17,40,44,45]. As indicated by the dashed lines in Figure 8.6, based on this model a mutation of tyrosine to phenylalanine (Y→F) or leucine (Y→L) is very likely to occur, whereas the tyrosine arginine transition (Y→R) is extremely rare. Another set of transition probabilities would result from a different σ value or a changed ordering of the residues. A great variety of substitution matrices have been suggested to measure distance between pairs of amino acid sequences, and it is not a trivial task to select the most appropriate for similarity searching or sequence design (for reviews, see [46–48]).

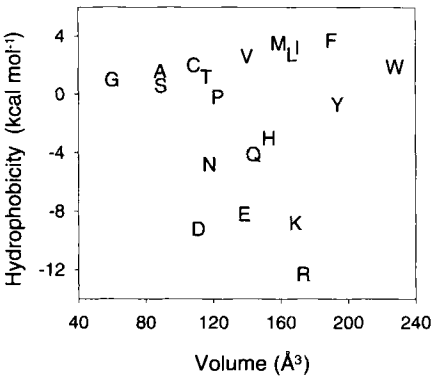


Figure 8.7. Distribution of 20 natural amino acids in a chemical space spanned by the properties volume [43] and hydrophobicity [42]. Several clusters of residues are formed.

8.4 Structure of Chemical Space and the "Principle of Strong Causality"

It is not sufficient to generate offspring around a parent structure in a systematic way. To enable a systematic search in chemical space, both the ordering of chemical space and the corresponding fitness landscape must be strongly related. This means that small steps in chemical space must be correlated with small changes in predicted fitness or experimentally determined bioactivity (*Principle of Strong Causality* [34]). Otherwise the whole concept of optimization by evolutionary search will be corrupted. Very likely it will be impossible to perform a systematic exploration of high-dimensional chemical space if a small structural modification of a molecule induces drastic fitness deviations between the parent and the offspring. There are several reports in the literature on seemingly small structural variations of active compounds leading to even complete loss of activity. For example, exchange of the phenolic hydroxy groups of isoproterenol with chlorine results in the conversion of an agonist (α-adrenergic, isoproterenol) to an antagonist (β-adrenergic blocker, dichloro-isoproterenol) [49]. Close structural relationships of gluco- and corticosteroids, e.g. also reflect the observation that apparently small structural variations can lead to significant loss of, or

change in, bioactivity. Obviously in these cases essential parts of the function-determining pharmacophore pattern were destroyed. As we do not know *a priori* which parts of a molecule are crucial determinants of bioactivity, we tend to believe that *any* "small" change of structure will only slightly affect molecular function. This way of thinking is often not appropriate. As a consequence, a reasonable fitness function must accurately account for important, less important, and even irrelevant molecular components.

Many computer-based techniques have been proposed for QSAR modelling during the past decades, and there probably is no generally best method. In particular, multiple partial-least-squares (PLS) regression and various types of artificial neural networks (ANN) emerged as useful approaches to develop artificial fitness landscapes (see Chapters 5 and 6) [44,50,51]. ANN can be used as function estimators [52,53] and classification systems [54]. They follow the principle of convoluting simple non-linear functions for approximation of complicated input-output relationships (*Kolmogorov's Theorem* [55]). Fourier transforms, for example, are based on a similar idea using super-positions of trigonometric functions. This general concept has been thoroughly discussed by Bishop [56]. As a consequence, ANN can be used to approximate arbitrary continuous functions that are suited for modelling fitness landscapes [57]. The architecture of a conventional three-layered, fully connected feed-forward network is depicted in Figure 8.8. This system takes the co-ordinates of a compound in chemical space, $\mathbf{x}$, as the input and computes the corresponding fitness value, y, as a non-linear function of the co-ordinates (Eq. 8.2):

$$y = f(\mathbf{x}) = T\left[\sum_k v_k\, T\left(\sum_i w_{i,k}\, x_i - \vartheta_k\right) - \theta\right] \tag{8.2}$$

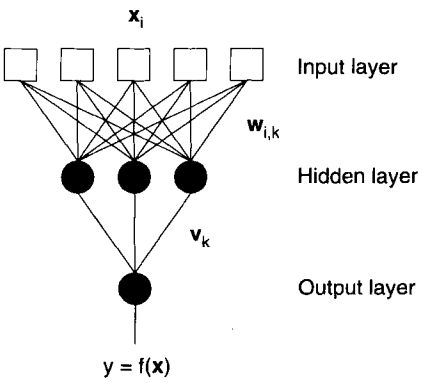


Figure 8.8. Schematic architecture of a three-layered artificial neural network system. The flow of information is unidirectional from the input layer to the output layer. Circles symbolize neurons, lines indicate weights. The input layer consists of "fan-out" neurons (squares) that distribute a numerical molecular representation to the hidden layer neurons. See the text for details.

A special feature of ANN is the introduction of a weighing scheme, $\mathbf{w}$, for the individual axes of chemical space, e.g. molecular descriptors, thereby connecting chemical space with a fitness landscape. Combinations of such weighted input schemes are biased by the hidden-to-output connection weights, $\mathbf{v}$. This ANN architecture makes possible the extraction of chemical features [50]. Input layer neurons simply distribute a data vector to the hidden-layer neurons without any calculation being performed. In the hidden and output layer neurons with

a sigmoidal activation function (also refered to as "squashing" or "transfer function", T) are usually employed. In Eq. 8.2, ϑ represents a hidden neuron's bias value, and Θ is the output neuron's bias. The more layers and neurons are present in a network the more complicated overall functions can be represented. At most two hidden layers with non-linear neurons are required to approximate arbitrary continuous functions [52,58]. However, the optimal number of layers and the number of neurons in a layer can vary, depending on the application and the accuracy of the fit required. Baum and Haussler addressed the question of what size of a network can be expected to generalize from a given number of training examples [59]. A "generalizing" solution derived from the analysis of a limited set of training data will be valid for any further data point leading to perfect predictions. This is possible only by the use of training data which is representative of the problem. Most solutions found by feature extraction from lifelike data sets will, however, be sub-optimal in this general meaning [44].

Peptide *de novo* design was the first successful combination of evolutionary design employing a neural network as the fitness function [17,40]. Figure 8.9 shows fitness landscapes, which were generated by a neural network system trained for the prediction of eubacterial signal peptidase I cleavage sites in amino acid sequences [17]. A set of known peptidase substrates served as the training data for feature extraction by ANN. It turned out that a chemical space spanned by the amino acid properties "hydrophobicity" and "volume" was suited for this particular application. The fitness functions for the three selected sequence positions shown are smooth and separate the set of natural amino acid residues into "low-fitness" and "high-fitness" candidates (see Figure 8.7). In a series of *in machina* design experiments, alanine was selected as best-suited in position –3 (numbered relative to the signal peptidase cleavage site; Figure 8.9a), tryptophan in position –2 (Figure 8.9b), and glycine in position –1 (Figure 8.9c). Due to the continuous nature of the fitness landscapes, evolutionary search for idealized substrates was straightforward. Indeed, the design run converged after only 52 optimization steps, converting the initial parent sequence FICLTMGYIC into the functional enzyme substrate FFFFGWYGWA*RE (the asterisk denotes the signal peptidase I cleavage site). Its activity is comparable to wild-type sequences, which was proven by an *in vivo* protein secretion assay and mass-spectrometric sequence analysis [45]. The X-ray structure of the catalytic domain of signal peptidase I from *Escherichia coli* was published after these ligand-based design experiments were completed [60]. It is evident from the structure of the active site that the model peptide excellently harmonizes with the structural and electrostatic realities of the enzyme.

In a further application of evolutionary search guided by neural networks antigen-mimicking peptides were developed "from scratch" [40]. This design approach included a round of bench experiments for data generation, and subsequent computer-assisted evolutionary optimization. The five-step procedure represents a special version of the design cycle shown in Figure 8.2:

1. Identification of a single compound with some desired activity, e.g. by expert knowledge, data base or random screening, combinatorial libraries, or phage display,
2. Generation of a focusing library taking the compound obtained in step 1 as a "seed structure". A limited set of variants is generated approximately Gaussian-distributed in some physicochemical space around the "seed peptide",
3. Synthesis and testing of the new variants for their bioactivity,

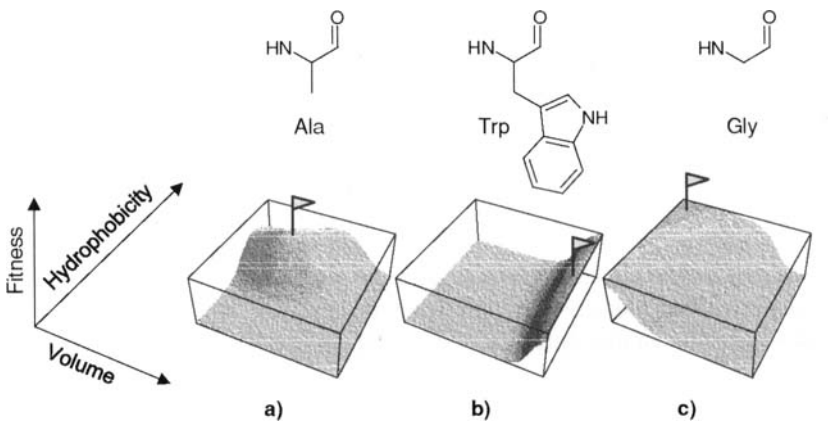


Figure 8.9. Artificial fitness landscapes generated by a neural network system. The neural network was trained to predict the value of individual amino acid residues in potential signal peptidase I substrates, based on the hydrophobicity and volume of the side-chains. **a** position –1 of the substrate, **b** position –2 of the substrate, and **c** position –3 of the substrate. The neural network response to varying hydrophobicity and volume values is plotted on the fitness axis. This was achieved by keeping all but the substrate position under investigation fixed [17].

4. Training of artificial neural networks providing simple heuristic QSAR based on the activities measured in step 3, and
5. Computer-based evolutionary search for highly active compounds taking the network models as the fitness function.

A novel peptide was identified fully preventing the positive chronotropic effect of anti-β_1-adrenoceptor auto-antibodies from the serum of patients with idiopathic dilated cardiomyopathy (DCM) [40]. In an *in vitro* assay, the designed active peptide showed more significant effects compared to the natural epitope. The idea was to test whether it is possible to derive artificial epitope sequences that might be used as potential immuno-therapeutical agents following the design strategy described above. The model peptide GWFGGADWHA exhibits an activity comparable to its natural counterpart (ARRCYNDPKC) but has a significantly different residue sequence. Selection of such antibody-specific "artificial antigens" may be regarded as complementary to natural clonal B-cell selection, leading to the production of specific antibodies. The peptide–antibody interaction investigated can be considered as a model of specific peptide–protein interactions. These results demonstrate that computer-based evolutionary searches can generate novel peptides with substantial biological activity.

Peptide design is a comparably simple task because of the inherent modular architecture of amino acid sequences, their easy synthetic accessibility, and the restricted size of sequence space. Of course, the ultimate goal is to generate arbitrary novel structures with desired properties. It is crucial for success to focus on the appropriate level of abstraction from the physical molecular structure, as illustrated in Figure 8.1b [61]. Coarse-grain or "bulk" descriptors of a molecular entity can be useful in early phases of virtual screening to roughly separate "potentially useful" structures from the vast majority of accessible compounds. In contrast,

special information about the relative spatial orientation of side-chains, for example, might be suitable in a later stage of lead optimization for fine-tuning of molecular properties like, e.g. enzyme specificity. The concept of pharmacophore patterns for representation of molecules proved to be successful when at least a single active molecule was identified, which could be used as the "seed" structure for similarity searching or library design (see Chapter 7). A special type of pharmacophore descriptors for this purpose, the "topological correlation of generalized atom types" [21,40], is described in the following Section. Appropriate pharmacophore models and similarity measures sometimes provide a means to design or find structurally diverse yet isofunctional compounds [62,63]. A worthwhile goal is to design peptide-analogue structures from bioactive, optimized peptides, thereby "hopping" from one optimum in chemical space to another optimum representing isofunctional non-peptide molecules.

8.5 Spanning a Topological Pharmacophore Space for Similarity Searching

As soon as some few lead structures have been identified, a very appealing approach is to derive a pharmacophore model from the known active structure (the "seed" or "query" structure) and perform a computer-based similarity search to speed up the process of lead-identification. It is known that ligands often bind to a receptor in a conformation other than their lowest energy conformations, where the degree of deformation can be correlated with the number of freely rotatable bonds in the molecule [64]. Thus it appears reasonable to develop pharmacophore models that are independent of 3-D structure, or accurately account for conformational flexibility. Several such approaches have been developed and successfully applied to similarity searching, database profiling, and design [65]. A simplistic alternative is to focus on special 2-D representations of molecules only. Topological correlation of generalized atom types builds on this ground. It is a special molecular descriptor that leads to a compact, molecular size-independent description of potential pharmacophores [21]. The general idea of this representation scheme is to measure distances between pairs of atoms and regard the histogram of pair counts as a simplifying but exhaustive pharmacophore fingerprint of the molecule. Distances are expressed as the number of bonds along the shortest path connecting two nodes (non-hydrogen atoms) in the molecular graph (Figure 8.10). Each node is checked as to whether it can be assigned one of the following generalized atom types: Hydrogen-bond donor (D), hydrogen-bond acceptor (A), positively charged (P), negatively charged (N), or lipophilic (L). The numbers of all 15 possible pairs of generalized atom types (DD, DA, DP, DN, DL, AA, AP, AN, AL, PP, PN, PL, NN, NL, LL) are determined, and the resulting histogram counts are divided by the total number of non-hydrogen atoms to obtain scaled vectors. Distances up to ten bonds proved to be relevant in most of the cases investigated until today, although the optimal path length varies in different applications (G. Schneider, unpublished). This leads to a $15 \times 10 = 150$-dimensional vector representation of a molecular compound.

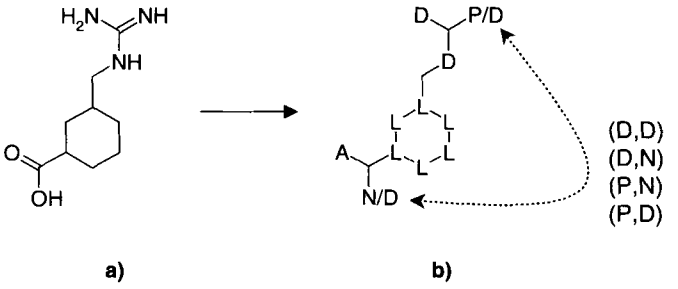


Figure 8.10. Principle of CATS topological pharmacophore calculation. The two-dimensional molecular structure **a** is converted to the molecular graph **b**. In **b** the assigned generalized atom-types are shown (D: hydrogen-bond donor, A: hydrogen-bond acceptor, L: lipophilic, P: positively charged, N: negatively charged). The four annotated pairs of generalized atom types are spaced eight bonds apart, as indicated by the arrow.

Given a molecule with a defined biological activity, large virtual compound libraries can be searched for similar structures, based on the correlation–vector representation. Each library molecule is compared to the query vector (derived from the query compound structure), e.g. using the Euclidian distance measure, $D(A,B)$, to express similarity between two molecules A and B (Eq. 8.3):

$$D(A,B) = \sqrt{\sum_{i=1}^{150}(x_i^A - x_i^B)^2} \tag{8.3}$$

In Eq. 8.3, $\boldsymbol{x}^A$ and $\boldsymbol{x}^B$ are the correlation vectors derived from molecule A and B, respectively. As a result, all library compounds are ordered by their distance-to-seed value, and a rank list of virtual hits is constructed (see Chapter 4 for an in-depth treatment of various similarity measures for this purpose.). This special technique was named CATS (Chemically Advanced Template Search) [21], and the corresponding *de novo* design strategy based on topological pharmacophores is referred to as TOPAS (Topology-Assigning System) in the following.

By defining this correlation vector scheme for similarity search we minimize the risk of using a misleading three-dimensional pharmacophore model which can easily happen if, e.g. the conformation of the receptor-bound ligand is unknown or several different conformations appear equally possible. An obvious downside of a 2-D approach like CATS is that important 3-D structural information is neglected, e.g. stereochemical or full topological information. Whenever such information is crucial for bioactivity the search or design algorithm will fail to produce useful results.

If information about the receptor structure is not available, an exhaustive 2-D descriptor can have practical advantage over the respective 3-D implementation for database searching, as shown in Figure 8.11. Validated ligands to nine different drug targets were collected from the MedChem database (version 1997, as distributed by Daylight Chemical Information Systems Inc., Irvine, CA, USA) and Derwent Word Drug Index, WDI (as at Nov. 1998; Derwent Information, London, UK). This set of known actives was added to a library of 8000 selected

"drug-like" substances from WDI. Taking each of the known ligands as the template structure CATS was able to retrieve significantly more actives than random testing of the identical number of compounds would have yielded (Figure 8.11). In some cases over 40-times enrichment was obtained. The 2-D version clearly outperformed the 3-D approach because only a single low-energy conformer was used in the latter. For one of the drug targets in the test, aldose reductase, pharmacophore searching retrieved a slightly even smaller number of actives than random screening. Obviously, in this case the molecular representation was not appropriate, probably due to the rather wide active site of aldose reductase and the very different ligand structures and binding sites. Restricting the maximal path length of a CATS pharmacophore to six bonds – thereby enabling a less restrictive search – leads to a slight enrichment of actives in the first percentiles of the database, as expected (data not shown). As a consequence of the above-mentioned findings, a 2-D topological pharmacophore concept is used in the current versions of the CATS and TOPAS software (vide infra) [21]. Nevertheless, one must be aware of the general limitations of topological descriptors [49], and there exist other techniques that are better suited to perform particular tasks. For example, the more sophisticated "feature tree" approach [66] proved to outperform the simple CATS method in several of our test cases (data not shown).

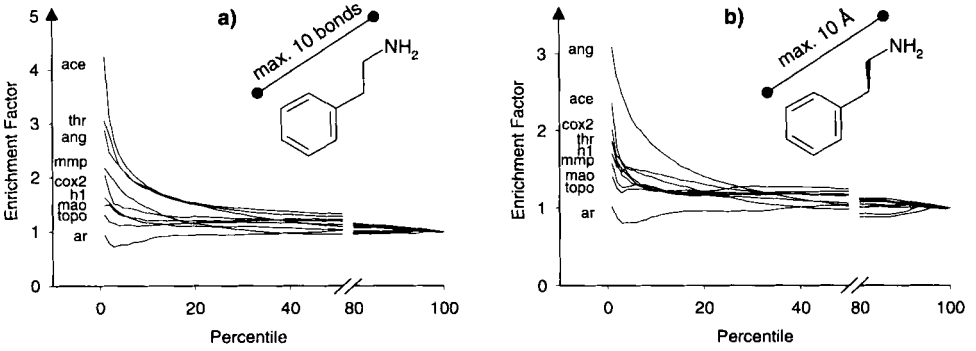


Figure 8.11. Enrichment of bioactive compounds by CATS similarity search for several drug targets. A value of 1 indicates an identical hit rate to random screening, and larger values reflect that the virtual screening method is superior to random selection. The enrichment factors were separately calculated for 100 fractions (percentiles) of the total test database: **a** two-dimensional CATS similarity, **b** three-dimensional CATS similarity. *ace:* Angiotensin-converting enzyme, *ang:* angiotensin II, *ar:* aldose reductase, *cox2:* cyclooxidase 2, *h1:* antihistamine H1 receptor, *mao:* monaminoxidase A, *mmp:* gelatinase, *thr:* thrombin, *topo:* topoisomerase II.

Two examples of "backbone-hopping" with CATS are shown in Figure 8.12. Taking known bioactive molecules as the query structures (left side), functional alternatives (right side) were retrieved from the Roche corporate database. Starting from the Ca^{2+}-T-channel inhibitor mibefradil (IC_{50} = 1.7 µM), 9 out of the 12 highest ranking structures exhibited significant Ca^{2+}-antagonist activity (IC_{50} = 1.7 µM, 2.2 µM, 3.2 µM, and 3.5 µM; structures not shown). Among this series of structurally different molecules the known Ca^{2+}-channel in-

hibitor clopimozid was found [67], yielding an $IC_{50} < 1$ μM in a cell-based fluorescence assay (Figure 8.12a) [21]. A second experiment aimed at testing the usefulness of the approach for finding isofunctional variants of active structures obtained in combinatorial chemistry projects. An optimized Ugi-reaction product, which has a K_i of 1.4 μM in a thrombin binding assay [68] (see Chapter 9 for further details about practical approaches to evolutionary optimization), led to the identification of a substance with K_i = 24 nM in the Roche corporate database (Figure 8.12b). Several other known potent thrombin inhibitors like PPACK and NAPAP were among the top-scoring hits (for review of thrombin inhibitors, see [69]). From these experimental results we conclude that the CATS approach to virtual screening can be useful for identifying novel molecular structures with substantial biological activity.

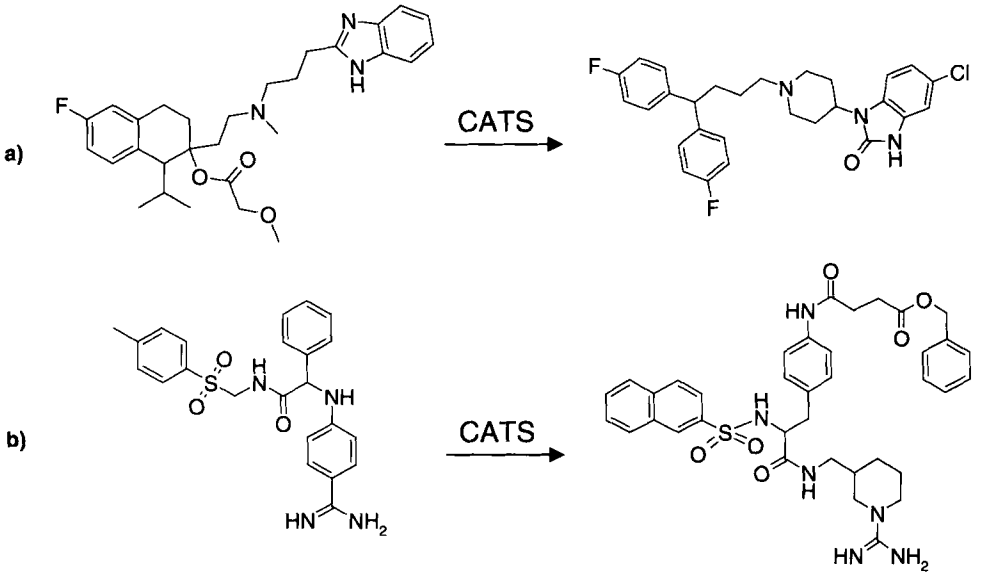


Figure 8.12. Active molecules retrieved by database searching with CATS. On the left the query structures ("templates") are shown, on the right isofunctional high-ranking hits are depicted. **a** Ca^{2+}–T channel inhibitors, **b** thrombin inhibitors.

Topological pharmacophore models proved their applicability in the CATS approach to database searching. It is impossible to enumerate all theoretically feasible molecules and store their structures in a database for virtual screening. Rather, a systematic structure optimization must be performed. As discussed above for the task of peptide design, a reasonable chemical space must be constructed on basis of a meaningful molecular representation scheme. The CATS results indicated that correlation vectors might be also useful for spanning a chemical space for evolutionary design. As depicted in Figure 8.13, the projections of a respective 150-dimensional chemical space filled with 1500 compounds show signs of a reasonable ordering of the molecules according to their activity in a thrombin binding assay. A pronounced optimum is visible in the two charts, a self-organizing map (SOM) and a projec-

tion onto the plane spanned by the two dominant principal components (PC) that were calculated by a principal component analysis (PCA). Furthermore, small steps in chemical space go along with small differences in activity, as demanded by the *Principle of Strong Causality*. If the ordering of molecules did not reflect bioactivity, the particular molecular representation used for spanning the chemical space would be ineffective for systematic optimization. The SOM represents a nonlinear projection, whereas the PCA leads to a linear projection of high-dimensional data. Such a graphical display is very useful to get an impression of the data distribution in chemical space. For further details on visualization of chemical data, see Chapters 5 and 6, and elsewhere [44,70,71].

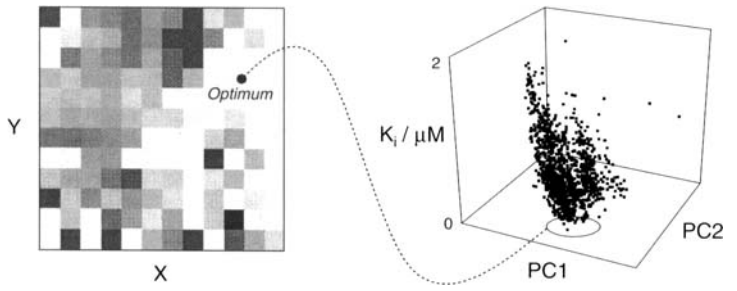


Figure 8.13. Visualization of a chemical space filled with 1500 thrombin inhibitors. Projections of a 150-dimensional space onto the plane by a self-organizing map (left) and principal component analysis (right) are shown. In the left plot, shading reflects bioactivity (*dark:* high K_i, *white:* low K_i).

8.6 Combinatorial Evolutionary Design of "Drug-Like" Molecules

Several evolutionary algorithms have successfully been applied to *de novo* design already, with a strong bias towards peptide structures and combinatorial chemistry [17,25,33,40,45,72–76]. TOPAS provides an algorithmic solution to the problem of evolutionary, template-based *de novo* design, i.e. novel molecular compounds are suggested in a cyclic process, taking a given structure as a reference point (template structure). In addition, instead of generating molecular architectures containing undesired structural features, or synthetically intractable compounds – a problem encountered by many *de novo* design procedures [14] – TOPAS is equipped with a limited set of drug-derived building blocks which were obtained from retro-synthetic fragmentation of the WDI compounds. The idea is that re-assembly of such pre-defined building blocks by a limited set of chemical reactions might lead to chemically feasible novel structures, from both the medicinal chemistry and the synthesis planning perspective.

To compile a stock of drug-like building blocks for evolutionary *de novo* design by TOPAS, all 36000 structures contained in the WDI, which had an entry related to "mechanism" or "activity", were subjected to retro-synthetic fragmentation. The reactions listed in Figure 2.9 [77] (see Chapter 2) were applied to exhaustive cleavage with the following restrictions:

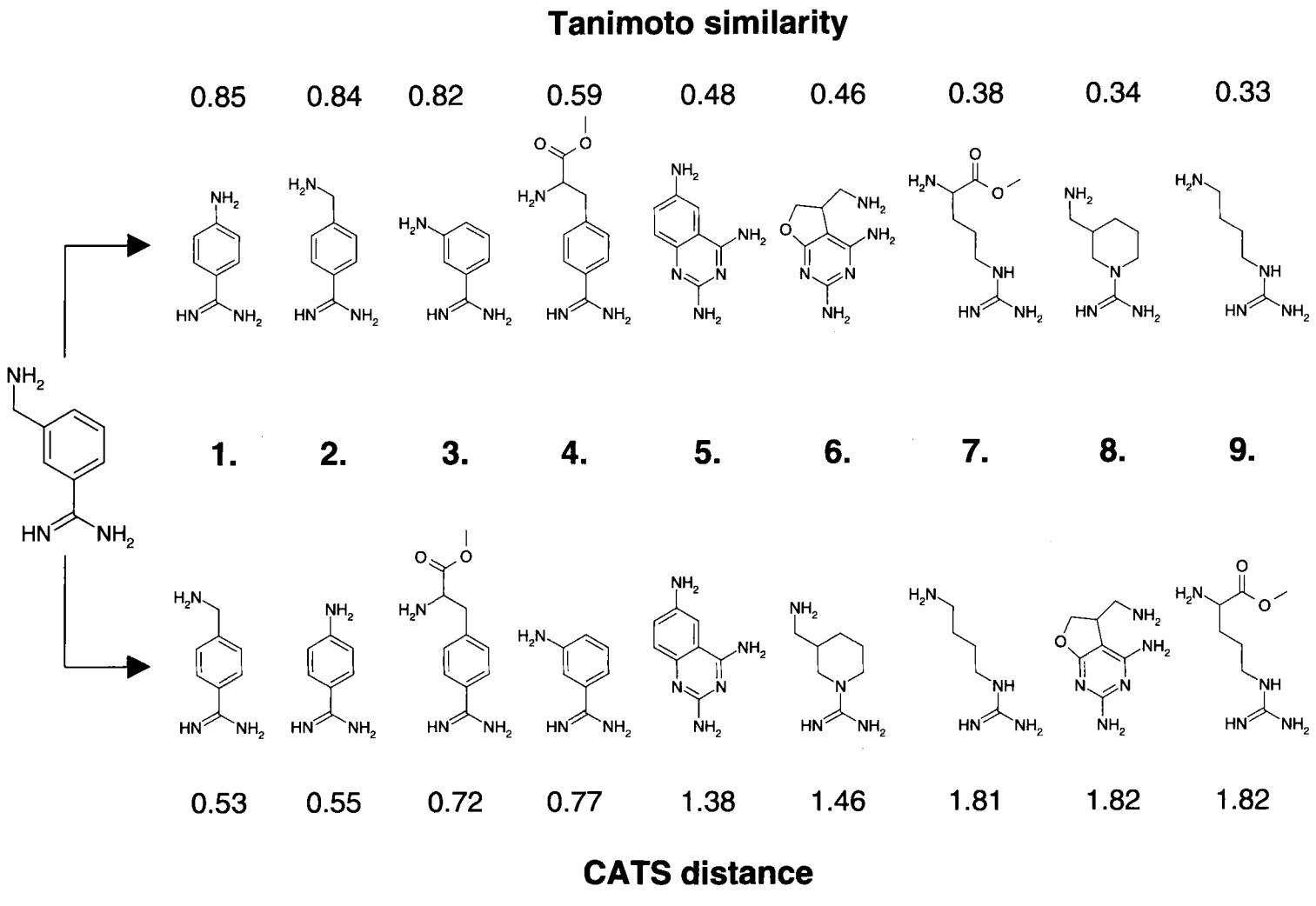


Figure 8.14. Ordering of molecular building blocks based on the Tanimoto similarity (upper row) and the CATS distance (lower row) to benzamidine. For small fragments there is no drastic difference in ranking resulting from either Tanimoto or topological pharmacophore similarity. Significantly different orderings are observable only for larger structures.

- Ring systems were not destroyed.
- Bonds between heteroatoms and ring carbons were not cleaved.
- If a terminal group was hydrogen, methyl, ethyl, propyl, or butyl, the reactions were not applied.

This approach is identical to the original RECAP procedure developed by Hann and coworkers [77]. We yielded a total number of 24563 unique building blocks for TOPAS ("stock of structures"). TOPAS is grounded on a $(1,\lambda)$ evolution strategy (see Section 8.3), strictly realizing the design cycle shown in Figure 8.2. Starting from an arbitrary point in search space, a set of λ variants are generated, satisfying a bell-shaped distribution centered in the chemical space co-ordinates of the parent structure. This means that most of the variants are very similar to their parent, and with increasing distance in chemical space the number of offspring decreases.

In TOPAS, fitness is defined as the pair-wise similarity between the template and the variant structures. Two different concepts are realized to measure similarity: 1. 2-D structural similarity as defined by the Tanimoto index on Daylight's 2-D fingerprints (see Chapter 4), and 2. 2-D topological pharmacophore similarity (see previous section). Tanimoto similarity varies between zero and one, where the value of one indicates structural identity. Topological pharmacophore similarity values vary between zero (indicating identical pharmacophore distribution in the two molecules) and positive values indicating varying degrees of pharmacophore similarity. Optimal fitness values are 1 for the Tanimoto measure, and 0 for the pharmacophore similarity measure (Figure 8.14). Additional penalty terms are added to the fitness function to avoid undesired structures if the total number of atoms exceeds 50, the sum of oxygen and nitrogen atoms is greater than 12, or if there are more than seven potential hydrogen-bond donors present in a given molecule.

Variant structures are derived from the parent molecule, S_P, in a four-step process, following the algorithm presented in section 8.3 (see above):

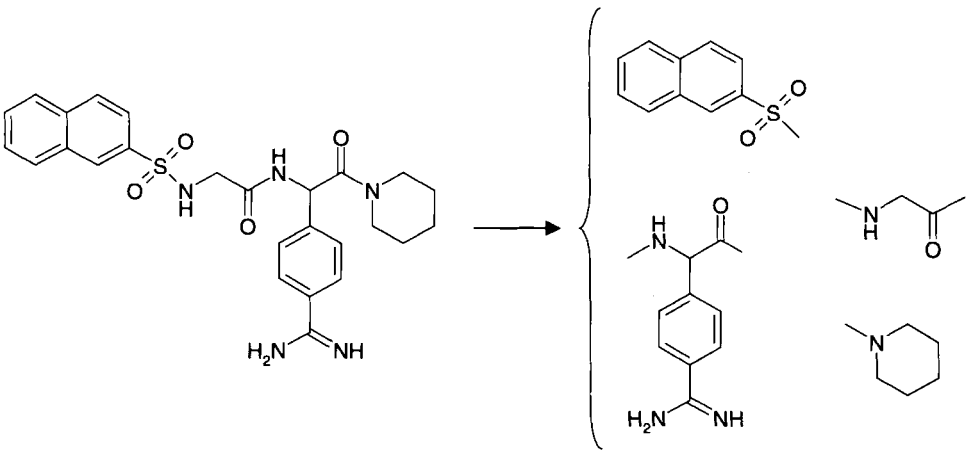


Figure 8.15. Fragmentation of the thrombin inhibitor NAPAP (left). Four fragments are generated by application of two retro-synthetic reaction schemes.

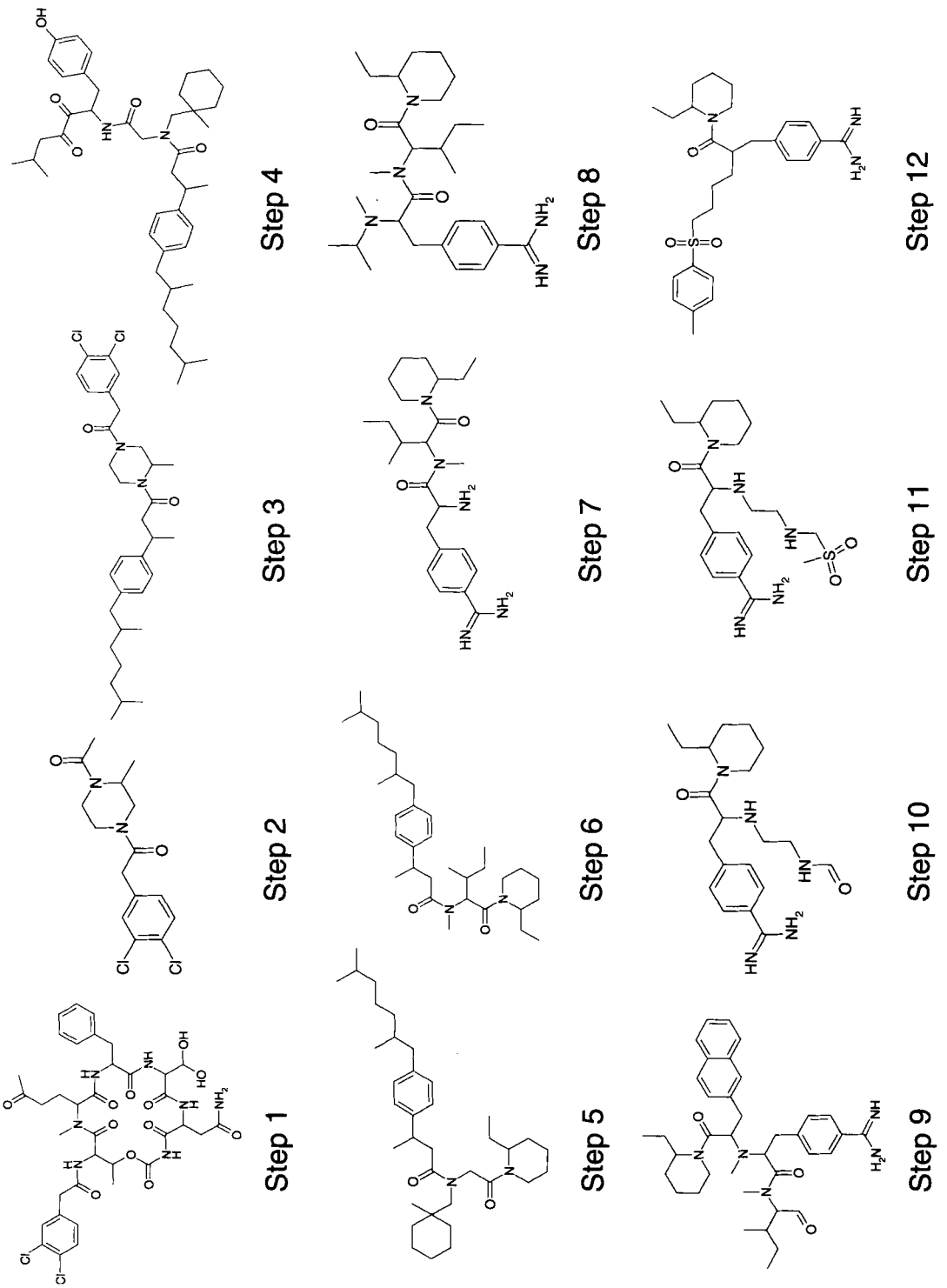


Figure 8.16. Evolution of a potential thrombin inhibitor by TOPAS, showing 12 subsequent parent structures of an evolutionary design experiment. NAPAP served as the template structure, and the Tanimoto index was used as fitness measure.

1. Exhaustive retro-synthetic fragmentation of S_P,
2. Random selection of one of the generated fragments,
3. Substitution of this fragment by the one from the stock of building blocks having the pairwise similarity index closest to the gaussian-distributed random number g (Eq. 8.1), and
4. Virtual synthesis to assemble the novel chemical structure.

To demonstrate step 1, the thrombin inhibitor NAPAP was subjected to fragmentation by TOPAS. Reaction scheme 1 (amide bond cleavage) was applied twice, and reaction 11 (sulfonamide bond cleavage) occurred once, resulting in four fragments (Figure 8.15). Depending on the similarity measure selected and the width of the variant distribution, offspring are generated, e.g. by subjecting the benzamidine residue to mutation (Figure 8.14). The other three fragments remain unchanged. Each of the re-assembled structures is compared to the template, and the most similar one becomes the parent of the new generation. This strategy offers the following advantages:

- An adaptive stochastic search is performed in chemical space,
- The type of molecules that are virtually generated is not restricted to a predefined combinatorial class (e.g. peptides, Ugi-reaction products),
- Novel structures are assembled from drug-derived building blocks using a set of "simple" chemical reactions, and
- A large diversity of molecular fragments can be explored.

An example of a TOPAS design experiment aiming at the generation of a NAPAP-like structure is shown in Figure 8.16. The Tanimoto index was used as the fitness measure. Initially, a random structure was generated from the stock of ~24000 available building blocks ("parent" of the first generation). The Tanimoto similarity to NAPAP was 0.31 reflecting a great dissimilarity, as expected. In each of the following generations, 100 variants were systematically generated by TOPAS, and the best of each generation was selected as the parent for the subsequent generation. Following this scheme, novel molecules were assembled which exhibited a significantly increased fitness (Figure 8.17). After only 12 optimization cycles the process converged at a high fitness level (approx. 0.86), and the standard deviation, σ, of the variant distributions around the parent structures decreased. The course of σ indicates that first comparably broad distributions were generated (large diversity), after some generations, however, a peak in the fitness landscape was climbed (restricted diversity). The parent structures of each generation are shown in Figure 8.16.

Obviously the resulting final design shares a significant set of substructure elements within the NAPAP template. Essential key features for thrombin binding included: the benzamidine moiety forming hydrogen bonds with Asp189 at the bottom of the thrombin P1 pocket, a sulfonamide moiety interacting with the backbone carbonyl of Gly216, and the lipophilic para-tolyl and piperidine rings filling a large lipophilic pocket of the thrombin active site cleft [69]. Automated docking by means of FlexX [78] essentially reproduced the NAPAP binding mode (PDB code 1dwd [79]) (Figure 8.18a). This "from scratch" TOPAS design experiment clearly demonstrated that the algorithm can be used for a fast guided search in a very large chemical space, ending up with rational proposals for novel molecular structures that are similar to a given template.

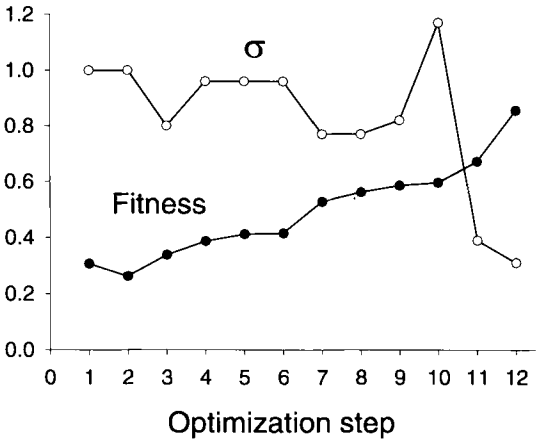


Figure 8.17. Course of fitness (Tanimoto similarity to NAPAP) and the standard deviation, σ, of the offspring distribution ("diversity") during a TOPAS design experiment. See Figure 8.16 for the corresponding parent structures.

The vast majority of the designs derived from NAPAP contain a reasonable P1 needle (Figure 8.19). We have not found any experimental support for needle **(a)** binding to the P1 pocket. Although this purine derivative has an appealing H-bond donor/acceptor pattern, the carbonyl might be detrimental to P1 pocket binding. However, the Roche corporate database comprises several potent thrombin inhibitors containing needles **(b)**, **(c)**, and **(d)** (K_i in the

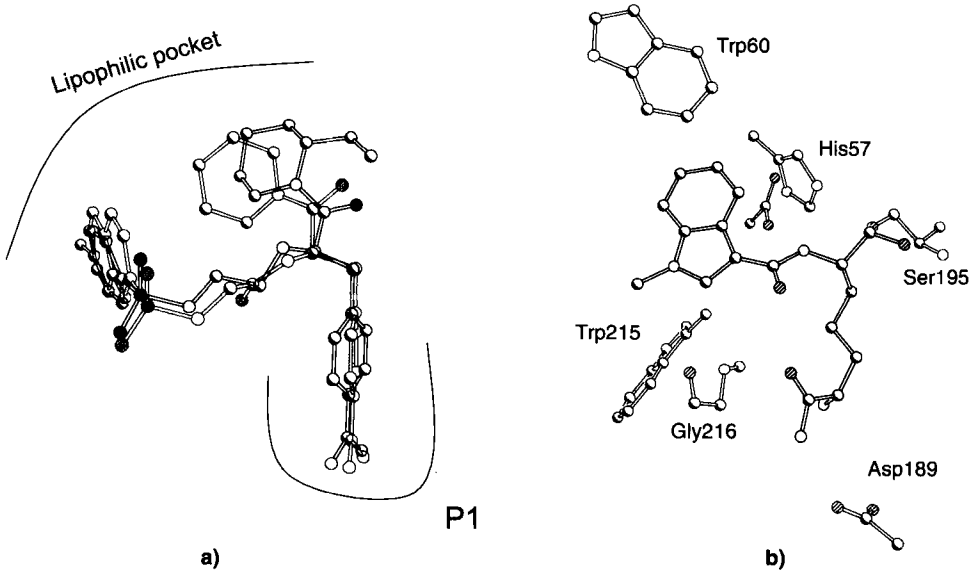
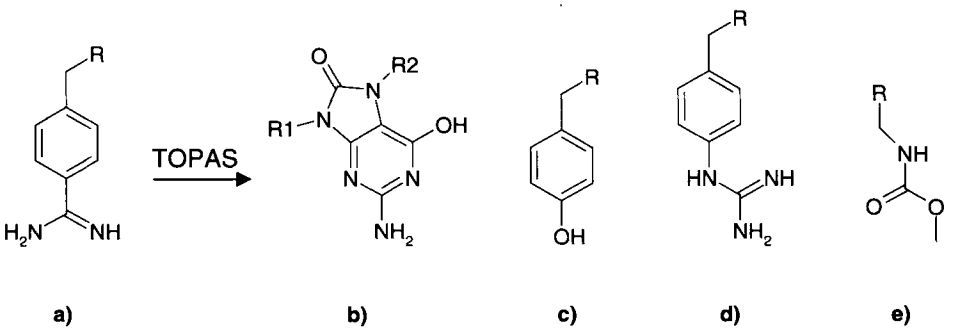


Figure 8.18. Structural models of potential thrombin inhibitors. **a** Superposition of NAPAP and a structure designed by TOPAS (Step 12 in Figure 8.16). The model was obtained by flexible docking into the thrombin active site (PDB code 1dwd) using the FlexX software. **b** Model structure of a peptide-derived potential thrombin inhibitor (*cf.* Figure 8.20) obtained by manual docking. Some relevant thrombin residues are shown. The potential ligand was covalently attached to Ser195.

nanomolar and low micromolar range, structures not shown). Several potent thrombin inhibitors with neutral H-bond donating phenols **(b)** and structures similar to **(d)** were also reported by others [69]. It is evident that TOPAS was able to derive interesting alternatives to benzamidine by evolutionary optimization.



a) b) c) d) e)

Figure 8.19. Proposed alternative "needle" structures (a–e) to the benzamidine moiety that might fit into the P1 pocket of thrombin. These residues were part of novel structures designed by TOPAS.

A further domain of algorithms like TOPAS is the design of peptide-analogues. Several TOPAS experiments were performed with the tripeptide D-Phe-Pro-Arg, an excellent natural thrombin ligand [80], as the template structure. The topological pharmacophore similarity was used to measure fitness to find structures exhibiting a distribution of functional groups that is similar to the peptide template, yet with a non-peptide backbone architecture. The design processes rapidly led to high-fitness structures ending up with surprising results (Figure 8.6). In many structures the original arginine side-chain was selected by TOPAS, whereas some structures contain interesting alternatives. Furthermore, lipophilic moieties are present in appropriate positions, possibly filling lipophilic pockets of the thrombin active site. Manual docking of the structure shown in Figure 8.20 into the thrombin active site (PDB code 1dwd) and subsequent energy minimization using the MAB force field of the MOLOC software package [81] led to a reasonable model (Figure 8.18). Although several peptide features are still present in the peptide-derived structures, some surprising alternatives were found. TOPAS clearly demonstrated its capability to evolve well-known features of small-molecule thrombin inhibitors from a peptide structure, e.g. the classical arginine plus aldehyde pattern for covalent binding to the catalytic Ser195 (as in efegatran) or the α-ketoamide derivative (as in CVS 863) [69]. For some further details about this experiment, see reference [82].

Currently, some of the TOPAS designs are being evaluated at F. Hoffmann-La Roche Ltd, Basel. A novel K^+-channel inhibitor, which was evolved by TOPAS and synthesized in one of our laboratories, has already proved to be biologically active in electrophysiological studies (structure not shown). Although its K^+-flux blocking potential (IC_{50} = 0.4 μM) is slightly below the activity of the template (IC_{50} = 0.2 μM), a novel promising structural class was found with help from the evolutionary design strategy. The new structure is ready to enter conventional medicinal chemistry programs for optimization.

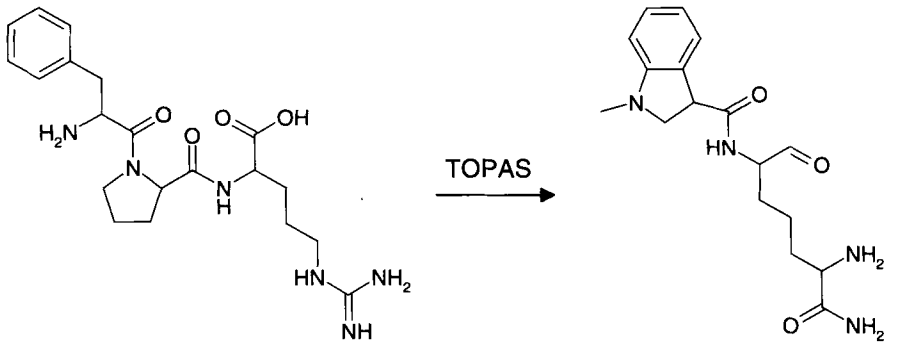


Figure 8.20. *De novo* design of a peptide-analogue by TOPAS. Taking the tripeptide Phe-Pro-Arg (left) as the template structure, the algorithm produced a non-peptide molecule that might represent a starting point for thrombin inhibitor development.

8.7 Conclusions

Most medicinal chemists excel at performing pattern recognition and feature extraction from molecular structures. Success in Drug Discovery projects is tightly coupled to the sensitivity of this process and the particular perception and association abilities of the individual chemists involved. It implicitly requires appropriate levels of abstraction from the molecular entity and usage of respective molecular representation schemes [83]. During the early stages of lead discovery, medicinal chemists are increasingly confronted with large amounts of screening data, and the concurrent demand for "follow-ups" and novel "back-up structures". To prosper in this situation, smart algorithms need to be formulated that mimic some simplified pattern recognition processes, eventually serving as a "chemist's guide through molecular space". The aim of such virtual screening approaches is twofold: sieving large data sets, and generating novel molecular architectures which might serve as the starting point for a medicinal chemistry program. Taking a known bioactive structure as a template, a cyclic optimization procedure can be applied:

1. Generate novel molecular scaffolds by computational design (e.g. with help from the TOPAS algorithm),
2. Synthesize and test individual molecules,
3. Create a virtual combinatorial library around the most active design,
4. Perform similarity searching to rank the library members (e.g. by using CATS),
5. Synthesize and test individual molecules, and
6. Proceed with Step 3 or terminate.

This procedure has already led to a number of novel molecular architectures showing significant desired bioactivity. Evolutionary design algorithms like TOPAS can help the medicinal chemist derive hypotheses about structure–activity relationships, and guide the required synthetic work. Furthermore, the proposed novel structures are excellently suited for subsequent evaluation by structure-based modelling, advanced virtual screening, and further prop-

erty prediction techniques [2,14]. A special appeal of TOPAS is its ability to generate synthetically accessible novel structures that have substantial predicted "drug-likeness" and are not restricted to a strictly combinatorial approach, although it is clearly based on "combinatorial thinking" [84]. Furthermore, due to its modular architecture several fitness functions can be combined, thus enabling "multi-dimensional optimization" (see Chapter 1). Gillet and Johnson pinpointed in a recent review on structure-based design [4]: *"**De novo** design is currently limited to the design of ligands to bind to receptors, and other factors that are important in the design of bioactive compounds are not considered, for example transport properties, toxicity, and stability."* By means of virtual screening it is comparably easy to find novel structures exhibiting some desired bioactivity, but the *de novo* design of real drugs is only very slowly emerging. Fragment-based evolutionary design approaches that are able to optimize for several fitness functions in parallel might lead to significant progress in this area.

Acknowledgements

Man-Ling Lee implemented the retro-synthetic fragmentation scheme of TOPAS and performed WDI fragmentation, Chiara Taroni compiled the CATS test data sets, and Martin Stahl performed the FlexX ligand-docking experiments and prepared Figure 8.18.

References

[1] H.-J. Böhm, *Curr. Opin. Biotechnol.* **1996**, *7*, 433–436.
[2] R. S. Bohacek, C. McMartin C., *Curr. Opin. Chem. Biol.* **1997**, *1*, 157–161.
[3] H. van de Waterbeemd, B. Testa, G. Folkers (Eds.) *Computer-Assisted Lead Finding and Optimization*, Wiley-VCH, Weinheim **1997**.
[4] V. J. Gillet, A. P. Johnson, in *Designing Bioactive Molecules – Three-dimensional Techniques and Applications*, Y. C. Martin, P. Willett (Eds.), American Chemical Society, Washington, DC **1998**, pp. 149–174.
[5] H. Kubinyi, *J. Recept. Signal Transduct. Res.* **1999**, *19*, 15–39.
[6] H.-J. Böhm, *J. Comput.-Aided Mol. Design* **1992**, *6*, 593–606.
[7] D. C. Roe, I. D. Kuntz, *J. Comput.-Aided Mol. Design* **1995**, *9*, 269–282.
[8] G. Lauri, P. A. Bartlett, *J. Comput.-Aided Mol. Design* **1994**, *8*, 51–66.
[9] D. E. Clark, D. Frenkel, S. A. Levy, J. Li, C. W. Murray, B. Robson, B. Waszkowycz, D. R. Westhead, *J. Comput.-Aided Mol. Design* **1995**, *9*, 13–32.
[10] C. W. Murray, D. E. Clark, T. R. Auton, M. A. Firth, J. Li, R. A. Sykes, B. Waszkowycz, D. R. Westhead, S. C. Young, *J. Comput.-Aided Mol. Design* **1997**, *11*, 193–207.
[11] S. H. Rotstein, M. A. Murcko, *J. Comput.-Aided Mol. Design* **1993**, *7*, 23–43.
[12] D. A. Pearlman, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 1651–1663.
[13] H.-J. Böhm, *J. Comput.-Aided Mol. Design* **1998**, *12*, 309–323.
[14] H.-J. Böhm, D. W. Banner, L. Weber, *J. Comput.-Aided Mol. Design* **1999**, *13*, 51–56.
[15] M. Stahl, H.-J. Böhm, *J. Mol. Graph. Model.* **1999**, *16*, 121–132.
[16] L. B. Kier, *Pure Appl. Chem.* **1993**, *35*, 509–520.
[17] G. Schneider, P. Wrede, *Biophys. J.* **1994**, *66*, 335–344.
[18] A. L. Lomize, I. D. Pogozheva, H. I. Mosberg, *J. Comput.-Aided Mol. Design* **1999**, *13*, 325–353.
[19] V. Tschinke, N. C. Cohen, *J. Med. Chem.* **1993**, *36*, 3863–3870.
[20] D. E. Clark, M. A. Firth, C. W. Murray, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 137–145.
[21] G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chemie Int. Ed.* **1999**, *38*, 2894–2896; *Angew. Chemie* **1999**, *111*, 3068–3070.
[22] D. Frenkel, D. E.Clark, J. Li, C. W. Murray, B. Robson, B. Waszkowycz, D. R. Westhead, *J. Comput.-Aided Mol. Design* **1995**, *9*, 213–225.

[23] J. B. Moon, W. J. Howe, *Proteins* **1991**, *11*, 314–328.
[24] G. Schneider, T. Todt, P. Wrede, *Comput. Appl. Biosci.* **1994**, *10*, 75–77.
[25] G. Schneider, J. Schuchhardt, P. Wrede, *Biophys. J.* **1995**, *68*, 434–447.
[26] B. Levitan, S. Kauffman, *Mol. Diversity* **1995**, *1*, 53–68.
[27] S. Kauffman, *The Origins of Order*, Oxford University Press, New York **1993**.
[28] I. Rechenberg, *Evolutionsstrategie '94*, Frommann-Holzboog, Stuttgart **1994**.
[29] M. Conrad, W. Ebeling, M. V. Volkenstein, *Biosystems* **1992**, *27*, 125–128.
[30] J. H. Mathews, *Numerical Methods for Mathematics, Science, and Engineering*, Prentice-Hall International, Englewood Cliffs, NJ **1992**.
[31] A. S. Schulz, D. B. Shmoys, D. P. Williamson, *Proc. Natl Acad. Sci. USA* **1997**, *94*, 12734–12735.
[32] J. R. Desjarlais, N. D. Clarke, *Curr. Opin. Struct. Biol.* **1998**, *8*, 471–475.
[33] D. E. Clark, D. R. Westhead, *J. Comput.-Aided Mol. Design* **1996**, *10*, 337–358.
[34] I. Rechenberg, *Evolutionsstrategie – Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*, Frommann-Holzboog, Stuttgart **1973**.
[35] J. H. Holland, *Adaptation in Natural and Artificial Systems*. M. I. T. Press, Cambridge, MA **1975**.
[36] P. Willett, *Trends Biotechnol.* **1995**, *13*, 516–521.
[37] V. M. Kolb, *Prog. Drug Res.* **1998**, *51*, 185–217.
[38] D. B. Schmoys, in *Combinatorial Optimization*, W. Cook, L. Lovász, L., P. D. Seymour (Eds.), Providence, RI, American Mathematical Society **1995**, pp. 355–397.
[39] D. S. Hochbaum (Ed.) *Approximation Algorithms for NP-Hard Problems*, Boston, PWS **1997**.
[40] G. Schneider, W. Schrödl, G. Wallukat, E. Nissen, G. Rönspeck, J. Müller, P. Wrede, R. Kunze, *Proc. Natl Acad. Sci. USA* **1998**, *95*, 12179–12184.
[41] N. Saravanan, D. B. Fogel, K. M. Nelson, *Biosystems* **1995**, *36*, 157–166.
[42] D. A. Engelmann, T. A. Steitz, A. Goldman, *Ann. Rev. Biophys. Biophys. Chem* **1986**, *15*, 321–353.
[43] A. A. Zamyatnin, *Prog. Biophys. Mol. Biol.* **1972**, *24*, 107–123.
[44] G. Schneider, P. Wrede, *Prog. Biophys. Mol. Biol.* **1998**, *70*, 175–222.
[45] P. Wrede, O. Landt, S. Klages, A. Fatemi, U. Hahn, G. Schneider, *Biochemistry* **1998**, *37*, 3588–3593.
[46] S. F. Altschul, *J. Mol. Biol.* **1991**, *219*, 555–565.
[47] S. Henikoff, J. G. Henikoff, *Proteins* **1993**, *17*, 49–61.
[48] G. Trinquier, Y.-H. Sanejouand, *Protein Engineering* **1998**, *11*, 153–169.
[49] H. Kubinyi, in *Computer-Assisted Lead Finding and Optimization*, H. Van de Waterbeemd, B. Testa, G. Folkers (Eds.), Wiley-VCH, Weinheim **1997**, pp. 9–28; and references therein.
[50] J. Zupan, J. Gasteiger, *Neural Networks for Chemists* (2nd extended edition), VCH, Weinheim **1993**.
[51] H. van de Waterbeemd (Ed.) *Advanced Computer-Assisted Techniques in Drug Discovery*, VCH, Weinheim **1994**.
[52] G. Cybenko, *Mathematics of Control, Signals, and Systems* **1989**, *2*, 303–314.
[53] J. Hertz, A. Krogh, R. G. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City **1991**.
[54] M. L. Minsky, S. A. Papert, *Perceptrons*, MIT Press, Cambridge, MA **1969**, expanded edition **1990**.
[55] A. N. Kolmogorov, *Dokl. Akad. Nauk SSSR* **1957**, *114*, 953–956.
[56] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford **1995**.
[57] G. Schneider, J. Schuchhardt, P. Wrede, *Biol. Cybernetics* **1995**, *73*, 245–254.
[58] K. Hornik, M. Stinchcombe, H. White, *Neural Networks* **1989**, *2*, 359–366.
[59] E. B. Baum, D. Haussler, *Neural Computation* **1989**, *1*, 151–160.
[60] M. Paetzel, R. E. Dalbey, N. C. J. Strynadka, *Nature* **1998**, *396*, 187–190.
[61] R. A. Lewis, A. C. Good, in *Computer-Assisted Lead Finding and Optimization*, H. van de Waterbeemd, B. Testa, G. Folkers (Eds.), Wiley-VCH, Weinheim **1997**, pp. 137–156.
[62] G. Sello, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 691–701.
[63] M. D. Miller, R. P. Sheridan, S. K. Kearsley, *J. Med. Chem.* **1999**, *42*, 1505–1514.
[64] M. C. Nicklaus, S. M. Wang, J. S. Driscoll, G. W. A. Milne, *Bioorg. Med. Chem.* **1995**, *3*, 411–428.
[65] S. D. Pickett, J. S. Mason, I. M. McLay, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214–1223.
[66] M. Rarey, J. S. Dixon, *J. Comput.-Aided Mol. Design* **1998**, *12*, 471–490.
[67] J. Qar, J. P. Galizzi, M. Fosset, M. Lazdunski, *Eur. J. Pharmacol.* **1987**, *141*, 261–268.
[68] L. Weber, S. Wallbaum, C. Broger, K. Gubernator, *Angew. Chemie Int. Ed.* **1995**, *34*, 2280–2202; *Angew. Chemie* **1995**, *107*, 2452–2454.
[69] M. R. Wiley, M. J. Fisher, *Exp. Opin. Ther. Patents* **1997**, *7*, 1265–1282.
[70] B. Bienfait, J. Gasteiger, *J. Mol. Graph. Model.* **1997**, *15*, 203–215.

[71] M. Stahl, D. Bur, G. Schneider, *J. Comput. Chem.* **1999**, *20*, 336–347.
[72] R. C. Glen, A. W. Payne, *J. Comput.-Aided Mol. Design* **1995**, *9*, 181–202.
[73] R. D. Brown, Y. C. Martin, *J. Med. Chem.* **1997**, *40*, 2304–2313.
[74] S. J. Cho, W. Zheng, A. Tropsha, *Pac. Symp. Biocomput.* **1998**, 305–316.
[75] L. Weber, *Curr. Opin. Chem. Biol.* **1998**, *2*, 381–385.
[76] V. Brusic, G. Rudy, M. Honeyman, J. Hammer, L. Harrison, *Bioinformatics* **1998**, *14*, 121–130.
[77] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
[78] M. Rarey, S. Wefing, T. Lengauer, *J. Comput.-Aided Mol. Design* **1996**, *10*, 41–54.
[79] D. W. Banner, P. Hadváry, *Biol. Chem.* **1991**, *266*, 20085–20093.
[80] M. A. Mohler, C. J. Refino, S. A. Chen, A. B. Chen, A. J. Hotchkiss, *Thromb. Haemost.* **1986**, *56*, 160–164.
[81] P. R. Gerber, K. Müller, *J. Comput.-Aided Mol. Design* **1995**, *9*, 251–268.
[82] G. Schneider, M.-L. Lee, M. Stahl, P. Schneider, *J. Comput.-Aided Mol. Design* **2000**, *14*, 487–494.
[83] G. Schneider, *Neural Networks* **2000**, *13*, 15–16.
[84] J. Ellman, B. Stoddard, J. Wells, *Proc. Natl Acad. Sci. USA* **1997**, *94*, 2779–2782.

Virtual Screening for Bioactive Molecules
Edited by Hans-Joachim Böhm and Gisbert Schneider

9 Practical Approaches to Evolutionary Design

Lutz Weber

9.1 Introduction

Medicinal chemistry is about understanding the relation between chemical structures and their biological activities. The creation of new drugs by medicinal chemists can be understood as an iterative process of selecting and synthesizing molecules, testing their properties and selecting new, better molecules in several rounds of feedback cycles (Figure 9.1). All parts of this cycle have undergone substantial change during the past decade. Biological testing has been changed by the implementation of high-throughput screening methods through the integration of the advances in molecular biology, automation and miniaturization. For the synthesis of new molecules, combinatorial chemistry has been introduced that allows the generation of molecules in numbers and with a speed unprecedented so far.

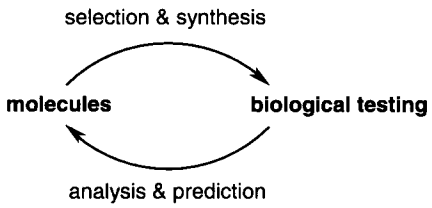


Figure 9.1. A typical feedback cycle.

Many methods have also been developed in the past to understand the relationship between chemical structures and their biological activities. Classical methods are for example quantitative–structure activity relationships (QSAR), comparative molecular field analysis and molecular modelling, to name only a few. In the beginning of medicinal chemistry, only relatively small numbers of molecules were considered in QSAR studies. High-throughput methods in screening and synthesis have shifted the attention and needs of medicinal chemists towards exploiting larger and larger chemical spaces in their search for molecules with useful properties. New analysis and prediction methods are thus required that enable the medicinal chemist to deal with the large volume of generated data faster and more efficiently than previously possible. However, the increase of such quantitative aspects of drug discovery has in parallel also induced an interest in the more qualitative, abstract understanding of the relation of chemical structures and their properties. This abstract understanding of chemical space should allow the development of methods that aid the selection of molecules out of this space, having desired properties with a higher likelihood than random selection.

The number of discrete and different molecules of interest that constitute the search space in medicinal chemistry is of practically unlimited size. For example, the number of all different proteins comprising just 200 amino acids from the possible 20 is 20^{200}, a number that is much larger than the number of particles in space. The same is true for the number of "drug-like" molecules with a molecular weight below 600. Using the tools of combinatorial chemistry, we are now able in principle to create compound libraries of such a large size that their members cannot be synthesized and tested individually. Even more, the recent term "very large library" refers to compound libraries that also exceed our computational capabilities for giving a number or description (enumeration) to each member of this library.

Thus, we find ourselves in a position quite similar to the evolution of molecules in living organisms – on the one hand we are in principle able to synthesize very many molecules by combining a rather limited number of starting materials. On the other hand, it is still physically impossible to obtain all molecules comprising all desired properties. How can we nevertheless select molecules out of this space of interest when we do not even know each member explicitly?

The iterative selection of useful molecules out of such very large spaces of possible molecular solutions is the basic principle of evolution in Nature. In the practice of medicinal chemistry, one observes a similar process: a variety of selection methods are applied in iterative cycles to arrive at useful drug candidates. The similarity of the drug design cycle with evolutionary selection processes has recently therefore inspired the development of evolutionary drug design methods.

In this Chapter, we present just such an evolutionary compound selection method. It tries to mimic Nature's molecular evolution within one integrated process, giving raise to a more automated drug discovery. Unlike other methods that include basically only one computer-aided molecule selection step and the subsequent testing of these molecules, this method selects such molecules rather in series of evolutionary cycles of computational selection, synthesis and biological screening. We would like to summarize the basic ideas and first successful examples of this practical approach to evolutionary design of new molecules.

9.2 The Structure of the Search Space

The nature of the search space, the space of accessible chemical entities and their biological, physico-chemical and pharmacological properties has a large influence on the optimal choice of appropriate selection methods. Structure–property relationships (SPRs) of various kinds have been considered in Chapter 8.

Obviously one would use different selection methods if dealing with a unimodal (Chapter 8, Figure 8.4a) or multimodal (Chapter 8, Figure 8.4b) SAR. But what do real SARs look like? Interestingly, this important question is still very much open to discussion in the community of medicinal chemists, mainly because very little knowledge is available about large SAR landscapes. Thus, we decided to get an impression of the problem by synthesizing a complete library of 256 chemically similar molecules of inhibitors that are biased towards the serine protease thrombin using the known Ugi-type four-component reaction. For this library four different aldehydes (A1–A4), amines (B1–B4), isonitriles (C1–C4), and carboxylic acids

(D1–D4) were used. All individual 256 molecules were tested for their inhibitory activity of the thrombin enzyme activity. These inhibitory concentrations are displayed in Figure 9.2a as equal height contour lines on the 4 x 4 x 4 x 4 = 256 compound library. The distribution scheme of the starting materials in the horizontal and vertical direction was C1D1, C1D2, C1D3, C1D4, C2D1, C2D2, C2D3, C2D4, C3D1, C3D2, C3D3, C3D4, C4D1, C4D2, C4D3, C4D4 and A1B1, A1B1, A1B2, A1B2, A2B3, A2B3, A2B4, A2B4, A3B1, A3B1, A3B2, A3B2, A4B3, A4B3, A4B4, A4B4, respectively. Thus, for example, position vertical-15/horizontal-14 represents the biological activity of the product of the reaction A4B4C4D2 (Figure 9.3).

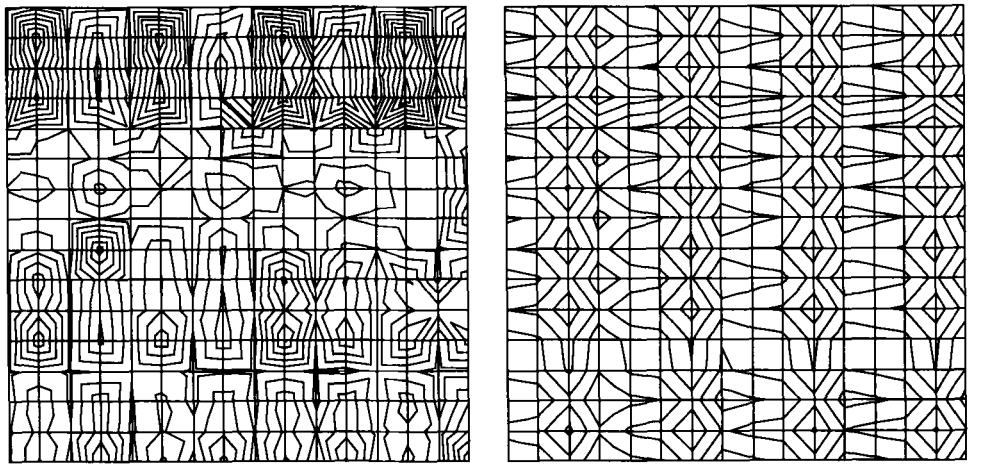


Figure 9.2. Structure–activity landscape of a combinatorial library of 256 Ugi-type four-component products. The inhibitory concentrations against the serine protease thrombin are displayed as equal height contour lines in **a**. Part **b** was generated by overlaying two sinusoidal functions, in the horizontal and the vertical direction respectively, and displaying the resulting amplitude as contour lines of equal height.

As a result, the experimental landscape 9.2a is quite similar to the multimodal landscape 9.2b, which was generated by an arbitrary overlay of sinusoidal functions in two dimensions.

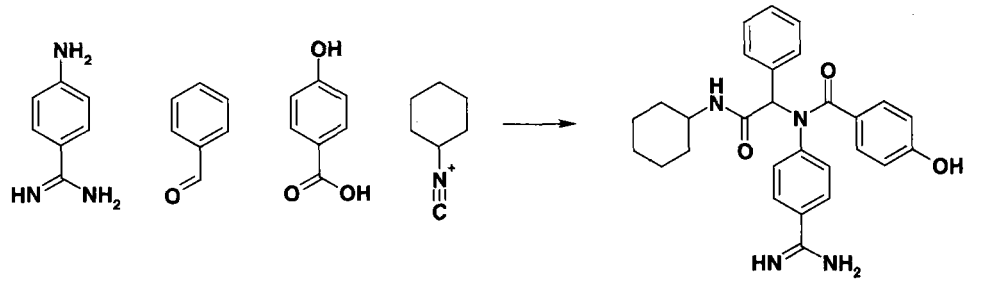


Figure 9.3. A typical Ugi four-component reaction as used in the 256-member compound library of Figure 9.2.

What are the lessons of this SAR? First of all it shows nicely the *Principle of Strong Causality* mentioned in Chapter 8 – small or conservative changes in structures cause small changes in biological activity. Secondly, the distribution scheme of the starting materials and hence their products results in a related distribution of their biological activities. Clearly, one could re-sort the starting materials in a way that a more unimodal SAR is achieved – but this would require knowledge of the SAR of interest *a priori*.

A model of a "universal" SAR is often understood as a landscape where most of the considered compounds are inactive for a given biological target, and active "islands" of similar molecules show a multimodal SAR behavior as in our preliminary experiment. Applying this model, the task of drug discovery is roughly divided into finding, firstly, these "islands" and, secondly, the most useful compound on this multimodal island.

This SAR model matches the recent work of W. Fontana and P. Schuster [1] who investigated the shape change of short RNA sequences depending on the changes in the sequence of those RNA aptamers. They give a definition of *continuity* (or, in other words, strong causality) between the RNA sequence space and the shape space (the property of the molecules): neighbors in sequence space have shapes that are also neighbors in the shape space. A change in sequence space that gives rise to a different shape is called *transition*. Two types of transitions are observed: *continuous*, that follow the principle of strong causality, and *discontinuous* (or in our context a different structure island), that do not follow this principle.

Selecting compounds in such a way that each island of the structure–property relationship is represented by one compound would give a library of maximally diverse compounds or *discontinuous transitions*. Selecting compounds from only one island of the structure–property relationship would give a library of similar compounds.

Current descriptions of SARs, and also Figure 9.2, often wrongly suggest that the chemistry space by itself is "smooth" or "continuous". It is important to understand that the space of possible molecules is discrete – each molecule is an assembly either of discrete atoms, or, as in the case of oligonucleotides and peptides, of a given set of monomers like nucleotides or amino acids. The chemical distance between the points or individual molecules of that space can be defined using the valence bond theory [2] providing a formal metric that has no implications whatsoever on SAR.

On the other hand, the property space, like biological activities or physico-chemical properties, is often continuous. This difference between the chemistry space and the property space may be understood in mathematical terms as that of the spaces of natural and rational numbers. A SAR is thus a unidirectional projection of the members of the discrete space of chemistry onto a continuous property space. The visualization of that relationship as in Figures 9.2 or 8.4 can be misleading, because it suggests that neighbors on the chemistry axis are also neighbors in chemistry space.

In summary, we need search strategies that find optimal solutions in the search space, which is composed of the discrete space of chemistry and the continuous property space.

9.3 Genetic Algorithms

Combinatorial optimization methods [3] have been introduced as general tools to solve problems in all kinds of discrete search spaces. One of the most prominent and widely ap-

plied combinatorial optimization methods has been inspired by the Darwinian principles of evolution and was termed "genetic algorithm" (GA) by Holland [4]. Since its inception, the number of drug discovery-related applications of GAs have grown from year to year [5].

GAs operate on the relation of two spaces – the space that encodes possible discrete solutions (genotype) and the property space (phenotype) of these solutions. In the language of GAs, the chemical structure of a compound can be understood as the genotype or chromosome which is constructed from genes – atoms or building blocks like amino acids or nucleotides – whereas the actual molecule with its physical and biological properties represents the phenotype.

GAs use several different genotypes or individuals at the same time – the population – and investigate their properties using a given selection function. After such evaluation of the population, the individual members are rank-ordered according to their fitness. The chromosomes of the rank-ordered individuals are then subject to changes (genetic operators) that generate new chromosomes according to predefined rules.

The mechanisms that drive evolution are reproduction, mutation, crossover, and the Darwinian principle of survival of the fittest. In Nature, these mechanisms enable life forms to adapt to a particular environment over successive generations. The fitness function provides a feedback system to measure each individual phenotype's fitness within this population. In drug discovery, new molecules are evaluated by a biological screening system providing the feedback for the selection of more fit molecules. Genetic operators give the rules governing how new genotypes are generated from parents. For example, the higher the fitness, the greater should be the probability of passing the genomic information onto the next generation. The choice of representation and the choice of genetic operators distinguish various types of evolution, natural as well as artificial.

The translation of this general method towards an evolutionary method for drug discovery requires the implementation of encoding, selection and genetic operators for molecules. Let's assume we would like to find a hexapeptide that inhibits the proteolytic activity of a given protease, using a GA. The chemistry space is defined by the 20 natural amino acids (genes) that can be encoded by either their single letter codes (alphanumeric coding) or any arbitrary number. Thus, for the six positions we may also use A1–A20 for position one, B1–B20 for position two and so forth. There are overall 20^6 different hexapeptides possible. The genotype (or chromosome) of each individual peptide is given by the assembly of six such genes, e.g.

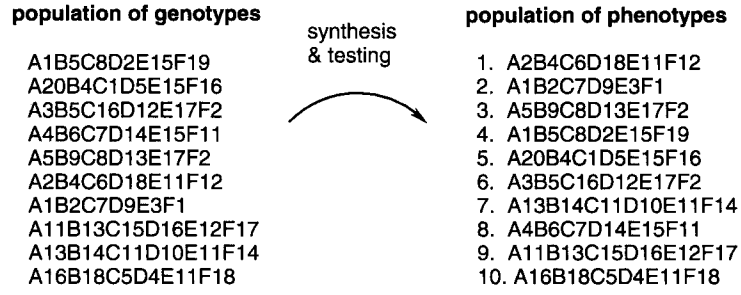


Figure 9.4. An initial population of chromosomes is synthesized and rank-ordered according to their biological activities.

A1B1C3D11E9F8 for the sequence of AALRKG in the alphanumeric coding. For other suitable coding schemes of molecules see [6].

A first population of such peptides can than be chosen by either random selection or by using appropriate rules. As a result, we may use a population size of ten peptides that are than synthesized and tested for their inhibitory activity in the laboratory. The ten peptides are than sorted according to their activity with the best peptide on the top (Figure 9.4).

In the next step of the genetic algorithm we apply genetic operators to the chromosomes of this list of ten peptides to generate new chromosomes. These chromosomes again may be subject to certain criteria, e.g. we could forbid the resynthesis of already known molecules or reject molecules with unwanted properties, such as having a molecular weight above 600.

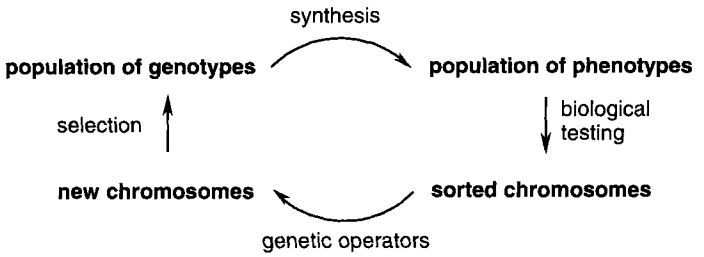


Figure 9.5. Examples of the steps in a genetic algorithm

The whole algorithm is than repeated several times over several generations until a peptide is found that satisfies our requirements (Figure 9.5). A pseudo-code for a GA could be described as follows [7]:

```
    // initialize a usually random population of individuals
    initpopulation P (t);
    // evaluate fitness of all initial individuals of population
    evaluate P (t);
    // test for termination criterion (time, fitness, etc.)
    while not criterion not satisfied do
    // select a sub-population for offspring production
    P' := selectparents P (t);
    // recombine the „genes" of selected parents
    recombine P' (t);
    // perturb the mated population stochastically
    mutate P' (t);
    // evaluate its new fitness
    evaluate P' (t);
    // select the survivors from actual fitness
    P := survive P,P' (t);
end
```

This GA procedure is similar to other evolutionary algorithms like evolutionary algorithms (see Chapter 8), evolutionary programming, genetic programming or cellular automata. GAs differentiate themselves from these methods by the nature of the genetic operators that are applied to the chromosomes. For example, evolutionary programming usually does not use the crossover operator. Also, in the GA approach, each chromosome represents a possible solution to the search problem – making GA different from evolutionary programming where the encoding is problem-dependent. Since several chromosomes or possible solutions are evaluated at the same time, GAs intrinsically are parallel search algorithms. They differ from other parallel algorithms, like simulated annealing or Monte-Carlo, because the parallel hypotheses are not independent from each other but take advantage of the acquired knowledge of all populations and individuals. This knowledge however is implicit since no explicit SAR model is established. Such implicit methods are also called heuristic stochastic search algorithms.

9.4 Genetic Operators and the Building Block Hypothesis

Various methods are possible for creating the new population of chromosomes from the ranked list of already evaluated chromosomes. These GA operators are inspired by those of DNA like genetics: death, replication, deletion, insertion, mutation, and crossover (Figure 9.6).

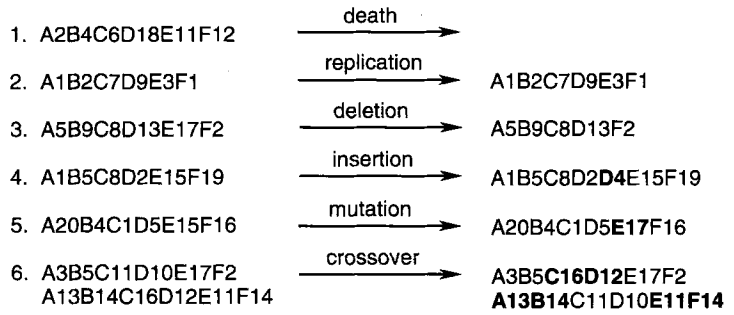


Figure 9.6. Examples of the genetic operators acting at a chromosome, using alphanumeric coding.

Replication regenerates an equivalent chromosome or individual. Mutation sets one or more elements in the parent gene to a different value, based on a predefined mutation rate between 0 and 100 percent. Crossover takes two or more chromosomes to build new chromosomes by mixing them according to various rules. Deletion deletes an element from the parent gene, insertion introduces new elements.

Contrary to evolution in Nature, we are completely free to define how these GA operators are applied in the computer: for example a new chromosome may be the result of mixing more than just two chromosomes. Each of the new chromosomes however has to have a chemical meaning – which is given by the chosen encoding scheme.

In the literature one finds various implementations of these genetic operators – while various combinations and variants are possible, it has not been shown that any specific version is superior. Many parameters can be set and varied during the course of a genetic algorithm experiment: for example the rate of mutation and crossover, size of population, number of replicated chromosomes, and finally the ranking function.

The good search power of GAs is believed to originate from the building-block hypothesis [8–11]. This model assumes that the combination of "fit" building blocks, or genes, or schemes of genes on the chromosome yields higher-order schemes of even better fitness. In the light of the above discussion on the properties of the SAR space, this hypothesis is equivalent to a continuous behavior.

The building-block hypothesis matches perfectly the discontinuous, discrete structure space of chemistry that is statistically analyzed by the GA. Such suitable building blocks for coding in GAs are as described above: atoms, substituents, reagents, starting materials, or even reactions. Combinatorial libraries are especially amenable for GA-based evaluation since they are generated from systematic arrays of real starting materials or building blocks.

For the above example in Figure 9.4, building block A1 appears in the chromosome with rank one and four, and B4 with one and five. It is therefore meaningful to assume that the combination A1B4 is a scheme of higher order. C8 appears with rank three and four, so A1B4C8 could therefore be part of an even better peptide. The task of the genetic operators is to observe such schemes and to propose new chromosomes using these schemes. Thus, in a typical implementation of GAs, the probability of applying a genetic operator towards a given chromosome depends on its rank or "fitness". For fitter chromosomes, the likelihood of replication and crossover should be higher while less fit chromosomes should be more likely to be eliminated by death. For a new book on the theory of GAs see [8].

9.5 Practical Examples

The idea of the experimental implementation of evolutionary chemistry for small molecules is based on the idea of mimicking Nature's evolution by encoding molecules in the computer and applying genetic algorithms towards these artificial chromosomes. Examples have been reported on the successful integration of genetic algorithms, organic synthesis, and biological testing in evolutionary feedback cycles to yield molecules with desired properties.

In a first example, hexapeptidic inhibitors of the serine protease trypsin have been selected [13]. The GA was started with a population of 24 randomly chosen peptides from the space of 64 million possible hexapeptides. The selection function was given by performing a chromogenic trypsin assay. Out of the evaluated 24 peptides, the best six inhibitory peptides were duplicated, the worst six were eliminated, and the resulting 24 peptides were then subjected to random crossover with 100% probability. Subsequent 3% mutation to the four residues F, I, K, or T produced a new population of 24 peptides for the next generation. The inhibitory activity was improved from the average of 16% of the first randomly chosen population to an average 50% in the sixth generation. In this generation, 13 peptides out of the 25 most active peptides showed a consensus sequence of Ac-XXXXKI-NH2, and eight had a Ac-XXKIKI-NH2 sequence. The identified best peptide was Ac-TTKIFT-NH2, which exhib-

ited an inhibition of 89% and was identical with a known trypsin inhibitor from a phage display library.

Substrates for the protease stromelysin were searched [14] in an other example out of the pool of possible hexapeptides. Only 300 peptides were synthesized in five generations to obtain such useful substrates. The peptides were synthesized on solid support and fluorescence-marked at the N-terminus. Non-fluorescent beads were analyzed after proteolytic cleavage of the beads with stromelysin. The starting population was biased by using 60 peptides with the sequence X1PX3X4X5X6, but removing the bias in all subsequent generations. The encoding of the amino acids was performed with binary bit strings of five bits for each gene. For example, amino acid number ten corresponded to the bit string 01010 in this binary gene representation.

From the population of 60 peptides, the best peptide was copied to the new generation, and the others were changed by a crossover rate of 60%. The new chromosomes were then subjected to mutation with a rate of 0.1% applied to each bit of the 30-bit string genome, giving a 3% overall mutation rate. The GA was terminated when 95% of the peptides in the population were identical. The hexapeptide MPQYLK was identified as the best substrate for stromelysin in the final generation, being cleaved between the amino acids tyrosine and lysine. The selectivity of the new substrates *versus* collagenase was also determined; and selective substrate sequences have been identified for both enzymes. Therefore, this method may help not only to find new substrates but also to obtain structure–activity and selectivity ideas for new inhibitors.

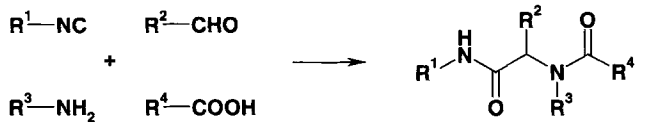


Figure 9.7. General reaction scheme for the Ugi four-component reaction.

The first example for a GA-driven selection and synthesis of non-peptide molecules has been published for the selection of thrombin inhibitors [15]. Using 10 isonitriles **A**, 10 amines **B**, 40 aldehydes **C**, and 40 carboxylic acids **D**, 160000 Ugi four-component reaction (Figure 9.7) combinations are possible. Binary encoding of the starting materials was used as in the example above, using four bits for the amines and isonitriles and six bits for the aldehydes and carboxylic acids, respectively. Whereas the isonitriles and aldehydes were selected based on their availability and to cover a broad range of chemical diversity, the amines were chosen to fill the thrombin arginine binding S1 pocket (Figure 9.8).

Some of these amines were already known for their affinity to thrombin in the high micromolar range. Due to this structural bias, we expect the molecules of this library to cover a broad range of affinities to thrombin as in our preliminary SAR example. Before starting real experiments, a virtual SAR of 160000 data points was constructed by an overlay of various four sinusoidal + linear functions to simulate a four-dimensional, multimodal space in order to determine optimal population sizes and mutation rates. The list of ranked individuals was processed as follows: the n best individuals were retrieved from the list of the chromosomes

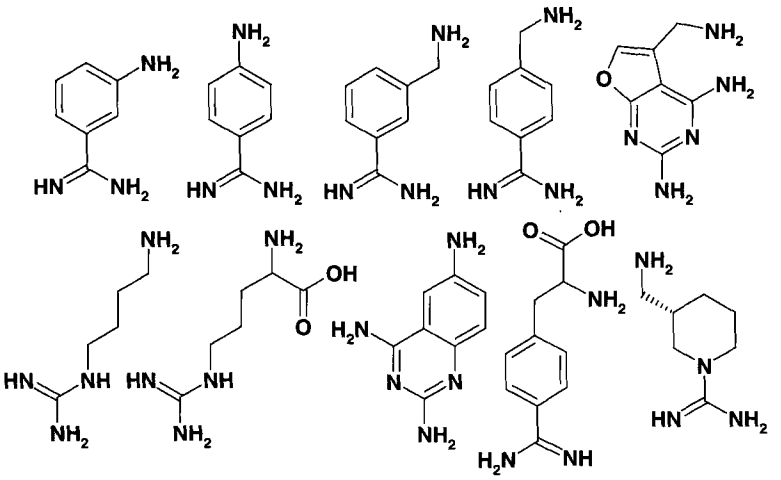


Figure 9.8. Amines used in the GA-based selection of thrombin inhibitors.

generated so far, and new individuals were generated using a crossover rate of 100% for each chromosome and a mutation rate of 1% per bit. Newly generated individuals were then looked up, to check whether they had been proposed already, and if so, they were rejected. The procedure was repeated until n new chromosomes were obtained for the next generation. The interesting result of these simulations was that the number of individuals needed to find the optimal solution in this rather limited space was nearly optimal for population sizes of $n = 20$. More extensive considerations on optimal GA parameters have been published recently, using real SAR data from larger combinatorial libraries [16,17].

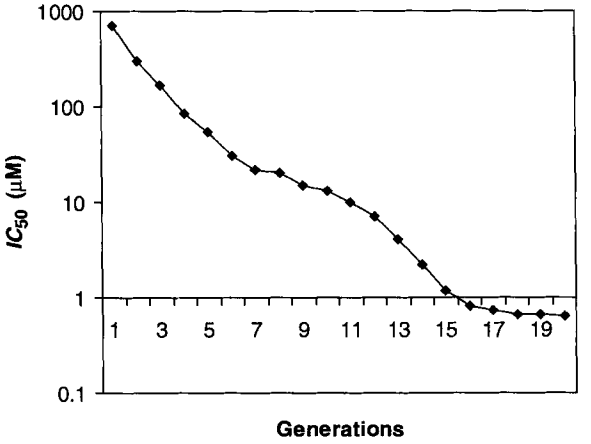


Figure 9.9. The activity increase of the best inhibitor found during the GA-driven compound selection experiment. In each generation, 20 compounds were synthesized and tested. Inhibitory activities are given for the best compound only, with its IC_{50} value.

Following this simulation, a real experiment was started by synthesizing the individual molecules, testing their biological activities against thrombin, and proposing new molecules for synthesis by the genetic algorithm. Using a population size of 20 Ugi products and the above procedure, the algorithm yielded sub-micromolar inhibitors after 14 cycles of synthesis and testing (Figure 9.9).

Whereas in the initial population the best reaction product exhibited an IC_{50} of about 300 μM, a thrombin inhibitor with a IC_{50} of 0.38μM was discovered after 20 generations of 20

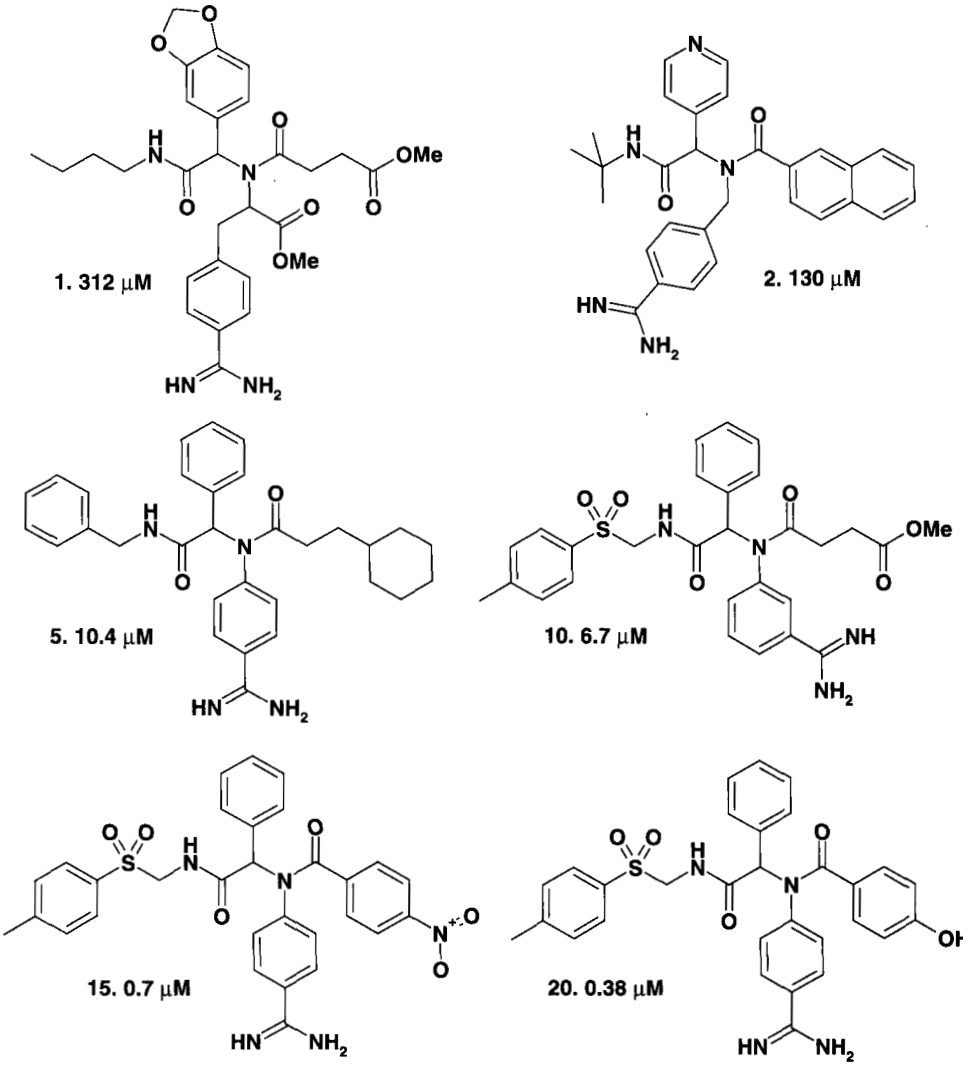


Figure 9.10. Examples of compounds and their biological activities that were generated during the GA-based selection of thrombin inhibitors.

single Ugi products per population. Figure 9.10 displays some representative structures that have been obtained in the different generations.

To our surprise, the most active reaction product contained two compounds, the expected product **A** and the *N*-aryl-phenylglycine amide derivative **B** (Figure 9.11). Resynthesis and purification of both products on a larger scale, and re-testing their biological activities, gave an IC_{50} value of 1.4 μM for **A** and 0.22 μM for **B**, respectively. **B** is the three-component side-product of the four-component reaction. While this result may appear disturbing, the coding for the GA describes rather the process of combining the four starting materials and not the final, expected products! The applied GA is obviously not product structure-based and the feedback function, the inhibitory effect of the crude reaction product, depends on various results of the overall process including varying the yield of the reaction, erroneous results in biological screening, and so forth.

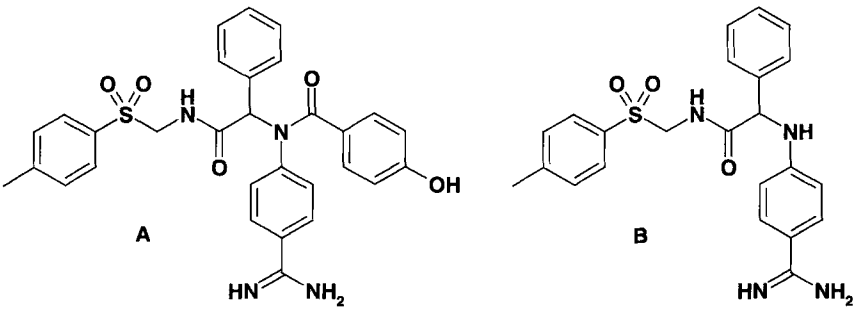


Figure 9.11. The two compounds in the most active reaction product found by GA.

In the light of these results, the introduction of a deletion and insertion function makes sense. Product **B** can thus be regarded as a product where the carboxylic acids are "deleted" from the final product. Moreover, in such a case the starting material would be active, for example the amine 4-amino-benzamidine, and the GA could find this by three successive deletions on the four-gene chromosome. *Vice versa*, a complete four-gene chromosome could be

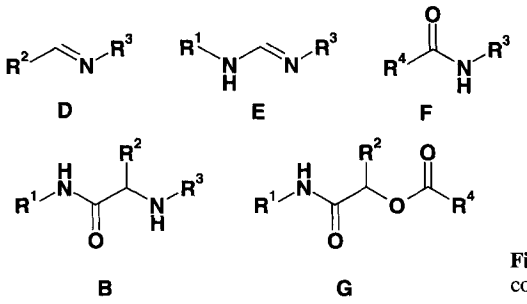


Figure 9.12. Possible side products of a Ugi four component reaction.

reconstructed from a starting material by insertions. Meaningful combinations of lower complexity are, for example, the known two-component products of amines and aldehydes **D**, amines and isonitriles **E**, amines and acids **F**, as well as the three-component products from amines, aldehydes and isonitriles **B**, or from aldehydes, isonitriles and carboxylic acids **G** (Figure 9.12).

This concept provides a more general framework for the application of genetic algorithms towards chemistry and SPRs. While the previous examples of practical applications of GA were designed to find molecules with a predefined backbone structure, like peptides or Ugi four-component products, we have recently introduced GAs that use several structural types in one genetic run [18]. This approach is especially efficient with isonitrile-based multi-component reactions where, for example, the variation of the acid-component and the amine gives easy access to a variety of backbone structures (Figure 9.13).

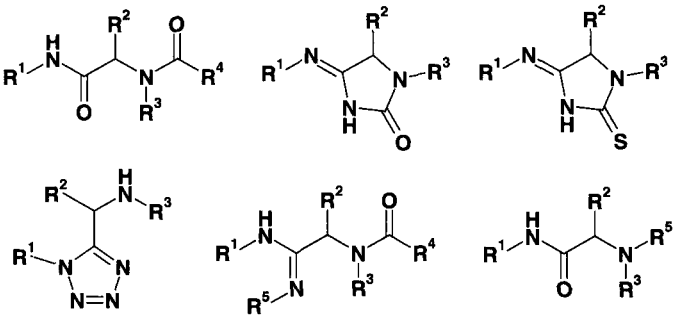


Figure 9.13. Typical backbone structures that are accessible by isonitrile-based four component reactions.

In this multi-backbone GA, the reaction type used and hence the resulting backbone of the reaction product was also encoded on the respective chromosome. The selection function of a GA can be any measurable quantity. Recently we have used such an extended GA to optimize the reaction conditions of a four-component reaction, where the yield of the chemical reaction provided the selection function [19]. The chromosomes in this evolutionary experiment included not only the starting materials but also the reaction conditions, such as various solvents or time points for adding the starting materials to the reaction mixture. Using this GA, the yield of the reaction of interest could be improved from 6% to 60% over only six generations.

9.6 Efficiency of Genetic Algorithms

The introduction of a GA-driven evolutionary selection of suitable molecules out of compound libraries poses the question of how does this algorithm compare with random selec-

tion or other selection methods? To answer this question, it is necessary to know all solutions *a priori*, such as for example the biological affinities of all members of the library against the target protein of interest. The resulting structure–activity landscape can than be used as a model, comprising the search space for GA-driven experiments. By running genetic algorithms within this search space, one could optimize various parameters of the genetic algorithms and evaluate their influence on the search efficiency of GAs in simulated molecular evolution experiments. In contrast to the previous Chapter, these evolution experiments will run on real landscapes rather than on virtual ones.

We reported recently the synthesis of a complete combinatorial library of 15360 parallel Ugi-type three-component reaction products **B**, having a structural bias towards arginine S1 and recognizing serine proteases as a test case [16]. The starting materials were 80 aldehydes (A1–A80), 12 amines (B1–B12), and 16 isonitriles (C1–C16). The respective isonitriles and aldehydes were selected to cover a broad range of chemical diversities by using large and small aliphatic, aromatic, heteroaromatic, benzylic, and both substituents with hydrogen bond donors, and acceptors. Amines B1–B12 were chosen to provide a structural bias towards the arginine-binding S1 pocket (Figure 9.14).

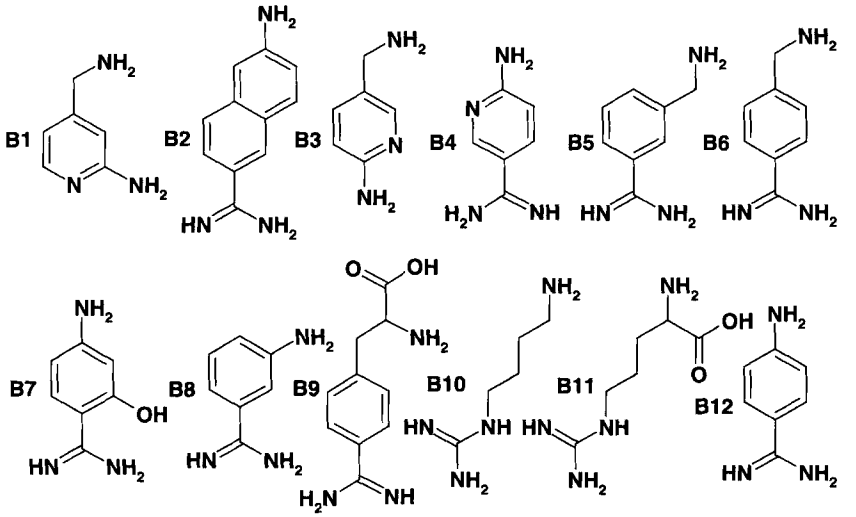


Figure 9.14. Amines used for the generation of the 15360-member library of parallel three component reaction products.

The biological test results, the inhibitory concentrations for thrombin, of the complete compound library are displayed in Figure 9.15. Each rectangle corresponds to one amine **B**, whereas the aldehydes and isonitriles are varied in the vertical and horizontal direction, respectively. Reaction products that exhibit an activity below 30 μM are marked. Thus, Figure 9.15 represents a two-dimensional projection of the three-dimensional SAR.

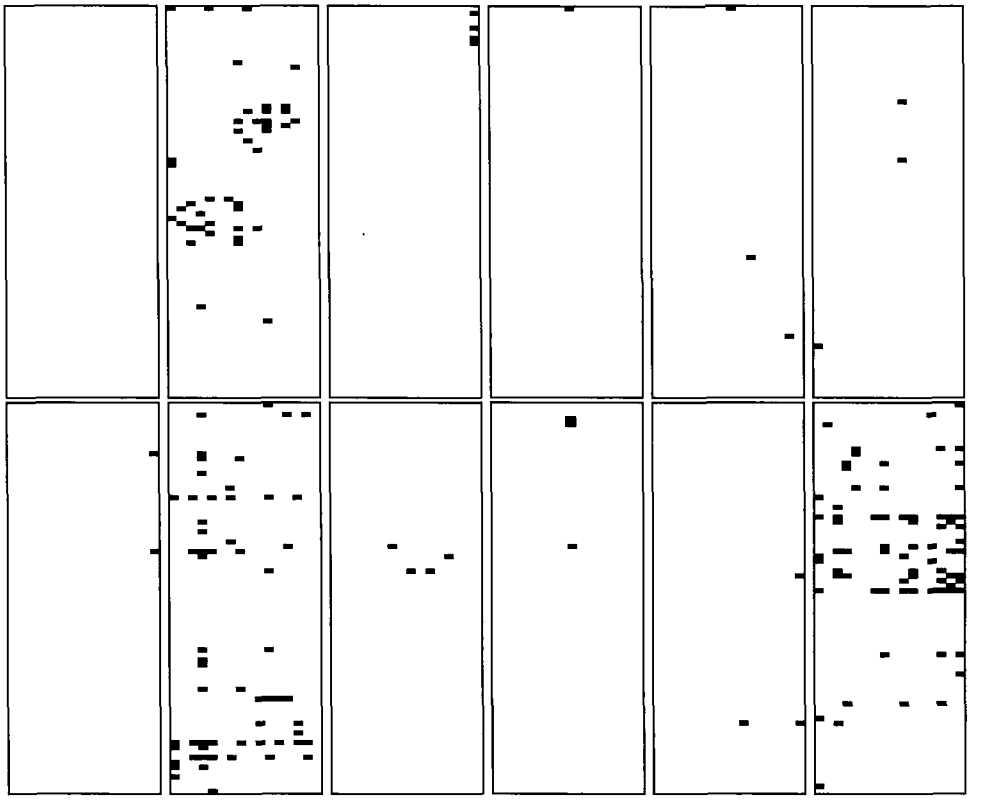


Figure 9.15. Display of the compounds in the 15360-member library exhibiting thrombin inhibitory activities below 30 μM (filled rectangles). Each panel corresponds to one amine from Figure 9.14. Isonitriles are varied in the horizontal, aldehydes in the vertical direction, respectively.

Running a GA on this landscape (using an algorithm as described above with a population size of 20, 100% crossover, and 1% mutation rate) was compared with the random selection of 20 compounds out of the space of this library in each round. Figure 9.16 compares the optimization of each procedure and shows the better efficiency of the GA *versus* random selection. Moreover, one can define a performance criterion that calculates this performance increase (Eq. 9.1):

$$P = (m_{GA}/N_{GA})/(m_{random}/N_{random}) \tag{9.1}$$

where m is the slope of the respective curves and N is the population size.

The "structure" of the search space has been reported to be a large influence on whether or not a genetic algorithm will be successful [4,20]. The structure of SAR landscapes can be described by three parameters: the dynamic range (e.g. the difference between the lowest and highest active compound), the distribution (e.g. how many active compounds in a certain activity interval), and the roughness of the landscape (multimodal or unimodal landscape, clus-

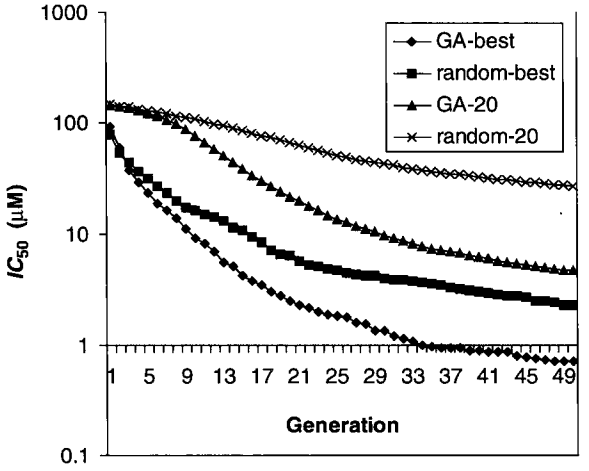


Figure 9.16. The evolution of the 20 best reaction products found with a GA (GA-20) and random selection (random-20), as well as the most active reaction product (GA-best and random-best, respectively) over the generations of 20 compounds per population. The SAR landscape of Figure 9.15 was used.

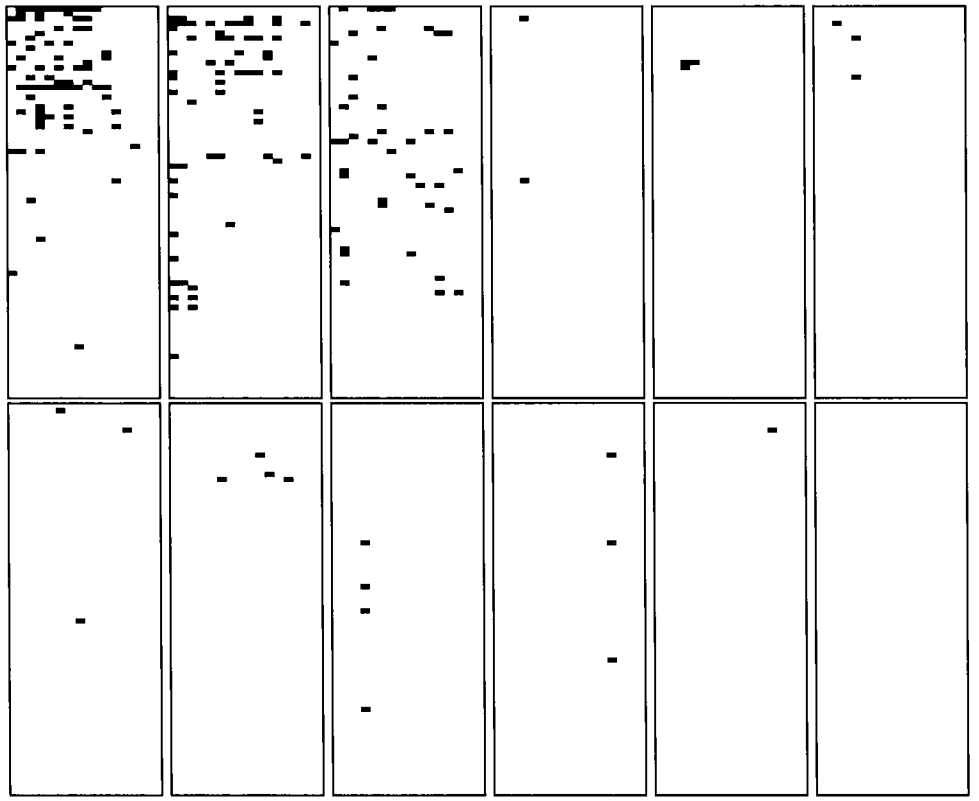


Figure 9.17. Display of the compounds in the 15360-member library exhibiting thrombin inhibitory activities below 30 μM (filled rectangles). Each panel corresponds to one amine from Figure 9.14. Isonitriles are varied in the horizontal, aldehydes in the vertical direction, respectively. The order of the amines, aldehydes and isonitriles from Figure 9.15 was re-sorted to achieve the best possible clustering of actives.

tering of active compounds). Thus, one could believe that a GA would be more efficient if the active compounds were clustered in one region. This was achieved in our test case by re-sorting the starting materials of Figure 9.15 in such a way that most active products are close to each other (Figure 9.17) and running the GA on this re-sorted landscape. Contrary to intuition, the experiment shows that the efficiency of the GA is almost the same as for the non-resorted SAR [17]. We believe that this behavior is advantageous since clustering of products and even the less effective clustering of starting materials based on biological results is not possible in real experiments.

A different distribution of active compounds was investigated by testing the same library against a different serine protease (Figure 9.18) which displays more active compounds below a 10-μM activity limit. The simulated GA on this landscape yielded sub-micromolar compounds faster, as did random selection. This is readily explained by the higher occurrence and hence probability of finding such compounds. However, the better efficiency of the GA-based selection as measured by P is still very similar to that of using the thrombin SAR land-

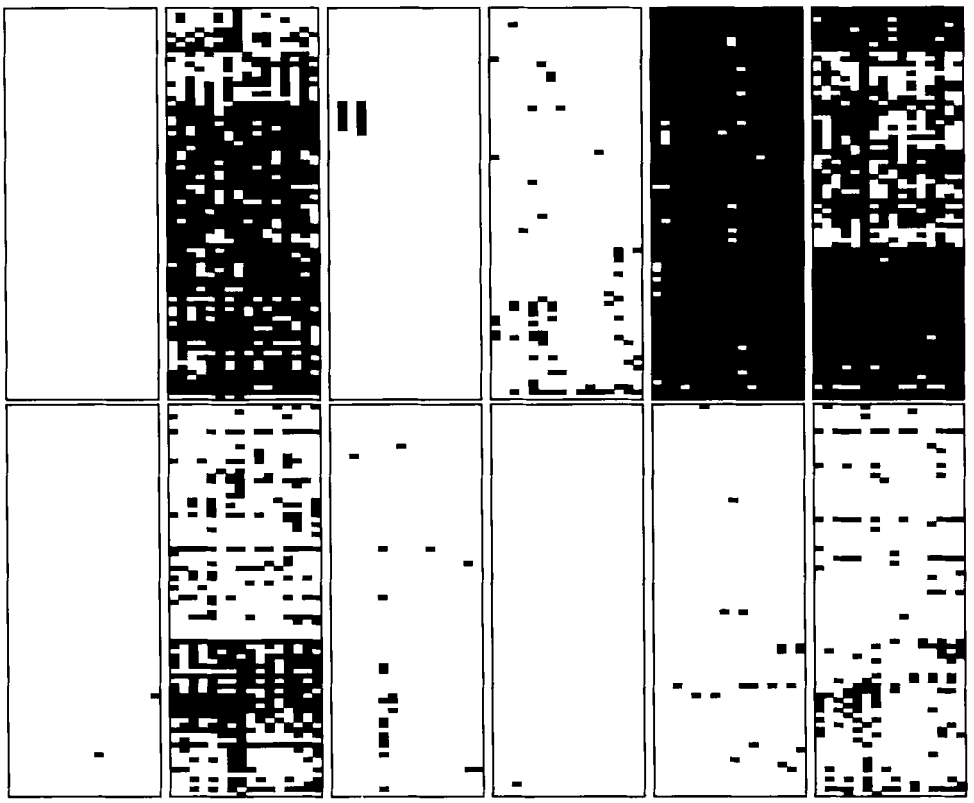


Figure 9.18. Display of the compounds in the 15 360-member library exhibiting protease inhibitory activities below 10 μM (filled rectangles) for a protease homologous to thrombin. Each panel corresponds to one amine from Figure 9.14. Isonitriles are varied in the horizontal, aldehydes in the vertical direction, respectively.

scape. These first results on the structure dependency of GA-driven evolutionary compoun selection methods are encouraging, since it allows the successful application of general G parameter sets to various types of SPR problems. A more general review on suitable par meters for GA-based compound selection has been given in [16,17] (see also Chapter 8 for discussion).

9.7 The Use of GA-Driven Evolutionary Experiments

As shown above, GAs are rather tolerant to experimental errors and may still yield good re sults even if the starting hypothesis is wrong. This behavior is especially true for false nega tive results, since these are simply eliminated in the selection step and are not remembered The elimination of misleading false positive results takes somewhat longer – depending on how often a good chromosome is allowed to replicate. This "fuzzy" and robust optimization property makes GAs especially attractive for real time experimental optimizations as described above. Furthermore, for the application of GA-based evolutionary methods, similar to Nature, one does not need to know or to describe chemical structures explicitly, or to develop a structure-based hypothesis about SARs.

On the other hand, GAs also have serious drawbacks which could render them unattractive for real experiments. First, GAs are stochastic which means that they do not necessarily give the same results in each run. By its very nature, a GA "learns" in a sequential, implicit way over several cycles of selection and testing which could be prohibitive for time-consuming, or expensive synthesis, or biological testing. Therefore, the application of GA-driven compound selection seems appropriate especially if [21,22]:

- the search space is very large and multi-dimensional,
- the sequential synthesis and testing can be performed rapidly, and
- the structure–property relationship is multimodal.

Keeping these criteria in mind, we believe that there are many SPR problems that could be solved with success using GA-based compound selection. Such problems could involve more complicated selection criteria that include not only biological activities but also other properties like selectivity or penetration through biological membranes, which would increase the dimensionality of the SPR landscape. Other examples may include finding entirely new multi-component reactions that yield products exhibiting biological activities simply by varying various starting materials. The area of material sciences where new materials or new catalysts are composed by mixing predefined starting materials or components, and explicit descriptions of SPR are rare, seems amenable to evolutionary selection methods. A genetic algorithm has recently been used to propose new polymer molecules that mimic a given target polymer [23], or exhibit a range of glass transition temperatures and hydrophobicities [24]. Although the number of polymers considered was not very large, this work represents a convincing example of diversity and similarity selection that was verified on the basis of experimental data in the field of materials discovery.

References

[1] W. Fontana, P. Schuster, *Science* **1998** *280*, 1451–1455.
[2] I. Ugi, M. Wochner, E. Fontain, J. Bauer, B. Gruber, R. Karl, in M. A. Johnson, G. M. Maggiora (Eds.) *Concepts and Applications of Molecular Similarity*, John Wiley & Sons Inc., New York **1990**, pp. 239–288.
[3] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, A. Schrijver, *Combinatorial Optimization*, Wiley, London **1997**.
[4] J. H. Holland, *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, Ann Arbor, MI **1975**.
[5] D. E. Clark, *MATCH* **1998** *38*, 85–98.
[6] L. Weber, *Drug Discovery Today* **1998**, *3*, 379–385.
[7] J. Heitkoetter, D. Beasley, *The Hitch-Hiker's Guide to Evolutionary Computation: A list of Frequently Asked Questions* (FAQ) 1999. USENET: comp.ai.genetic. Available via anonymous FTP from: rtfm.mit.edu/pub/usenet/news.answers/ai-faq/genetic/
[8] T. Baeck, D. B. Fogel, Z. Michalewicz (Eds.) Handbook of Evolutionary Computation, IOP Publishing and Oxford University Press, Bristol/New York 1997.
[9] J. H. Holland, *Hidden Order – How Adaptation Builds Complexity*, Addison-Wesley, Reading, MA **1996**.
[10] M. Forrest, M. Mitchell, in *Relative building-block fitness and the building-block hypothesis, in Foundations of Genetic Algorithms 2*, D. Whitley (Ed.), Morgan Kaufmann, San Mateo, CA **1993**, pp. 109–126.
[11] C. Stephens, H. Waelbroeck, *Evolutionary Computation* **1999**, *7*, 109–124.
[12] R. Wehrens, E. Pretsch, L. M. C. Buydens, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 151–157.
[13] Y. Yokobayashi, K. Ikebukuro, S. McNiven, I. Karube, *J. Chem. Soc. Perkin Trans. I* **1996**, 2435–2437.
[14] J. Singh, M. A. Ator, E. P. Jaeger, M. P. Allen, D. A. Whipple, J. E. Soloweij, S. Chowdhary, A. M. Treasurywala, *J. Am. Chem. Soc.* **1996**, *118*, 1669–1676.
[15] L. Weber, S. Wallbaum, C. Broger, K. Gubernator, *Angew. Chem. Int. Ed. Engl.* **1995** *107*, 2453–2454.
[16] K. Illgen, T. Enderle, C. Broger, L. Weber, *Chemistry & Biology*, **2000**, *7*, 433–441.
[17] L. Weber, M. Almstetter, *Diversity in very large libraries, in Molecular Diversity*, R. Lewis (ed.), Kluwer, Amsterdam **1999**.
[18] L. Weber, C. Broger, Hoffmann-La Roche AG, unpublished results.
[19] L. Weber, K. Illgen, M. Almstetter, *SYNLETT* **1999**, *3*, 366–374.
[20] I. Rechenberg, *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*, Frommann-Holzboog, Stuttgart **1973**. 2nd edition **1994**.
[21] G. Schneider, P. Wrede, *Mathematical Res.* **1993**, *81*, 335–346.
[22] G. Schneider, W. Schrödl, G. Wallukat, E. Nissen, G. Rönspeck, J. Müller, P. Wrede, R. Kunze, *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 12179–12184.
[23] V. Venkatasubramanian, K. Chan, J. Caruthers, J., J. Chem. Inf. Comput. Sci. 1995, 35, 188-195.
[24] C. H. Reynolds, *J. Comb. Chem. Sci.* **1999**, *1*, 297–306.

Virtual Screening for Bioactive Molecules
Edited by Hans-Joachim Böhm and Gisbert Schneider

10 Understanding Receptor–Ligand Interactions as a Prerequisite for Virtual Screening

Gerhard Klebe, Ulrich Grädler, Sven Grüneberg, Oliver Krämer, Holger Gohlke

10.1 Introduction

Lead identification is the key element and starting point for any drug discovery project. At present, a dual approach is followed: experimental high throughput screening (HTS) of large compound libraries and, alternatively or as a complement, computational methods for virtual screening and *de novo* design [1–4]. Experimental HTS involves highly sophisticated technologies and advanced engineering know-how. It typically produces a tremendous amount of data on ligand binding, e.g. in the range of several million compounds tested per HTS assay. However, it hardly contributes to our understanding of why and how the detected candidate molecules ("hits") bind to the target. Any increase in knowledge is only produced once molecular modelling comes into play and tries to correlate and compare the obtained hits, which often enough turn out to be quite diverse in chemical structure. Even so, the observed hit rates of about 1% appear quite scarce, and a comparison of several thousand hits in structural terms implies already an elaborate task. As a key prerequisite in contrast, Virtual Screening requires knowledge about the binding criteria responsible for binding to a particular target, e.g. knowledge about the binding-site geometry, the binding modes of several ligands to the protein under investigation, criteria for molecular similarity, ideas about possible mutual ligand superpositions, or putative bioisosteric functional group replacements. Only if these features are sufficiently understood can virtual screening be applied successfully to screen either compound libraries of existing substances, or samples of computer-generated molecules that might be detected as prospective leads and potential candidates for subsequent synthesis.

In the present overview we want to summarize some strategies and actual results for virtual screening applications, based on the given structure of a target protein. We will also describe briefly the software tools used at the different levels of our search strategies.

10.2 The Structure of the Target Protein: Starting Point for Virtual Screening Experiments

For all of the projects described in the following, our starting point was the active site of a given target protein, whose structure had previously been determined by X-ray crystallography to a sufficiently high resolution (≤ 2Å). Usually co-crystallization with one or several already

known inhibitors significantly contributes to a better characterization of the binding-site properties. As experience shows, structural changes due to the flexibility of the binding-site are more pronounced between the uncomplexed (e.g. *apo*) and complexed form than among several complexed structures [5]. In this contribution, we will discuss three case studies. The first two examples, tRNA-guanine transglycosylase and human carbonic anhydrase II, belong to our best knowledge to the former class. The third example, aldose reductase, undergoes substantial changes upon ligand binding. These changes can differ from inhibitor to inhibitor and depend on their structural properties. This example highlights the importance of induced fit effects governed by ligand binding.

10.3 Thermodynamic Parameters Determining Ligand Binding

The starting point for any structure-based virtual screening procedure is the detailed analysis of the binding features defined by the spatial arrangement of the amino acid residues that compose the binding site. Ligand binding is achieved through non-bonded interactions such as hydrogen bonding, or lipophilic aromatic or aliphatic contacts [6]. However, the correlation between structural features and the binding affinity of a ligand toward its protein receptor is complex. Assuming the conditions of equilibrium thermodynamics are given, the affinity can be related to the binding constant. The latter value is a Gibbs free enthalpy composed of enthalpic and entropic contributions. At first glance, it is tempting to establish simple and intuitive correlations such as: the more hydrogen bonds a ligand forms with its receptor, the stronger its affinity will be, or an increasing lipophilic surface portion of a ligand that becomes buried upon complex formation will progressively enhance binding. Although systematically analyzing these parameters unravels their general importance for binding, a simple correlation is not displayed. Binding results between soluted partners, and accordingly the ubiquitously present water molecules play an important role in the binding process. On going from the soluted to the bound state, interactions formed by the functional groups of the binding partners have to rearrange. The energetic parameters determining these interactions have to be considered, but importantly, any breaking or reforming of interactions will be related to changes of the ordering parameters of the entire system. The latter properties are expressed by entropic terms. One of the reasons for the complexity of the resulting correlation is the mutual compensation of enthalpic and entropic contributions. Strong enthalpic interactions usually reduce the internal flexibility of a system, thus resulting in a higher ordering or an entropically less favorable state. In contrast, an enthalpically loosely coupled system allows more degrees of freedom to be activated. In consequence this system corresponds to a state of larger entropy.

Let us imagine a typical binding process as sketched in Figure 10.1. Before complex formation, a ligand moves around in the solvent (presumably in a cage formed by solvent molecules) and several of its conformational degrees of freedom are accessible. The ligand forms interactions with neighboring solvent molecules, usually water molecules. Some of these neighboring waters will be strongly bound, others loosely associated. The macromolecular receptor also tumbles around in water, and its binding site will be filled by water molecules. Some of them will be fairly fixed and positionally well defined (those usually detected by

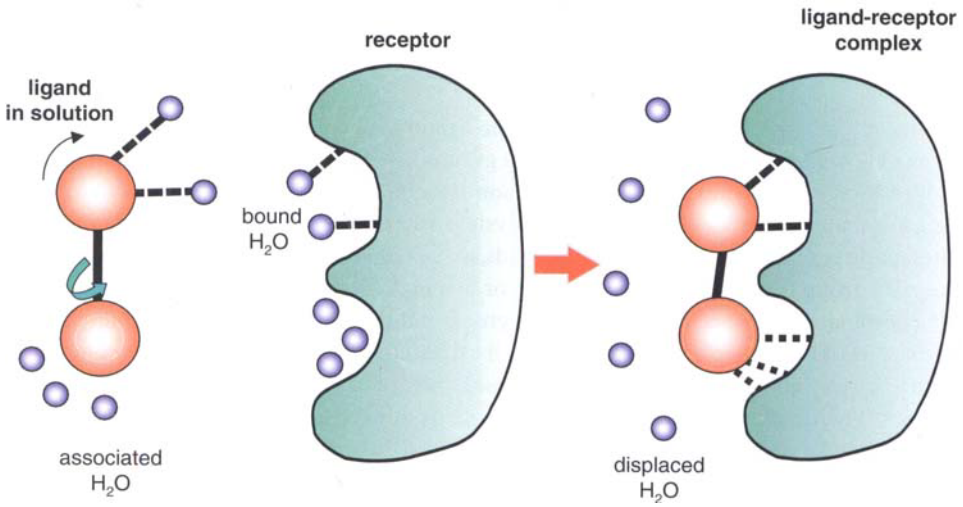


Figure 10.1. The binding of a drug-type ligand toward a macromolecular receptor is determined by enthalpic and entropic factors. In the uncomplexed state, receptor and ligand are solvated and interact with neighboring water molecules. In this situation, several degrees of conformational flexibility are accessible but become immobilized upon complex formation. Water molecules are released from the binding site and the local water structure around the previously solvated ligand is changed. Both effects contribute to the entropic portion of ligand binding.

protein crystallography), others will float around and associate in structurally varying water clusters. Upon complex formation, these waters will be released from the binding site. Some of them contribute to a gain in entropy, others lose in enthalpy. The ligand is released from the solvent. It loses interactions with the neighboring water molecules and gains new interactions at the binding site. Of these, only those stronger at the binding site, as compared to the water environment, will contribute to binding affinity. Many will just be compensated and do not matter in the enthalpic inventory. The ligand is immobilized at the binding site. Compared to the situation in the solvent, this might involve a loss in entropy. On the other hand, the release of the ligand from a water cavity results in a local reorganization of the water structure. Assuming that the water molecules at the interface of lipophilic ligand portions have to be arranged in a more ordered structure, the release of the ligand from the solvent can imply a gain in entropy. Ligand and protein form a complex. Two species, moving separately before binding, assemble and thus lose some degrees of molecular flexibility. This again changes the entropic contributions.

In virtual screening, putative leads are selected by estimating their binding properties to a target protein. The thermodynamic aspects summarized above have to be taken into account if a reliable prediction of binding is to be achieved.

10.4 Spatial Location of Putative Interaction Sites between Ligands and Proteins

In order to obtain a first idea about a possible ligand molecule, information must be provided as to how the exposed protein functional groups favorably interact with ligand functional groups. First insights into the geometry of non-bonded interactions can be derived from the intermolecular packing in small molecule crystal structures [7]. If we are, e.g. interested in the interaction geometries about carboxylic acids, we can analyze all structures containing a carboxylate group being involved as an acceptor in a hydrogen bond. If we store together both the coordinates of the central carboxylate group and those of the hydrogen-bonded neighboring NH or OH groups, and superimpose all extracted fragments in terms of the coordinates of the carboxylate group, we reveal a composite picture of all experimentally observed geometries (Figure 10.2). This composite picture maps out the possible interaction geometries found around a carboxylate group. It helps to estimate the spatial area in a binding-site where, e.g. an aspartate or glutamate can favorably interact with a putative ligand. Similar analyses can be performed for all functional groups of interest. Finally, they can be used to derive a set of rules to generate putative interaction sites around the exposed amino acid residues in a binding site. A valuable collection of such non-bonded contact geometries about a large collection of different functional groups has been compiled in the database IsoStar [8].

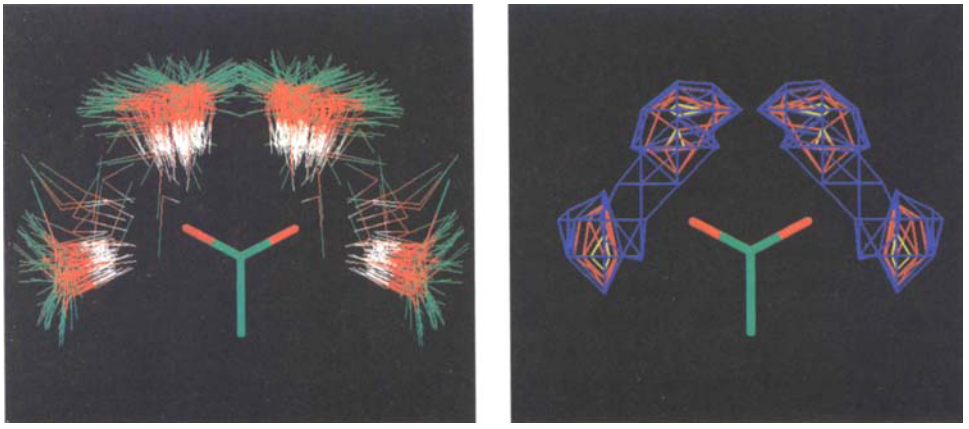
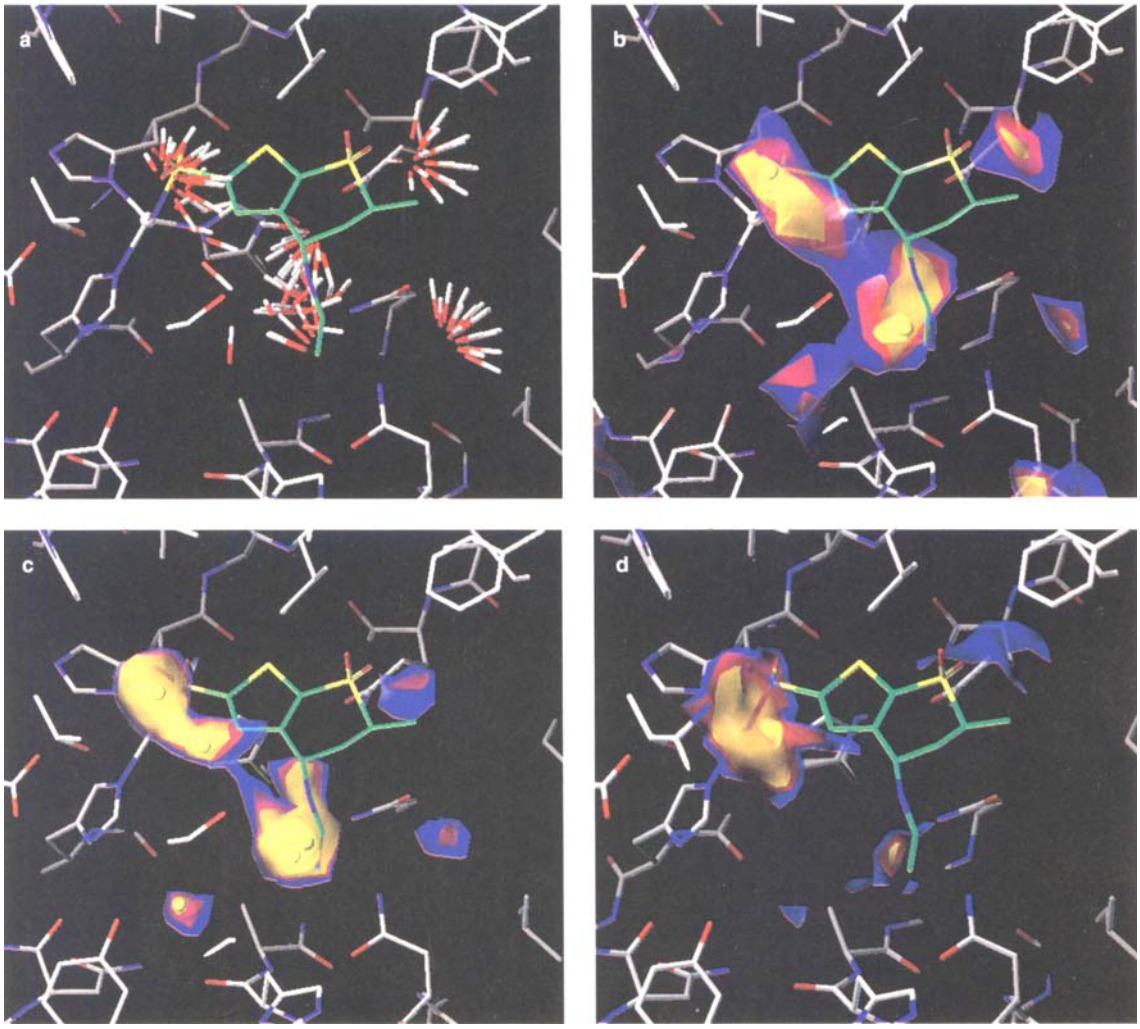


Figure 10.2. Composite picture of contact geometries of hydrogen-bond donors with a carboxylate group to map the spatial orientation of such interactions. Data were retrieved from small molecule crystal structures. The coordinates of the probe functional group were extracted, together with those of the contacting groups (OH). A common orientation of all considered examples was achieved by a least-squares superposition of the atoms of the carboxylate group
left: individual geometries superimposed by showing all compiled contact groups (C: green, O: red, H: white);
right: translation into a contoured propensity distribution.

The program LUDI makes use of this information [9]. After identifying the exposed functional groups in a binding site, it generates putative interaction sites in the binding pocket according to the above-mentioned rules. These sites can subsequently be exploited for screening purposes (Figure 10.3a).



Dorzolamide (Trusopt)

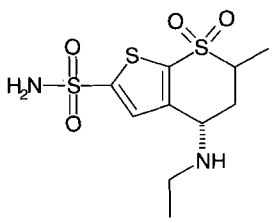


Figure 10.3. Analysis of the binding pocket of human carbonic anhydrase II: **a** putative interaction sites generated by LUDI, or by probing the pocket with, e.g. an $O_{carbonyl}$ atom for the most favorable binding regions (contoured in blue–red–yellow at three consecutive levels, and in b and c centers indicated by balls), using **b** GRID (based on a force-field), **c** SuperStar (based on experimentally evidenced propensities), and **d** DrugScore (based on knowledge-based contact preferences). Superimposed, the crystallographically determined binding mode of dorzolamide, a potent carbonic anhydrase inhibitor, is given.

Another strategy for learning about putative binding sites is provided by the program GRID [10]. GRID is based on a sophisticated force-field approach. Different probe atoms or functional groups have been parameterized to analyze a particular binding pocket for local minima in terms of interaction potentials. A regularly spaced grid is generated inside the binding pocket. Subsequently, for each intersection of the grid, the potential energy between a probe and the surrounding protein is computed according to the implemented parametrization and functional form. Various types of probes can be analyzed. Finally, the potential energy values obtained are contoured for the different probes, to reveal the energetically most favorable regions in the binding pocket. They can be used as a guideline to define the requirements for a putative ligand (Figure 10.3b).

A similar analysis of hot-spots for binding can be performed again using the information of composite crystal-field environments stored in IsoStar. The program SuperStar automatically detects the amino acids exposed to the binding pocket [11]. It selects and superimposes the most appropriate distributions from IsoStar. Subsequently, it combines and scales them in order to transform them into propensities. Finally, for different contacting groups such as OH, NH, $O_{carbonyl}$, $C_{aromatics}$, or $C_{aliphatic}$, the spatial propensity distribution can be contoured. Again, very similar to GRID, the local minima of such maps can be used to define the binding requirements for possible ligands (Figure 10.3c).

Statistics on the occurrence frequency of non-bonded contacts between ligand– and protein–functional groups, observed in crystallographically determined protein–ligand complexes provide another way to predict favorable binding sites [12]. A regularly spaced grid is embedded into the binding site. At each intersection the contribution derived from a knowledge-based potential (DrugScore, see below) between a particular ligand atom type and the protein is computed. Subsequently, the assigned grid values are contoured and highlight the regions most favorable for particular ligand functional groups (Figure 10.3d). For the binding site of human carbonic anhydrase II in Figure 10.3, a "hot-spot" analysis is shown using LUDI-generated sites (**a**), and contour maps generated by GRID (**b**), SuperStar (**c**), and DrugScore (**d**).

10.5 Using Putative Interaction Sites for the Placement of Possible Ligands

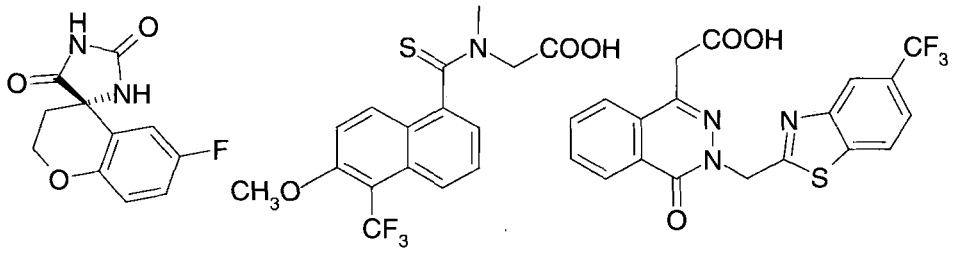
LUDI automatically uses the generated interaction sites for further analysis. It operates on a database assembled from small and fairly rigid molecules. In this database, the stored molecules are classified in terms of their hydrogen-bonding properties and their lipophilic functional groups. LUDI then tries to fit each database entry into the binding site. Beside a good superposition with the previously generated interaction sites, a minimal clash with the protein and a good affinity score (see below) are important criteria for the selection of an entry as putative hit. To our experience, the best results are obtained with LUDI using a sample of rather small and rigid fragments compiled from data in the Available Chemicals Directory (ACD) [13]. The retrieved hits can be purchased and directly tested or co-crystallized with the protein of interest.

10.6 Consecutive Hierarchical Filtering as a Strategy for Virtual Screening of Larger Ligands

The search for larger ligands as putative leads containing several rotatable bonds involves a more elaborate screening strategy. It is usually performed in a step-wise fashion using a series of filters and rejection criteria of increasing complexity. As a first step, a fast database search engine is applied, e.g. UNITY developed by Tripos [14]. UNITY selects molecules according to their 2-D connectivity and/or predefined 3-D pharmacophore requirements. Molecules can be screened for likely conformations using a very fast tweak-search algorithm [15]. A sophisticated pharmacophore model can be derived from the binding mode and actual chemical composition of known protein ligands, and/or from the properties of the surrounding binding-site residues. The latter information is translated into the pharmacophore model *via* the above-described LUDI-, GRID-, SuperStar-, or DrugScore-associated descriptors. However, a prerequisite to reliably exploit the latter information is a rather rigid protein-binding site. It has virtually to remain unchanged upon ligand binding. In cases where ligand binding involves pronounced effects resulting from induced fits of the binding pocket, or where the actual binding modes of a considerable number of structurally diverse ligands is known, this information about the ligands should be incorporated into the definition of the pharmacophore.

A generally applicable procedure to generate a pharmacophore from several superimposed ligands involves the following steps. First of all, each of the superimposed molecules is described by a set of atom-based Gaussian functions associated with a vector describing the steric, electrostatic, hydrophobic, and hydrogen-bonding properties of each molecule [16–18]. All considered molecules are embedded into a regularly-spaced grid. Subsequently, at each grid point the contributions of the Gaussians of all molecules are summed, either forming the arithmetic or geometric mean. Finally, the resulting values at the different grid points are again translated into a set of associated Gaussian functions [19]. The field-based pharmacophore thus derived exhibits the superimposed ligands in terms of associated Gaussian functions and highlights the properties that the molecules share in common. The two alternatives for calculating the mean resemble either the properties that all molecules allot similarly in the same spatial area (AND-pharmacophore), or the properties that are represented by at least one molecule of the superimposed set at a particular region in space (OR-pharmacophore). Figure 10.4 shows the field-based pharmacophore derived from three superimposed ligands binding to the enzyme aldose reductase. The features expressed in such a field-based pharmacophore can either be translated into a query for UNITY, or can be directly used for a database search based on molecular similarity. The latter search compares molecules in terms of their associated Gaussian-type descriptors. The programs SEAL [16–18] or FlexS [20] operate on this information.

After the initial filtering based on UNITY, further data reduction is following using either molecular superpositioning or docking techniques. For the superposition, we apply two alternative methods, SEAL and FlexS, that are based on the same similarity concepts [16–18, 20]. However, they follow different strategies. Both approaches compare a test and a reference molecule in terms of associated Gaussian functions. These functions can either be localized at the atomic positions in a molecule or, to reduce computational effort, they can be at-



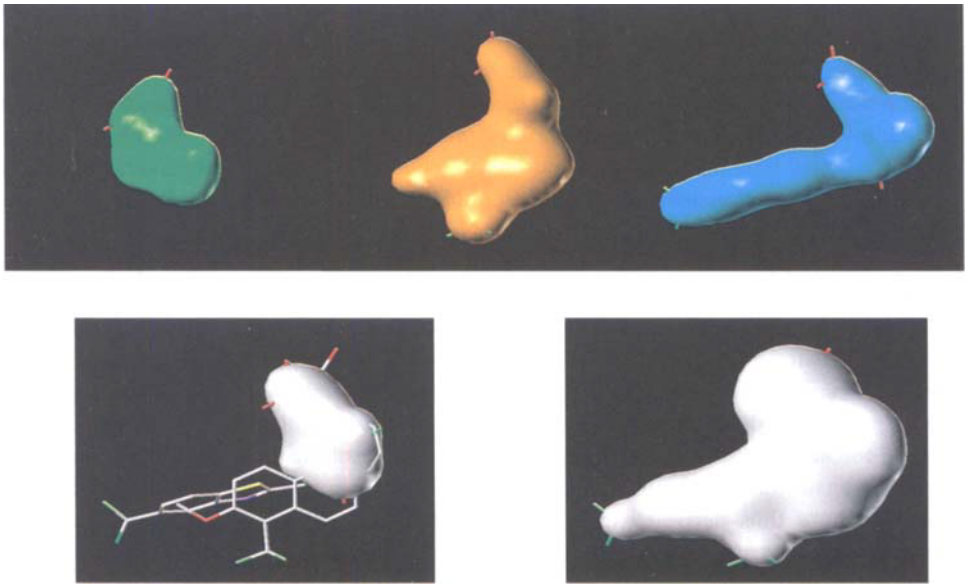


Figure 10.4. Field-based pharmacophores derived from superpositioning the three potent inhibitors of aldose reductase: sorbinil, tolrestat, and zopolrestat. The shape of the inhibitors is approximated by a set of Gaussian functions associated with a vector considering steric, electrostatic, hydrophobic, and hydrogen-bond donor and acceptor properties. As an example, the steric properties of the three inhibitors are shown. This property of the three molecules is combined to represent either those areas in space where all molecules allot the same property (AND-pharmacophore, bottom left) or where at least one of the molecules displays this particular property (OR-pharmacophore, bottom right).

tributed to entire fragments. They represent the shape of the molecules in space. In addition, they are loaded with a vector describing five different physicochemical properties (steric, electrostatic, hydrophobic, and hydrogen-bond donor or acceptor features). The mutual similarity is measured by spatially superimposing the molecules and computing the scalar product of the property vectors, weighted by the integral of the overlapping Gaussian functions.

Our extended and modified version of SEAL [17] first performs a superposition of rigid multiple conformers of a test molecule onto a reference molecule. The multiple conformers are generated by MIMUMBA [21], a knowledge-based approach, to produce well distributed conformers in that part of conformational space which is relevant for ligands in protein-bound situations. In a final step, the best ranked rigidly superimposed test molecules are conformationally relaxed by optimizing the similarity condition with the reference. Simultaneously their conformations are kept as close as possible to the local minima obtained from the MIMUMBA approach [21]. FlexS follows a reverse strategy [20]. It first decomposes the test ligand into several rigid portions. After selecting a base fragment, the program as an incipient step superimposes this fragment with the most similar part in the reference molecule. Two procedures are available. Either a superimposition in terms of Gaussian functions is performed in Fourier space, or patterns of putative interaction sites, generated similarly to the LUDI approach about the functional groups in the test and reference ligand, are mutually matched. After the base fragment has been placed, the additional fragments are added in a step-wise fashion. The local conformations about the bonds to be connected are selected following the same rules as applied in MIMUMBA. The incremental build-up procedure follows a tree-type structure. At each step, the volume overlap with the reference is checked and a similarity ranking in terms of the above-described Gaussian functions is performed to score the achieved similarity in space.

The program suggests several solutions ranked by similarity scoring. In our virtual screening assays we focus on the best solutions for further analysis. The data sample thus reduced from SEAL or FlexS is either directly analyzed by visual inspection, or submitted to further filtering by docking to the active site of the protein under consideration. This final step is computationally the most demanding one. In our studies we apply the program FlexX [22]. Similarly to FlexS, FlexX decomposes the ligand to be docked into several rigid portions. It then places a thoroughly selected base fragment into the binding site. In this step, similarly to the concepts in LUDI, triangular patterns of pre-generated interaction sites in the binding pocket are matched with appropriate functional groups in the base fragment. Subsequently, the entire ligand is incrementally constructed in the binding site. The hook-up of additional fragments follows the rules on conformational preferences originally implemented into MIMUMBA. At each step, clashes with the protein are avoided and the binding geometry obtained is analyzed in terms of its expected binding affinity. This ranking is based on an empirical regression-based scoring function. As in FlexS, a tree-type build-up procedure is followed, leading to multiple docking solutions. Again, in virtual screening usually only the best-ranked solution is considered for further analysis. Accordingly, the reliability of the applied scoring function is essential for the entire approach. Normally, as the final step of a virtual screening procedure, the computationally selected hits are visually inspected. However, this step requires a substantial reduction of the data entries, since more than approximately 200 hits are hardly manageable. Docking with FlexX requires some minutes on a presently available single processor workstation, and accordingly between 500 and 2000 entries can be processed routinely. The SEAL and in particular the FlexS approaches are faster, and data samples of about 5000 candidate molecules can be handled. The filtering on the first incipient UNITY-based level is critical since it achieves the most pronounced data reduction. Too stringent search conditions might discard potential hits. In contrast, too soft restrictions will overwhelm the subsequent and more sophisticated search techniques with a rather inade-

quate pre-selection. Since these initial searches can be performed quite fast, a problem-oriented tailoring to select the best-adapted conditions can be easily achieved.

10.7 Of Ultimate Importance: A Discriminative and Reliable Scoring Function

Crucial in all virtual computer screening experiments is the relative ranking of the suggested hits. In docking applications [22–24], used as the final step in the described strategy, the binding affinity has to be predicted correctly [25,26]. As mentioned above, this is a Gibbs free energy composed of enthalpic and entropic terms. It has to be remembered that only differences in the inventory matter between the commonly bound state and a situation where all interacting partners are individually solvated [27].

Most satisfactorily, such a required scoring function is developed from first principles, resulting in a master equation thus considering per se all contributing effects [28,29]. Although being clearly the most satisfactory approach, up to now no method has been reported that is reliable enough and computationally affordable. Scoring functions that are based on a regression analysis of experimentally determined binding affinities and crystallographically resolved protein–ligand complexes are explicitly incorporated into the above-mentioned docking tools LUDI and FlexX [30,31]. These approaches try to partition the free enthalpy of binding into several terms following physical concepts. Accordingly, they provide some insight into the fundamentals of the binding process. They are fast and achieve a precision in reproducing the binding constant of about 1.5 orders of magnitude if binding geometries, in agreement with experiment, are ranked. However, as with any regression analysis, they suffer from the fact that their conclusions are only as precise and generally valid as the data used are relevant and complete when considering all contributing and discriminating effects important in protein–ligand complexes.

The binding geometries, used for ranking by such a scoring function, are most critical. As long as the evaluated geometries are close to the experimentally determined crystal structures, the regression-based scoring function produces reliable predictions [30,31]. However, as experience shows, docking programs also produce geometries strongly deviating from the experimentally observed binding mode. Frequently, scoring functions rank these obviously artificial solutions equally well or even better than the experimental situation. This clearly indicates a weakness of the regression-based scoring functions. Apparently, their discriminative power to rank best ligand poses closely approximating the observed binding geometries is not sufficient.

Meanwhile three approaches, developed in parallel, to derive alternative scoring functions based on the ideas of a so-called "inverse Boltzmann" law have been described [32–34]. Only those binding modes suggested by a docking algorithm are assumed as favorable, that agree with the maxima of the distributions of occurrence frequencies among interatomic contacts between particular atom pairs in experimentally determined structures. Such a scoring function is believed to rank as best all the ligand poses closely approximating the experimentally determined structure. The derived statistical preferences for typical protein–ligand contacts are supposed to reflect implicitly the favorable interaction patterns between functional

groups, according to their probability of occurring in protein–ligand complexes. Any binding feature not in agreement with the most frequently observed contact preferences is penalized by collecting potential contributions described as less likely and therefore less favorable. In our approach DrugScore, we sample contact distances between ligand and protein atom types up to 6 Å using the database system ReLiBase [35]. Subsequently, the occurrence frequencies have been translated into statistical potentials. These distance-dependent pair potentials have been calibrated to the total distance distribution, considering all atom types. Significant deviations to shorter contacts from the mean all-atom distribution are observed for hydrogen-bonding groups whereas contacting groups of a hydrophobic nature show reduced frequency, and accordingly unfavorable potentials at short distances. Furthermore, we have incorporated for each atom type a solvent-accessible surface-dependent potential considering ligand- and protein-to-solvent interactions. This potential punishes the exposure of hydrophobic groups to the solvent, or polar functional groups to nonpolar counter parts. In contrast, it favors mutual contacts between polar groups or tolerates unchanged solvation of polar ligand functional groups carried over from the solvated to the bound state.

The developed scoring function is fast to compute and demonstrates a significantly improved discriminative power to render prominent binding modes closely approximating the experimentally determined structure [34]. It thus can be used to discard computer-generated artifacts. This latter aspect is of utmost importance for virtual screening since, due to the immense data flow produced for searches in large database, a subsequent detailed investigation of only the best scoring hits will be feasible.

Besides an improved discrimination of different binding geometries, knowledge-based scoring functions have been successfully applied to predict absolute affinities [12,32,36]. For this purpose the scores assembled from the different potential contributions are scaled and directly related to the experimentally observed affinities. As the comparison with several data sets shows, affinity predictions with a precision of 1–1.5 orders of magnitude can be achieved.

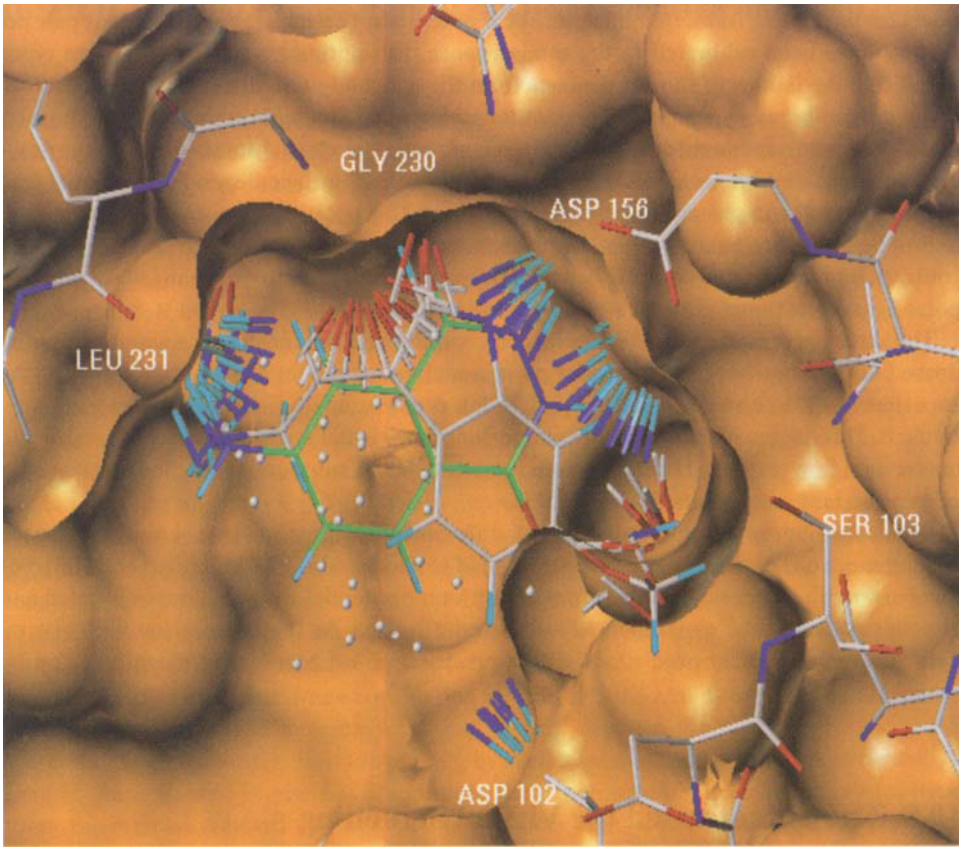
10.8 The Targets used for Virtual Screening

Three different targets have been used in our virtual screening assays. According to the given knowledge base, the previously collected experience, and the reported structural properties of the three targets, different search strategies have been applied.

10.8.1 First Leads for tRNA-Guanin-Transglycosylase by Searches with LUDI

The first target, tRNA-guanine transglycosylase (TGT) catalyzes the initial step of the post-transcriptional modification in the anticodon loop of cognate tRNA, resulting in an exchange of guanine-34 at the wobble position by the queuine precursor 7-aminomethyl-7-deazaguanine (preQ1) [37,38]. The exact biological role of queuine in tRNA is not yet fully understood, however it seems to fine-tune protein biosynthesis in eubacteria [39]. In Shigellae,

which cause dysentery and affect lethally some 500000 infants per year, the enzymatic activity of TGT is a prerequisite for pathogenicity [40]. A selective inhibitor for this enzyme could provide a new therapeutic principle against Shigellosis. The crystal structure of TGT from



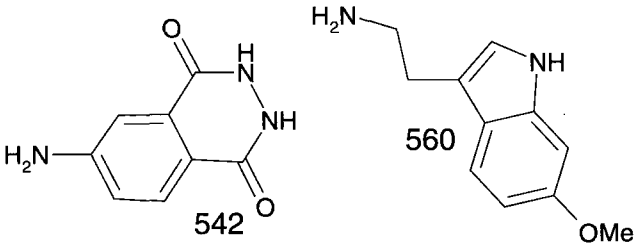


Figure 10.5. Active site of *Z. mobilis* TGT, together with the pre-calculated interaction sites for hydrogen-bond donor, acceptor, and lipophilic properties, generated by LUDI. Two possible lead structures, retrieved by LUDI from the ACD, are superimposed onto the interaction sites.

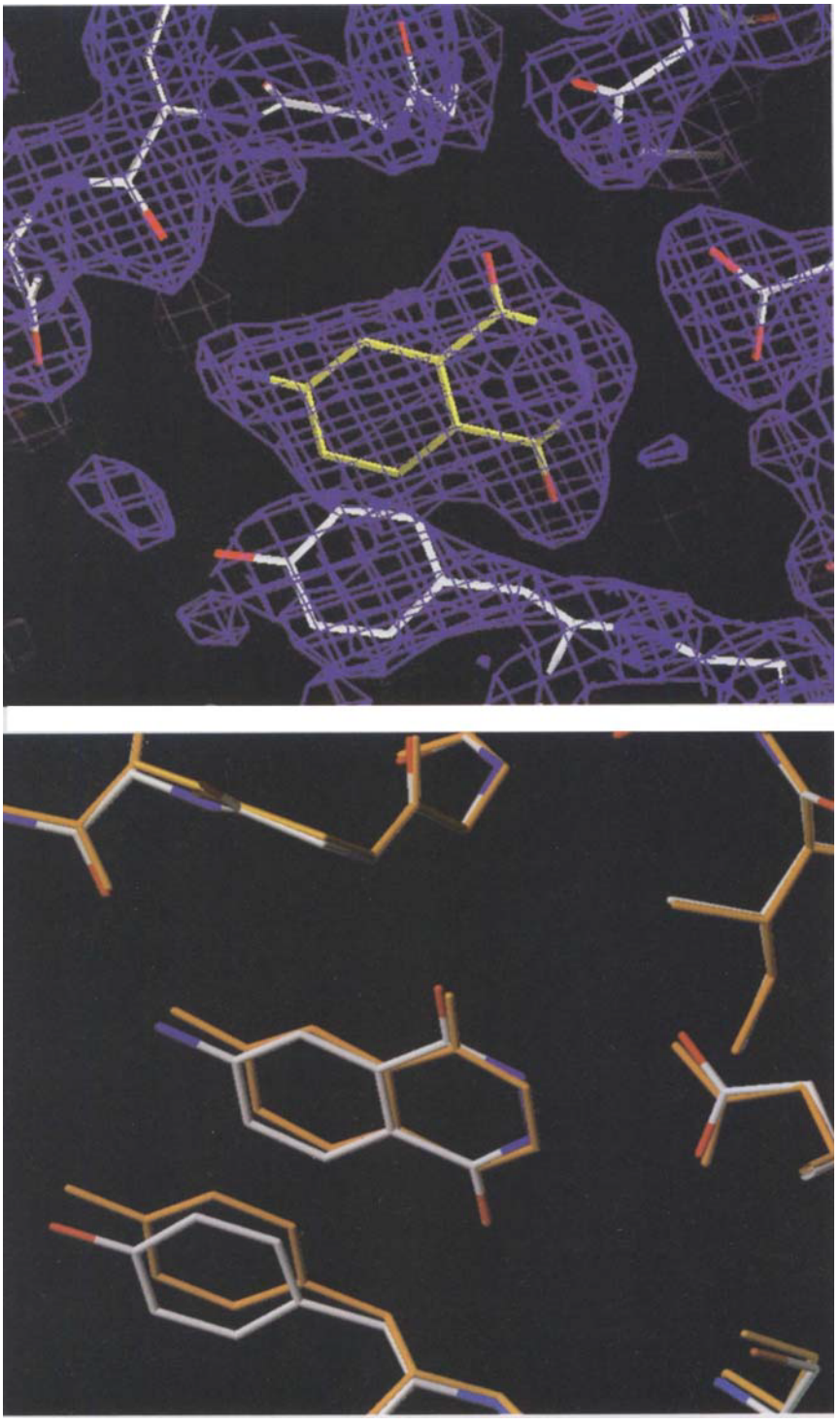


Figure 10.6. Crystal structure at 1.95 Å resolution of 4-aminophthalhydrazide soaked into crystals of *Z. mobilis* TGT. This compound has been retrieved by LUDI. A comparison of the suggested binding mode (orange) with the subsequently determined crystal structure (atom-type coded) shows excellent agreement.

Zymomonas mobilis (besides one residue in the active site identical with the TGT *Shigella flexneri* enzyme) has been solved at 1.85Å resolution. Furthermore, soaking TGT with $preQ_1$ and other substrate analogue inhibitors indicated a specific binding pocket at the C-terminal face of the overall $(\beta/\alpha)_8$-barrel fold of the enzyme.

Detailed analysis of the binding pocket revealed a polar recognition site providing directional interactions with the ligands and an extended region yet unoccupied by the known substrate analogue inhibitors and filled by four water molecules. The given binding-site conditions appeared ideally suited for an initial search with LUDI for putative lead structures (Figure 10.5). To obtain a rough estimate for the scoring range of possible ligands, we computed with LUDI the relative affinity prediction for a substrate analogue inhibitor. The obtained score of 700 corresponds to an experimentally determined inhibition constant of 0.5μM for the *E. coli* enzyme (with identical binding site composition). In Figure 10.6 the active site of TGT is shown together with the pre-calculated interaction sites. Two putative hits retrieved from the ACD [13] were fitted onto these sites. These placements obtained an affinity scoring of 542 and 560, respectively. Subsequently, the inhibitory power of the suggested leads has been studied experimentally. In Figure 10.6 the crystallographically determined binding mode of 4-aminophthalhydrazide is given. As a comparison shows, the LUDI-predicted binding mode is in full agreement with the subsequently determined crystal structure. This molecule meanwhile serves as a first lead for structurally extended ligands that exploit unoccupied space in the binding pocket.

10.8.2 Commercially Available Candidates with Carbonic Anhydrase Inhibitory Potency Discovered by a Structure-Based Pharmacophore Hypothesis

Glaucoma is a general eye disease, in which intraocular pressure rises [41]. In open-angle glaucoma, the inner eye fluid drains too slowly from the front chamber of the eye. As a secondary event, it can also affect the optic nerve causing slow progressive loss of vision. It finally leads, if left untreated, to blindness. Inhibition of human Carbonic Anhydrase II (HCA II) in the eye lowers the intraocular pressure by decreasing aqueous production. The crystal structure of the key enzyme has been solved and it was used for elaborate design studies [42–44]. Accordingly our knowledge base in terms of structural data of this enzyme is significant. Furthermore, to our present knowledge the enzyme remains fairly rigid and unchanged upon ligand binding [42]. It consists of a single polypeptide chain of 260 amino acid residues and a zinc at the bottom of a cone-shaped amphiphilic catalytic center (15Å deep, 15Å wide). The zinc ion is chelated by three histidine residues and in the ligand-free form a water molecule occupies the fourth coordination site. Inhibitors bind at or near the metal atom forming, in addition, hydrogen bonds to Thr199. Nearly all known leads are exclusively heterocyclic sulfonamides, although recently some hydroxamates have been discovered (Figure 10.7).

In light of the broad structural knowledge about HCA II, we embarked on a virtual screening study built on a structure-based pharmacophore hypothesis. To obtain a first insight into the essential requirements for binding, we consulted the different maps produced by LUDI, GRID, SuperStar, and DrugScore (Figure 10.3). The combined information from these programs was translated into a 3-D pharmacophore model ready to be exploited in a

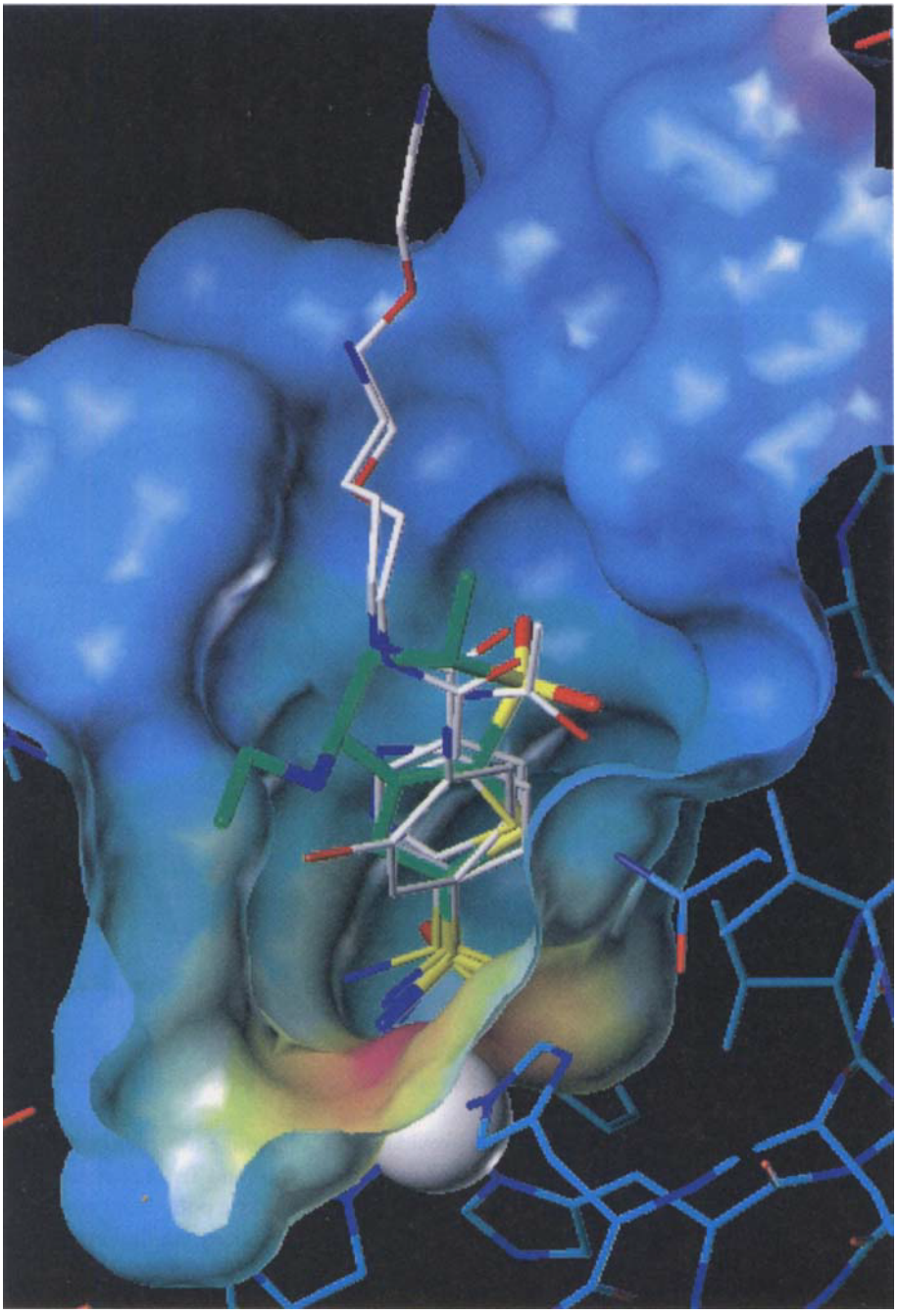


Figure 10.7. Binding site of human carbonic anhydrase II together with several co-crystallized inhibitors. The reference compound dorzolamide (green) is highlighted in bold, and the catalytic zinc displayed as a white ball. Local charges have been mapped onto the solvent-accessible surface (red: positive, blue: negative).

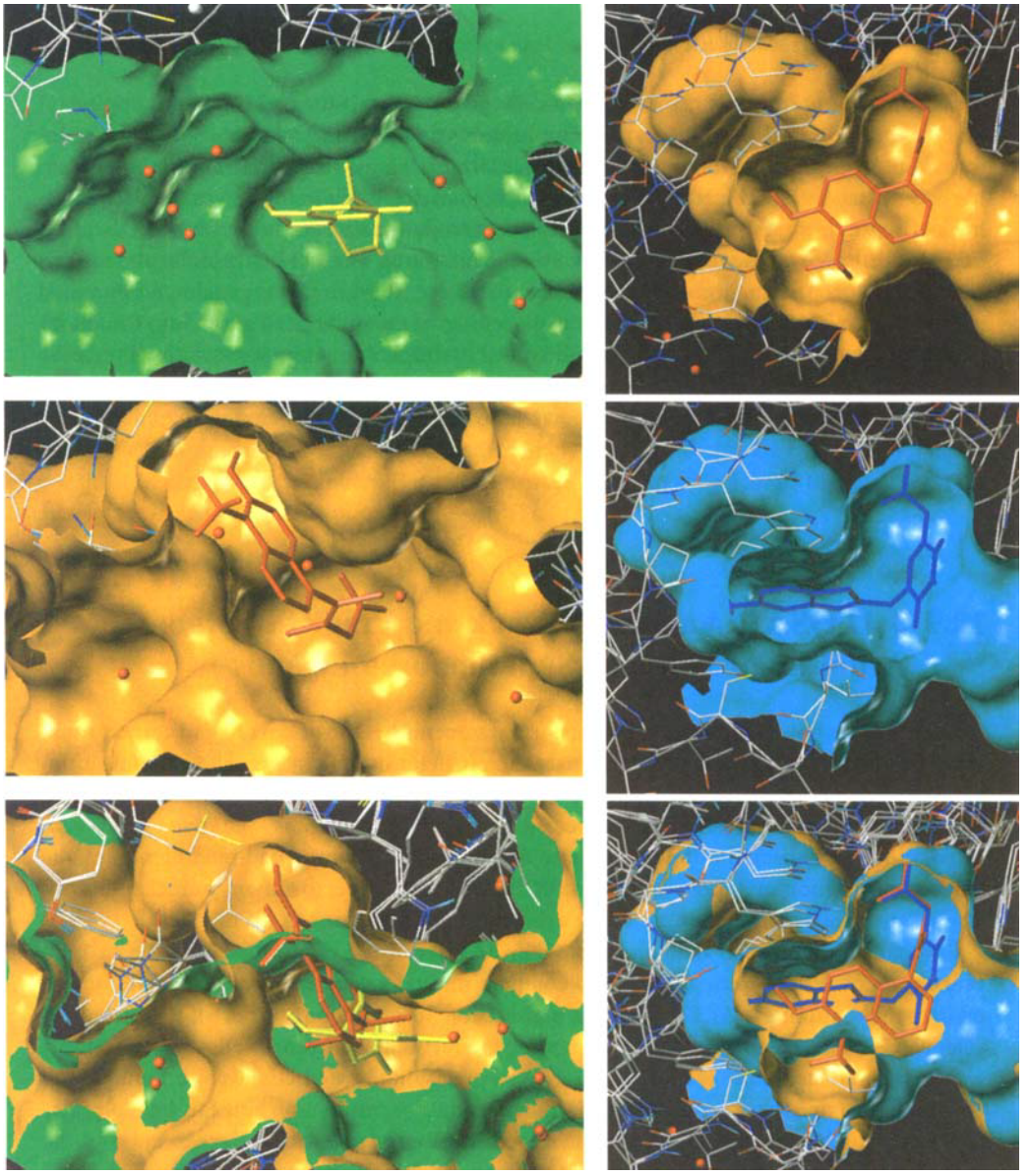
3-D query in UNITY [14]. Since all presently known potent HCA inhibitors bind to zinc, a 2-D connectivity-based pre-selection of candidate molecules containing at least one zinc-anchoring group such as carboxylate, amide, phosphonamide, sulfonamide, or hydroxamate was performed. All searches were accomplished with the Maybridge and Lead Quest (Tripos Inc.) database [45] containing about 60 000 compounds. This data sample has been complemented by 35 known HCA II inhibitors, to control and monitor the performance of our search at all steps.

Subsequent to the data reduction using 2-D and 3-D queries with UNITY, we submitted the remaining set of compounds to FlexS. The well established and potent inhibitor dorzolamide [46] was used as reference ligand in FlexS. The hundred best scoring hits from FlexS were submitted to a docking analysis with FlexX. Finally, a visual inspection of the FlexX-generated binding modes was achieved for the best ranked solutions before purchasing and experimentally testing the most promising candidates from our virtual screening. Interestingly enough, among a small set of prospective hits from the search, several were in the milli-up to micromolar range, and three in the nanomolar range. The best compounds discovered turned out to possess sub-nanomolar inhibitory potency.

10.8.3 Virtual Screening with Aldose Reductase, an Enzyme Performing Pronounced Induced Fit upon Ligand Binding

Diabetes mellitus is presently a major health problem, in particular with respect to an increasingly ageing population [47]. Most treatments are based on insulin administration, but they are far from solving the issue. Through the search for alternative treatments, the enhanced flux of glucose via the polyol pathway has been causally linked to diabetic complications [48]. The rate limiting step along this pathway is the reduction of glucose to sorbitol by Aldose Reductase (AR). Clinical studies showed that compounds inhibiting AR are bene-ficial to diabetic complications. AR is capable of reducing a broad range of structurally diverse aliphatic and aromatic aldehydes. This remarkable substrate promiscuity is presumably achieved via a pronounced induced fit adapting a hydrophobic area of the flexible binding pocket to the actual size and shape of the substrates. Likewise, this capability to adapt its binding pocket has been crystallographically observed in several AR-inhibitor complexes. Figure 10.8 shows the binding modes of sorbinil, tolrestat and zopolrestat with the enzyme [19,49]. Whereas the binding pocket remains virtually unchanged next to the catalytic center, the part of the pocket accommodating the hydrophobic portion of the ligands undergoes dramatic structural changes.

Due to these induced fit adaptations, AR provides a real challenge to structure-based virtual screening. The catalytic center is composed of Tyr48, His110 and Trp111. In most of the known inhibitors this so-called anion binding pocket is occupied by a polar head group of the ligand, e.g. a carboxylate, hydantoin, or formamide group. Accordingly, we limited our initial search in the Maybridge database [45] to carbonic acids, acid amides, formylamines, hydantoins and tetrazoles [19]. Nearly 3800 compounds passed these search criteria. Subsequent to the 2-D connectivity search with UNITY, we performed a 3-D search with the same tool based on a predefined pharmacophore hypothesis. However, in contrast to the HCA case study, we now derived the pharmacophore requirements from the OR field-based pharma-



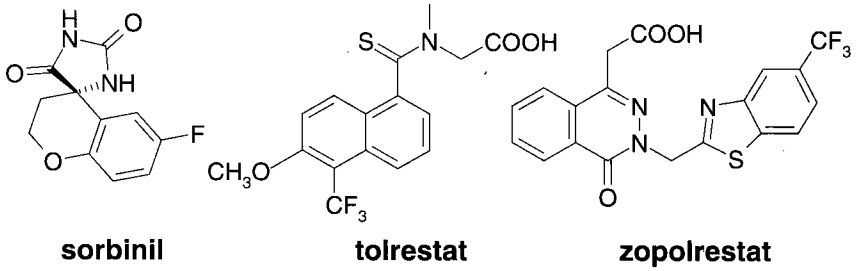


Figure 10.8. The enzyme aldose reductase performs significant adaptations of the binding site upon ligand binding. **On the left**, the binding of sorbinil (yellow) and tolrestat (red) are shown, and below, the two structures are superimposed. The latter inhibitor, tolrestat, opens a binding pocket that is totally closed in the structure with sorbinil. **On the right**, the binding geometry of tolrestat (red, in a different orientation compared to the left) is compared with that of zopolrestat (blue). Zopolrestat opens another, even larger binding pocket in the enzyme.

cophore resulting from the crystallographically given superposition of tolrestat, zopolrestat and sorbinil (Figure 10.4). Due to the complications arising from the pronounced ligand-induced binding-site adaptations, we favored this latter strategy. The accordingly defined flexible 3-D search query in UNITY reduced the original data sample to 222 entries. In the following, a mutual superposition with SEAL has been performed on the combined OR-pharmacophore, based on the three above-mentioned inhibitors. The candidate molecules ranked best with respect to the normalized SEAL similarity score were further evaluated by visual inspection. Considering criteria such as the placement of the polar head group into the anion-binding pocket or the filling of one of the known hydrophobic sub-pockets with a lipophilic ligand portion, a set of compounds was selected for purchase and biological testing. Out of the compounds thus tested, one example actually showed AR inhibition in the micromolar range [50].

10.9 3 D QSAR Analysis to Rank and Predict Binding Affinities of Mutually Superimposed Ligands

As mentioned above, the ultimate goal in virtual screening is the discovery and correct ranking of candidate molecules from a large data sample in terms of their predicted binding affinity. As selection criteria on an intermediate or final stage, we have described docking and molecular superposition methods. Whereas the former approach ranks putative ligands with respect to their expected affinities, directly derived from the computed binding modes, the latter superposition methods can only score ligands with respect to their mutual similarity. This relative comparison does not allow a direct estimate of absolute affinity data.

Comparative molecular field analyses (CoMFA) have been established over the past ten years as a powerful tool to analyze and compare mutually superimposed molecules and to predict their affinities [51,52]. Usually, a precision to about one order of magnitude is achieved. Novel molecules added to an existing model obtained from a well-selected training set can be predicted quite reliably. The achieved precision for prediction is also found to fall into a range of one to two orders of magnitude, strongly dependent on how far the molecules being predicted deviate in shape and physicochemical properties from the examples of the training set. Comparative molecular field analyses are based on a relative comparison, and accordingly they can only reliably interpolate. They cannot predict beyond the model to anything it has never experienced. This is an important difference to docking methods where the constraints from the surrounding binding pocket still allow predictions for molecules structurally deviating from previously known examples.

The comparative field analysis approach maps gradual changes in the potential interaction properties of the aligned ligands by evaluating the potential energy at regularly spaced grid points. Either Lennard-Jones and Coulomb potentials are determined [51], or – in a more recent approach – we make use of molecular similarity indices (CoMSIA, comparative molecular similarity indices analysis) enumerated via a common probe and considering a distance-dependence in terms of a Gaussian-type functional form [52].

An accordingly developed CoMFA or CoMSIA model can also be used for virtual screening. Once a reliable model has been established and the rule as to how to superimpose new molecules onto the examples from the training set is defined, these novel candidates can be ranked according to their CoMFA or CoMSIA affinity predictions. The above-described superposition methods, either based on SEAL or FlexS, provide the required superposition. We have used this approach to score possible members of a combinatorial library constructed for thermolysin as a target [53]. Initially, we derived a CoMSIA model for this enzyme, based on 61 known inhibitors. Then, different members of a tripeptidylphosphonate library of the sequence Cbz–X^{P}–^{O}Y–Z were matched onto this model. The attempted variations at the P_1, P_1' and P_2' positions are shown in Figure 10.9. This library had actually been synthesized by Campbell *et al.* [54], and affinity data for several members of the library had been reported. Our CoMSIA model reproduces quite convincingly the actually observed trends in affinity for the different library members. It can thus assist the process of selecting the most prospective hits for further consideration.

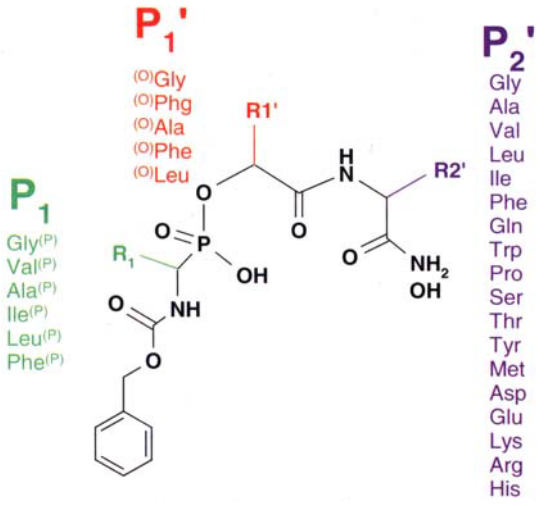


Figure 10.9. Combinatorial library of different peptidylphosphonates Cbz-X^{P}-^{O}Y-Z. The different library members are composed at the positions P_1, P_1', and P_2' by the side chains R_1, R_1', and R_2', schematically indicated in this diagram.

10.10 Summary and Conclusions

In this Chapter we describe different schemes for performing a virtual screening study. A key point is a consecutive hierarchical filtering strategy to decompose a large database into a small sample of prospective hits ready for testing and co-crystallization with the target protein. Three successful case studies, using tRNA-guanine transglycosylase, human carbonic anhydrase, and aldose reductase are described in this paper, to suggest database entries as new lead structures.

In a structure-based virtual screening program, a sophisticated and detailed analysis of the binding site, the real target for the database search, is of utmost importance. In principle, the features exposed toward the binding pocket provide a blue-print of a putative ligand that can bind to this site. Accordingly, the underlying distribution of binding-site features has to be characterized and translated into a pattern convenient for database searches. In addition, already known ligands provide further information important for this approach. This has to be combined with the search criteria given by the binding-site features.

We describe the latter features via interaction sites either generated by a rule-based approach or derived from an analysis with GRID, SuperStar, or DrugScore. Furthermore, we use the knowledge-based contact potentials from the latter approach to map out the binding site and to discriminate and rank computer-generated ligand poses. Usually as incipient step, a crude 2-D connectivity search, followed by a fast pharmacophore-based database screen, is performed in order to reduce the original huge data sample. Subsequently, actual ligand poses are computed by mutual superposition techniques or docking methods. At this stage, a reliable affinity prediction comes into play. The superposition techniques compare ligands in terms of their relative similarity. Accordingly, affinity prediction requires a previously trained model based on a set of reference molecules. Comparative molecular field methods provide such models and can be exploited in due course. In molecular docking, affinity prediction is attempted, directly exploiting the binding-site features given by the protein.

Most of the methods involved in the presently applied virtual screening searches are still under development. Substantial improvements in our understanding of protein–ligand interactions and the definition of molecular similarity are required together with an appropriate translation of this knowledge into fast and reliable computer algorithms to mature virtual screening into a routine technique applied in the lead discovery process.

Acknowledgements

The studies described in this contribution have been funded by grant 0311619 (ReLiMo) of the German Federal Ministry for Education, Science, Research and Technology (BMB+F) and the Deutsche Forschungsgemeinschaft (DFG, grants KL1204/1 and KL1204/3). Furthermore, we are grateful to Tripos GmbH (Munich) for their support and for providing software tools and the Lead Quest Library. The ACD was kindly provided by Molecular Design Ltd. (San Leandro, USA). Fruitful collaboration with M. Rarey, C. Lemmen and T. Lengauer (GMD-SCAI, Bonn, Germany), Thomas Mietzner and Jens Sadowski (BASF, Ludwigshafen, Germany) is gratefully acknowledged.

References

[1] H. Kubinyi, *Curr. Opin. Drug Discov. Develop.* **1998**, *1*, 4–15.
[2] K. Müller, in *Perspectives in Drug Discovery and Design Vol. 3*, P. S. Anderson, G. L. Kenyon, G. R. Marshall (Eds.), ESCOM, Leiden **1995**.
[3] J. H. Van Drie, M. S. Lajiness, *DDT* **1998**, *3*, 274–283.
[4] W. P. Walters, M. T. Stahl, M. A. Murcko, *DDT* **1998**, *3*, 160–178.
[5] M. Hendlich, **1999**, unpublished results.

[6] H.-J. Böhm, G. Klebe, *Angew. Chem.* **1996**, *108*, 2750–2778.
[7] G. Klebe, *J. Mol. Biol.* **1994**, *237*, 212–235.
[8] I. J. Bruno, J. C. Cole, J. P. Lommerse, R. S. Rowland, R. Taylor, M. L. Verdonk, *J. Comput.-Aided Mol. Des.* **1997**, *11*, 525–537.
[9] H. J. Böhm, *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61–78.
[10] P. J. Goodford, *J. Med. Chem.* **1985**, *28*, 849–857.
[11] M. L. Verdonk, J. C. Cole, R. Taylor, *J. Mol. Biol.* **1999**, *289*, 1093–1108.
[12] H. Gohlke, K. Hendlich, G. Klebe, *Perspect. Drug Design Discov.* **2000**, in press
[13] Database ACD (Available Chemicals Directory), MDL Information Systems, Inc., San Leandro, CA.
[14] Program UNITY Chemical Information Software, v. 4.0.3, Tripos Associates Inc., St. Louis, MO.
[15] T. Hurst, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190–196.
[16] S. K. Kearsley, G. M. Smith, *Tetrahedron Comp. Methodol.* **1990**, *3*, 615–633.
[17] G. Klebe, T. Mietzner, F. Weber, *J. Comput.-Aided Mol. Des.* **1994**, *8*, 751–778.
[18] G. Klebe, T. Mietzner, F. Weber, *J. Comput.-Aided Mol. Des.* **1999**, *13*, 35–49.
[19] O. Krämer, diploma thesis, *Rationales Wirkstoff-Design am Beispiel der Aldose-Reduktase*, Philipps-University of Marburg, **1999**.
[20] C. Lemmen, T. Lengauer, *J. Comput.-Aided Mol. Des.* **1997**, *11*, 357–368.
[21] G. Klebe, T. Mietzner, *J. Comput.-Aided Mol. Des.* **1994**, *8*, 583–606.
[22] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, *J. Mol. Biol.* **1996**, *261*, 470–489.
[23] G. Jones, P. Willet, R. C. Glen, A. R. L. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727–748.
[24] T. J. A. Ewing, I. D. Kuntz, *J. Comput. Chem.* **1997**, *9*, 1175–1189.
[25] I. D. Kuntz, E. C. Meng, B. K. Shoichet, *Acc. Chem. Res.* **1994**, *27*, 117–123.
[26] T. Lengauer, M. Rarey, *Curr. Opin. Struct. Biol.* **1996**, *6*, 402–406.
[27] H.-J. Böhm, G. Klebe, *Angew. Chem. Int. Ed. Engl.* **1996**, *35*, 2566–2587.
[28] D. L. Beveridge, F. M. DiCapua, *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431–492.
[29] P. Kollman, *Chem. Rev.* **1993**, *93*, 2395–2417.
[30] H.-J. Böhm, *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243–256.
[31] H. J. Böhm, *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309–323.
[32] I. Muegge, Y. C. Martin, *J. Med. Chem.* **1999**, *42*, 791–804.
[33] J. B. O. Mitchell, R. A. Laskowski, A. Alex, J. M. Thornton, *J. Comput. Chem.* **1999**, *20*, 1165–1176.
[34] H. Gohlke, K. Hendlich, G. Klebe, *J. Mol. Biol.* **2000**, *295*, 337–356.
[35] M. Hendlich, *Acta Crystallogr. D.* **1998**, *54*, 1178–1182.
[36] J. B. O. Mitchell, R. A. Laskowski, A. Alex, M. J. Forster, J. M. Thornton, *J. Comput. Chem.* **1999**, *20*, 1177–1185.
[37] N. Okada, S. Nishimura, *J. Biol. Chem.* **1979**, *254*, 3061–3066.
[38] U. Grädler, R. Ficner, G. A. Garcia, M. T. Stubbs, G. Klebe, K. Reuter, *FEBS Lett.* **1999**, *454*, 142–146.
[39] B. C. Persson, *Mol. Microbiol.* **1993**, *8*, 1011–1016.
[40] J. M. Durand, N. Okada, T. Tobe, M. Watarai, I. Fukuda, T. Suzuki, N. Nakata, K. Komatsu, M. Yoshikawa, C. Sasakawa, *J. Bacteriol.* **1994**, *176*, 4627–4634.
[41] T. H. Maren, *Drug Development Research* **1987**, *10*, 255–276.
[42] G. M. Smith, R. S. Alexander, D. W. Christianson, B. M. McKeever, G. S. Ponticello, J. P. Springer, W. C. Randall, J. J. Baldwin, C. N. Habecker, *Protein Sci.* **1994**, *3*, 118–125.
[43] L. R. Scolnick, A. M. Clements, J. Liao, L. Crenshaw, M. Hellberg, J. May, T. R. Dean, D. W. Christianson, *J. Am. Chem. Soc.* **1997**, *119*, 850–851.
[44] P. A. Boriack-Sjodin, S. Zeitlin, H. H. Chen, L. Crenshaw, S. Gross, A. Dantanarayana, P. Delgado, J. A. May, T. Dean, D. W. Christianson, *Protein Sci.* **1998**, *7*, 2483–2489.
[45] Database MAYBRIDGE, Maybridge Chemical Company Ltd., Tintagel, Cornwall, England.
[46] M. F. Sugrue, A. Harris, I. Adamson, *Drugs of Today* **1997**, *33*, 283–298.
[47] J. I. Wallace, *Clinical Diabetes* **1999**, *17*, 19–26.
[48] P. F. Kador, *Med. Res. Rev.* **1988**, *8*, 325–352.
[49] A. Urzhumtsev, F. Tete-Favier, A. Mitschler, J. Barbanton, P. Barth, L. Urzhumtseva, J. F. Biellmann, A. Podjarny, D. Moras, *Structure* **1997**, *5*, 601–612.
[50] O. Krämer, G. Klebe, **1999**, unpublished results.
[51] R. D. Cramer, III, D. E. Patterson, J. D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
[52] G. Klebe, U. Abraham, T. Mietzner, *J. Med. Chem.* **1994**, *37*, 4130–4146.
[53] G. Klebe, U. Abraham, *J. Comput.-Aided Mol. Design* **1999**, *13*, 1–10.
[54] D. A. Campbell, J. C. Bermate, T. S. Burkoth, D. V. Patel, *J. Am. Chem. Soc.* **1995**, *117*, 5372.

Virtual Screening for Bioactive Molecules
Edited by Hans-Joachim Böhm and Gisbert Schneider

11 Structure-Based Library Design

Martin Stahl

11.1 Introduction

Modern medicinal chemistry relies heavily on 3-D structural information about therapeutic targets. Wherever such information is available, it is used to understand target mechanisms and experimental ligand binding data to guide the search for new ligands. In many cases, structural information has been the key to successful and efficient drug design [1–8]. Structure-based design methodology can roughly be divided into interactive modelling and automated techniques. Interactive modelling work is essential to visualize structures, to gain an understanding of the target structure, and to generate, modify and select individual molecules as potential inhibitors in a user-biased way. In contrast, automated techniques can objectively assemble and assess large libraries of molecules. Over the past decade, interactive modelling in lead finding as well as lead optimization has increasingly been augmented by automated techniques. Reasons for this development are manifold. Firstly, a wealth of structural knowledge about receptor–ligand interactions is hidden in structural databases of proteins and small molecules. Manual work alone cannot make full use of this information. Secondly, combinatorial and parallel syntheses as well as high-throughput assays have made it possible to test the validity of many more structure-based ideas in far shorter time. Finally, since computational techniques to predict receptor–ligand binding are not yet accurate enough to replace experiment, structure-based design is still prone to high error rates. On average, however, error rates can be expected to be lower than in any other approach which does not take into account available structural information. To deal with libraries instead of single molecules therefore means to replace success or failure in individual cases by the notion of *enrichment* of active compounds.

The primary goal of structure-based library design is to assemble collections of molecules that are potential ligands of a given target. One can approach this goal from two sides: Ligand design tools *generate* new molecules within the boundaries of a binding site, while docking methods *select* them from larger libraries. Docking is "virtual screening" in its truest sense. At the borderline between docking and *de novo* design, there exist methods for generating and docking molecules that can be made with specific chemical reactions, most importantly by application of combinatorial chemistry. An introduction to each of these methods will be given here (see Table 11.1 for a list of selected docking and *de novo* design programs). Other methods exist that make use of structural information by translating it and using it in a different context. Among these is the generation of pharmacophores from binding sites (see Chapters 7 and 10). The present Chapter is exclusively devoted to technology that makes direct use of receptor 3-D structures.

Table 11.1. Selected examples of docking and *de novo* design programs.

Docking	
Autodock [9–12]	Simulated annealing / GA, AMBER scoring
DOCK [13–18]	Matching receptor sphere centers to ligand atoms, multiconformer rigid-body docking or flexible docking, combinatorial docking, AMBER scoring
FlexX [19–24]	Incremental construction, combinatorial docking, adapted LUDI scoring function
FLOG [25, 26]	Multiconformer docking, empirical scoring function
GOLD [27–29]	GA, empirical scoring function
PRO_SELECT [30]	Combinatorial docking, empirical scoring function
De *novo* design	
GrowMol [31]	Fragment-based, sequential growth, stochastic search
SMoG [32, 33]	Fragment-based, sequential growth, stochastic search
GroupBuild [34]	Fragment-based, sequential growth, combinatorial search
SPROUT [35–37]	Fragment-based, sequential growth, combinatorial search
LUDI [38–40]	Fragment-based, separate placement and linking, combinatorial search
PRO_LIGAND [41, 42]	Fragment-based, separate placement and linking, combinatorial search
LEGEND [43]	Atom-based, stochastic search
CONCEPTS [44]	Atom-based, starting from a lattice of atoms, stochastic search
MCDNLG [45]	Atom-based, starting from an initial molecular lattice, stochastic search
MCSS [46]	Simultaneous minimization of multiple probe fragments
HOOK [47]	Linker database search for fragments placed by MCSS
CAVEAT [48]	Database search for fragments fitting onto two bonds
Builder [49, 50]	Recombination of docked ligands, combinatorial search
SPLICE [51]	Recombination of ligands retrieved by a 3-D database search

All methods for structure-based library design are based on a simplified computational description of a binding site (Figure 11.1) and techniques to search the translational, rotational, and conformational space of small organic molecules within that binding site. This space can be immensely large and is usually associated with a very complicated energy hy-

persurface. In the case of *de novo* design, parts of a chemical space of even larger size must be searched in addition. All methods have therefore in common that they alleviate the search through some degree of simplification, heuristic rules or stochastic elements.

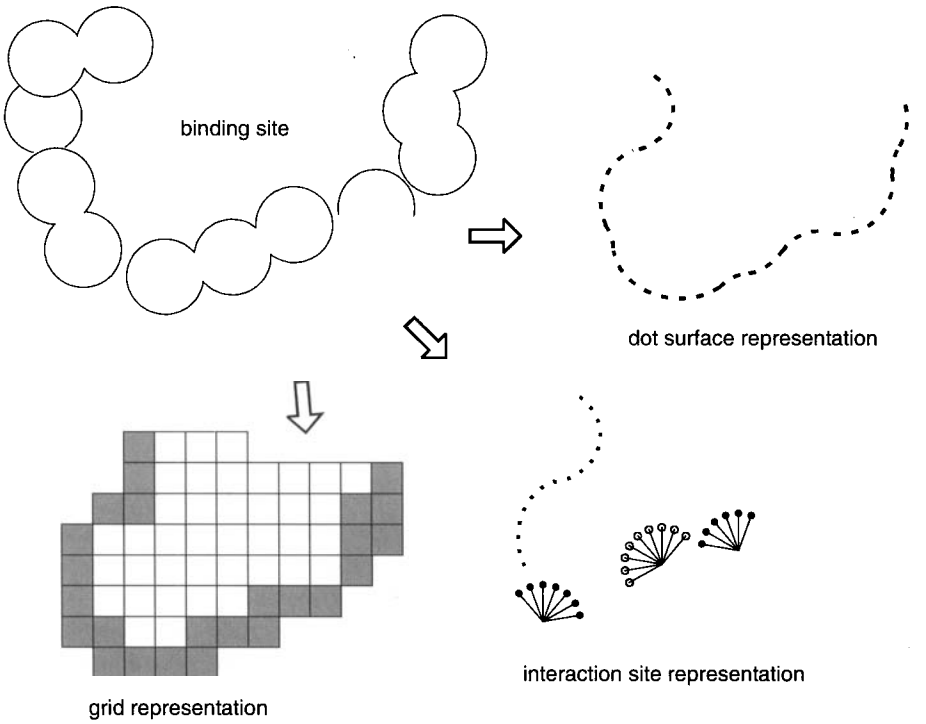


Figure 11.1. Different representations of binding sites for computational purposes. A rectangular grid can be used to define the spatial extensions of the cavity as well as to store interaction potentials. Dot surfaces define the borders of the cavity in a less memory-intensive way, but are less straightforward to process. Interaction sites are potential positions of ligand atoms that form specific interactions (hydrogen bonds, salt bridges, lipophilic interactions) with individual receptor atoms and are used in both docking and *de novo* design.

Another common feature of all structure-based library design tools is a so-called *scoring function* that gives an estimate of the free energy of binding of a molecule or molecular fragment in a given orientation and conformation (called a *pose*) in the binding site. Scoring functions are necessary to predict binding modes, to rank libraries of molecules with respect to a given target, to estimate the selectivity of molecules, and to guide the design of new molecular scaffolds in *de novo* design. Scoring functions will be discussed with special emphasis here, because their lack of accuracy is the main critical issue in current library design. The final Section of this Chapter contains a representative selection of examples from the literature, demonstrating the power of structure-based library design in practical applications.

11.2 Scoring Functions for Receptor–Ligand Interactions

A very broad range of methods exists to estimate how strongly a molecule in a given pose will bind to a macromolecular target. Apart from simple measures of steric and electrostatic complementarity that are often used in the initial stages of docking or *de novo* design, most scoring functions estimate the free energy of binding of a receptor–ligand complex in aqueous solution. Only few methods, however, address the full thermodynamic cycle [52] involved in the binding process (see Chapter 10). They require large-scale simulations of the systems involved [53–56] and are therefore only suitable for small sets of compounds [57]. Those scoring functions that can be evaluated quickly enough to be valuable in virtual screening invariably decompose the free energy of binding into a sum of terms. In a strict physical sense, this is not allowed, since the free energy of binding is a state function, but its components are not [58]. Nevertheless, receptor–ligand binding can often be understood in an additive fashion [59, 60] and reasonable estimates of binding free energy are in this way available at very low computational cost [61–66]. Fast scoring functions can be categorized in three main classes, i.e. force field-based methods, empirical scoring functions and knowledge-based methods.

11.2.1 Force Field-Based Methods

An obvious idea to circumvent parametrization efforts for scoring is to use non-bonded energies of existing, well established molecular mechanics force fields for the estimation of binding affinity. In doing so, one substitutes estimates of the free energy of binding in solution by an estimate of the gas phase enthalpy of binding. Even this crude approximation can lead to satisfying results. A good correlation was obtained between non-bonded interaction energies calculated with a modified MM2 force field and IC_{50} values of 33 HIV-1 protease inhibitors [67]. Similar results were reported in a study of 32 thrombin-inhibitor complexes with the CHARMM force field [68]. In both studies, however, experimental data represented rather narrow activity ranges and little structural variation.

The AMBER [69, 70] and CHARMM [71] non-bonded terms are used as a scoring function in several docking programs, often in a slightly simplified grid-based version that can be quickly evaluated [15]. Distance-dependent dielectricity constants are usually employed [72]. Better approximations of binding free energies are obtained when additional desolvation energy terms are calculated. This has several beneficial effects for virtual screening applications. A surface-based hydrophobic desolvation term automatically reduces the number of high-scoring compounds not filling the cavity. When electrostatic interactions are complemented by a solvation term calculated by the Poisson-Boltzmann equation [73] or faster continuum solvation models(e.g. reference [74]), compounds with high formal charges no longer receive extremely high scores due to overestimated ionic interactions. In a validation study on three protein targets, correction for ligand solvation significantly improved the ranking of known inhibitors [75]. In this context, the van der Waals term is mainly responsible for penalizing poses with overlap between receptor and ligand atoms. It is often omitted when only the binding of experimentally determined complex structures is analyzed [76–78].

The calculation of ligand strain energy also traditionally lies in the realm of molecular mechanics force fields. Usually, ligand strain is calculated during the construction of poses, but not added to the final binding free energy, because it is assumed that strained conformations can be weeded out before the final score is calculated and because better correlation with experimental binding data is observed. Effects of strain energy have rarely been determined experimentally [2], but it is generally accepted that high-affinity ligands bind in low-energy conformations [79, 80]. Estimation of ligand strain energy based on force fields can be time-consuming and therefore alternatives are often employed, such as empirical rules derived from small-molecule crystal structure data [81]. Poses generated by such programs are, however, often not strain-free, because only one torsional angle is regarded at a time.

11.2.2 Empirical Scoring Functions

The term "empirical scoring function" [66] collectively describes scoring schemes that approximate the free energy of binding, $\Delta G_{\text{binding}}$, as a sum of interactions multiplied by weighting coefficients ΔG_i (Eq. 11.1):

$$\Delta G_{\text{binding}} \approx \Sigma \Delta G_i f_i(r_l, r_p) \tag{11.1}$$

where each f_i is a simple geometrical function of the ligand coordinates r_l and the receptor coordinates r_p. The individual terms of this sum are chosen such that they intuitively cover important contributions of the total binding free energy. Most empirical scoring functions are derived by evaluating the functions f_i on a set of protein–ligand complexes and fitting the coefficients ΔG_i to experimental binding affinities of these complexes by multiple linear regression or supervised learning. The relative weight of the individual contributions to each other depends on the training set. Usually, between 50 and 100 complexes are used to derive the weighting factors. In a recent study it has been shown that many more than 100 complexes were necessary to achieve convergence [82]. A general problem of empirical scoring functions is the fact that the experimental binding energies necessarily stem from many different sources and therefore form inconsistent datasets containing systematic experimental errors.

Empirical scoring functions usually contain individual terms for hydrogen bonds, ionic interactions, hydrophobic interactions and binding entropy. Each contribution has been calculated in a number of ways as outlined here:

1. **Hydrogen bonds** are often scored by simply counting the number of donor–acceptor pairs that fall in a given distance and angle range favorable for hydrogen bonding, weighted by penalty functions for deviations from preset ideal values [83–85]. The amount of error-tolerance in these penalty functions is critical. When large deviations from ideality are tolerated, the scoring function cannot suficiently well differentiate between poses, whereas small tolerances lead to situations where many structurally similar poses obtain very different scores. Attempts have been made to reduce the localized nature of such interaction terms by using continuous modulating functions on an atom-pair basis [86]. Other workers have avoided the use of penalty functions and introduced separate regression coefficients for strong, medium and weak hydrogen bonds [82]. The Agouron group has

used a simple four-parameter potential that is a piecewise linear approximation of a potential well without angular terms [87]. The functions mentioned above treat all types of hydrogen bond interactions equally. Some attempts have been made to distinguish between different donor–acceptor functional group pairs. Hydrogen bond scoring in the docking program GOLD [28, 29] is based on a list of hydrogen bond energies for all combinations of 12 donor and 6 acceptor atom-types derived from *ab initio* calculations of model systems incorporating these atom-types. A similar differentiation of donor and acceptor groups is made in the hydrogen bond functions in the program GRID [88] for the characterization of binding sites [89–91]. The inclusion of such look-up tables in scoring functions might help to avoid errors originating from the over-simplification of individual interactions.

2. **Ionic interactions** are usually scored in a similar manner as hydrogen bonds. Purely electrostatic charge–charge interactions are usually disregarded, and so it is more appropriate to refer to salt bridges or charged hydrogen bonds here. The Chemscore function by Protherics [85] differs from the original work of Böhm [83] mainly in that it does not contain a separate term for ionic interactions. On their training set of complexes, the Protherics group obtained similar regression coefficients to Böhm. In a later publication, the group reported a study of thrombin inhibitors where Bayesian regression was used to adapt the scoring function to new data under the restraints of the old training set [92]. It was found that the model could only be improved by including an indicator variable noting the presence or absence of a charged group binding to the recognition pocket. This indicated that it is indeed justified to use an explicit ionic interaction term. However, when the weight of ionic interactions is too high, there is the danger that highly charged molecules receive unreasonably high scores. When using the Böhm scoring function [83], for example, even a single ionic interaction can drastically improve the rank of a molecule.
3. **Hydrophobic interactions** are the major driving force for complex formation. They are usually estimated by the size of the contact surface at the receptor–ligand interface. Various approximations have been used, such as grid-based methods [83] and volume-based methods (see the discussion in reference [12]). Many functions employ distance-scaled sums over neighboring receptor–ligand atom pairs with small distance cutoffs [20] or relatively long cutoffs including atom pairs that do not form direct van der Waals contacts [85, 87]. The weighting factor ΔG_i of the hydrophobic term depends strongly on the training set. It might have been underestimated in most derivations of empirical scoring functions [93], because most training sets contain an overly large proportion of ligands with many donor and acceptor groups (many peptide and carbohydrate fragments). Another approach to estimating the size of the hydrophobic effect is the use of atomic solvation parameters derived from experimental data [94–97], which has the advantage that not all hydrophobic surfaces are treated equally.
4. **Entropy terms** account for the restriction of conformational degrees of freedom of the ligand upon complex formation. A crude but useful estimate of this entropy contribution is the number of freely rotatable bonds of a ligand. This measure has the advantage of being a function of the ligand only [83, 84]. More elaborate estimates try to take into account the nature of each ligand half on either side of a flexible bond and the interactions they form with the receptor [85, 96], since it is argued that purely hydrophobic contacts allow more residual motion in the ligand fragments. Such penalty terms are also robust with respect

to the distribution of rotatable bonds in the ligands of the training set. In addition, the group at Agouron has used an entropy penalty term proportional to the score [98] to account for the phenomenon of entropy–enthalpy compensation [99].

11.2.3 Knowledge-Based Methods

The steadily increasing number of experimentally solved protein–ligand complexes in the PDB has triggered renewed interest in computational techniques to exploit this structural data for the generation of scoring functions. Such knowledge-based approaches have their foundation in the inverse formulation of the Boltzmann law (Eq. 11.2):

$$E_{ijk} = -kT \ln(p_{ijk}) + kT \ln(Z) \tag{11.2}$$

where the energy function E_{ijk} is called a potential of mean force (PMF) for a state defined by the variables i, j, and k, p_{ijk} is the corresponding probability density, and Z is the partition function. The second term of the sum is constant at constant temperature T and does not have to be regarded, since Z=1 can be chosen by definition of a suitable reference state leading to normalized probability densities p_{ijk}. The inverse Boltzmann technique has been applied to derive potentials for protein folding from databases of protein structures [100]. For the purpose of deriving scoring functions, the variables i, j, and k can be chosen to be protein atom-types, ligand atom-types and their inter-atom distance. The statistical distribution of these distances should display preferences for specific contacts and the absence of repulsive interactions. These frequencies can be converted to sets of atom-pair potentials that are easy to evaluate.

The first applications in drug research [101–103] were restricted to small datasets of HIV protease–inhibitor complexes and did not result in generally applicable scoring functions. Recent publications [32, 33, 104–108] have shown that robust and useful general scoring functions can be derived with this method. The *de novo* design program SMoG [32, 33] contains a very coarse-grained implementation of such a potential that nevertheless gave good correlations between experimental and calculated scores in several datasets. It was able to select the native ligand out of a large number of *de novo* designed molecules for several targets.

The knowledge-based scoring function by Muegge and Martin [104] (Eq. 11.3) consists of a set of distance-dependent atom-pair potentials $E_{ij}(r)$ that are written as:

$$E_{ij}(r) = -kT \ln[f_j(r)\, \rho^{ij}(r)/\rho^{ij}] \tag{11.3}$$

Here, r is the atom pair distance, $\rho^{ij}(r)$ is the number density of pairs ij that occur in a certain radius range around r. This density is calculated in the following manner: first, a maximum search radius is defined. This radius describes a reference sphere around each ligand atom j, in which receptor atoms of type i are searched, and which is divided into shells of a specified thickness. The number of receptor atoms i found in each spherical shell is divided by the volume of the shell and averaged over all occurrences of ligand atoms i in the database of protein–ligand complexes. The term ρ^{ij} in the denominator is the average density of receptor atoms j in the whole reference volume. It is argued that the spherical reference vol-

ume around each ligand atom needs to be corrected by eliminating the volume of the ligand itself, because ligand–ligand interactions are not regarded. This is done by the volume correction factor $f_j(r)$, which is a function of the ligand atom only and gives a rough estimate of the preference of atom j to be solvent-exposed rather than buried within the binding pocket. A large reference radius of 12 Å was chosen in order to implicitly include solvent effects. A correlation coefficient of r2 = 0.61 was obtained in the comparison of calculated and experimental binding energy of 77 complexes. For docking calculations, it is evaluated in a grid-based manner and combined with a repulsive van der Waals potential at short distances.

The potential by Gohlke, Hendlich and Klebe [107] is based on roughly the same formalism, albeit with several differences in the derivation leading to different potential forms. Most notably, the statistical distance distributions $\rho^{ij}(r)/\rho^{ij}$ for the individual atom pairs *ij* are related to a common reference state that is simply the average of the distance distributions of all atom pairs $\rho(r) = \Sigma\Sigma\rho^{ij}(r)/ij$. Furthermore, no volume correction term is used and the sampling cutoff (the radius of the reference sphere) is set to only 6 Å. The individual potentials have the form (Eq. 11.4):

$$E_{ij}(r) = -kT\,(\ln[\rho^{ij}(r)/\rho^{ij}] - \ln[\rho(r)]) \tag{11.4}$$

These pair potentials are used in combination with potentials depending on one (protein or ligand) atom-type only that express the propensity of an atom-type to be buried within a lipophilic protein environment upon complex formation. Contributions of these surface potentials and the pair potentials are weighted equally in the final scoring function. This scoring function was developed with the primary goal of differentiating between correctly docked ligand poses *versus* decoy poses for the same protein-ligand pair. For this purpose, it has proven to be superior to standard FlexX scoring as well as over the "chemical scoring" in DOCK. Promising results were also obtained in predicting free energy differences of different complexes [108].

Mitchell and co-workers choose a different type of reference state [105] (Eq. 11.5). The pair interaction energy is written as:

$$E_{ij}(r) = -kT\ln[1 + m^{ij}\sigma] - kT\ln[1 + m^{ij}\sigma\,\rho^{ij}(r)/\rho(r)] \tag{11.5}$$

Here, the number density $\rho^{ij}(r)$ is defined as above, but it is normalized by the number density of all atom pairs at this same distance instead of by the number of pairs *ij* in the whole reference volume. The variable m^{ij} is the number of pairs *ij* found in the database, and σ is an empirical factor that defines the weight of each observation. This potential is combined with a van der Waals potential as a reference state to compensate for the lack of sampling at short distances and for certain under-represented atom pairs. A correlation coefficient of $r^2 = 0.55$ was obtained with this function on a test set of 90 protein–ligand complexes.

The development of knowledge-based scoring functions is a very active research field. It is therefore difficult to draw valid conclusions about the individual approaches. Results depend not only on the functional form, but also on the definition of the protein and ligand atom-types, cutoff radii and other parameters. Experience at Roche indicates that the PMF scoring function by Muegge and Martin does not penalize unfavorable interactions strongly enough and has a tendency to prefer loosely bound complexes over tightly bound ones. The aromat-

ic carbon–carbon potential is partially responsible for this effect, because it becomes repulsive at relatively long inter-atom distances already. This scoring function will nevertheless serve to illustrate a number of general points about scoring functions in the next Section.

11.2.4 Assessment of Current Scoring Functions for Virtual Screening

Due to the simplicity of theoretical models employed in fast scoring functions, it cannot be expected that they provide accurate measures of free energy. For most virtual screening applications, however, crude estimates are sufficient. Criteria for useful scoring functions are the ability to differentiate not only between plausible and nonsensical poses of the same ligand but also between weak ligands of a target and molecules that do not bind at all. The focus is on elimination of non-binders, whereas the correct rank order of weakly, medium and strongly binding molecules is of secondary interest.

Weeding out wrong binding modes and non-binding molecules is difficult, because it requires penalty terms for binding situations that do not occur in nature. Scoring functions are derived from experimentally determined crystal structures, where certain "unnatural" and energetically unfavorable orientations of the ligand within the receptor cavity will never be observed and therefore cannot be accounted for by the scoring function. Knowledge-based scoring functions try to capture such effects indirectly by making repulsive those interactions which are not observed in crystal structures. It seems, however, that the statistical difference between what is not observed and what is to be expected on average is often not significant enough to form reliable repulsive interactions. In the derivation of regression-based empirical scoring schemes, on the other hand, there is no room for penalty terms. Some situations like obvious electrostatic and steric clashes can be avoided by guessing reasonable penalty terms or by importing them from molecular mechanics force fields. An example of this is the "chemical scoring" function available in the docking program DOCK [13, 15–17, 109, 110], which is a modified van der Waals potential made attractive or repulsive between individual groups of donor, acceptor, and lipophilic receptor and ligand atoms [111, 112]. Other situations cannot be avoided by simple clash terms. Among these are imperfect steric fit of the ligand within the cavity, an unnaturally high degree of solvent-accessible ligand surface in the complex or the formation of voids at the receptor–ligand interface. Potential solutions are empirical filters that measure such fit parameters and remove poses above a user-specified threshold [113]. A promising approach is the inclusion of artificially generated decoy poses in the optimization of scoring functions, as reported for the scoring function of a flexible ligand superposition algorithm [114, 115]. Reducing the weight of hydrogen bonds formed at the outer surface of the binding site is another useful measure for reducing the number of false positives in virtual screening applications. This can be done by reducing the charge of surface residues when explicit electrostatic terms are used [72] or by multiplying the H-bond with a factor that depends on the accessibility of the protein-bonding partner [116] in empirical scoring functions.

The simple additive nature of empirical and force field-based methods often leads to large molecules obtaining high scores, which is of course an unrealistic effect. In some virtual screening applications, a constant penalty value has been added to the score for each heavy atom [18] or a penalty term proportional to the molecular weight has been used [117]. The

scoring function of the docking program FLOG, which contains force field and empirical terms, has been normalized to remove the linear dependence of the crude score from the number of ligand atoms found in a docking study of a 7500 compound database [25]. Entropy terms designed to estimate the restriction of conformational mobility upon ligand binding also help to eliminate overly large and flexible molecules, although they were originally introduced to improve the correlation between experimental and calculated affinities. The size of the solvent-accessible surface of the ligand within the binding pocket is also a useful penalty term to avoid excessively large ligands.

On average, empirical scoring functions seem to lead to better correlation between experimental and calculated binding energies than force field-based approaches because the nonbonded interactions in general purpose force fields are not optimized to reproduce individual intermolecular binding phenomena, but depend highly on the bond, angle, and torsional parameters employed. The addition of solvation corrections is computationally demanding, at least in the preparatory stage of the calculations. Finally, two examples may serve to illustrate the comparative performance of a potential of mean force and an empirical scoring function. The PMF scoring function by Muegge and Martin described in Section 11.2.3 was re-implemented and used to score poses generated by the docking program FlexX [20, 21]. PMF scores are compared to FlexX scores. The FlexX scoring function is a slightly modified version of the original function by Böhm [83]. The first example compares data for several poses generated by FlexX for the ligand maltotetraose binding to beta-amylase (PDB code 1byb). This carbohydrate ligand forms many hydrogen bonds with the receptor. FlexX generates poses within about 1 Å of the crystal structure pose, but gives the best score to a pose of about 2 Å root mean square deviation (rmsd) from the coordinates of the experimentally determined structure (Figure 11.2). The PMF scoring function performs better, in that it gives the highest rank to the closest pose. Most strikingly, however, the FlexX score varies greatly

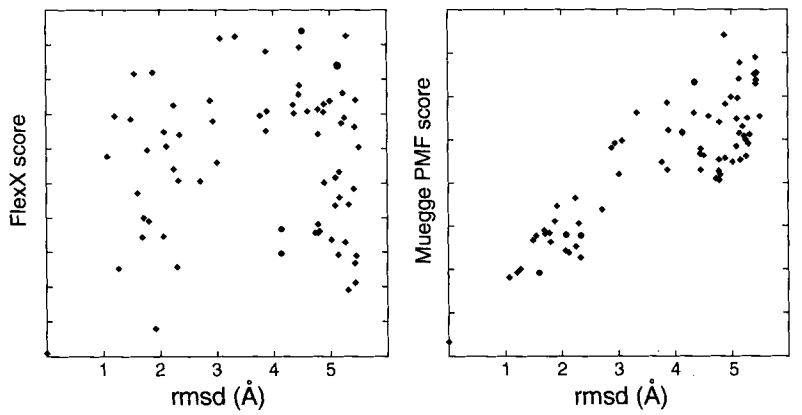


Figure 11.2. Comparison of two scoring functions on poses generated by the docking program FlexX for the ligand maltotetraose binding to beta-amylase (PDB code 1byb). In both cases, the X-ray structure scores best (points on the y-axis). The FlexX scoring function is highly sensitive to small changes in ligand conformation and orientation, as can be seen from the large score variation for poses with small root mean square deviation (rmsd). The Muegge PMF function is not sensitive to small coordinate changes. Score values are in arbitrary units.

for several structures close to the crystal structure pose, whereas the PMF is obviously less error-sensitive. The FlexX score results are typical of highly localized functions; a similar picture is obtained with many other PDB complexes. PMF scores consist of many small contributions that also include more distant interactions, which leads to total scores that are less sensitive towards small changes in ligand pose.

The second example compares FlexX and PMF scores for a library of 470 diaminopyrimidines from the Roche compound collection that were docked into the active site of *Staphylococcus aureus* DHFR [118]. For each compound, only the pose ranking highest with the FlexX scoring function was retained and re-scored with the PMF scoring function. Since in this case the orientation of the diaminopyrimidine fragment was known and could be kept in a fixed position for docking, the docked poses are rather accurate. Figure 11.3 demonstrates that a decent correlation between IC_{50} data and calculated scores can be established with both scoring functions (r^2 = 0.61 in both cases). There is an even better correlation between the two calculated scores (r^2 = 0.67). Two reasons can be responsible for this fact. Firstly, systematic errors in the determination of IC_{50} values can affect the correlation. Secondly, both scoring functions assess only static interactions between protein and ligand and omit many solvation and conformational aspects of ligand binding. It is certainly encouraging to see that protein–ligand interaction terms calculated by two very different methods correlate well. On the other hand, it becomes clear that accurate affinity predictions cannot be based on such interaction terms alone.

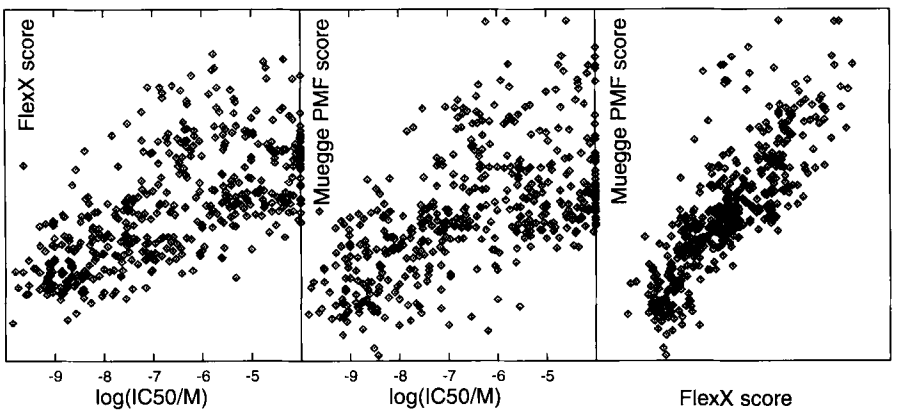


Figure 11.3. Plots of calculated FlexX score and Muegge PMF score *versus* experimentally determined log(IC_{50}) values for diaminopyrimidines binding to *Staphylococcus aureus* DHFR. Docking was done with FlexX using a fixed orientation of the diaminopyrimidine moiety. Correlation coefficents are r^2=0.61 for both scoring functions. The correlation between the two calculated scores is r^2=0.67.

Different scoring schemes emphasize different aspects that are important for ligand binding. Differences between scoring schemes might not be visible in the calculation of binding affinities for active compounds (as in the case of the DHFR inhibitors in Figure 11.3), but can

be very pronounced in the assessment of non-binding molecules. The computational group at Vertex has reported good experience with a concept called "consensus scoring", where libraries of molecules are docked and assessed with several scoring functions and only those molecules are retained that score well with the majority of them. This leads to a significant decrease in false positives. In a recently published validation study [119], sets of several hundred active molecules for three different targets, p38 MAP kinase, inosine monophosphate dehydrogenase, and HIV protease, were docked into the corresponding active sites together with 10000 randomly chosen commercial compounds. Three scoring functions performed consistently well in enriching active compounds, namely the Chemscore function [85, 120], the DOCK AMBER force field score and the piecewise linear potential developed at Agouron [87]. These three functions have in common that they allow a rigid-body minimization of ligand poses due to the inclusion of repulsive terms, which is obviously beneficial for the improvement of orientations as well as scores.

11.3 Receptor–Ligand Docking

Docking methods are computational algorithms developed to predict the three-dimensional structures of receptor–ligand complexes and to evaluate the relative affinity of these bound ligands [27, 121–125]. Consequently, they are used for two purposes: Prediction of receptor–ligand complex structures and ranking of compound libraries with respect to binding affinity towards a given receptor. The correct prediction of the geometry of receptor–ligand complexes has been a long-standing topic in biostructural research (for example, see the discussion of the docking section of the CASP2 contest [126]). Those docking programs fast enough for library design applications can correctly reassemble PDB complexes for about 70% of representative selections from the PDB [23, 29, 127]. Finding correct poses of ligands within their experimentally determined protein structure is of course a prerequisite for successful virtual screening applications. In virtual screening, however, it is necessary to locate probable binding modes of weakly binding ligands within a binding pocket of fixed geometry belonging to a different complex or even the *apo* structure of the protein. This task is far more difficult to solve.

Preparatory calculations for all docking runs include the generation of at least one 3-D conformation for each compound to be docked. For large libraries, 2-D to 3-D conversion programs [128] such as Concord [129] or Corina [130, 131] are invaluable tools. The estimation of protonation states is another critical issue. Only a few docking programs search several protonation states of a ligand simultaneously [110]. Usually, a probable protonation state is estimated for the ligand by simple calculated pK_a criteria. Furthermore, many molecules exist in the form of several stereoisomers that often cannot be handled separately. In such cases, an arbitrary stereoisomer and its mirror image are often chosen to cover at least part of configurational space. In some docking algorithms, docking of two mirror images can be done at little additional computational cost [17]. The current version of FlexX [20, 21] can handle stereocenters as additional degrees of freedom during the construction of ligand poses (M. Rarey, unpublished results).

11.3.1 Docking of Rigid Molecules

When the docking problem is reduced to fitting a rigid molecule into a rigid binding site, there remain six degrees of freedom to be searched. This search space is still too large to be scanned in a systematic way. The usual approach has been to generate a discrete model of the binding site (Figure 11.1) consisting of a set of points that define its spatial extensions with special emphasis on surface properties [14, 19, 25, 109, 132–140]. Atoms of the ligand are then matched onto these receptor points. Various heuristic strategies are used to guide the matching process, since even this discrete formulation of the docking problem is usually too large for an exhaustive search.

The first docking program that used such a technique was DOCK, developed in the group of I. D. Kuntz. Since the first publication almost two decades ago [13], it has been widely applied. The underlying philosophy has proven to be robust and has remained essentially identical over the years. In a first step, the binding cavity is filled with spheres of varying radii, so placed that they contact the Connolly surface of the cavity at a minimum of two points. After several post-processing steps, the number of remaining spheres is in the order of tens to one

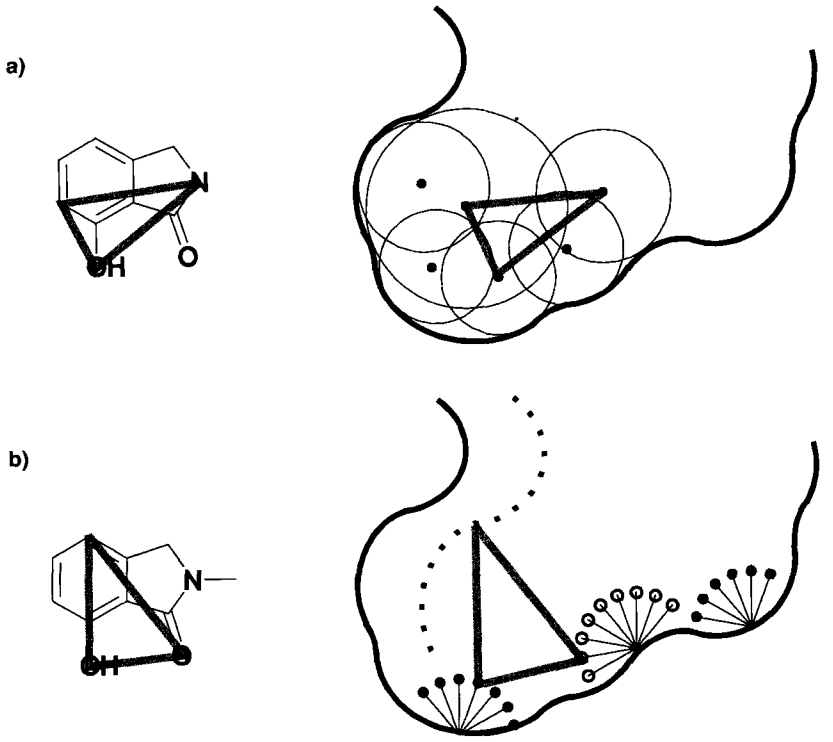


Figure 11.4. Schematic illustration of two alternative methods to dock rigid molecules or molecular fragments into binding sites. a The program DOCK fills the binding site with spheres and matches ligand atoms with sphere centers. b The program FlexX matches interaction site points onto complementary ligand atoms.

hundred, depending on the size of the cavity. Sphere centers are potential positions of ligand atoms (Figure 11.4a). The matching is done by first forming all sphere center–ligand atom pairs. Starting from each pair, new sphere center–ligand atom pairs are searched such that the distance between the sphere centers is approximately equal to the distance between the ligand atoms. Early versions of DOCK used a bipartite graph-matching algorithm [109]. The current version of DOCK employs a single graph representation of the search space together with clique detection techniques, which is approximately twice as fast [17]. A minimum of four ligand atoms must be assigned to sphere centers to define a valid match for the whole ligand. Then the transformation is calculated that optimally superimposes the ligand atoms onto the sphere centers. The resulting pose is tested for overlap with binding site atoms and scored. Since orientational sampling is still rather coarse due to the sparse distribution of receptor spheres and non-exhaustive matching, additional rigid-body minimization is useful to bring poses closer to their minimum orientations [16]. The sampling of sphere center–ligand atom matches is guided by various heuristic rules and parameters, leading to a *meta*-optimization problem that has only been addressed recently in a systematic way [141].

The matching of ligand atoms and sphere centers can be made more efficient by taking into account chemical information in addition to shape considerations. Receptor spheres can be "colored" by a type of interaction a ligand atom must be able to form if it is located at this point of the binding site. The number of ligand atom–sphere center matches is thus reduced to chemically compatible pairs only [132]. Database screening using colored spheres is also more efficient, since ligands not matching a predefined pharmacophore pattern are quickly eliminated [133] (*cf.* Chapters 7 and 10).

The idea of matching atoms or functional groups in order to form energetically favorable local binding situations forms the basis of several docking algorithms. The program FlexX [20, 21] analyzes the binding pocket by placing a distribution of so-called interaction sites around receptor atoms that can potentially be in contact with a ligand. An interaction site is a point in space at which a suitable ligand atom can make favorable interactions with one or more receptor atoms: hydrogen bonds, salt bridges, metal contacts, or lipophilic interactions. The interaction sites in FlexX are used in a similar manner as the receptor spheres in DOCK. The search proceeds as follows: all pairs of interaction sites are formed and stored in a hash table according to the distance they span. Query triangles are formed from three ligand atoms. The hash table is searched for pairs of interaction sites that fit onto two edges of the query triangle, and the resulting lists of compatible edges are merged to form triangles of interaction sites. When a query triangle has been successfully matched onto a triangle of interaction sites, the corresponding pose is generated and scored (Figure 11.4b). This procedure, while conceptually analogous to the procedure used in DOCK, illustrates that very different receptor representations, data structures, and algorithms can be used to solve the rigid-body docking problem.

11.3.2 Conformationally Flexible Docking

Docking single conformations is fast but highly inaccurate for ligands with rotatable bonds. Even for moderately flexible molecules, there are so many diverse and significantly populated low-energy conformations that they cannot be represented by one conformer. The con-

sideration of ligand flexibility is therefore essential to obtain a minimum level of accuracy in docking.

11.3.2.1 Multi-Conformer Docking

The most straightforward approach to the incorporation of ligand flexibility is to sequentially dock several conformers per molecule. The success of this approach naturally depends on the quality and number of pre-computed conformers. Ideally, they should represent all low-energy conformations adopted by the ligand in aqueous solution and a more lipophilic environment, since the local environment of each ligand fragment cannot be reasonably guessed in advance. The use of databases of pre-computed conformers has been reported by a group at Merck. Their docking program FLOG [25, 26] generates diverse conformations by distance geometry and docks each conformer separately. Attempts have been made to speed up the search by docking a large rigid fragment of the ligand first and expanding each of the partial poses to full conformations within the cavity. For this purpose, the rigid fragment must be held fixed during conformational analysis [142]. This idea has been taken one step further for the conformational analysis of reagents for combinatorial synthesis [143]. Reagents are grouped according to their backbone structures, each group being represented by a template reagent representing the group's largest common substructure. Conformational analysis is performed for the template only, and additional degrees of freedom of the individual group members are searched at a later stage. Such techniques can speed up docking calculations by one to two orders of magnitude.

11.3.2.2 Incremental Construction Algorithms

In this class of algorithms, the generation of ligand conformers is no longer an external preprocessing step. Conformers are generated within the boundaries of the binding site by growing the ligand from an initially placed "anchor fragment". Incremental construction algorithms thus consist of three steps:

1. An arbitrary conformer of the ligand is dissected at its rotatable bonds. A large rigid fragment – preferably one that can form several hydrogen bonds – is selected as an anchor fragment.
2. The anchor fragment is docked by one of the methods in Section 11.3.1. Significant computation time can be saved when the position of the anchor fragment is known and can be held fixed in space.
3. Fragments are added to each pose of the anchor fragment in a stepwise manner (Figure 11.5). Bond lengths and angles are usually taken from the input conformer, whereas low-energy positions of torsional angles are searched incrementally. The conformational space of each flexible substituent of the anchor fragment is thus represented by a tree whose nodes are discrete values of torsional angles, the torsional angles being ordered by their distance from the anchor fragment. This tree can be searched in a depth-first [110, 144], or a breadth-first [20, 137] manner, or combinations thereof [145]. Breadth-first searches

have the advantage that the total number of conformations can be controlled by cluster ing and greedy strategies, whereas depth-first searches lead to better solutions in difficu regions of binding sites, where either narrow pockets must be penetrated or zones mus be crossed that contribute little to the binding energy.

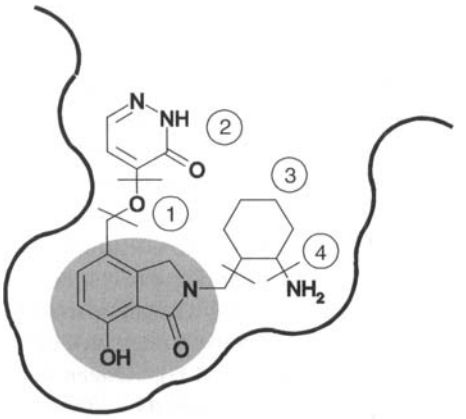


Figure 11.5. Incremental construction algorithms can be used to generate poses of flexible molecules when an anchor fragment has been placed in the binding site (shaded part of the molecule). Rigid fragments of the molecule are added in a stepwise manner, thereby taking into account several positions of dihedral angles at single bonds. Subsequent steps are indicated by numbers.

11.3.2.3 Stochastic Search Algorithms

The algorithms discussed so far employ combinatorial search strategies to solve discrete formulations of the docking problem. Incremental construction algorithms very elegantly separate the orientational and conformational search into two combinatorial search steps. Also, it is an appealing idea to search the whole orientational and conformational space in one process. Docking is then tantamount to an optimization on a very complex multi-dimensional energy hypersurface, for which stochastic methods are suitable search strategies. Advantages of stochastic search algorithms are that they require less complicated data structures and are readily implemented. A pose of a molecule is usually represented by a string of real-value variables describing translation, rotation and variable torsion angles. Random changes in these variables form the basis of stochastic search algorithms.

The program Autodock [9] employs a Monte Carlo Simulated Annealing technique for generating low-energy poses of molecules within binding sites. Autodock has mainly been used to explore possible binding modes of ligands and has been proven to correctly reproduce crystal structures of known complexes [10, 11]. Many other applications of Monte Carlo techniques in docking have been published (for a recent example and literature summary, see reference [146]), and some techniques seem to be fast enough for database docking [147]. Monte Carlo or Molecular Dynamics-based simulated annealing can also be used as a final refinement step in docking calculations. A recent study successfully employed a three-step procedure consisting of rigid body docking of low-energy conformers, torsional sampling in the binding site and final refinement by simulated annealing [148].

Genetic algorithms (GA) and other evolutionary optimization techniques (see Chapters 8 and 9 for a discussion) have been applied by several groups [12, 28, 29, 87, 120, 149–151]. The

typical element of GAs, the encoding of candidate solutions in a chromosome, is usually not employed. Most algorithms operate directly on real-value variables. The program GOLD [28, 29], however, uses a chromosome representation of poses consisting of four strings, two of which contain conformational information. The other two strings map donor atoms of the receptor to acceptor atoms of the ligand and vice versa, such that each pose must be generated by a least-squares fitting step of hydrogen bonding groups. It has repeatedly been found that GA-based docking methods are more efficient in small search spaces, while for global searches they are outperformed by Monte Carlo techniques [12, 120]. Classical genetic algorithms are not efficient in the final stages of optimization. A modified version of the Autodock program, which uses GA instead of simulated annealing, therefore adds intermediate steps of local optimization to find minima more quickly [12]. Another stochastic search algorithm that has been used is tabu search, a method that keeps track of regions of conformational space that have already been visited and that can thus search large search spaces rather efficiently [120].

11.3.3 Current Status of Docking Methods

Which docking method should one choose for virtual screening applications? The answer depends on the systems being investigated. Using an incremental construction algorithm, drug-size molecules can today be flexibly docked in an average CPU time of about one minute on a typical workstation. Most of this time is needed to locate poses of the anchor fragment. If the position of the anchor fragment is known and can be fixed for docking, docking time typically reduces to a few seconds, which is especially useful when libraries containing a common, tightly binding motif are docked. Since this is often the case in drug design projects (e.g. the hydroxamic acid moiety as a central binding motif in metalloprotease libraries, or a common heterocycleforming backbone hydrogen bonds in kinase libraries), incremental construction algorithms are currently the method of choice for many applications (Figure 11.6). Stochastic algorithms are significantly slower and must be run several times per ligand to achieve good sampling. On the other hand, they have the advantage that they operate on the

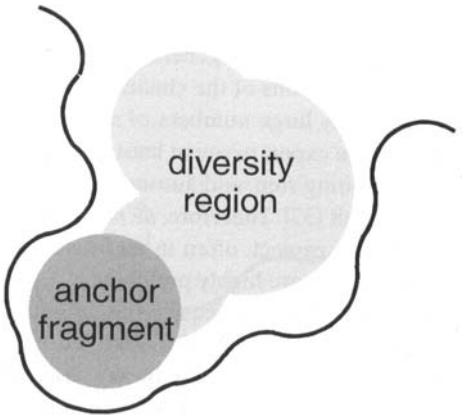


Figure 11.6. A common strategy in drug research is the use of a chemical motif that binds strongly in a fixed position on the binding site (an anchor fragment) and that is substituted in various ways to optimize the binding properties of the whole molecule. Substituents can be thought of as exploring the "diversity region" of the binding site. Incremental construction algorithms are ideally suited for this strategy.

ligand as a whole and treat torsional space as a continuum, leading to better solutions in the case of long chain ligands binding into narrow cavities.

All docking methods that have been mentioned so far share two deficiencies: they cannot take account of protein flexibility [152], and they cannot recognize interactions mediated by water molecules. The induced fit observed in many protein–ligand complexes can be separated into local movements – conformational changes in individual amino acid side chains – and global movements of the receptor, like the "breathing" of a cavity or the reorientation of domains with respect to each other. Force fields in conjunction with appropriate search techniques can predict local induced fit phenomena very accurately [153–155], but can only be applied to selected complexes because of their high computational cost. Such methods are useful in the lead optimization phase of a project. Small movements of individual side chains can be accounted for in database docking [156]. When larger, concerted movements of side chains are known or anticipated, the use of several side chain conformations [157, 158] or docking to ensembles of protein structures [159] provide potential solutions. Such approaches can take account of the flexibility of the receptor when there is already some knowledge about alternative conformations – the *prediction* of conformational changes and induced fit phenomena remains an extremely difficult problem.

Water-mediated interactions play a large role in inhibitor binding, as exemplified by the "flap water" in HIV protease or the arrangement of water molecules in the specificity pocket of thrombin with different positively charged side chains [6]. Careful analysis of water structure is often essential both for the success of docking calculations [160, 161], and for an understanding of the thermodynamics of related protein–ligand complexes [162]. An extension of the docking program FlexX can recognize potential positions of water molecules during docking [22], but so far there is still little experience with this algorithm in virtual screening applications. For most applications, it is still safer to neglect all but very highly conserved water molecules in docking calculations than to model individual water structures around each potential ligand.

11.4 Ligand Design

Ligand design tools create new molecules within the confines of a binding site of known structure. The term *de novo* design comprises methods that attempt to generate compounds starting from an empty binding site and with few or no restrictions of the chemical space to be searched. Straight *de novo* design can produce extremely large numbers of suggestions within little computer time, most of which look absurd to the expert or are at least difficult to synthesize. Post-processing can easily become the rate-limiting step, and automated estimation of synthetic accessibility is time-consuming and difficult [37]. Therefore, *de novo* design tools are often used as idea generators in early stages of a project, often in an interactive modelling process, and chemical libraries generated in this way are highly preliminary. To restrict the search to critical regions of the binding site, many programs require the definition of a set of pharmacophore points or an interaction map of the binding site obtained from programs such as GRID [88], or from a CoMFA analysis (see Chapter 10). In any case, the search for new ligands or ligand fragments must be guided by the expert. The following sec-

tion gives a very condensed overview of *de novo* design strategies. More detailed reviews have appeared recently (see references [122, 163–165] and especially reference [166]). The problem of synthetic accessibility can be solved by restricting the chemical search space to molecules that can be made with a series of specified reactions. Structure-based design of combinatorial libraries will be discussed in a second subsection.

11.4.1 *De Novo* Design Techniques

The large number of existing *de novo* design techniques makes any categorization into subclasses difficult and rather subjective. The following list of criteria can help to judge the pros and cons of individual methods:

- **Sequential growth *versus* fragment placement and linkage.** The order in which ligands are assembled in the active site is of central importance. Either small fragments or functional groups are placed at energetically favorable positions of the binding site and then connected by means of linker fragments, or the search is started at one specified point in the binding site, growing the ligand to other regions of the binding site by sequentially adding

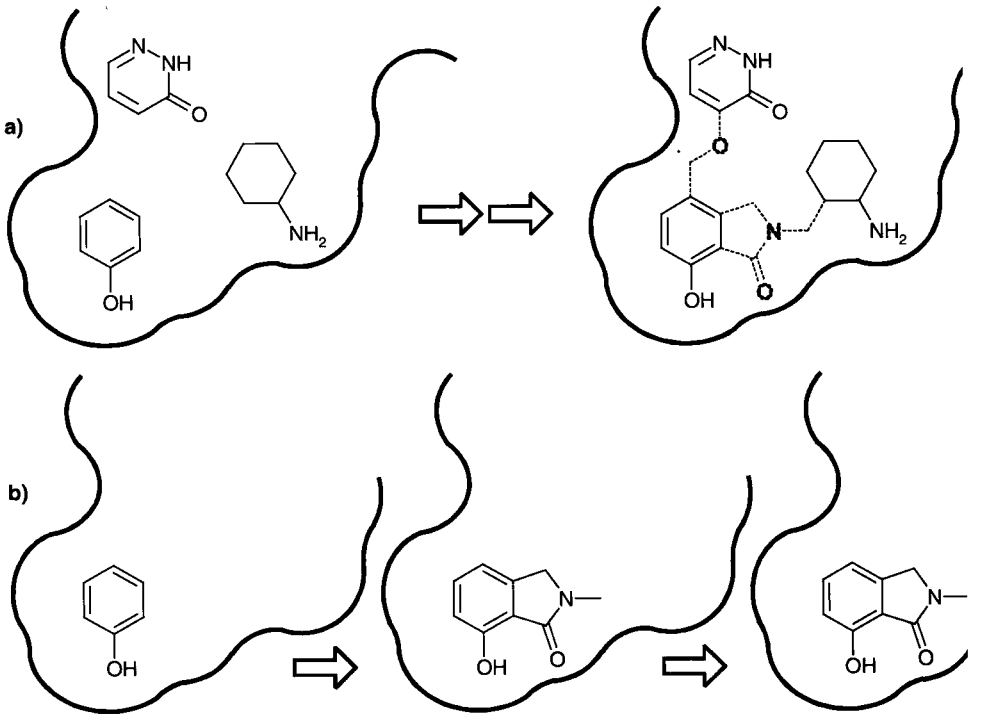


Figure 11.7. Illustration of two alternative strategies in *de novo* design: **a** fragment placing and linking, **b** sequential growth.

fragments (Figure 11.7). The latter strategy has been called an "inside-out" strategy, as c posed to the "outside-in" strategy of fragment placement and linkage [167]. Sequent growth algorithms are highly related to incremental construction algorithms in dockir the main difference being that each fragment of the growing ligand is not fixed but chos from a library. In *de novo* design, however, the goal is not to generate the *best* pose for o ligand but to come up with many solutions of reasonable quality. Therefore, the combin chemical and conformational tree is usually searched in a depth-first manner to arrive solutions quickly, whether or not random steps are involved in the search. In the progra SPROUT by Gillet and co-workers [35, 36], a dynamic balance of depth-first and breadt first search is maintained by a cost function weighing the probability of arriving at a sol tion soon against an estimate of the complexity of the structure that has already been bui For sequential growth algorithms, it can become difficult to cross regions of the bindin site where few favorable interactions can be formed. Fragment-placing and -linking alg rithms overcome this problem of "dead zones" by focusing on the important interaction from the beginning, albeit with the danger that the resulting structures are larger, mor strained, and more difficult to synthesize.

- **Size of the smallest ligand fragment.** Methods that build up ligands from atoms have th advantage that no large libraries of fragments need to be generated and stored. Further more, atom-based techniques can in theory provide more diverse structures. Since th number of potential solutions that can be generated by such methods is extremely high only stochastic search algorithms have been applied. The program LEGEND [43] use atoms to construct ligands in a sequential manner, while the programs CONCEPTS [44] and MCDNLG [45] fill the active site with atoms and employ random moves such as bond formation or breakage, coordinate changes, and atom-type mutations in conjunction with Monte Carlo criteria to gradually evolve ligands. Small, functional-group-sized organic fragments have also been used. The program MCSS [46] places multiple copies of such fragments, each representing one common functional group, into the binding site and performs a simultaneous minimization of all copies. The advantage of this method is that optimal positions of potential interaction groups are easily located. The program HOOK [47] has been written to combine the resulting placements with larger skeletons to form complete ligands. Most other programs use fragments with one or more functional groups. The programs LUDI [38, 39] and PRO_LIGAND [41, 42] make use of libraries containing 1000–100000 small and medium-sized organic molecules for placement as well as linkage steps. This leads more directly to realistic suggestions for screening and synthesis. Placement algorithms are not as exhaustive as docking, and are therefore very fast. A library of small molecules can be docked in its own right to select small molecules for screening or as anchor fragments for manual ligand design [168]. Alternatively, consecutive steps of placement and linking can be applied to generate completely new proposals for ligands or to suggest modifications of existing ones. Such tools are thus applicable in all stages of drug discovery projects.
- **Stochastic search *versus* combinatorial search.** The immense size of search space and the fact that only local minima of search space must be found makes *de novo* design an ideal field for stochastic algorithms. In fragment-based sequential buildup approaches, stochastic [32, 33] and combinatorial [35, 36] algorithms have both been applied, but results are difficult to compare. Stochastic algorithms are especially useful with libraries of small,

non-specific fragments leading to many possible combinations. Fragment placing and linking algorithms are inherently combinatorial and deterministic.

- **Fragment combination *versus* ligand recombination.** Most *de novo* design methods build large ligands from small fragments. Another class of methods starts with complete ligands that are already docked to the binding site and generates new combinations of fragments or atoms of these structures. The program SPLICE [51] by Ho and Marshall joins components of ligands retrieved by the program FOUNDATION [169] at overlapping bonds in a combinatorial fashion. Genetic algorithms have also been applied for this purpose [170]. The "interbreeding" of ligands can easily be realized by crossing over chromosomes representing ligands. This idea is carried further in the program BUILDER [49, 50]. A collection of compounds is first docked into the active site. A molecular lattice is then generated from these structures by interconnecting all atoms that are within covalent bond distance. This lattice is then searched in a breadth-first manner to generate the shortest possible linker between starting fragments within the binding site. An attractive feature of such strategies is the fact that they provide a way to "recycle" existing knowledge about binding fragments.
- **Replacement of ligand segments.** Several tools aim at replacing a segment of a ligand in a complex structure rather than at creating completely new ligands. The program CAVEAT [48] is one such method. It uses pairs of bond vectors and their relative orientation in space to identify molecules in a database. Such methods are easily applicable to library design if the segment to be replaced is identical to a chemical building block that can easily be exchanged. For example, the program LUDI can search for amino acids or peptide mimetic building blocks that could replace a specified segment in a ligand [40].

A study by Bohacek and McMartin employing the program GrowMol [31] has proven that a wealth of design ideas can be obtained with rather simple means: a grid representation of the active site is used to break down the binding site into forbidden, neutral and contact zones. A starting atom of the ligand or receptor is specified. Then, two steps are repeated until a specified molecular weight is reached:

1. A member of a small list of functional groups, atoms and ring systems is selected and added to a randomly chosen growth point by means of a set of chemical rules.
2. A crude complementarity score in conjunction with a Monte Carlo sampling criterion is used to decide whether the new fragment is accepted.

This method was applied to generate a library of roughly 12500 suggestions for the binding site of thermolysin using a sulfur atom binding to the catalytic zinc atom as a starting point. Several criteria were then applied to reduce the number of suggestions: a minimum number of hydrophobic and hydrogen bond contacts, a strain energy threshold after minimization, and a score threshold with a crude empirical scoring function. After a clustering step, 300 suggestions were obtained that varied greatly in structure. Shakhnovich and coworkers have written the program SMoG [32], which proceeds in a similar manner as GrowMol. Their study of Src SH3 domain shows how *de novo* design suggestions can be refined in an interactive process [33]. Out of an initial set of 1000 suggestions, a promising candidate was selected and modified by cycles of manual modification, force field minimization, and additional SMoG growth steps.

11.4.2 Design of Combinatorial Libraries

Structure-based design of active and easy-to-synthesize compounds is a challenging task. Since the *a posteriori* estimation of synthetic accessibility is time-consuming and can only insufficiently be solved by computer, it is reasonable to observe synthetic constraints as early as possible in the design process. Simple examples are programs written for the design of peptide ligands. The program GROW does this in a sequential manner [171]. Alternatively, the program MCSS has been used to place acetamide probes into the active site of HIV-1 protease. A second algorithm was used to connect placed fragments to peptide backbones which were then decorated with appropriate side chains [172].

Since combinatorial chemistry gives access to many diverse compounds under the same reaction conditions, the design of combinatorial libraries is an especially attractive topic for structure-based design. Genetic algorithms have been used to identify individual members of combinatorial libraries that display an especially high degree of complementarity with the binding site [117]. In this application, the genetic algorithm guides the search through the chemical space of the library, allowing its members to interbreed and to mutate. The docking score of individual compounds serves as their fitness value. A similar procedure is used in the CONJURE program at Vertex [173], among others. Such approaches are faster than complete enumeration and sequential docking of combinatorial libraries and are able to locate multiple alternative "good" solutions. They are the virtual counterparts of the evolutionary design techniques discussed in Chapters 8 and 9.

Products of combinatorial libraries can be written as consisting of a common core structure and several groups of substituents (the *closed form* of a combinatorial library). Many combinatorial docking methods are based on the approximation that the orientation of the core remains approximately constant for many members of the library. This is a valid assumption in those cases where the core forms essential interactions with the receptor – where it is identical to an anchor fragment (Figure 11.8a). The program PRO_SELECT [30] can

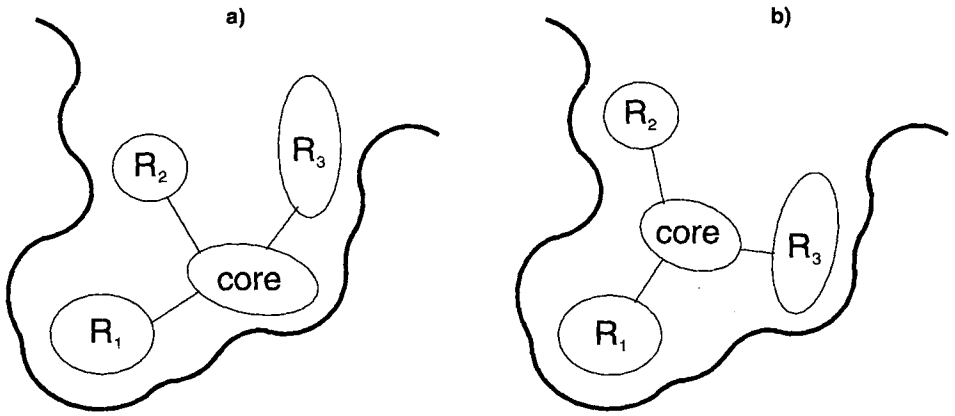


Figure 11.8. Schematic illustration of two combinatorial libraries, whose members bind to the receptor in different ways. **a** The common core of the library forms specific interactions with the receptor. **b** The core merely serves as a linker to keep the R groups in well defined relative orientations. Situations **a** and **b** require different docking strategies.

consider several poses of the core structure. Substituents are added to these poses by a "directed tweak" algorithm [174]. The highest-scoring conformation of each substituent is maintained and can be used to enumerate the library. Similarly, modified versions of the DOCK program [18, 145] first place the core structure into the binding site and then attach substituent fragments independently at each position. Library enumeration starts with the highest-scoring fragments such that it can be terminated when all solutions above a user-specified score threshold have been built.

When the core structure of a combinatorial library does not form specific contacts with the receptor or no contacts at all (Figure 11.8b), it cannot be used as an anchor fragment. The program FlexX [20, 21] has been extended to handle combinatorial libraries in a more general fashion [24]. The structure of the combinatorial library is written as a tree, the core being the root of the tree. The tree can have more than one level, since substituents can be added not only to the core, but also to other substituents. The data structure allows for simultaneous handling of several core instances. Core and substituent definitions can be interchanged by the user. These properties of the tree allow for compact and flexible library definitions. The first step in library docking is the placement of the core fragments. Further fragments are grown to each core pose by a recursive algorithm. This algorithm is called for each partially constructed molecule and tries to extend it by considering all instances of one substituent, using the incremental construction algorithm. Naturally, the docking results depend on the choice of the core and the order in which the substituents are added. As in the choice of the base fragment for single molecule docking, the core fragments should be able to make well defined, specific contacts with the receptor. As a test, the method has been applied to a large UGI library with thrombin as a target. Computing times could be reduced by a factor of about 30 compared to sequential docking of each library member.

11.5. Practical Applications of Structure-Based Library Design

11.5.1 Database Ranking

How many more hits than from a random selection can be expected from library ranking by means of docking calculations? Objective criteria for assessing and improving docking and scoring algorithms can only be found when suitable test libraries are available. However, few libraries are available to the public where experimental data have been measured under uniform conditions for all members.

When consistent datasets are not available, test libraries can be generated by adding known inhibitors to a set of otherwise randomly chosen compounds and calculating the enrichment of active compounds by database ranking. For this purpose, enrichment factors can be defined as the ratio of the percentage of active compounds found in a given subset of the database and the percentage of active compounds that are to be expected when library members are randomly picked (Figure 11.9). A recent study of a database of 100 diverse compounds and two cAMP-dependent protein kinase inhibitors demonstrated the inappropriateness of single conformer rigid body docking for database ranking [175]: only the crystal structure conformation of one ligand scored among the top 20% of the database. The Merck

group has reported a study of 14 HIV protease inhibitors and 15 other protease inhibitors added to an 8000-compound subset of the Merck index. Multiconformer docking with FLOG [25] resulted in excellent enrichment of the HIV protease inhibitors: all but one were among the top 500 library members. The other protease inhibitors were also considerably enriched [176].

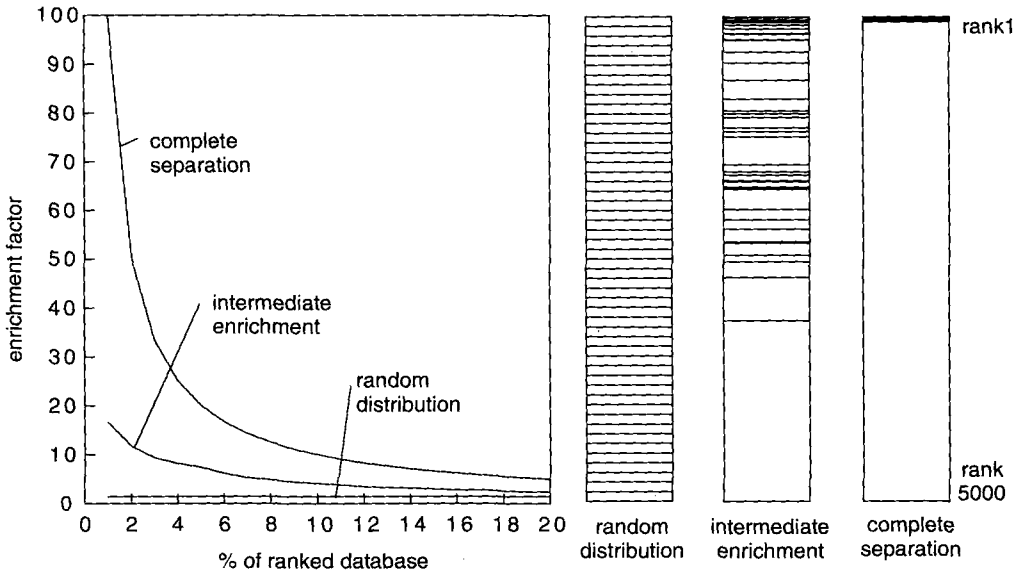


Figure 11.9. Analysis of a hypothetical library of 50 active and 4950 inactive compounds. The three columns on the right show the positions of the active molecules in three different rankings. The corresponding enrichment factors are plotted on the left. The enrichment factor for a given subset of the library is defined as the fraction of active compounds in this subset divided by the fraction of active compounds in the whole library. In a random distribution of the actives, the enrichment factor is always equal to one, and the maximum achievable enrichment factor is in this case 100.

For many targets it is possible to identify potent inhibitors (low nanomolar affinity) from among a number of random molecules, provided that a significant number of hydrogen bonds are formed between the receptor and the ligand. It is, however, very difficult to distinguish weakly binding inhibitors (high micromolar affinity) from non-binders. A very interesting recent study showed that library ranking can successfully be applied to enrich even very weak ligands. A database of approximately 4000 commercially available compounds had been screened against FKBP by means of the SAR-by-NMR technique [177] and had been found to contain 31 compounds active below 2mM. Three examples are shown in Figure 11.10. The compounds **1a, 1b,** and **1c** have measured dissociation constants of 0.1, 0.13, and 0.5 mM, respectively. This set of structures was flexibly docked into the FKBP binding site using DOCK 4.0 and the Muegge PMF scoring function [178]. At 20% of the ranked database, enrichment factors of 2–3 were achieved. Enrichment factors were twice as large as those obtained with the standard AMBER score implemented in DOCK.

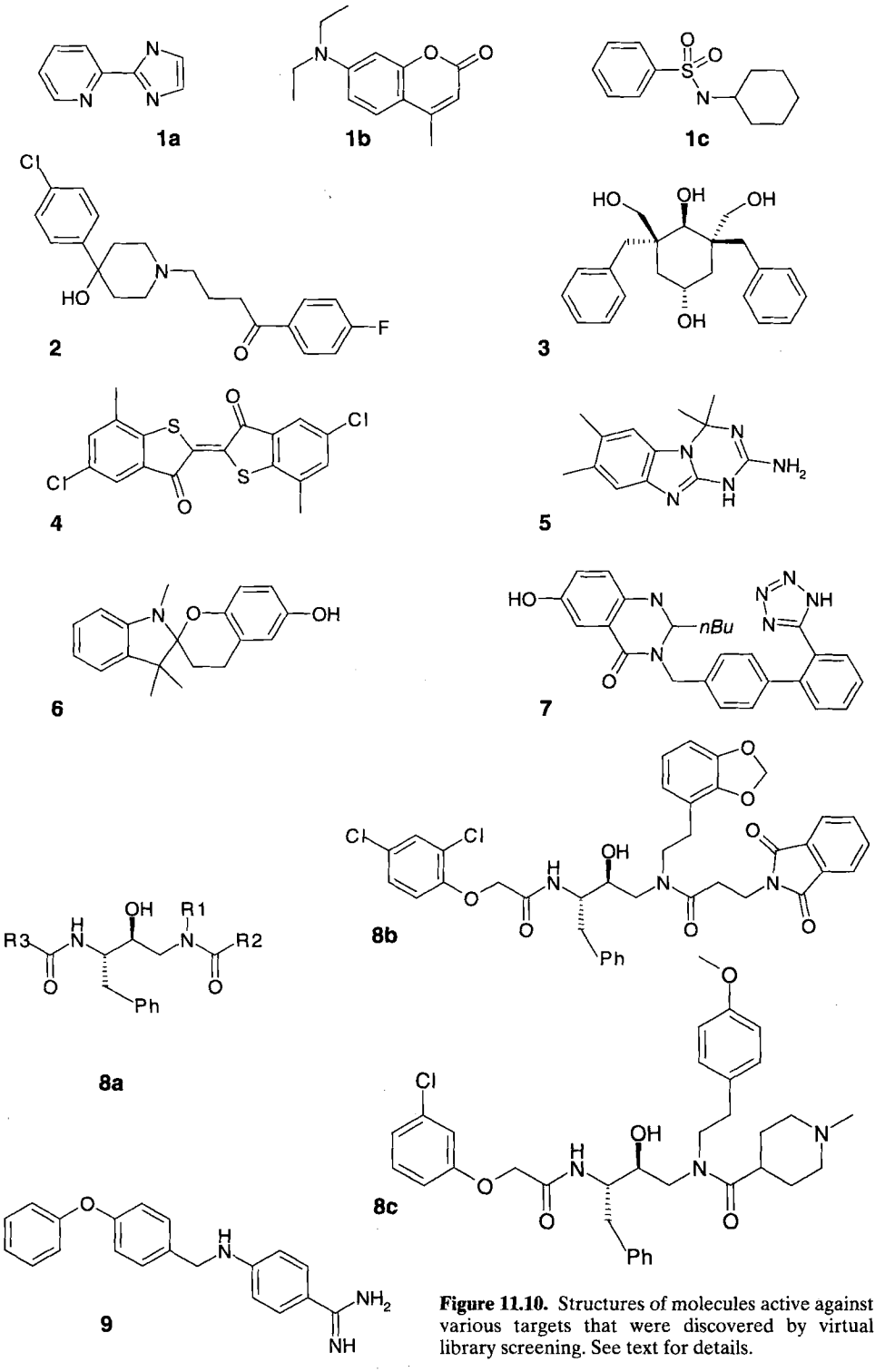


Figure 11.10. Structures of molecules active against various targets that were discovered by virtual library screening. See text for details.

Several publications have proven the capability of the program DOCK for lead finding [133, 179–184]. An early success was the identification of haloperidol (compound **2**, Figure 11.10) as a 100 μM inhibitor of HIV protease through docking of a 10000-molecule subset of the Cambridge Structural Database [179]. Only shape complementarity was used for scoring in this application, and so it was by chance that the docked structure of haloperidol also displayed a chemically reasonable orientation in the active site. A modified version of DOCK matching chemical groups in the sphere matching process already was later used at SmithKline Beecham to dock an in-house database into HIV protease with the restriction that oxygen atoms formed interactions with the Asp25 and Asp25′ carboxylate groups as well as with the Ile50 and Ile50′ backbone nitrogens [133]. This resulted in the identification of 1,4-disubstituted six-membered rings, leading to the 48 μM inhibitor **3** that resembles the cyclic urea inhibitors by DuPont [185]. In many applications of DOCK, a single-conformer database of the Available Chemicals Directory (ACD, reference [186]) has been docked. Typically, around 500 top scoring ligands are then optimized, re-scored and subjected to manual selection processes. An interesting study aimed at the identification of DHFR inhibitors selective for *Pneumocystis carinii* DHFR [182]. For this purpose, an ACD subset was docked into the structure of *P. carinii* DHFR. The 2700 top-scoring structures were then separately rigid-body optimized in human and *P. carinii* DHFR. Subsequent filtering steps included the removal of compounds with small differences in score. Eleven inhibitors were found to have IC_{50} values of 100 μM or less. IC_{50} values of 13 compounds were eventually determined for both enzymes, ten of which showed some selectivity for *P. carinii* DHFR. The most potent compound, the dye **4** (Figure 11.10), showed 25-fold selectivity. A study of thymidylate synthase underlines the importance of structure verification by means of crystal structure analysis [180]. An initial hit from docking was co-crystallized with the target and found to bind in a different region than predicted. A second round of docking led to the identification of phenolphtalein derivatives that bound as predicted by DOCK.

Alternatives to DOCK have started to emerge, and successful applications have been reported with these programs. The docking program GREEN was used to identify the submicromolar inhibitor **5** of *Plasmodium falciparum* DHFR [187]. The program SANDOCK [139] led to the identification of several novel FKBP inhibitors [188], one of which was the spirocyclic compound **6**. In a recent study, the rigid body docking program EUDOC was successfully employed to screen the ACD for compounds binding to farnesyl transferase [189]. The program FLOG has been applied to the search of inhibitors for *Bacteroides fragilis* metallo-β-lactamase [160] within the Merck in-house database. This study is of special interest because docking was performed with three different configurations of bound water in the active site. The top-scoring compounds showed an enrichment in biphenyl tetrazoles, several of which were found to be active below 20 μM. A crystal structure of the tetrazole **7** (IC_{50} = 1.9 μM) not only confirmed the predicted binding mode of one of the inhibitors, but also displayed the water configuration that had been the most predictive one of the three models.

Parallel to the increasing availability of powerful compute servers, increasingly larger virtual libraries or databases by commercial vendors are being screened by means of docking programs [190, 191]. While the ACD has been the main source for virtual screening, docking of in-house compound collections of pharmaceutical companies is equally worthwhile. Since the error-rate of experimental high-throughput screening is often high, the likelihood of finding alternative leads through focused screening is not negligible.

11.5.2 Design of Combinatorial Libraries

As illustrated by the examples above, docking and ranking of large commercially available or proprietary databases can be successfully applied in lead finding. Conversely, structure-based design of combinatorial libraries is most efficient in the lead optimization phase, when an active molecular scaffold has already been identified. Structure-based methods can then be applied to focus on the library's most promising members. When the type of combinatorial library has not yet been identified, suitable scaffolds are also more quickly identified when the receptor structure is known. The importance of structural information in scaffold selection was highlighted in a recently published study on combinatorial libraries for estrogen–receptor ligands [192]. The design was initially based on the structural motifs found in the known ligands benzestrol, raloxifene, and tamoxifen. In the final phase of this work, the designed scaffolds were fitted into the newly released crystal structures of the estrogen receptor with estradiol and raloxifene, and it became clear that a structure-based selection process would have allowed better focusing on those scaffolds that optimally fit into the binding pocket.

An outstanding example of structure-based combinatorial library design has been given by the Ellman and Kuntz groups [193]. In this study, cathepsin D inhibitors were designed, based on a common hydroxyethylamine backbone, a known isostere of the transition state of aspartic proteases (**8a**, Figure 11.10). In a first step, the backbone was manually modelled into the active site. R groups for the variable positions R1, R2 and R3 were selected from the ACD by choosing acids and amines with a molecular weight below 275. Each R group was separately added to the backbone pose and its conformation optimized by a systematic conformational analysis. The best 50 R groups were selected at each position based on force field score. Finally, application of cost and diversity criteria led to a library of $10 \times 10 \times 10$ members. A second library of identical size was constructed based solely on diversity criteria. Both libraries were synthesized and tested against cathepsin D. Of the structure-based library, 23 members displayed at least 50% inhibition at 330 nM, the diverse selection contained only three such compounds. A small second generation library yielded several inhibitors in the low nanomolar range, one of which was the 9 nM inhibitor **8b** (Figure 11.10). A combinatorial docking strategy based on DOCK was later tested on the structure-based library, showing that further enrichment would have been possible, based on calculated scores. Among the top 20% of the ranked library, approximately 80% of the active compounds were retrieved [18].

The compounds synthesized in this study were later tested against plasmepsin II. Two compounds were shown to be active at 200–300 nM [194]. These compounds were optimized by exploring each of the R group sites separately. Docking led to a substituent with improved activity at the R_3 position. For the R_1 and R_2 sites, it was expected that docking would not lead to reasonable results, because both sites were not occupied in the crystal structure used and adopted a collapsed conformation. For one of these sites, a selection of R groups by diversity criteria led to a compound with improved activity. The combined exploration of all three sites finally led to potent inhibitors with K_i values as low as 2 nM, e.g. compound **8c** (Figure 11.10). This study shows that structure-based and diversity-based compound selections can be complementary to each other and allow the identification of highly potent compounds with very small library sizes. Similar success with structure-based design of hydroxyethylamine libraries targeting cathepsin D and plasmepsin II was reported by another

group [195]. A group at MDS Panlabs has reported the design of combinatorial libraries in the search for new heterocycles binding to the ATP binding site of kinases [196].

A two-step design process at Roche quickly lead to potent, non-peptide thrombin inhibitors. In a first step, 5300 commercially available primary amines were docked into the recognition pocket of thrombin, leading to the identification of *p*-amino-benzamidine as the top-scoring compound. An extension of the program LUDI that can differentiate between several kinds of chemical link sites was employed to add 540 benzaldehydes to this anchor fragment through reductive amination. Of the top 100 compounds, ten were synthesized. The most potent compound (**9** in Figure 11.10) had a K_i of 95 nM [197].

In another study at Roche, library design through docking and library design based on diversity criteria were compared in a study starting from the same virtual library targeting DHFR enzymes (both selection processes are summarized in the flowchart in Figure 11.11). The full virtual library contained roughly 4500 diaminopyrimidines, which were accessible from secondary amines as starting materials. Since all members of the library contain a diaminopyrimidine anchor, it was anticipated that this library would contain many compounds displaying some activity against DHFR enzymes. For docking, single conformers of the library were generated with Corina [130, 131]. Carboxylic acids were deprotonated, and aliphatic amines were protonated. The library was then flexibly docked into the active site of *Staphylococcus aureus* DHFR [118] with a fixed position of the diaminopyrimidine fragment, which was taken from an in-house crystal structure. The program FlexX [20, 21] was used with default parameter settings except for the scoring function, which was changed to reduce the weight of hydrogen bonds formed at the surface of proteins ("accessibility scaling" [116]). The docking run gave solutions for about half of the library. For each compound, the pose with the lowest score was selected. Compounds were then sorted according to score and the 150 top-ranking ones were selected for synthesis. Of this selection, 104 compounds were synthesized by parallel chemistry. The compounds were directly tested for antibacterial activity against *S. aureus* (SA), resistant strain *S. aureus* (S1), *Streptococcus pneumoniae* (SPN) and resistant strain *S. pneumoniae* (SP1). The subset of 16 compounds that displayed significant *in vitro* activity against at least one of these organisms was re-synthesized for structure verification and determination of enzyme IC_{50} values.

In an alternative selection process, the whole virtual library was clustered according to chemical similarity. The similarity measure applied was an in-house method based on single conformers. Each pair of library members was superimposed at the newly formed C-N bond and rotated around this bond to generate conformers with maximum volume, H-bond donor and H-bond acceptor overlap. The matrix of similarity scores was used to cluster the compounds. It was found that approximately 400 compounds were needed to adequately represent the chemical space spanned by the library. A sub-library of 375 compounds was tested and found to contain 32 active compounds, of which IC_{50} values were also determined.

In this study, structure-based selection of library members resulted in a hit rate twice as high as in the diverse sub-library, even though antibacterial activity and not enzyme inhibition was used as the primary selection criterion. This corresponds to the enrichment factors obtained with historical collections of diaminopyrimidine libraries for which IC_{50} data are available (such as the data plotted in Figure 11.3). The docking selection covered six out of the eight structural classes of active molecules discovered by the diversity-based selection within only 2% of the total virtual library. Furthermore, the active compounds from the dock-

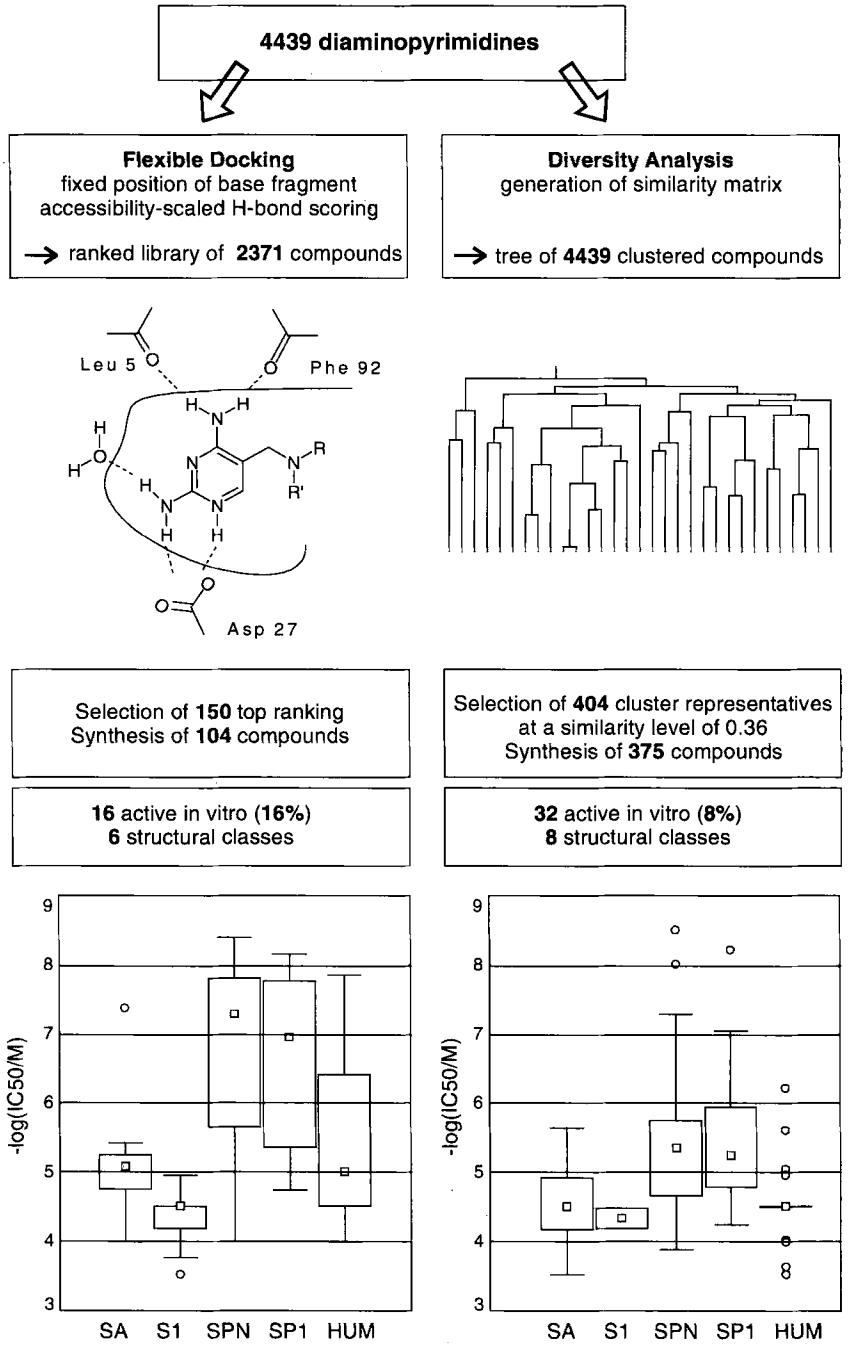


Figure 11.11. Flowchart showing the two alternative procedures applied to select subsets from a large diaminopyrimidine library. On the left, the fixed orientation of the diaminopyrimidine fragment as used in the docking calculations is drawn. The tree structure on the right stands for the diversity-based selection process, where cluster representatives (thick lines) are selected at a given similarity level. The Box plots at the bottom illustrate the activity distributions of the active compounds identified with both procedures. SA: *S. aureus* DHFR, S1: resistant strain *S. aureus* DHFR, SPN: *S. pneumoniae* DHFR, SP1: resistant strain *S. pneumoniae* DHFR, HUM: human DHFR.

ing selection are on average significantly more potent than the active compounds identified by the diversity-based selection (Box plots in Figure 11.11). The active sites of all DHFR enzymes share all essential features, as can be deduced both from crystal structures of human, SA and S1 DHFR, and from homology models of SPN and SP1 DHFR. Differences concern the width of the main cavity and the size of individual hydrophobic patches. It is therefore not surprising that active compounds were found for all five enzymes. In a second control experiment, 150 compounds that had received low ranks in the docking calculation and 150 compounds for which no docking solution was obtained were synthesized and tested. Only one of these compounds showed some antibacterial activity. This is another proof for successful enrichment by means of docking calculations.

When the structure of an inhibitor–receptor complex has been experimentally determined, attempts can be made to optimize its binding properties by proposing modifications based on this structure. However, such optimization procedures can easily fail because induced fit phenomena cannot be predicted (which can lead to dramatic changes in the active site structure, for example as observed with a class of Roche Renin inhibitors [198]), or simply because one does not consider non-conventional solutions. It is then useful to subject part of the inhibitor to random modifications, keeping the remainder of it constant. The fruitful interplay between diversity-based and structure-based library design has been documented by many groups [199, 200], for example in the discovery of non-basic moieties binding to the thrombin recognition pocket [201, 202] or new enalapril-related MMP-inhibitors [203]. More examples have been compiled by others [204–210]. Mixed structure-based and diversity-based approaches are also useful in the design of focused libraries for target classes rather than individual targets. Most general protease libraries are founded on the large body of structural knowledge that is available on serine, aspartic, cysteine and zinc proteases [211]. As a caveat it should be mentioned that in most published examples of combinatorial library design there was a lack of attention to keep the physical properties of the library members in the range required for drug candidates [212]. This observation stresses the importance of multidisciplinary work in drug research even at very early stages in the drug discovery process (see Chapters 1–3).

11.6 Conclusions

This Chapter has given an introduction to computational structure-based design methods aiming at the assembly of libraries rather than single compounds. Over the last decade, structure-based library design has become an established and indispensable field in drug research through many newly developed techniques. These techniques are mainly tested and applied, if not developed, in the pharmaceutical industry, which makes it difficult to present a detailed picture of all current activities. Nevertheless, their potential as well as their limitations are obvious.

The docking problem has been solved to the extent that a multitude of poses can be generated (within seconds to a few minutes of computer time) on modern workstations, whereby coverage of conformational and orientational space is sufficient to find poses close to the experimental complex structure in the majority of cases. Flexible docking of libraries with

some 10^4 members is computationally feasible. Calculation times become minimal when a fragment of the ligand can bind in a known position and can be used as a fixed anchor fragment. However, both prediction of individual complex structures and library ranking require scoring functions to rank solutions according to some estimate of the free energy of binding. This is the still a weak point of all automated structure-based design methods. Still, considerable enrichment has been achieved in library ranking experiments. A respectable number of published examples have proven that docking is a powerful virtual screening method in lead finding as well as lead optimization. Improvements in scoring functions could be achieved by a variety of approaches, such as knowledge-based potentials of mean force or the refinement of empirical filters to weed out unlikely binding modes.

De novo design methods have mostly been used as interactive or semi-automated tools to accelerate the process of collecting new ideas in structure-based design. Its use in the design of libraries is limited by the synthetic accessibility of each library member. An active area of research has therefore been the development of techniques that allow the assembly of novel structures, with the constraint that the new molecules must be synthesizeable with defined chemical reactions. The design of combinatorial libraries is of especially high interest, as combinatorial chemists no longer focus on the sheer size of libraries, but on fast synthesis of pure and well characterized compounds with good activity and physicochemical profiles.

For the purpose of a clear representation, it was necessary to present the spectrum of methods in a well defined order. This does not mean that each method can only be applied under specific circumstances and in only one way. On the contrary, the creative combination of various tools is of fundamental importance for structure-based design. The right combination of automated and interactive tools and the experience and imagination of the expert is still the key to successful structure-based design.

Acknowledgements

During the past two years in the Computational Chemistry group at Roche, I have enjoyed a scientifically and personally stimulating environment, for which I would like to thank all my colleagues, especially Hans-Joachim Böhm, Daniel Bur, Paul Gerber, Frank Grams, Man-Ling Lee, Manfred Kansy, Gisbert Schneider and Pierre Wyss. I would also like to thank Matthias Rarey and Thomas Lengauer for a very fruitful collaboration.

References

[1] M. A. Navia, M. A. Murcko, *Curr. Op. Struct. Biol.* **1992**, *2*, 202–210.
[2] J. Greer, J. W. Erickson, J. J. Baldwin, M. D. Varney, *J. Med. Chem.* **1994**, *37*, 1035–1054.
[3] C. L. M. J. Verlinde, W. G. J. Hol, *Structure* **1994**, *2*, 577–587.
[4] R. S. Bohacek, C. McMartin, W. C. Guida, *Med. Res. Rev.* **1996**, *16*, 3–50.
[5] R. E. Hubbard, *Curr. Op. Biotechnology* **1997**, *8*, 696–700.
[6] R. E. Babine, S. L. Bender, *Chem. Rev.* **1997**, *97*, 1359–1472.
[7] H. Kubinyi, *Curr. Op. Drug Disc. Dev.* **1998**, *1*, 4–15.
[8] H. Kubinyi, *J. Recept. Signal Transduction Res.* **1999**, *19*, 15–39.
[9] D. S. Goodsell, A. J. Olson, *Proteins* **1990**, *8*, 195–202.
[10] D. S. Goodsell, G. M. Morris, A. J. Olson, *J. Mol. Recognition* **1996**, *9*, 1–5.

[11] G. M. Morris, D. S. Goodsell, R. Huey, A. J. Olson, *J. Comput.-Aided Mol. Design* **1996**, *10*, 293–304.
[12] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson, *J. Comp. Chem.* **1998**, *14*, 1639–1662.
[13] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, T. Ferrin, *J. Mol. Biol.* **1982**, *161*, 269–288.
[14] R. DesJarlais, R. P. Sheridan, G. L. Seibel, J. S. Dixon, I. D. Kuntz, R. Venkataraghavan, *J. Med. Chem.* **1988**, *31*, 722–729.
[15] E. C. Meng, B. K. Shoichet, I. D. Kuntz, *J. Comp. Chem.* **1992**, *13*, 505–524.
[16] E. C. Meng, D. A. Gschwend, J. M. Blaney, I. D. Kuntz, *Proteins* **1993**, *17*, 266–278.
[17] T. J. A. Ewing, I. D. Kuntz, *J. Comp. Chem.* **1997**, *18*, 1175–1189.
[18] Y. Sun, T. J. A. Ewing, A. G. Skillman, I. D. Kuntz, *J. Comput.-Aided Mol. Design* **1998**, *12*, 579–604.
[19] M. Rarey, S. Wefing, T. Lengauer, *J. Comput.-Aided Mol. Design* **1996**, *10*, 41–54.
[20] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, *J. Mol. Biol.* **1996**, *261*, 470–489.
[21] M. Rarey, B. Kramer, T. Lengauer, *J. Comput.-Aided Mol. Design* **1997**, *11*, 369–384.
[22] M. Rarey, B. Kramer, T. Lengauer, *Proteins* **1999**, *34*, 17–28.
[23] M. Rarey, B. Kramer, T. Lengauer, *Bioinformatics* **1999**, *15*, 243–250.
[24] M. Rarey, *Recent Developments of FlexX: Parallel Virtual Screening and Processing of Combinatorial Libraries, Talk given at the International Workshop on Virtual Screening*, Schloss Rauischholzhausen, Germany, March 15–18, **1999**
[25] M. D. Miller, S. K. Kearsley, D. J. Underwood, R. P. Sheridan, *J. Comput.-Aided Mol. Design* **1994**, *8*, 153–174.
[26] S. K. Kearsley, D. J. Underwood, R. P. Sheridan, M. D. Miller, *J. Comput.-Aided Mol. Design* **1994**, *8*, 565–582.
[27] G. Jones, P. Willett, *Curr. Op. Biotechnology* **1995**, *6*, 652–656.
[28] G. Jones, P. Willett, R. C. Glen, *J. Mol. Biol.* **1995**, *245*, 43–53.
[29] G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727–748.
[30] C. W. Murray, D. E. Clark, T. R. Auton, M. A. Firth, J. Li, R. A. Sykes, B. Waskowycz, D. R. Westhead, S. C. Young, *J. Comput.-Aided Mol. Design* **1996**, *11*, 193–207.
[31] R. S. Bohacek, C. McMartin, *J. Am. Chem. Soc.* **1994**, *116*, 5560–5571.
[32] R. S. DeWitte, E. I. Shakhnovich, *J. Am. Chem. Soc.* **1996**, *118*, 11733–11744.
[33] R. S. DeWitte, A. V. Ishchenko, E. I. Shaknovich, *J. Am. Chem. Soc.* **1997**, *119*, 4608–4617.
[34] S. H. Rotstein, M. A. Murcko, *J. Med. Chem.* **1993**, *36*, 1700–1710.
[35] V. J. Gillet, A. P. Johnson, P. Mata, S. Sike, P. Williams, *J. Comput.-Aided Mol. Design* **1993**, *7*, 127–153.
[36] V. J. Gillet, W. Newell, P. Mata, G. Myatt, S. Sike, Z. Zsoldos, A. P. Johnson, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 207–217.
[37] V. J. Gillet, G. Myatt, Z. Zsoldos, A. P. Johnson, in *De Novo Design, Vol. 3*, K. Müller (Ed.) Escom, Leiden **1995**, p. 34–50.
[38] H.-J. Böhm, *J. Comput.-Aided Mol. Design* **1992**, *6*, 61–78.
[39] H.-J. Böhm, *J. Comput.-Aided Mol. Design* **1992**, *6*, 593–606.
[40] H.-J. Böhm, *J. Comput.-Aided Mol. Design* **1996**, *10*, 265–272.
[41] B. Waszkowycz, D. E. Clark, D. Frenkel, J. Li, C. W. Murray, B. Robson, D. R. Westhead, *J. Med. Chem.* **1994**, *37*, 3994–4002.
[42] D. E. Clark, D. Frenkel, S. A. Levy, J. Li, C. W. Murray, B. Robson, B. Waszkowycz, D. R. Westhead, *J. Comput.-Aided Mol. Design* **1995**, *9*, 13–32.
[43] A. Itai, Y. Nishibata, *Tetrahedron* **1991**, *47*, 8985–8990.
[44] D. A. Pearlman, M. A. Murcko, *J. Comp. Chem.* **1993**, *14*, 1184–1193.
[45] D. K. Gehlhaar, K. E. Moerder, D. Zichi, C. J. Sherman, R. C. Ogden, S. T. Freer, *J. Med. Chem.* **1995**, *38*, 466–472.
[46] A. Miranker, M. Karplus, *Proteins* **1991**, *11*, 29–34.
[47] M. B. Eisen, D. C. Wiley, M. Karplus, R. E. Hubbard, *Proteins* **1994**, *19*, 199–221.
[48] G. Lauri, P. A. Bartlett, *J. Comput.-Aided Mol. Design* **1994**, *8*, 51–66.
[49] R. A. Lewis, D. C. Roe, C. Huang, T. E. Ferrin, R. Langridge, I. D. Kuntz, *J. Mol. Graphics* **1992**, *10*, 66–78.
[50] D. C. Roe, I. D. Kuntz, *J. Comput.-Aided Mol. Design* **1995**, *9*, 269–282.
[51] C. M. W. Ho, G. R. Marshall, *J. Comput.-Aided Mol. Design* **1993**, *7*, 623–647.
[52] M. K. Gilson, J. A. Given, B. L. Bush, J. A. McCammon, *Biophys. J.* **1997**, *72*, 1047–1069.
[53] T. P. Straatsma, in *Reviews in Computational Chemistry, Vol. 9*, K. B. Lipkowitz, D. B. Boyd (Eds.) VCH Publishers, New York **1991**, pp. 81–98.

[54] P. A. Kollman, *Acc. Chem. Res.* **1996**, *29*, 461–469.
[55] J. Aquist, C. Medina, J.-E. Samuelsson, *Prot. Eng.* **1994**, *7*, 385–391.
[56] T. Hansson, J. Marelius, J. Aquist, *J. Comput.-Aided Mol. Design* **1998**, *12*, 27–35.
[57] M. K. Gilson, J. A. Given, M. S. Head, *Chem. Biol.* **1997**, *4*, 87–92.
[58] A. E. Mark, W. F. van Gunsteren, *J. Mol. Biol.* **1994**, *240*, 167–176.
[59] P. R. Andrews, D. J. Craik, J. L. Martin, *J. Med. Chem.* **1984**, *27*, 1648–1657.
[60] T. J. Stout, C. R. Sage, R. M. Stroud, *Structure* 1998, *6*, 839–848.
[61] Ajay, M. A. Murcko, *J. Med. Chem.* **1995**, *38*, 4953–4967.
[62] J. D. Hirst, *Curr. Op. Drug Disc. Dev.* **1998**, *1*, 28–33.
[63] R. M. A. Knegtel, P. D. J. Grootenhuis, in *3D QSAR in drug design: ligand protein interactions and molecular similarity, Vol. 9/10/11*, H. Kubinyi, G. Folkers, Y. C. McMartin (Eds.) Kluwer/Escom, Dordrecht **1998**, pp. 99–114.
[64] T. I. Oprea, G. R. Marshall, in *3D QSAR in drug design: ligand protein interactions and molecular similarity, Vol. 9/10/11*, H. Kubinyi, G. Folkers, Y. C. McMartin (Eds.) Kluwer/Escom, Dordrecht **1998**, pp. 3–17.
[65] J. R. H. Tame, *J. Comput.-Aided Mol. Design* **1999**, *13*, 99–108.
[66] H.-J. Böhm, M. Stahl, *Med. Chem. Res.* **1999**, *9*, 445–462.
[67] M. K. Holloway, J. M. Wai, T. A. Halgren, P. M. D. Fitzgerald, J. P. Vacca, B. D. Dorsey, R. B. Levin, W. J. Thompson, L. J. Chen, S. J. deSolms, N. Gaffin, T. A. Lyle, W. A. Sanders, T. J. Tucker, M. Wiggins, C. M. Wiscount, O. W. Woltersdorf, S. D. Young, P. L. Darke, J. A. Zugay, *J. Med. Chem.* **1995**, *38*, 305–317.
[68] P. D. J. Grootenhuis, P. J. M. van Galen, *Acta Cryst.* **1995**, *D51*, 560–566.
[69] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, P. Weiner, *J. Am. Chem. Soc.* **1984**, *106*, 765–784.
[70] S. J. Weiner, P. A. Kollman, D. T. Nguyen, D. A. Case, *J. Comp. Chem.* **1986**, *7*, 230–252.
[71] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, *J. Comp. Chem.* **1983**, *4*, 187–217.
[72] M. Vieth, J. D. Hirst, A. Kolinski, C. L. Brooks, *J. Comp. Chem.* **1998**, *19*, 1612–1622.
[73] A. Nicholls, B. Honig, *Science* **1995**, *268*, 1144–1149.
[74] N. Majeux, M. Scarsi, J. Apostolakis, C. Ehrhardt, A. Caflisch, *Proteins* **1999**, *37*, 88–105.
[75] B. K. Shoichet, A. R. Leach, I. D. Kuntz, *Proteins* **1999**, *34*, 4–16.
[76] T. Zhang, D. E. Koshland, jr., *Protein Sci.* **1996**, *5*, 348–356.
[77] M. Schapira, M. Trotov, R. Abagyan, *J. Mol. Recognition* **1999**, *12*, 177–190.
[78] P. H. Hünenberger, V. Helms, N. Narayana, S. S. Taylor, J. A. McCammon, *Biochemistry* **1999**, *38*, 2358–2366.
[79] J. Boström, P.-O. Norrby, T. Liljefors, *J. Comput.-Aided Mol. Design* **1998**, *12*, 383–396.
[80] M. Vieth, J. D. Hirst, C. L. Brooks, III, *J. Comput.-Aided Mol. Design* **1998**, *12*, 563–572.
[81] G. Klebe, T. Mietzner, *J. Comput.-Aided Mol. Design* **1994**, *8*, 583–606.
[82] R. Wang, L. Liu, L. Lai, Y. Tang, *J. Mol. Model.* **1998**, *4*, 379–394.
[83] H.-J. Böhm, *J. Comput.-Aided Mol. Design* **1994**, *8*, 243–256.
[84] H.-J. Böhm, *J. Comput.-Aided Mol. Design* **1998**, *12*, 309–323.
[85] M. D. Elridge, C. W. Murray, T. R. Auton, G. V. Paolini, R. P. Mee, *J. Comput.-Aided Mol. Design* **1997**, *11*, 425–445.
[86] A. N. Jain, *J. Comput.-Aided Mol. Design* **1996**, *10*, 427–440.
[87] D. K. Gehlhaar, G. M. Verkhivker, P. A. Rejto, C. J. Sherman, D. B. Fogel, L. J. Fogel, S. T. Freer, *Chem. Biol.* **1995**, *2*, 317–324.
[88] P. J. Goodford, *J. Mol. Graphics* **1985**, *3*, 107–108.
[89] D. N. A. Boobbyer, P. J. Goodford, P. M. McWhinnie, R. C. Wade, *J. Med. Chem.* **1989**, *32*, 1083–1094.
[90] R. C. Wade, K. J. Clark, P. J. Goodford, *J. Med. Chem.* **1993**, *36*, 140–147.
[91] R. C. Wade, P. J. Goodford, *J. Med. Chem.* **1993**, *36*, 148–156.
[92] C. W. Murray, T. R. Auton, M. D. Elridge, *J. Comput.-Aided Mol. Design* **1999**, *12*, 503–519.
[93] A. M. Davis, S. J. Teague, *Angew. Chem. Int. Ed.* **1999**, *38*, 736–749.
[94] S. Vajda, Z. Weng, R. Rosenfeld, C. DeLisi, *Biochemistry* **1994**, *33*, 13977–13988.
[95] E. C. Meng, I. D. Kuntz, D. J. Abraham, G. E. Kellogg, *J. Comput.-Aided Mol. Design* **1994**, *8*, 299–306.
[96] R. D. Head, M. L. Smythe, T. I. Oprea, C. L. Waller, S. M. Green, G. R. Marshall, *J. Am. Chem. Soc.* **1996**, *118*, 3959–3969.
[97] C. Zhang, G. Vasmatzis, J. L. Cornette, C. DeLisi, *J. Mol. Biol.* **1997**, *267*, 707–726.

[98] P. Rose, *Scoring Methods in Ligand Design*, UCSF Computer-Assisted Molecular Design Course, Jan. 16–18, **1997**, San Francisco.
[99] P. Gilli, V. Ferretti, G. Gilli, P. A. Brea, *J. Phys. Chem.* **1994**, *98*, 1515–1518.
[100] M. J. Sippl, *J. Comput.-Aided Mol. Design* **1993**, *7*, 473–501.
[101] G. Verkhivker, K. Appelt, S. T. Freer, J. E. Villafranca, *Prot. Eng.* **1995**, *8*, 677–691.
[102] A. Wallquist, R. L. Jernigan, D. G. Covell, *Protein Sci.* **1995**, *4*, 1181–1903.
[103] A. Wallquist, D. G. Covell, *Proteins* **1996**, *25*, 403–419.
[104] I. Muegge, Y. C. Martin, *J. Med. Chem.* **1999**, *42*, 791–804.
[105] J. B. O. Mitchell, R. A. Laskowski, A. Alex, M. J. Forster, J. M. Thornton, *J. Comp. Chem.* **1999**, *20*, 1177–1185.
[106] J. B. O. Mitchell, R. A. Laskowski, A. Alex, J. M. Thornton, *J. Comp. Chem.* **1999**, *20*, 1165–1177.
[107] H. Gohlke, M. Hendlich, G. Klebe, *J. Mol. Biol.* **1999**, *295*, 337–356.
[108] H. Gohlke, M. Hendlich, G. Klebe, *Using Empirical Potentials to Predict Protein-Ligand Interactions, Talk given at the International Workshop on Virtual Screening*, Schloss Rauischholzhausen, Germany, March 15–18, **1999**
[109] B. K. Shoichet, D. L. Bodian, I. D. Kuntz, *J. Comp. Chem.* **1992**, *13*, 380–387.
[110] S. Makino, I. D. Kuntz, *J. Comp. Chem.* **1997**, *18*, 1812–1825.
[111] Author, DOCK user manual, Version 4.0, Regents of the University of California, San Francisco **1997**.
[112] R. M. A. Knegtel, D. M. Bayada, R. A. Engh, W. von der Saal, V. J. van Geerestein, P. D. J. Grootenhuis, *J. Comput.-Aided Mol. Design* **1999**, *13*, 167–183.
[113] M. Stahl, H.-J. Böhm, *J. Mol. Graphics Mod.* **1998**, *16*, 121–132.
[114] C. Lemmen, T. Lengauer, G. Klebe, *J. Med. Chem.* **1998**, *41*, 4502–4520.
[115] C. Lemmen, A. Zien, R. Zimmer, T. Lengauer, in *Proceedings of the Pacific Symposium on Biocomputing*, R. B. Altman, K. Lauderdale, A. K. Dunker, L. Hunter (Eds.) World Scientific Publishing Co., Singapore **1999**, pp. 482–493.
[116] M. Stahl, *Filters for Empirical Scoring Functions to Improve the Scoring in FlexX, Talk given at the International Workshop on Virtual Screening*, Schloss Rauischholzhausen, Germany, March 15–18, **1999**.
[117] E. J. Martin, R. E. Critchlow, D. C. Spellmeyer, S. Rosenberg, K. L. Spear, J. M. Blaney, *Pharmacochem. Libr.* **1998**, *29*, 133–146.
[118] G. Dale, C. Broger, A. D'Arcy, P. Hartmann, R. DeHoogt, S. Jolidon, I. Kompis, A. M. Labhardt, H. Langen, H. Locher, M. G. P. Page, D. Stüber, R. L. Then, B. Wipf, C. Oefner, *J. Mol. Biol.* **1997**, *226*, 23–30.
[119] P. S. Charifson, J. J. Corkery, M. A. Murcko, W. P. Walters, *J. Med. Chem.* **1999**, *42*, 5100–5109.
[120] C. A. Baxter, C. W. Murray, D. E. Clark, D. R. Westhead, M. D. Eldridge, *Proteins* **1998**, *33*, 367–382.
[121] P. M. Colman, *Curr. Op. Struct. Biol.* **1994**, *4*, 868–874.
[122] I. D. Kuntz, E. C. Meng, B. K. Shoichet, *Acc. Chem. Res.* **1994**, *27*, 117–123.
[123] T. P. Lybrand, *Curr. Op. Struct. Biol.* **1995**, *5*, 224–228.
[124] R. Rosenfeld, S. Vajda, C. DeLisi, *Ann. Rev. Biophys. Biomol. Struct.* **1995**, *24*, 677–700.
[125] T. Lengauer, M. Rarey, *Curr. Op. Struct. Biol.* **1996**, *6*, 402–406.
[126] J. S. Dixon, *Proteins* **1997**, *Suppl. 1*, 198–204.
[127] B. Kramer, M. Rarey, T. Lengauer, *Proteins* **1999**, *37*, 228–241.
[128] J. Sadowski, J. Gasteiger, *Chem. Rev.* **1993**, *93*, 2567–2581.
[129] Concord, Tripos Associates Inc., St. Louis.
[130] J. Sadowski, C. Rudolph, J. Gasteiger, *Tetrahedron Comput. Methodol.* **1990**, *3*, 537–547.
[131] Corina, 2.1, Molecular Networks GmbH Computerchemie, Erlangen **1998**.
[132] B. K. Shoichet, I. D. Kuntz, *Prot. Eng.* **1993**, *6*, 723–732.
[133] R. L. DesJarlais, J. S. Dixon, *J. Comput.-Aided Mol. Design* **1994**, *8*, 231–242.
[134] N. Tomioka, A. Itai, *J. Comput.-Aided Mol. Design* **1994**, *8*, 347–366.
[135] D. Fischer, S. L. Lin, H. J. Wolfson, R. Nussinov, *J. Mol. Biol.* **1995**, *248*, 459–477.
[136] B. Sandak, R. Nussinov, H. J. Wolfson, *CABIOS* **1995**, *11*, 87–99.
[137] W. Welch, J. Ruppert, A. N. Jain, *Chem. Biol.* **1996**, *3*, 449–462.
[138] C. M. Oshiro, I. D. Kuntz, *Proteins* **1998**, *30*, 321–336.
[139] P. Burkhard, P. Taylor, M. D. Walkinshaw, *J. Mol. Biol.* **1998**, *277*, 449–466.
[140] D. K. Hendrix, T. E. Klein, I. D. Kuntz, *Protein Sci.* **1999**, *8*, 1010–1022.
[141] J.-P. Salo, A. Yliniemelä, J. Taskinen, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 832–839.
[142] D. M. Lorber, B. K. Shoichet, *Protein Sci.* **1998**, *7*, 938–950.

[143] S. Makino, I. D. Kuntz, *J. Comp. Chem.* **1998**, *19*, 1834–1852.
[144] A. R. Leach, I. D. Kuntz, *J. Comp. Chem.* **1992**, *13*, 730–748.
[145] S. Makino, T. J. A. Ewing, I. D. Kuntz, *J. Comput.-Aided Mol. Design* **1999**, *13*, 513–532.
[146] M. Liu, S. Wang, *J. Comput.-Aided Mol. Design* **1999**, *13*, 435–451.
[147] D. J. Diller, C. L. M. J. Verlinde, *J. Comp. Chem.* **1999**, *16*, 1740–1751.
[148] J. Wang, P. A. Kollman, I. D. Kuntz, *Proteins* **1999**, *1999*, 1–19.
[149] R. S. Judson, Y. T. Tan, E. Mori, C. Melius, E. P. Jaeger, A. M. Treasurywala, A. Mathiowetz, *J. Comp. Chem.* **1995**, *16*, 1405–1420.
[150] C. M. Oshiro, I. D. Kuntz, J. S. Dixon, *J. Comput.-Aided Mol. Design* **1995**, *9*, 113–130.
[151] K. P. Clark, Ajay, *J. Comp. Chem.* **1995**, *16*, 1210–1220.
[152] C. W. Murray, C. A. Baxter, A. D. Frenkel, *J. Comput.-Aided Mol. Design* **1999**, *13*, 547–562.
[153] J. Apostolakis, A. Plückthun, A. Caflisch, *J. Comp. Chem.* **1998**, *19*, 21–37.
[154] G. M. Keserü, I. Kolossváry, *Molecular mechanics and conformational analysis in drug design*, Blackwell Science, Oxford **1999**.
[155] I. Kolossváry, W. C. Guida, *J. Comp. Chem.* **1999**, *20*, 1671–1684.
[156] V. Schnecke, C. A. Swanson, E. D. Getzoff, J. A. Tainer, L. A. Kuhn, *Proteins* **1998**, *33*, 74–87.
[157] A. R. Leach, *J. Mol. Biol.* **1994**, *235*, 345–356.
[158] L. Schaffer, G. M. Verkhivker, *Proteins* **1998**, *33*, 295–310.
[159] R. M. A. Knegtel, I. D. Kuntz, C. M. Oshiro, *J. Mol. Biol.* **1997**, *266*, 424–440.
[160] J. H. Toney, P. M. D. Fitzgerald, N. Grover-Sharma, S. H. Olson, W. J. May, J. G. Sundelof, D. E. Vanderwall, K. A. Cleary, S. K. Grant, J. K. Wu, J. W. Kozarich, D. L. Pompliano, G. G. Hammond, *Chem. Biol.* **1998**, *5*, 185–196.
[161] W. E. Minke, D. J. Diller, W. G. J. Hol, C. L. M. J. Verlinde, *J. Med. Chem.* **1999**, *42*, 1778–1788.
[162] T. G. Davies, R. E. Hubbard, J. R. H. Tame, *Protein Sci.* **1999**, *8*, 1432–1444.
[163] L. M. Balbes, S. W. Mascarella, D. B. Boyd, in *Reviews in Computational Chemistry, Vol. 5*, K. B. Lipkowitz, D. B. Boyd (Eds.) VCH Publishers, New York **1994**, pp. 337–379.
[164] A. C. Good, J. S. Mason, in *Reviews in Computational Chemistry, Vol. 7*, K. B. Lipkowitz, D. B. Boyd (Eds.) VCH Publishers, New York **1996**, pp. 67–117.
[165] D. E. Clark, C. W. Murray, J. Li, in *Reviews in Computational Chemistry, Vol. 11*, K. B. Lipkowitz, D. B. Boyd (Eds.) Wiley-VCH, New York **1997**, p. 67–126.
[166] M. A. Murcko, in *Reviews in Computational Chemistry, Vol. 11*, K. B. Lipkowitz, D. B. Boyd (Eds.) Wiley-VCH, New York **1997**, p. 1–66.
[167] R. A. Lewis, A. R. Leach, *J. Comput.-Aided Mol. Design* **1994**, *8*, 467–475.
[168] H.-J. Böhm, *J. Comput.-Aided Mol. Design* **1994**, *8*, 623–632.
[169] C. M. Ho, G. R. Marshall, *J. Comput.-Aided Mol. Design* **1993**, *7*, 3–15.
[170] R. C. Glen, A. W. R. Payne, *J. Comput.-Aided Mol. Design* **1995**, *9*, 181–191.
[171] J. B. Moon, W. J. Howe, *Proteins* **1991**, *11*, 314–320.
[172] A. Caflisch, M. Karplus, in *De Novo Design, Vol. 3*, K. Müller (Ed.) Escom, Leiden **1995**, pp. 51–84.
[173] W. P. Walters, M. T. Stahl, M. A. Murcko, *Drug Discovery Today* **1998**, *3*, 160–178.
[174] T. Hurst, *J. Chem. Inf. Comput. Sci.* **1994**, *1994*, 190–196.
[175] J. W. Godden, F. Stahura, J. Bajorath, *J. Mol. Graphics Mod.* **1999**, *16*, 139–143.
[176] M. D. Miller, R. P. Sheridan, S. K. Kearsley, D. J. Underwood, in *Methods in Enzymology, Vol. 241*, L. C. Kuo, J. A. Shafer (Eds.) Academic Press, San Diego **1994**, pp. 354–370.
[177] S. B. Shuker, P. J. Hajduk, R. P. Meadows, S. W. Fesik, *Science* **1996**, *274*, 1531–1534.
[178] I. Muegge, Y. C. Martin, P. J. Hajduk, S. W. Fesik, *J. Med. Chem.* **1999**, *42*, 2498–2503.
[179] R. L. DesJarlais, G. L. Seibel, I. D. Kuntz, P. S. Furth, J. C. Alvarez, P. R. Ortiz de Montellano, D. L. DeCamp, L. M. Babé, C. S. Craik, *Proc. Natl Acad. Sci. USA* **1990**, *87*, 6644–6648.
[180] B. K. Shoichet, R. M. Stroud, D. V. Santi, I. D. Kuntz, K. M. Perry, *Science* **1993**, *259*, 1445–1450.
[181] L. R. Hoffman, I. D. Kuntz, J. M. White, *J. Virology* **1997**, *71*, 8808–8820.
[182] D. A. Gschwend, W. Sirawaraporn, D. V. Santi, I. D. Kuntz, *Proteins* **1997**, *29*, 59–67.
[183] I. Massova, P. Martin, A. Bulychev, R. Kocz, M. Doyle, B. F. P. Edwards, S. Mobashery, *Bioorg. Med. Chem. Lett.* **1998**, *8*, 2463–2466.
[184] D. Tondi, U. Slomczynska, M. P. Costi, D. M. Watterson, S. Ghelli, B. K. Shoichet, *Chem. Biol.* **1999**, *6*, 319–331.
[185] P. Y. S. Lam, P. K. Jadhav, C. J. Eyerman, C. N. Hodge, Y. Ru, L. T. Bacheler, J. L. Meek, M. J. Otto, M. M. Rayner, Y. N. Wong, C.-H. Chang, P. C. Weber, D. A. Jackson, T. R. Sharpe, S. Erickson-Viitanen, *Science* **1994**, *263*, 380–384.
[186] Available Chemicals Directory, , MDL Information Systems Inc., San Leandro, CA.

[187] T. Toyoda, R. K. B. Brobey, G.-I. Sano, T. Horii, N. Tomioka, A. Itai, *Biochem. Biophys. Res. Comm.* **1997**, *235*, 515–519.
[188] P. Burkhard, U. Hommel, M. Sanner, M. D. Walkinshaw, *J. Mol. Biol.* **1999**, *287*, 853-858.
[189] E. Perola, K. Xu, T. M. Kollmeyer, S. H. Kaufmann, F. G. Prendergast, Y.-P. Pang, *J. Med. Chem.* **2000**, *43*, 401–408.
[190] A study of 100,000 compounds docked to the estrogen receptor: http://www.protherics.com/crunch/
[191] J. W. Godden, F. L. Stahura, J. Bajorath, *J. Comp. Chem.* **1999**, *20*, 1634–1643.
[192] B. E. Fink, D. S. Mortensen, S. R. Stauffer, D. A. Zachary, J. A. Katzenellenbogen, *Chem. Biol.* **1999**, *6*, 205–219.
[193] E. K. Kick, D. C. Roe, A. G. Skillman, G. Liu, T. J. A. Ewing, Y. Sun, I. D. Kuntz, J. A. Ellman, *Chem. Biol.* **1997**, *4*, 297–307.
[194] T. S. Haque, A. G. Skillman, C. E. Lee, H. Habashita, I. Y. Gluzman, T. A. Ewing, D. E. Goldberg, I. D. Kuntz, J. A. Ellman, *J. Med. Chem.* **1999**, *42*, 1428–1440.
[195] C. D. Carroll, T. O. Johnson, S. Tao, G. Lauri, M. Orlowski, I. Y. Gluzman, D. E. Goldberg, R. E. Dolle, *Bioorg. Med. Chem. Lett.* **1998**, *8*, 3203–3206.
[196] F. L. Stahura, L. Xue, J. W. Godden, J. Bajorath, *J. Mol. Graphics Mod.* **1999**, *17*, 1–9.
[197] H.-J. Böhm, D. W. Banner, L. Weber, *J. Comput.-Aided Mol. Design* **1999**, *13*, 51–56.
[198] C. Oefner, A. Binggeli, V. Breu, D. Bur, J.-P. Clozel, A. D'Arcy, A. Dorn, W. Fischli, F. Grüninger, R. Güller, G. Hirth, H. P. Märki, S. Mathews, M. Müller, R. G. Ridley, H. Stadler, E. Vieira, M. Wilhelm, F. K. Winkler, W. Wostl, *Chem. Biol.* **1998**, *6*, 127–131.
[199] F. R. Salemme, J. Spurlino, R. Bone, *Structure* **1997**, *5*, 319–324.
[200] P. L. Myers, *Curr. Op. Biotechnology* **1997**, *8*, 701–707.
[201] W. C. Lumma, K. M. Witherup, T. J. Tucker, S. F. Brady, J. T. Sisko, A. M. Naylor-Olsen, S. D. Lewis, B. J. Lucas, J. P. Vacca, *J. Med. Chem.* **1998**, *41*, 1011–1013.
[202] S. F. Brady, K. J. Stauffer, W. C. Lumma, G. M. Smith, H. G. Ramjit, S. D. Lewis, B. J. Lucas, S. J. Gardell, E. A. Lyle, S. D. Appleby, J. J. Cook, M. A. Holahan, M. T. Stranieri, J. J. Lynch, jr., J. H. Lin, I.-W. Chen, K. Vastag, A. M. Nayor-Olsen, J. P. Vacca, *J. Med. Chem.* **1998**, *41*, 401–406.
[203] A. Rockwell, M. Melden, R. A. Copeland, K. Hardman, C. P. Decicco, W. F. DeGrado, *J. Am. Chem. Soc.* **1996**, *118*, 10337–10338.
[204] H. Kubinyi, *Curr. Op. Drug Disc. Dev.* **1998**, *1*, 16–27.
[205] R. A. Fecik, K. E. Frank, E. J. Gentry, S. R. Menon, L. A. Mitschler, H. Telikepalli, *Med. Res. Rev.* **1998**, *18*, 149–185.
[206] R. E. Dolle, *Mol. Diversity* **1998**, *3*, 199–233.
[207] M. Leibl, *J. Comb. Chem.* **1999**, *1*, 3–24.
[208] R. E. Dolle, K. H. Nelson Jr, *J. Comb. Chem.* **1999**, *1*, 235–282.
[209] D. L. Kilpatrick, S. Watson, S. Ulhaq, *Comb. Chem. High-Throughput Screening* **1999**, *2*, 211–221.
[210] C. D. Floyd, C. Leblanc, M. Whitaker, *Prog. Med. Chem.* **1999**, *36*, 91–168.
[211] M. Whittaker, *Curr. Op. Chem. Biol.* **1998**, *2*, 386–396.
[212] R. A. Fecik, K. E. Frank, E. J. Gentry, S. R. Menon, L. A. Mitscher, H. Telikepalli, *Med. Res. Rev.* **1998**, *18*, 149–185.

Virtual Screening for Bioactive Molecules
Edited by Hans-Joachim Böhm and Gisbert Schneider

12 The Measurement of Molecular Diversity

Dimitris K. Agrafiotis, Victor S. Lobanov, Dmitrii N. Rassokhin, Sergei Izrailev

12.1 Introduction

After nearly a decade of frantic development, the dream of an "ideal" library remains elusive. Traditionally, combinatorial chemistry has been used primarily for lead generation, and molecular diversity has been the main guiding principle for the design of combinatorial and high-throughput screening experiments. However, there seems to be little agreement as to what molecular diversity is, let alone how it should be measured. This debate has been fueled by a series of so-called validation studies which advocate that diversity measures should be assessed on their ability to increase the hit-rate of high-throughput screening experiments. While no one argues that relevance is an important factor, the root of the problem stems from a lingering confusion between two largely unrelated concepts: molecular diversity and experimental design. Just like the notions of distance and similarity, the former has a clear intuitive interpretation albeit with multiple methods of measurement. On the contrary, the latter relates to the art of contemplating experiments, and represents a vague and multi-faceted process that involves a heterogeneous mix of chemistry, mathematics, experience, intuition, and corporate dynamics.

In this Chapter, we attempt to eliminate some of the ambiguity by providing a formal framework for the analysis of molecular diversity as it applies to library design. We do so by dissecting the problem into its constituent parts and addressing each one as an independent subject. Three main components are identified and discussed: 1. *diversity metrics*, which quantify diversity and are tightly coupled to the concept of molecular similarity and chemical distance, 2. *diversity spaces*, which refer to the precise molecular characteristics that are used to define chemical distance, and 3. *diversity sampling*, which refers to the methods used to identify a maximally diverse and representative subset of compounds from a large compound collection.

The remainder of this Chapter is organized as follows. Section 12.2 describes diversity metrics and methods to compute chemical distance, Section 12.3 discusses diversity spaces and reviews the descriptors that are currently used for diversity profiling, and Section 12.4 addresses diversity sampling and compound selection. Finally, Section 12.5 addresses the issue of experimental design and describes a general methodology for designing complex experiments based on the principles of multi-objective optimization developed by our group. While we have strived to provide a thorough account of what we consider the most important work in the field, the reader should be aware that the emphasis of this review is on methodology and not its applications in drug design. For additional information, the reader is referred to other recently published reviews by Agrafiotis [1], Agrafiotis *et al.* [2], and Martin *et al.* [3].

12.2 Diversity Metrics

A diversity metric is a function used to quantify the diversity of a set of compounds in some predefined chemical space. The term was originally introduced in an objective function-based approach for compound selection, as detailed in the previous Section of this Chapter [4,5], and in a sense represents a generalization of the concept of molecular similarity from individuals to collections. Note that herein the term *metric* refers to a general measure or function, and is not used in the strict mathematical sense. (In mathematics, a *metric* refers to a non-negative symmetric distance function that satisfies the triangular inequality.) In the following discussion, the terms diversity metric, diversity measure, and diversity function are used interchangeably.

Diversity metrics fall into three distinct classes: 1. *distance-based* methods which express diversity as a function of the pairwise molecular dissimilarities defined through measurement or computation, 2. *cell-based* methods which define it in terms of the occupancy of a finite number of cells that represent disjoint regions of chemical space, and 3. *variance-based* methods which quantify diversity based on the degree of correlation between the molecules' pertinent features. In the vast majority, these metrics encode the ability of a given set of compounds to sample chemical space in an even and unbiased manner, and are used to produce space-filling designs that minimize the size of unexplored regions known as "diversity voids".

12.2.1 Distance-Based Diversity Metrics

Distance-based metrics quantify the diversity of a set of compounds as a function of their pairwise dissimilarities. In most cases, these dissimilarities are derived indirectly by computing (or measuring) a set of characteristic features, and then combining them using some form of dissimilarity or distance measure [6,7]. Distance measures are divided into two main classes depending on whether the underlying features are continuous or binary (for further details on similarity measures and related topics, see also Chapter 4).

The most common distance measure for continuous properties is the Euclidean distance (Eq. 12.1), defined as:

$$d_{ij} = \sqrt{\sum_{k=1}^{K} |x_{ik} - x_{jk}|^2} \tag{12.1}$$

where x_{ik} represents the k-th feature of the i-th molecule, and K is total number of features. Other frequently used distance measures include the L_1 norm or Manhattan metric (Eq. 12.2), which represents the sum of the absolute parametric differences:

$$d_{ij} = \sum_{k=1}^{K} |x_{ik} - x_{jk}| \tag{12.2}$$

and the L_∞ norm or ultrametric which represents the maximum absolute parametric difference:

$$d_{ij} = \max_{k=1}^{K} |x_{ik} - x_{jk}| \tag{12.3}$$

In fact, all three of these distances represent special cases of the generalized Minkowski metric, defined as:

$$d_{ij} = \left[\sum_{k=1}^{K} |x_{ik} - x_{jk}|^r \right]^{1/r} \tag{12.4}$$

Indeed, the Manhattan, Euclidean, and ultrametric distances are derived from Eq. (12.4) by substituting r with 1, 2, and ∞, respectively. A somewhat different class of distance measures used predominantly with descriptors that represent counts or sums include:

$$d_{ij} = 1 - \frac{2 \cdot \sum_{k=1}^{K} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{K} x_{ik} + \sum_{k=1}^{K} x_{jk}} \tag{12.5}$$

and:

$$d_{ij} = 1 - \frac{\sum_{k=1}^{K} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{K} \max(x_{ik}, x_{jk})} \tag{12.6}$$

where x_{ik} is, again, the value of the k-th descriptor in the i-th compound, and K is the total number of descriptors. Both indices range from 0 to 1, with 0 indicating complete identity and 1 indicating that the two structures have nothing in common. Eq. (12.5) has been used by Kearsley [8] to compare molecules based on atom pairs and topological torsion descriptors, and Eq. (12.6) by Martin [9] who used it to compute the similarities of combinatorial side chains based on atom layers of increasing radius emanating from the side chain attachment points on the scaffold (see Section 12.3). In the latter case, the molecular features were actually fixed-size matrices rather than vectors, but this representation can be easily converted to vectorial form (Eq. 12.6) by concatenating the columns of the descriptor matrix.

For binary features, the most frequently used distance functions are the normalized Hamming distance:

$$H = \frac{|XOR(x, y)|}{N} \tag{12.7}$$

where x and y are two binary sets (encoded molecules), XOR is the bitwise "exclusive or" operation (a bit in the result is set if the corresponding bits in the two operands are different), and N is the number of bits in each set. The result, H, is a measure of the number of bits that are dissimilar in x and y; the Tanimoto or Jaccard coefficient:

$$T = \frac{|AND(x, y)|}{|IOR(x, y)|} \tag{12.8}$$

where *AND* is the bitwise "and" operation (a bit in the result is set if both of the corresponding bits in the two operands are set) and *IOR* is the bitwise "inclusive or" operation (a bit in the result is set if the either of corresponding bits in the two operands are set). The result, T, is a measure of the number of features shared by two molecules relative to the one they *could* have in common, and the Dice coefficient:

$$D = \frac{2|AND(x,y)|}{|x| + |y|} \tag{12.}$$

Another popular metric is the Euclidean distance (Eq. 12.1), which, in the case of binary sets, can be recast in the form of Eq. (12.10):

$$E = \sqrt{N - |XOR(x, NOT(y))|} \tag{12.10}$$

where $NOT(x)$ denotes the binary complement of x, and the expression $|XOR(x, NOT(y))|$ represents the number of bits that are identical in x and y (either 1 or 0). The Euclidean distance is a good measure of similarity when the binary sets are relatively rich, and is mostly used in situations in which similarity is measured in a relative sense.

Once a distance function is defined, the diversity of a compound collection can be computed in a variety of ways. Perhaps the most commonly used diversity measure is the minimum pairwise distance defined as:

$$D_1(C) = \min_{i<j} d_{ij} \tag{12.11}$$

where d_{ij} is the distance between the i-th and j-th compounds in C. The first implicit application of this function in library design was reported by Lajiness *et al.* [10] and formed the basis of a greedy selection algorithm known as maxmin (see Section 12.4). This function belongs to a general family of functions [11] which also include the power sum:

$$D_2(C) = \left[\frac{2}{N(N-1)} \sum_{i<j} d_{ij}^p \right]^{1/p} \tag{12.12}$$

where N is the number of compounds in C, and p is a user-defined exponent, and the product:

$$D_3(C) = \left[\prod_{i<j} d_{ij} \right]^{2/N(N-1)} \tag{12.13}$$

Both functions are normalized to allow meaningful comparisons between sets of different cardinality. As is evident from Eq. (12.11), $D1$ does not explicitly depend on the number of compounds in C.

The major disadvantage of D_1 is that it depends on a single inter-molecular distance. As illustrated in Figure 12.1, even a single close contact or duplicate is sufficient to destroy the diversity of the entire set (the two subsets highlighted in Figure 12.1a,b are equally diverse according to this metric). This is counter-intuitive, and makes this function hard to optimize during subset selection, particularly when additional constraints are imposed on the design.

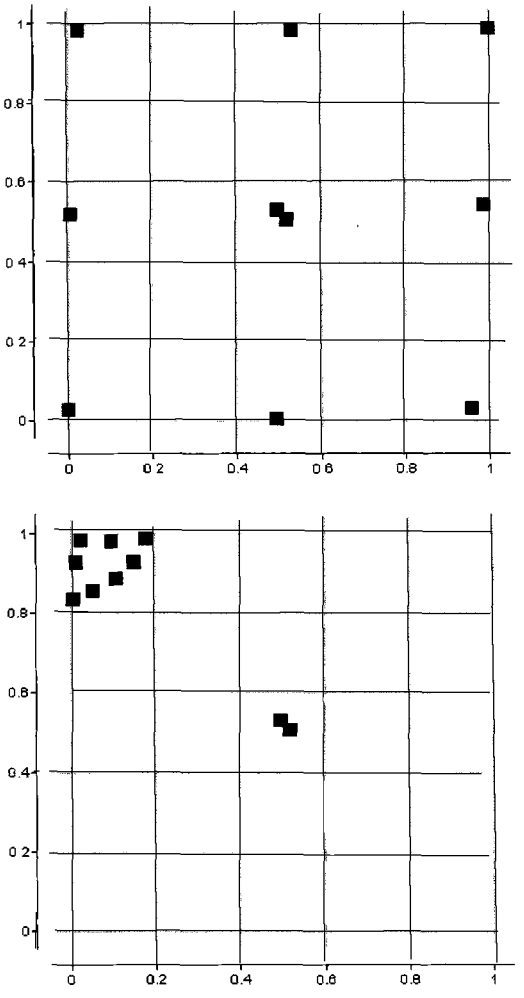


Figure 12.1. Minimum dissimilarity as a measure of molecular diversity. According to Eq. (12.11), the two data sets are equally diverse.

To address this problem, we proposed an alternative measure based on the average nearest neighbor distance [4] defined as:

$$D_4(C) = \frac{1}{N}\sum_i \min_{j \neq i} d_{ij} \tag{12.14}$$

This function is much less sensitive to outliers, can discriminate more effectively between the various ensembles, and is much easier to optimize in the context of compound selection (see below). The type of designs it produces is illustrated in Figure 12.2, which shows a selection of the 100 most diverse points from a hypothetical library of 10000 points uniformly distributed in the unit square. We have found this dataset to be very useful for testing the "san-

ity" of the various diversity algorithms. When D_4 is used, the space is covered in an even and unbiased manner, and the selection is consistent with our general notion of spread.

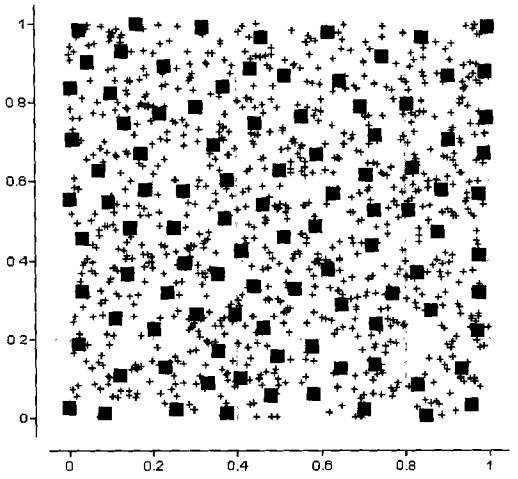


Figure 12.2. Diversity selection based on the average nearest neighbor distance (Eq. 12.14).

However, the metric is not flawless, particularly for datasets involving multiple, well separated clusters. In such cases, D_4 measures the intra-cluster (nearest neighbor) separations and does not take into account the inter-cluster distances, that is, the relative separation between them. This problem is illustrated in Figure 12.3. According to Eq. (12.14), the two sets depicted in Figure 12.3a,b are equally diverse. However, this problem manifests itself only in extreme, pathological situations, and is rarely encountered in subset selection due to the vast number of possibilities [4].

The second major disadvantage of these functions is their quadratic dependence on the number of compounds in C, which renders them virtually useless for the analysis of large collections. However, we recently demonstrated that when the dimensionality of the space is relatively small (<10), the nearest neighbor computation in D_4 can be carried out in an efficient manner using a combinatorial data structure known as a k-dimensional (or k-d) tree, resulting in an algorithm of $O(N \log N)$ [12]. This algorithm achieves computational efficiency by first organizing all the points in C into a k-d tree, and then performing a nearest neighbor search for each point in the set using a branch-and-bound approach. A k-d tree is a generalization of a binary tree used for sorting and searching. The tree partitions a set of multi-dimensional points into smaller subsets using the coordinates of the points as discriminators. Every node in the tree represents a subset of C and a partitioning of that subset. Each non-terminal node partitions the points associated with that node according to their position with respect to a k-dimensional hyper-plane; the points that lie to the left of that hyper-plane are stored on the left sub-tree, while those that lie on the right are stored on the right sub-tree (Figure 12.4).

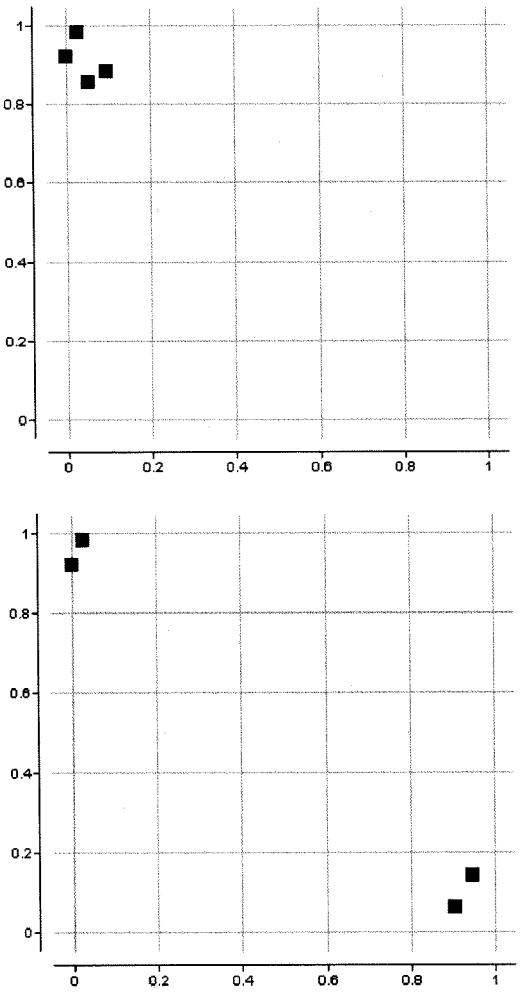


Figure 12.3. Average nearest neighbor distance as a measure of molecular diversity. According to Eq. (12.14), the two data sets are equally diverse.

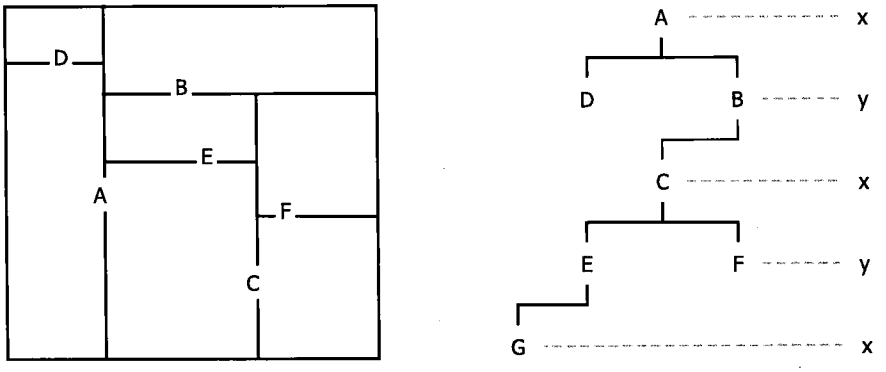


Figure 12.4. Data space partitioning using a *k*-dimensional tree.

The tree can be elaborated to any depth, ending in terminal nodes which can hold up to a predetermined number of points. In our implementation, the partition plane associated with a particular node is perpendicular to a coordinate axis, which is determined by the depth of that node in the tree. The discriminator is chosen by cycling through the coordinates in a strictly alternating sequence. The tree is constructed recursively in a manner similar to a binary tree. Once the construction of the tree is completed, nearest neighbor searching is conducted by descending the tree, choosing at each node to investigate either the left or right child according to whether the query lies to the left or to the right of the respective partition plane. Some backtracking may be necessary, but the process completes in a time that scales logarithmically with the size of the dataset. This algorithm is not applicable to the power sum and product functions (Eqs. 12.12 and 12.13), but it is applicable to the minimum pairwise dissimilarity (Eq. 12.11) which can be recast in the form:

$$D_1(C) = \min_i \min_{j \neq i} d_{ij} \tag{12.15}$$

An alternative way to reduce the quadratic complexity of the problem is to use the mean pairwise inter-molecular dissimilarity [13] defined as:

$$D_5(C) = \frac{1}{N^2} \sum_i \sum_j d_{ij} \tag{12.16}$$

which is closely related to the power sum when p=1. Although with certain distance measures the use of this function can be algorithmically beneficial, it is perhaps one of the least effective means of measuring spread. The reason is the fact that the metric is not really a function of all the intermolecular dissimilarities, but a function of the distances of the samples from their center, $\frac{1}{N}\sum_i(\boldsymbol{x}_i)$ [14]. In the case of squared Euclidean distances:

$$d_{ij} = \sum_k (x_{ik} - x_{jk})^2 \tag{12.17}$$

the sum in Eq. (12.16) can then be written as:

$$\begin{aligned} & \sum_i \sum_j d_{ij} \\ = \quad & \sum_i \sum_j \sum_k (x_{ik} - x_{jk})^2 \\ = \quad & 2N \sum_i \sum_k \left[x_{ik} - \frac{1}{N} \sum_j (x_{jk}) \right]^2 \\ = \quad & 2N \sum_i d(x_i, - \frac{1}{N} \sum_j (x_j)) \end{aligned} \tag{12.18}$$

where x_{ik} represents the k-th value of the property vector associated with the i-th molecule. This property has dire consequences in diversity profiling, as it tends to favor designs that contain a large number of duplicates located at the extremes of the feature space. In fact, this is a general tendency that is not limited to Euclidean distances. The same problem arises when the cosine coefficient is used as a measure of molecular similarity, as proposed by the original authors [13]. The cosine coefficient is defined as the cosine of the angle formed between two molecular property vectors:

$$\cos(i,j) = \frac{\sum_{k=1}^{K} x_{ik} \cdot x_{jk}}{\sqrt{\sum_{k=1}^{K} x_{ik}^2 \sum_{k=1}^{K} x_{jk}^2}} \tag{12.19}$$

Accordingly, the distance or dissimilarity between two compounds is defined as:

$$d_{ij} = 1 - \cos(i,j) \tag{12.20}$$

When Eq. (12.20) is substituted in Eq. (12.16), D_5 can be expressed as a dot product between the centroids of the property vectors, x_i, which requires a single pass through the dataset. However, this linear order comes at a high price. Based on simple trigonometry, it was demonstrated that this method has a strong tendency to sample the principal axes of the feature space and produce highly redundant designs [12]. This is illustrated in Figure 12.5 using the same artificial dataset of 10000 points uniformly distributed in the unit square. The squares represent the 100 most "diverse" points according to Eq. (12.16), selected using an annealing optimization procedure described in Section 12.5.

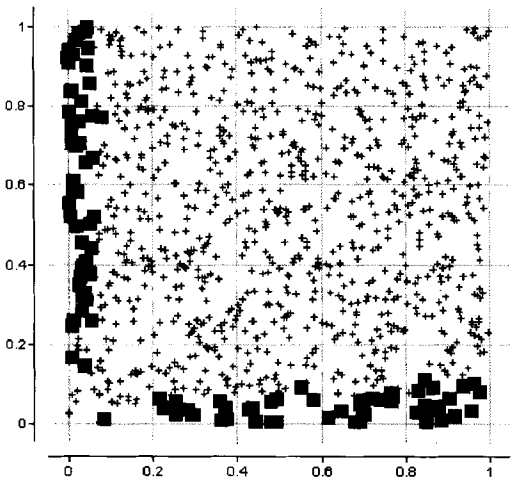


Figure 12.5. Diversity selection based on the mean intermolecular distance and the cosine similarity coefficient (Eqs. 12.16 and 12.19).

Mount *et al.* [14] have recently proposed a diversity metric based on the concept of a minimum spanning tree (The term "minimum spanning tree" is short for "minimum-weight spanning tree".). Given a connected undirected graph, $G = (V, E)$, where V is a set of vertices and E is the set of all possible pairwise interconnections between them, a spanning tree is any acyclic subset $T \subseteq E$ that interconnects all of the vertices in V. If every edge $(u, v) \in E$ is associated with a weight $w(u, v)$, a minimum spanning tree is a spanning tree with the least possible weight, as measured by:

$$w(T) = \sum_{(u,v)\in T} w(u, v) \tag{12.21}$$

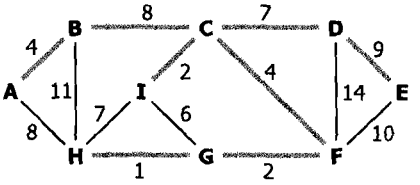


Figure 12.6. Minimum spanning tree.

An example of a minimum spanning tree is shown in Figure 12.6. If the vertices V correspond to a set of compounds, C, and the weights $w(u, v)$ are their corresponding dissimilarities, then the diversity of C can be defined as the weight of the minimum spanning tree:

$$D_6(C) \min_T(w(T)) \tag{12.22}$$

The minimum spanning tree can be determined using Kruskal's or Prim's algorithms which can run in time $O(E \log V)$ using conventional binary heaps, or in the case of Prim's algorithm in order $O(E+V \log V)$ using Fibonacci heaps, which is better if $|V| << |E|$. Thus, in the case of molecular diversity where we are dealing with fully connected graphs, the algorithm scales to the square of the number of compounds considered, and its use is therefore limited to datasets of moderate size. Moreover, while the minimum spanning tree addresses some of the problems associated with the other distance-based metrics, it is not monotonic and tends to produce results that are very similar to those obtained with D_4, albeit at a higher computational cost.

The last method discussed in this section is based on Shannon's information theory. This method is rooted on the fundamental premise that diversity is synonymous to information content, which can be quantified using Shannon's classical entropy equation [15]. Under this formalism, every collection of compounds represents a finite number of distinguishable species whose "distinguishability" can be described as a function of their mutual dissimilarity (hence its classification as a distance-based technique). The more distinguishable the species, the greater their information content. To cast this idea in the form of an equation, it was proposed that the diversity (or information content) of a system be described as a function of the entropy:

$$D_7(C) = S_{max} - S \tag{12.23}$$

where:

$$S = -\sum_{i=1}^{N}\sum_{j=1}^{N} p_{ij} \ln p_{ij} \tag{12.24}$$

and N is the total number of compounds, p_{ij} is the probability of finding the i-th individual in the j-th species (given by some function of their dissimilarity), and S_{max} is the maximum entropy of the system. The probabilities, p_{ij}, are computed from a molecular similarity table.

While the use of information theory seems like a natural choice, the actual implementation suffers from a number of drawbacks. In a recent article [16] we reported that a strict application of this approach produced extremely unbalanced designs, and clustered points at maximum separation along the diagonal of the feature space. This is illustrated in Figure 12.7, which shows a selection of 100 points that maximize D_7 in our uniformly distributed dataset.

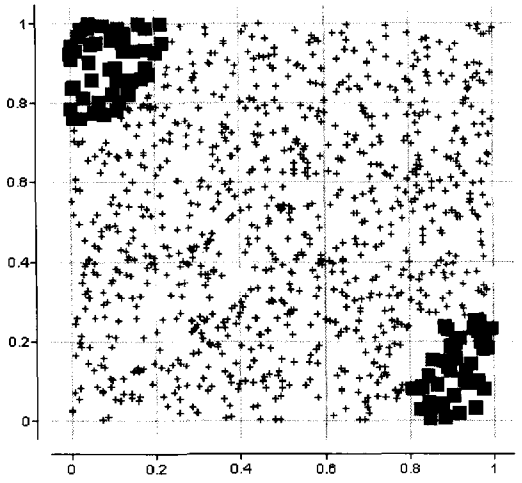


Figure 12.7. Diversity selection based on Eq. (12.23).

We believe that this is due to the use of the wrong type of "information", and to the implicit assumption that ideal designs should be equiprobable (i.e. that the pairwise intermolecular dissimilarities should be as uniform as possible). In a private communication, the author suggested that our results could be an artifact of the similarity measure used in our study, but a detailed response has yet to appear in print.

12.2.2 Cell-Based Diversity Metrics

Cell-based methods attempt to quantify diversity by dividing chemical space into hyper-rectangular regions and measuring the occupancy of the resulting cells. A putative advantage of these methods is that, unlike distance-based approaches, they can encode absolute position

in space in addition to inter-molecular distance [17]. However, a more practical advantage is the fact that diversity estimation and library comparisons can be carried out significantly faster, and are not plagued by the quadratic complexity of distance-based algorithms.

The most intuitive cell-based diversity measure is simply the number of cells occupied by a design:

$$D_8(C) = \sum_{i=1}^{M} \delta_i \tag{12.25}$$

where δ_i is 1 if the i-th cell is occupied and 0 if it is not, and M is the total number of cells. This is a measure of absolute diversity, and can be used to determine the diversity of any library regardless of size. However, this measure does not take into account the clustering of the dataset, and can be very poor in discriminating collections with similar span but very different distributions (Figure 12.8).

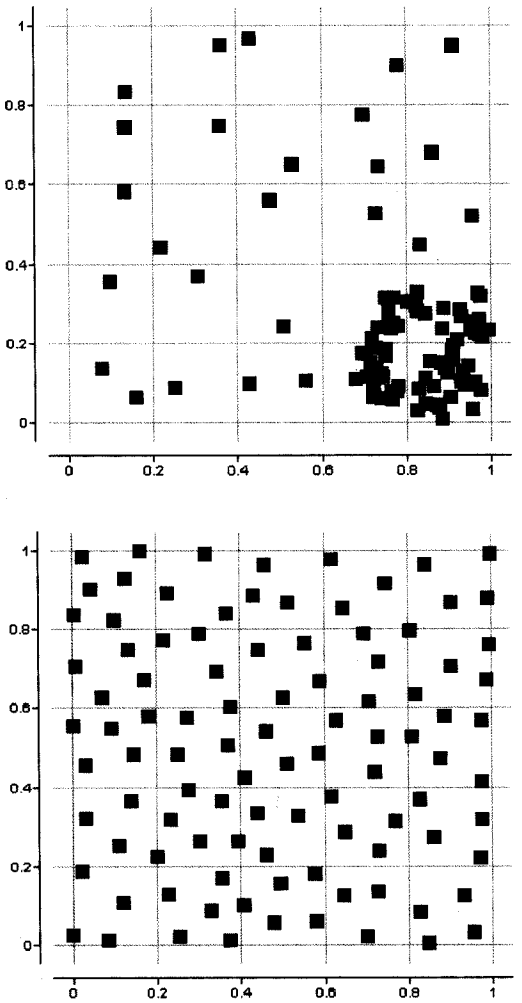


Figure 12.8. Bin occupancy as a measure of molecular diversity. According to Eq. (12.25), the two datasets are equally diverse.

Four different diversity metrics have been proposed for measuring diversity in a relative sense [18]. These include the cell-based fraction:

$$D_9(C) = \frac{N_C}{N_R} \tag{12.26}$$

where N_C is the number of cells occupied by C (typically, a subset of a library), and N_R is the number of cells occupied by the reference set (typically, the entire library); the cell-based χ^2:

$$D_{10}(C) = \sum_i (N_i - N^*)^2 \tag{12.27}$$

the cell-based entropy:

$$D_{11}(C) = -\sum_i (N_i \cdot \log(N_i)) \tag{12.28}$$

and the cell-based density:

$$D_{12}(C) = -\sum_i (N_i \cdot \log(N_i / M_i)) \tag{12.29}$$

where N_i is the number of compounds in the i-th cell for the subset C, M_i is the number of compounds in the i-th cell for the reference set, N^* is the average number of compounds per cell expected for the subset, and i is an index over all cells occupied by the subset. The cell-based fraction is independent of density and measures what fraction of the data space is occupied by the subset relative to the entire library. The χ^2 and entropy metrics are maximized when the distribution of compounds is uniform among the occupied cells, while the density metric favors designs that exhibit the same pattern of occupancy as the entire library.

Unfortunately, cell-based approaches are applicable to data spaces of very modest dimensionality (typically no more than five or six). Although there is no algorithmic reason why they shouldn't be used with higher-dimensional spaces (cell-based methods can be efficiently implemented using a sorted data structure and do not require sparse matrices which are memory intensive), one should keep in mind that the number of hypercube partitions created by binning each axis into r intervals is r^d, where d is the number of dimensions. For a mere ten dimensions and five bins per dimension, the number of hypercube partitions is $\sim 10^7$, while with 20 dimensions that number increases to $\sim 10^{14}$. Thus, when the dimensionality of the space is high, only a tiny fraction of the feature space will be occupied by any given library, even with the most moderate bin resolution (Pearlman recommends that the minimum cell occupancy should be approximately 10–15% [19]). Moreover, the results can be very sensitive to grid resolution and the presence of outliers. If the bins are too large, the method loses its discriminating value; if they are too small, it tends to focus on local differences and lose sight of more important global trends in the data, and boundary effects can become more of a problem. Cell-based methods were introduced by Cummins [20] and were subsequently popularized by Pearlman [17]. A thorough discussion of the effect of bin resolution and an elegant approach for the removal of outliers can be found in [20].

A special case of cell-based methods are the diversity measures proposed for binary fingerprints. Unlike continuous properties, binary descriptors such as structural keys and hashed fingerprints can be compared using fast binary operations to render quick estimates of molecular similarity, diversity and complementarity. The most common example of a diversity measure applied to binary descriptors is the binary union ("inclusive OR"). This can be exploited in a number of different ways, elegant examples of which can be found in the work of Shemetulskis [21], Pickett [22], and Davies [23].

12.2.3 Variance-Based Diversity Metrics

In this Section, we describe a series of variance-based methods rooted on the principles of optimal experimental design [24]. According to these methods, a selection of molecules can be thought of a series of experiments designed to probe some predefined molecular property space. The concepts are illustrated with a simple example involving four hypothetical observations and two independent variables. Consider the matrix:

Run	x_1	x_2
1	1	1
2	1	1
3	−1	−1
4	−1	−1

where x_1 and x_2 represent two different design variables. Because the two columns are identical to each other, this design does not allow one to test, independently, the statistical significance of the two variables' contribution to the prediction of the dependent variable (which, in this case, would be some measure of the biological activity of the respective compounds). Indeed, the correlation matrix of the two variables is 1:

	x_1	x_2
x_1	1	1
x_2	1	1

Ideally, one would prefer to run the experiment so that the two variables are varied independently from each other:

Run	x_1	x_2
1	1	1
2	1	−1
3	−1	1
4	−1	−1

In this case, the two variables are uncorrelated (i.e. orthogonal) and their correlation matrix becomes:

	x_1	x_2
x_1	1	0
x_2	0	1

A simple method to determine the redundancy between the rows and columns of a square matrix is to compute the value of the determinant of that matrix. In the example above, the determinant of the first correlation matrix computed from completely redundant variable settings is 0, while that of the second matrix where the variables are orthogonal is 1. This basic relationship extends to larger design matrices, that is, the more redundant the columns of the design matrix, the closer to 0 is the determinant of the correlation matrix of these variables. This brings us to the definition of the *D*-optimal criterion as a measure of molecular diversity:

$$D_{13}(C) = |X'X| \tag{12.30}$$

where X is the design matrix, X' is the transpose of that matrix, and $|.|$ denotes the determinant. For a collection of N compounds, the design matrix is an $N \times M$ matrix where each of the M columns represents a distinct molecular feature (descriptor). Thus, a maximally diverse set is one that maximizes the value of D_{13}.

A closely-related measure is the *A*-optimal or trace-optimal criterion which is based on the trace of the inverse of the cross-product matrix (also known as information matrix), $X'X$:

$$D_{14}(C) = trace(X'X)^{-1} \tag{12.31}$$

where *trace* stands for the sum of the diagonal elements. In effect, this criterion maximizes the diagonal elements of $X'X$, while minimizing the off-diagonal elements. Other criteria have also been used such as *I*-optimality, *S*-optimality, etc. [24].

The use of *D*-optimality for experimental design has a very long history in chemometrics. Martin [9] has adopted this criterion in the design of combinatorial libraries, but the measure has not been widely used outside the Chiron group. Unlike the distance-based methods, the *D*-optimal criterion maximizes volume in *covariance* space. While the method is statistically sound, it is not quite clear just how much spread is sacrificed to achieve a higher rank. Hassan *et al.* [11] have recently shown that the loss may in fact be quite substantial. Their analysis suggested that the *D*-optimal criterion favors the extremes of the feature space and tends to ignore the central region. This is particularly true when the number of compounds far exceeds the dimensionality of the space. Moreover, *D*-optimality is model-dependent, that is, it assumes a linear response surface of the dependent variable. While the underlying model can be tailored by introducing additional columns that represent higher order terms (such as the squares, cubes or cross terms of the original variables), it is not at all clear how these additional parameters should be chosen. Hassan's analysis was based on a linear model which may partially account for the redundancy that he observed (when the model contains fewer terms than the number of observations, duplicates introduce an estimate of uncertainty in the response surface). It is possible that the inclusion of higher-order terms will partially compensate for this redundancy, but this remains to be demonstrated. So far, the use of *D*-optimal design has been limited to reagent selection.

12.3 Diversity Spaces

As is evident from the preceding discussion, the computation of molecular diversity requires a numerical representation of chemical space. This can be defined using a set of molecular properties that are pertinent to the application at hand, or by means of a pairwise similarity coefficient relating the molecules to each other. For the sake of completeness, we provide a brief overview of the most popular descriptors used for diversity analysis, but we must stress that the account is by no means exhaustive. A more thorough discussion can be found in references [1] and [2].

The abundance of descriptors proposed for diversity analysis fall under three broad categories: 1. *two-dimensional* descriptors which encode the topology of the molecular graph and are derived directly from the connection table, 2. *three-dimensional* descriptors which are based on the three-dimensional structure of the molecule, and 3. *physicochemical* descriptors which represent physicochemical and electronic properties of interest.

12.3.1 Two-Dimensional Descriptors

Two-dimensional descriptors are derived directly from the molecular graph. The most prominent members of this family are the molecular connectivity indices popularized by the availability of Kier and Hall's MolconnX software suite [25]. These indices capture the constitutional, branching, and ring character of a molecule, and are attractive for quantifying molecular diversity because they are inexpensive to compute, and have been validated through years of use in the field of structure-activity correlation [26]. Most of these descriptors are based on the adjacency matrix (e.g. the total adjacency index, the Zagreb group indices, the Randic connectivity index, the Platt index, the compatibility code, the largest eigenvalue index, etc.), the topological distance matrix (e.g. the Wiener index, the polarity number, the distance sum, the Altenburg polynomial, the mean square distance, the Hosoya index, the distance polynomial, etc.), or information theory (the Shannon index, the chromatic information index, the orbital information index, the topological information superindex, the electropy index, the Merrifield and Simmons indices, etc.).

Another frequently referenced set of descriptors are the atom pairs and topological torsions [8,27]. Atom pairs are patterns of the form a_i–d–a_j, where a_i and a_j are the types of atoms i and j, respectively, and d is the topological distance between them (the number of bonds along the shortest path connecting these atoms). Topological torsions are of the form a_i–a_j–a_k–a_m, where i, j, k, and m are sequentially bonded atoms, and a_i is again the type of the i-th atom. Atoms can be classified into a relatively small number of types based on their atomic number, number of neighbors, number of π-electrons, binding properties, atomic log$P$ contribution, partial atomic charge, hydrogen-bonding potential, polar and hydrophobic character, and many others.

Moreau [28] has proposed the use of an autocorrelation function defined as:

$$A(d) = \sum_{i,j} p_i p_j \tag{12.32}$$

where p_i and p_j represent the values of an atomic property of the i-th and j-th atoms, respectively, and d is the topological distance between them measured in bonds along the shortest path. The function has the desirable property that the molecule can be encoded in a fixed-length vector of small rank, not matter how large it may be. Typically, only paths of length 2–8 are considered. Atomic properties that have been encoded using this method include volume, electronegativity, hydrogen bonding character and hydrophobicity. Topological autocorrelation vectors were also used by Gasteiger [29] as input to a Kohonen network which successfully separated dopamine from benzodiazepine receptor agonists, even when these compounds were embedded in a large and diverse set of chemicals extracted from a commercial supplier catalog. Gasteiger [30,31] has also extended the concept to three dimensions, by introducing a spatial autocorrelation vector based on properties measured on the molecular surface (see below).

A closely related set of descriptors are atom layers developed by Martin *et al.* [9] in an attempt to account for the topological distribution of chemical features around a combinatorial core. Atom layers are based on the assumption that atoms of a side chain which are close to the point of attachment to the core contribute differently to binding than those that are more distant. They are constructed by summing a given property over all atoms in a substituent at a given number of bonds away from the attachment point. Any set of atomic properties like the ones described above can be considered. Of all the descriptors presented, these are the least general and can only be used for reagent evaluation.

A more poorly understood set of descriptors are the B-Cut values developed by Pearlman [17]. These descriptors represent an extension of Burden's molecular identification numbers [32] which are based on the two lowest eigenvalues of a symmetric $N \times N$ matrix representing the connection table of a molecule. This matrix is constructed by placing the atomic numbers along the diagonal, and assigning to the off-diagonal elements a value that is related to the bond order of the respective atoms. Pearlman extended this approach to include atomic properties deemed significant in protein-ligand binding, such as atomic charge, polarizability and hydrogen bonding character, etc. Optimal combinations of on- and off-diagonal properties were selected by their ability to produce a uniform distribution of molecules in the property space, and were used to construct a low-dimensional chemical space that was amendable to cell-based analysis.

Finally, a very common set of descriptors are substructure keys and hashed fingerprints which are both of binary nature. Substructure keys encode molecular structures as bit-strings, where each bit indicates the presence or absence of a particular structural feature or pattern. Typical features might include the number of occurrences of a particular element (e.g. the presence of at least 1, 2, or 3 nitrogen atoms), electronic configurations or atom types (e.g. sp^2 nitrogen or aromatic carbon), common functional groups such as alcohols, amines etc., and ring systems. Substructure keys were originally developed for rapid database searching, but have proven very effective in similarity and diversity applications. Hashed fingerprints are bit-strings derived directly from the connection table and were also developed primarily for database searching. They differ from structural keys in that they do not depend on pre-selected structural fragments to perform the bit assignment. Instead, every pattern in the molecule up to a predefined path length is systematically enumerated and serves as input to a hashing algorithm that turns "on" a small number of bits at pseudo-random positions in the bit-string. Experience has shown that the similarity between two fingerprints is a good indi-

cator of the similarity between the respective structures. A number of studies have shown that fingerprints and substructure keys are equally effective for diversity profiling [33,34].

12.3.2 Three-Dimensional Descriptors

The origins of biological specificity of receptors and enzymes lie in their recognition of shapes and electronic properties rather than specific atom types or substructures, and many have argued that these properties should be directly incorporated into the diversity analysis. Naturally, to obtain such information one needs a three-dimensional structure of the target molecule. Earlier methods relied on a single low-energy conformation, determined either crystallographically or computationally using a fast model-building utility [35]. These methods were subsequently modified to account for conformational flexibility.

One method of recording 3-D information extends the substructure key definition. 3-D structural keys are binary sets designed to function as screens for fast 3-D database searching applications [36,37]. Most screens involve two-center keys that represent the distances or angles between two structural features such as particular atom types, centroids of aromatic rings, ring normals, and attachment points of functional groups. Each pair of features is assigned a fixed number of bits in the key, each bit corresponding to a particular distance or angle range. The key is computed by mapping the features of interest onto the target molecule and recording the pairwise distances by turning on the corresponding bits in the bitmap. While capturing some of the structural information, 3-D structural keys are limited by the finite number of features that can be effectively stored in a bitmap, as well as by poor representation of shape and chirality.

3-D pharmacophore keys is a related class of descriptors introduced by Sheridan and coworkers [38]. Pharmacophore keys are 3-D structural keys whose structural features include macromolecular recognition sites, e.g. charged centers, hydrogen bond donors and acceptors, hydrophobic centers, and aromatic ring centers. The pharmacophore itself is a set of three or four centers forming a triangle or tetrahedron. To generate the key, the pharmacophores matching a particular conformation or an ensemble of conformations are mapped onto appropriate bits in the bitmap. These descriptors have been used extensively by the groups at Chemical Design [37], Rhone-Poulenc [22], and Abbott [24].

Another class of descriptors, developed by Cramer *et al.* [39], is based on the steric fields of single side chain conformers "topomerically" aligned around a common combinatorial core. This alignment tries to find a representative conformation for each side chain attached to a particular variation site on a combinatorial template. First, a model-building routine generates a low energy conformation, which is then fitted as a rigid body onto the template using least-squares minimization. After that, the bond torsions are adjusted one at a time starting from the bond closest to the template, according to a simple set of topological precedence rules. Once the alignment is complete, the steric field of the side chain is calculated using a CoMFA-like approach. These fields can be used to compute a similarity index between two compounds from the root of the squares of the differences in steric field values summed over all lattice points in the CoMFA region, or another equivalent distance function. While the assumption that template-based combinatorial compounds can be aligned with respect to their common core is a reasonable one, it is not as clear how this technique could be used to com-

pare multiple libraries or heterogeneous compound collections, or even libraries where the template is relatively small compared to the side chains, or is itself a variable site.

Spatial autocorrelation vectors constitute a condensed representation of the distribution of physico-chemical properties on the molecular surface. These descriptors were introduced by Gasteiger *et al.* [30] and are defined as:

$$A(d_l, d_u) = \frac{1}{N}\sum_{i,j} p_i p_j, \quad d_l \leq d_{ij} < d_u \tag{12.33}$$

where p_i and p_j are the values of the electrostatic potential at two randomly chosen points, i and j, on the molecular surface, d_{ij} is the distance between the points, and N is the total number of distances in the interval $[d_l, d_u]$. In fact, any property, p, that can be mapped on the molecular surface can be used in the equation. Thus, auto-correlation vectors compress shape and electronic information into a fixed-length vector, resulting in a compact, informative molecular descriptor with direct bearing on biological activity.

12.3.3 Physicochemical and Electronic Descriptors

Molecular property descriptors capture steric, electronic, and lipophilic characteristics that play a critical role in the transport and binding of a drug to its target. These properties can be calculated using standard molecular modeling and quantum mechanical packages and include molecular weight, octanol–water partition coefficient (log$P$), HOMO and LUMO energies, total energy, heat of formation, ionization potential, number of filled orbitals, standard deviation of partial atomic charges and electron densities, the dipole moment, van der Waals volume and surface area, molar refractivity, and many others. Molecular property descriptors have been extensively reviewed by Kubinyi [40], and have been used for diversity profiling by Willett *et al.* [7], Martin *et al.* [9], Lewis *et al.* [41], Brown *et al.* [34,42], and many others.

12.3.4 Dimensionality Reduction

Despite their diverse origin, many of the descriptors described above are highly correlated. In fact, the more descriptors are used to describe the data, the greater the likelihood that they are correlated. Redundant variables affect chemical distance and have the tendency to overemphasize certain molecular characteristics at the expense of others. Moreover, high-dimensional representations can have a significant impact on speed of computation and can even limit the number of available analysis options (the declining performance of k-d trees and the increasing complexity of cell-based analysis represent two such examples). It is thus desirable to reduce the dimensionality of the space by eliminating dimensions that add very little to the overall picture. Four methods have traditionally been used to perform this task: 1. principal component analysis, 2. factor analysis, 3. multi-dimensional scaling, and 4. non-linear mapping.

Principal component analysis takes as input a set of vectors described by partially cross-correlated variables and transforms it into one described by a smaller number of orthogonal

variables (principal components) without a significant loss in the variance of the data [43]. Principal components correspond to the eigenvectors of the covariance matrix, m_{ij}, a square symmetric matrix that contains the variances of the variables in its diagonal elements and the covariances in its off-diagonal elements:

$$m_{ij} = m_{ji} = \frac{1}{N}\sum_{k=1}^{N}(x_{ki} - \mu_i)(x_{kj} - \mu_j) \tag{12.34}$$

where μ_i is the mean value of the i-th variable:

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij} \tag{12.35}$$

and N is the number of points in the dataset. The eigenvalues of this matrix represent the variances of the principal components. PCA reduces the dimensionality by filtering out the principal components which contribute the least to the variance of the data (i.e. those with the smallest eigenvalues). Once the covariance matrix is diagonalized, the original data can be transformed by:

$$\boldsymbol{x}' = \boldsymbol{V}^T\boldsymbol{x} \tag{12.36}$$

where $\boldsymbol{V}^T$ is the transpose of the filtered eigenvector matrix, $\boldsymbol{x}$ is the input vector in the original coordinate frame, and $\boldsymbol{x}'$ are the coordinates of that sample in the transformed frame. Thus, the components of $\boldsymbol{x}'$ are linear combinations of the original, cross-correlated variables.

Factor analysis is a closely related technique that attempts to explain the correlations between variables in the form of underlying factors which are themselves not directly observable and which are thought to be representative of the underlying process that has created these correlations. On the surface, factor analysis and principal component analysis are very similar. Both rely on an eigenvalue analysis of the covariance matrix, and both use linear combinations of variables to explain a set of observations. However, in PCA the quantities of interest are the observed variables themselves; the combination of these variables is simply a means for simplifying their analysis and interpretation. Conversely, in factor analysis the observed variables are of little intrinsic value; what is of interest is the underlying factors. This method has been used by Cummins *et al.* [20] to reduce a set of 61 molecular properties to four factors, which were then used to compare the diversity of five chemical databases using a cell-based approach described above. It was also explored by Gibson *et al.* [44] in a comparative study of 100 different heterocyclic aromatic systems, but they concluded that factor analysis did not reduce the complexity of the analysis, and did not offer any significant advantages over PCA.

Unlike PCA and factor analysis which attempt to de-correlate the data, multi-dimensional scaling (MDS) [45,46] attempts to construct a configuration of points in a low-dimensional space from information about the distances between these points. In particular, given a set of k data points in the input space $\{x_i, i = 1, 2, \ldots k\}$, a symmetric matrix d_{ij} of the observed dissimilarities between these points, and a set of images of x_i on a d-dimensional display plane

$\{\xi_i, i = 1,2,..,k; \xi_i \in \Re^d\}$, the objective is to place ξ_i onto the plane in such a way that their Euclidean distances $\delta_{ij} = \|\xi_i - \xi_j\|$ approximate as closely as possible the corresponding values d_{ij}. A sum-of-squares error function can be used to decide the quality of the embedding. The most commonly used criterion is Kruskal s stress:

$$S = \sqrt{\frac{\sum_{i<j}(\delta_{ij} - d_{ij})^2}{\sum_{i<j}\delta_{ij}^2}} \tag{12.37}$$

The actual embedding is carried out in an iterative fashion. The process starts by: 1. generating an initial set of coordinates ξ_i, 2. computing the distances δ_{ij}, 3. finding a new set of coordinates ξ_i using a steepest descent algorithm such as Kruskal's linear regression or Guttman's rank-image permutation, and 4. repeating steps 2 and 3 until the change in the stress function falls below some predefined threshold. Multi-dimensional scaling was used by Martin *et al.* [9] to reduce the 2048-bit Daylight fingerprints associated with 721 commercially available primary amines to only five continuous variables that reproduced all 260000 original pairwise dissimilarities (distances) with a standard deviation of only 10%. Similarly, only seven dimensions were necessary to represent the 642000 pairwise similarities among a set of 1133 carboxylic acids and acid chlorides to the same precision. Unfortunately, MDS algorithms are $O(N^2)$ which makes them unsuitable for large combinatorial datasets.

Non-linear mapping is a closely related technique developed by Sammon [47]. Just like MDS, non-linear mapping attempts to approximate local geometric relationships on a low-dimensional space. Although an exact projection is only possible when the distance matrix is positive definite, meaningful projections can be obtained even when this criterion is not satisfied. The embedding is carried out in an iterative fashion by minimizing an error function, E, which measures the difference between the distance matrices of the original and projected vector sets:

$$E = \frac{\sum_{i<j}^{k} \frac{[d_{ij} - \delta_{ij}]^2}{d_{ij}}}{\sum_{i<j}^{k} d_{ij}} \tag{12.38}$$

E is minimized using a steepest-descent algorithm similar the one used in multi-dimensional scaling. While Sammon mapping is also $O(N^2)$, we have devised an algorithm of $N\log N$ time complexity which can handle hundreds of thousands to millions of items [48]. Most recently, we extended this algorithm by combining conventional non-linear mapping techniques with feed-forward neural networks, allowing the processing of datasets orders of magnitude larger than those accessible with conventional methodologies [49]. Rooted on the principle of probability sampling, the method employs a classical algorithm to project a small random sample, and then "learns" the underlying non-linear transform using a multi-layer neural network trained with the back-propagation algorithm. Once trained, the neural network can be used in a feed-forward manner to project the remaining members of the population as well as new, unseen samples with minimal distortion. Using several examples from the fields of image processing and combinatorial chemistry, we demonstrated that this

method is widely applicable and can generate projections that are virtually indistinguishable from those derived by conventional approaches.

12.4 Diversity Sampling

Once a diversity space and a diversity metric have been chosen, an algorithm must be devised for selecting a small subset of compounds from a typically much larger collection for physical synthesis and biological evaluation. This can be carried out at the reagent or product level. In either case, an efficient algorithm is required to sort through the large number of possibilities.

12.4.1 Selection Algorithms

The most well known algorithm for subset selection is maxmin. The algorithm begins by selecting a compound at random and building up the selection one compound at a time. At each step, the compound that is most dissimilar to the compounds already selected is identified and added to the selection, and the process continues until the desired number of compounds has been reached. Maxmin is a greedy algorithm that attempts to maximize the minimum inter-molecular distance, encoded by the D_1 metric. The method is trivial to implement, but is $O(NK)$ where N is the size of the virtual library and K is the number of compounds selected, and thus is unsuitable for very large datasets. Maxmin, which is a special case of Dykstra's optimal design algorithm (see below), was first applied to diversity selection by Lajiness [10], and has been adopted by Polinsky [50], Chapman [51], and many others.

With the exception of the initial choice, maxmin is a deterministic algorithm. Clark has proposed a stochastic variant known as K-dissimilarity selection or OptiSim [52]. The method starts by selecting a compound at random, and comparing it to K other compounds chosen at random from the dataset, excluding from consideration any compounds which are too similar to the initial selection. From among these K compounds, the one which is the most dissimilar to the initial selection is added to the curremt selection. At each iteration thereafter, a fresh sample of K candidates is compared to the compounds which have already been selected, and the most dissimilar among them is added to the selection; all compounds are considered once before any candidate is considered twice. The algorithm is $O(K \times M^2)$, where M is the number of compounds being selected. The smaller the sample size K, the more "representativeness" is favored in the selection set. As K approaches N, OptiSim becomes more and more similar to maxmin. At modest values of K (2–10), the balance between representativeness and diversity is very similar to that seen for selection based on hierarchical clustering.

Taylor *et al.* [53] have developed a stepwise elimination algorithm that works in the opposite direction. Starting with a symmetric $N \times N$ similarity matrix, the two most similar compounds are identified, and one of them is eliminated. This process continues until a single compound is left in the set. The compounds are then sorted, and the most diverse molecules

are placed at the top of the list. A related technique developed by the same group is cluster sampling. The method begins by generating a list of nearest neighbors for each compound in the dataset using a minimum similarity threshold of 0.8. These lists are then merged to form the nearest neighbor table (NNT) for the entire set. During each iteration, the procedure selects the compound which occurs most often in the NNT, which corresponds to the compound situated at the center of the most densely populated region (cluster) of property space. All the nearest neighbors of this molecule are then flagged as unavailable for subsequent cycles of selection. The procedure terminates when all of the compounds in the set are either selected or flagged as unavailable. Note that despite its name, this method does not explicitly partition the compounds into clusters. Both cluster sampling and stepwise elimination are intuitive and robust procedures, but require an exhaustive enumeration of all pairwise distances which makes them impractical for large datasets.

A very popular methodology for subset selection is to cluster the compounds into disjoint sets and select a number of representatives from each cluster [7]. The clustering algorithm attempts to ensure that the resulting clusters are internally homogenous (compounds within each cluster are similar to each other) and externally heterogeneous (compounds within each cluster are dissimilar to members of other clusters). Clustering algorithms can be classified as hierarchical or non-hierarchical based on the way in which the clusters are formed. Hierarchical cluster analysis constructs a tree, or dendogram, the structure of which reflects the organization of all the members of the collection. The dendogram may be created from the top down beginning with a single cluster which is recursively sub-divided into increasingly smaller groups until each member is a cluster by itself (a "singleton"). Alternatively, one could proceed in the opposite direction, starting with singletons which are gradually combined into larger and larger clusters until all the data points belong to a single group. Non-hierarchical methods employ a user-defined heuristic for the cluster assignment. In the case of the Jarvis–Patrick algorithm, this heuristic is the existence of a minimum number of common nearest neighbors. This method is fast but has the tendency to generate either too many singletons or too few very large clusters depending on the stringency of the clustering criteria. A thorough evaluation of several clustering methodologies can be found in [33] and [34]. These studies conclude that hierarchical methods such as the Ward and group-average hierarchical agglomerative methods and the minimum diameter polythetic hierarchical divisive method should be preferred over the Jarvis–Patrick algorithm for similarity applications.

The extensive use of clustering for diversity sampling is due to a very large extent to the availability of clustering codes for similarity analysis of large chemical databases. Once the representatives from each cluster have been selected, they are examined for possible co-linearities. If the sample is found to be non-orthogonal, suspect compounds are replaced with other members of the same cluster, and the new solution is re-evaluated. This cycle continues until a quasi-orthogonal set is identified.

A conceptually close relative to cluster analysis are the partitioning methods developed by Cummins *et al.* [20] and later by Pearlman [17]. Once the space is partitioned in an array of hyper-rectangular cells, a diverse design is constructed by extracting representative samples from each cell, in a manner similar to the clustering approach described above. The advantages of cell-based methods are described in greater detail in Section 12.2.2.

The last family of algorithms described in this section is rooted on the principles of statistical experimental design. In Section 12.2 we reviewed *D*-optimality and *A*-optimality as di-

versity metrics; this section describes a set of well-established algorithms for producing *D*-optimal and other related designs. Given a list of points that specify which regions of the design are valid or feasible, and given a user-specified number of points, these algorithms select points that optimize the respective criterion. The most common algorithms include Dykstra's sequential method, the single exchange (or Wynn–Mitchell) method, the DetMax algorithm, and the Fedorov and modified Fedorov simultaneous switching algorithms [24].

Starting with an empty design, Dykstra's method searches through the list of candidate points, choosing in each step the one that maximizes the chosen criterion. The method does not iterate, and terminates after the desired number of points is selected. This is a fast algorithm, and is used to construct the initial designs for many of the other methods described below.

The Wynn–Mitchell algorithm starts with an initial design of the requested size (typically derived by Dykstra's sequential method) and replaces one point at a time with another point from the candidate set. The point that is dropped is the one that contributes the least to the *D*-optimal or *A*-optimal criterion, and the point it is replaced with is the one that maximizes that criterion. This algorithm is terminated when additional exchanges render no further improvement.

DetMax was proposed by Mitchell and is probably the best known and most widely used optimal design algorithm. Like the simple exchange method, it starts from an initial design and carries out a single exchange in the manner described above. However, if the respective criterion does not improve, the system will undertake *excursions*, that is, it will add or subtract more than one point at a time so that during the search the number of points in the design may vary between $K - \Delta K$ and $K + \Delta K$, where K is the requested design size and ΔK is the maximum allowed excursion defined by the user. The iterations stop when the chosen optimality criterion no longer improves within the maximum excursion.

The modified Fedorov algorithm is a modification of the original algorithm proposed by Fedorov. Just like all the preceding methods, it starts with an initial design derived with the sequential method, and during each iteration it exchanges each point in the design with a point from the candidate set so as to optimize the design according to the chosen criterion. Unlike the simple exchange algorithm, the exchange is not sequential but simultaneous. During each iteration, each point in the design is compared to every point in the candidate list, and the exchange is made for every pair that maximizes the design criterion. In Fedorov's original algorithm, only a single exchange is performed, that is, only the best among all possible exchanges is accepted. The iterations continue until there is no further improvement in the optimality criterion. It is easy to see that the original algorithm can be significantly slower since in each iteration $K \times N$ comparisons are performed in order to exchange a single point, where K is the number of points selected and N is the number of candidates.

A more thorough review of algorithms for constructing optimal designs is presented in [24]. It should be clear from the above, that all these methods are impractical for datasets containing hundreds of thousands to millions of items.

12.4.2 Reagents *versus* Products

For large datasets, the techniques described above can be computationally very demanding. This has forced several groups to consider a divide-and-conquer approach in which each sub-

stitution site is treated independently of the others. This approach is based on the assumption that a diverse set of reagents would lead to a diverse set of products, and has three compelling advantages. First, it is intuitive to a chemist and selections can be easily augmented if there are problems with synthesis or reagent availability. Second, the number of reagents to choose from is always relatively small and grows linearly, whereas the number of products is enormous and grows combinatorially (e.g. a $100 \times 100 \times 100 \times 100$ four-component library contains 10^8 products but involves only 400 reagents). Finally, any reagent-based selection constitutes by definition a full combinatorial array and, hence, requires minimal synthetic resources to execute.

However, the simplicity of the reagent-based design comes at a price in terms of effectiveness. Gillet *et al.* [54] have demonstrated that maximizing the diversity of the reagent pool almost always results in libraries that are noticeably less diverse at the product level. In their study, three types of virtual libraries were considered: libraries without a common substructural core, libraries with a small core substructure, and libraries with a large core substructure where the substituents were generally relatively small compared to the core itself. Jamois *et al.* [18] have also reached the conclusion that product-based design is more effective, although they found that the benefit depends to a large extent on the types of descriptors used and the nature of the library. They reported that the difference between reagent and product-based designs is more pronounced with Daylight fingerprints, less so with physicochemical properties, and only marginal with ISIS keys. This is a natural consequence of the fact that, unlike ISIS keys, fingerprints involve an exhaustive enumeration of paths of up to eight atoms which can span multiple side chains, and thus encode information that cannot be captured in a simple substructure key. These authors also reported that the efficiency of reagent-based selection is compromised when there is a degeneracy in the reagent lists. This was demonstrated using a tripeptide library built from three identical sets of amino acids, and a benzodiazepine library constructed from a heterogeneous set of acid chlorides, amino acids, and alkylating agents. Since in the reagent-based approach the selections of R_1, R_2 and R_3 were made independently, there was a significant overlap between the selected R groups in the case of the tripeptide library. Of course, this problem can be partially eliminated by combining all the reagents into a single list and performing the analysis on the combined pool.

12.5 Advanced Techniques

Selecting subsets of compounds from larger collections is a combinatorial problem of enormous proportions. Given an n-member collection and a number k, the number of different k-member subsets of an n-member set is given by the binomial:

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \tag{12.39}$$

This problem is NP-complete, and the cardinality of that space is enormous even for the most conservative cases encountered in combinatorial design. Most of the sampling algorithms described in the previous section are greedy in nature and specific to the problem of maximizing diversity. Our group recognized early on that diversity is only one of many dif-

ferent selection criteria that one may wish to employ in library design [55], and developed a more general and extensible method based on Monte-Carlo sampling [4,5]. Our approach, which is rooted on the principles of multi-objective optimization, is to employ an objective function that encodes all the desired selection criteria, and then use a simulated annealing or evolutionary approach to identify the optimal (or a nearly optimal) subset from among the vast number of possibilities.

Simulated annealing is a global, multivariate optimization technique based on the Metropolis Monte-Carlo search algorithm. The method starts from an initial random state, and walks through the state space associated with the problem of interest by a series of small, stochastic steps. In the problem at hand, a state represents a particular subset of compounds from the virtual collection, and a step is a small change in the composition of that set (i.e. replacement of a small fraction of the points comprising the set). An objective function, f, maps each state to a real value which represents its energy or fitness. While downhill transitions are always accepted, uphill transitions are accepted with a probability that is inversely proportional to the energy difference between the two states. This probability is controlled by a parameter called temperature, which is adjusted in a systematic manner during the course of the simulation. This general optimization scheme can be implemented in a serial or parallel manner.

The system developed in our group supports two different types of designs [56]. The first is called "singles" and refers to a subset of products that is not constrained by the number or types of reagents involved. The second is called an "array" and represents the products derived by combining a given subset of reagents in all possible combinations as prescribed by the reaction scheme. Note that in this context, the term "array" is basically equivalent to reagent selection at the product level [9] and does not refer to the physical layout and execution of the experiment. These two types of designs are illustrated in Figure 12.9.

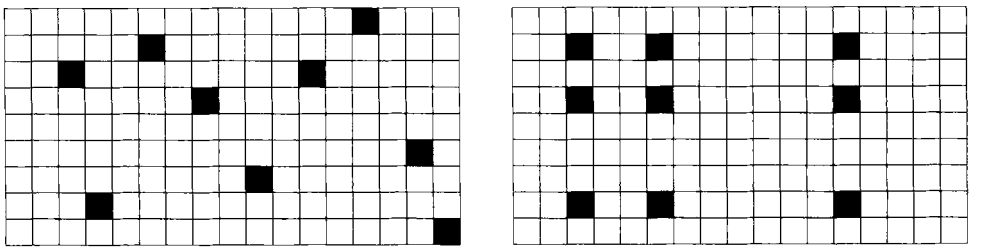


Figure 12.9. Singles *versus* arrays.

The combinatorial nature of the two problems is vastly different. For singles, the number of states that one has to consider (the number of different k-subsets of an n-set) is given by Eq. (12.39), while for arrays the number of possibilities (i.e. the number of different $k_1 \times k_2 \times \ldots k_R$ arrays derived from an $n_1 \times n_2 \times \ldots n_R$ R-component combinatorial library) is:

$$\prod_{i=1}^{R} \frac{n_i!}{(n_i - k_i)!k_i!} \tag{12.40}$$

For a 10 × 10 two-component combinatorial library, there are 10^{25} different subsets of 25 compounds, and only 63504 different 5 × 5 arrays. For a 100 × 100 library and a 100/10 × 10 selection, those numbers increase to 10^{241} and 10^{26} for singles and arrays, respectively. In our implementation, these two types of designs are encoded using two different internal state representations, each with its own mutation protocol. In a singles selection, a step represents the substitution of a few compounds comprising the current state, whereas in an array selection a step represents the substitution of a single reagent in the combinatorial array. Although arrays are generally inferior in terms of meeting the design objectives, they require fewer reagents and are much easier to synthesize in practice.

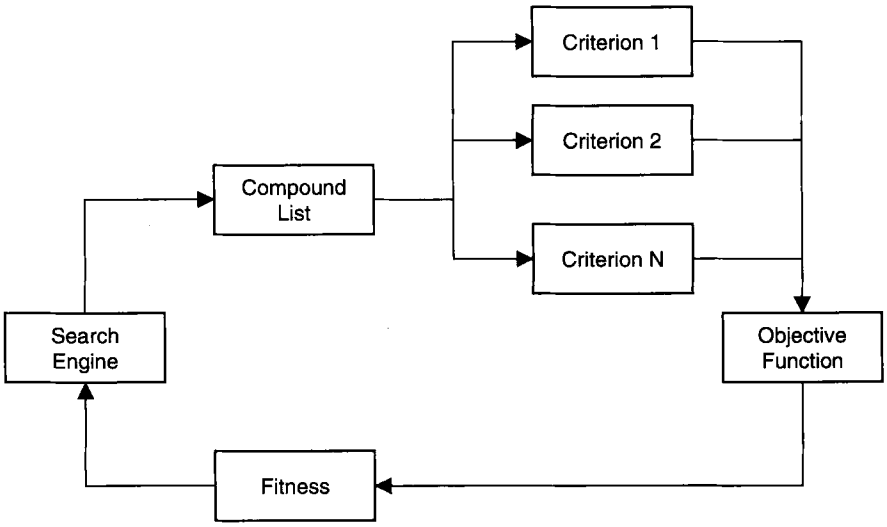


Figure 12.10. Process flow of the multi-objective selector.

The major advantage of this approach is that the search algorithm is completely independent of the performance measure, and can be applied on a wide variety of selection criteria and fitness functions [4,12,16,56]. More importantly, this approach allows one to combine more than one selection criteria, and create designs that simultaneously satisfy several, often conflicting, design objectives. The system is outlined in Figure 12.10. Unlike many of the algorithms described above which are tailored to a particular application, this approach is completely general, programmatically simple, and easily extensible. The remaining paragraphs describe how this multi-objective approach can be used to design diverse libraries which are biased towards more pharmacologically relevant regions of chemical space, and represents a brief overview of some recent work that we published in a special issue of the *Journal of Molecular Graphics and Modeling* devoted to combinatorial library design [56].

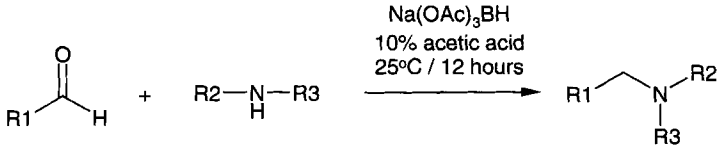


Figure 12.11. Synthetic sequence for the generation of the reductive amination library.

Experience has shown that selecting compounds from a virtual library solely on the basis of maximum diversity, no matter what diversity measure is used, can often result in the design of combinatorial libraries with poor pharmacokinetic properties or other undesirable characteristics. Consider, for example, the combinatorial library shown in Figure 12.11. This library is based on the reductive amination reaction and is utilized for the construction of structurally diverse drug-like molecules with useful pharmacological properties, particularly in the GPCR super-family [57]. For demonstration purposes, 300 primary and secondary amines and 300 aldehydes were selected at random from the Available Chemicals Directory [58] and were used to generate a virtual library of 90000 products using the library enumeration classes of the DirectedDiversity® toolkit [59]. Each compound in the 90000-member library was characterized by an established set of 117 topological descriptors [26], which were subsequently normalized and de-correlated using principal component analysis, resulting in an orthogonal set of 23 latent variables which accounted for 99% of the total variance in the data. To simplify the analysis and interpretation of results, this 23-dimensional dataset was further reduced to two dimensions using the non-linear mapping algorithm described in Section 12.3.4 [49], and formed the basis of all subsequent diversity calculations. The resulting non-linear map is shown in Figure 12.12a.

In addition to the 117 topological descriptors, the molecular weight and octanol–water partition coefficient (log$P$) of each compound was computed independently using the Ghose–Crippen approach [60] as implemented in the DirectedDiversity® toolkit [59], and were used to assess the drug-likeness of the resulting designs. Given the probabilistic nature of the problem, we proposed that the drug-likeness of a given collection of compounds be measured by how well the distribution of a certain property in that collection matches the distribution of the same property in a large set of known drugs. As a numerical measure of the dissimilarity between two property distributions, we used the Kolmogorov–Smirnov statistic, K^* (Figure 12.13), which is defined as [61]:

$$K^* = \max_{-\infty < x < \infty} |P(x) - P^*(x)| \tag{12.41}$$

where $P(x)$ is an estimator of the cumulative distribution function of the actual probability distribution from which it is drawn, and $P^*(x)$ is the target cumulative distribution function.

For a set of n points x_i, $i = 1, \ldots, n$, $P(x)$ is the fraction of data points to the left of (and including) a given value of x. The Kolmogorov–Smirnov statistic is applicable to unbinned distributions that are functions of a single independent variable, and represents the maximum value of the absolute difference between two cumulative distribution functions. The measure

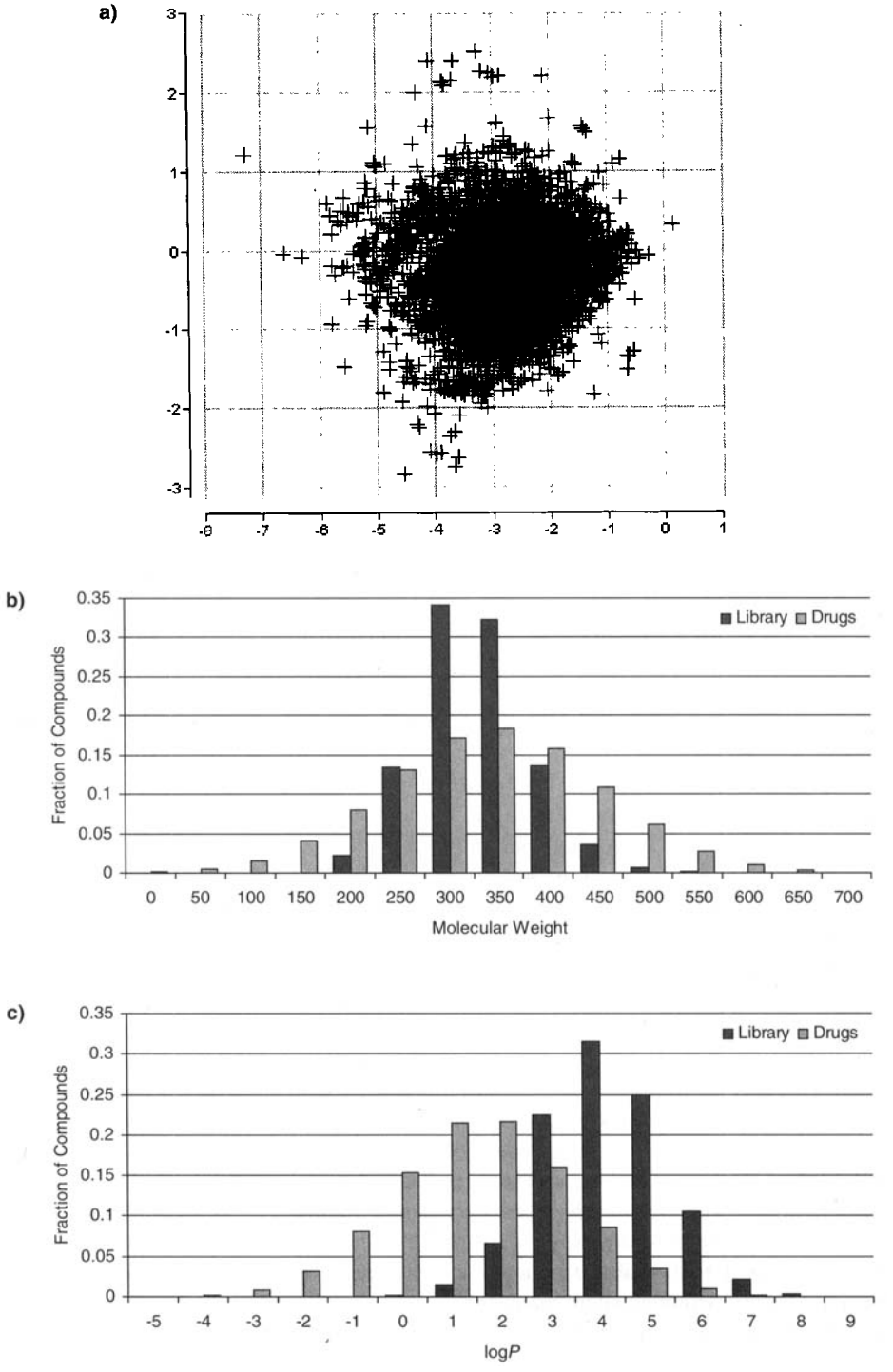


Figure 12.12. The reductive amination dataset: **a** non-linear projection, **b** molecular weight distribution, **c** log*P* distribution. Reprinted from [56].

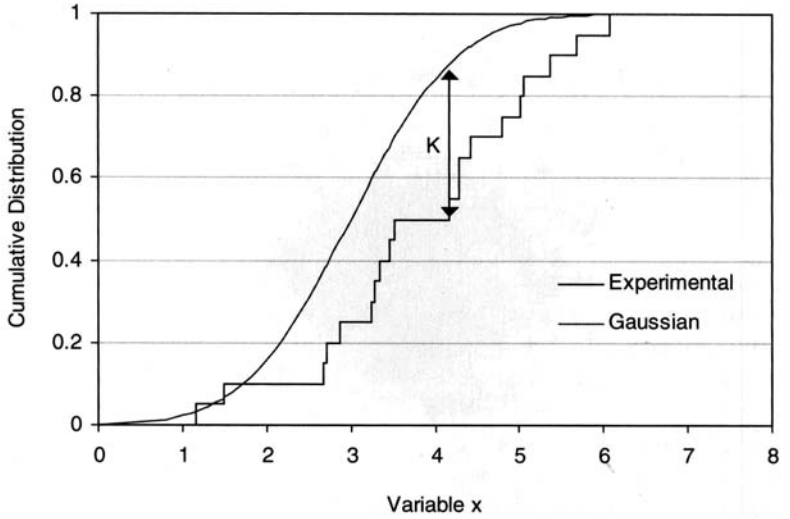


Figure 12.13. The Kolmogorov–Smirnov statistic K. Reprinted from [56].

is very fast to compute, and unlike the more commonly used χ^2 test, it does not require binning of the data, which is arbitrary and leads to loss of information. K^* takes values in the interval [0,1] and can be recast as a similarity index using Eq. (12.42):

$$K = 1 - K^* \tag{12.42}$$

The "ideal" molecular weight and log$P$ distributions of known drugs were derived by analyzing a subset of 7484 compounds from the World Drug Index [62]. This set consisted of drugs that had an INN or USAN number and were approved for marketing in at least one country. Since the computation of the Kolmogorov–Smirnov criterion becomes significantly faster if the target cumulative distribution function is known analytically, the raw data were fitted to a normal distribution using a least-squares fitting procedure. The mean and sigma of the fitted gaussians (shown in Figure 12.14a and Figure 12.14b along with the original distributions) were 314.3 and 108.3 for molecular weight, and 1.04 and 1.78 for log$P$, respectively.

The corresponding distributions for the reductive amination library are shown in Figure 12.12b, c. It is evident that while the molecular weight of these compounds is within acceptable limits, the log$P$ distribution is shifted upward by more than three units compared to the WDI set. Thus, statistically speaking, one would expect that most leads identified from this library (or any randomly chosen sample thereof) would reflect the physicochemical characteristics of the collection as a whole, and could prove problematic for SAR development. Unfortunately, maximizing the diversity of the design does not improve the situation. As shown in Figure 12.15b, c, the most diverse 20 x 20 array as determined by Eq. (12.14) and the annealing algorithm described above exhibits an unfavorable log$P$ distribution similar to that of the entire library. Although many chemists believe that undesirable physicochemical properties can be eliminated through classical medicinal chemistry approaches, experience sug-

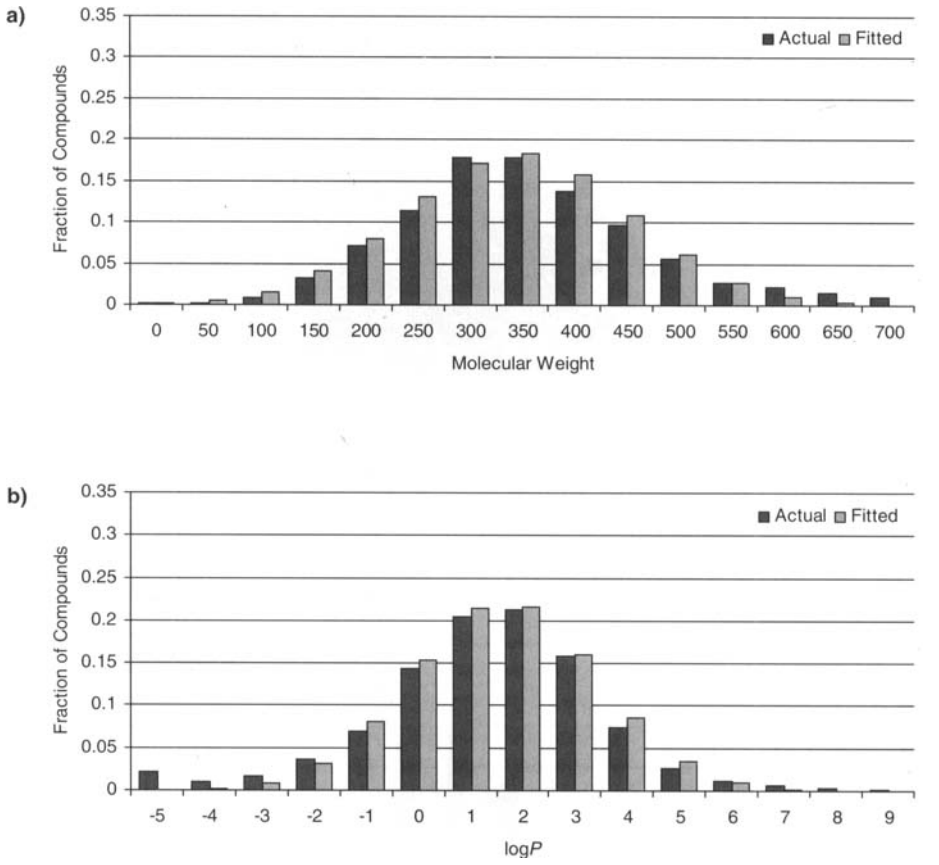


Figure 12.14. **a** Molecular weight, and **b** log$P$ distributions of drug-like molecules based on 7484 marketed drugs from the World Drug Index. Two series are shown for each property. The one on the left is the true distribution, while the one on the right is a normal approximation determined by a least-squares fitting procedure. Reprinted from [56].

gests that correcting solubility and permeability remains the rate-limiting step in the development of a drug candidate. It is clear that in order to produce more practically useful selections, it is necessary to optimize not only molecular diversity, but also the overall pharmacokinetic profile of the resulting compounds.

Our solution to this problem was to combine diversity with the Kolmogorov–Smirnov similarity measure into a single unifying function of the form:

$$f = D + 0.2 \cdot K(\log P) + 0.2 \cdot K(mw) \tag{12.43}$$

where D is the diversity criterion defined in Eq. (12.14), and $K(\log P)$ and $K(mw)$ are the Kolmogorov–Smirnov similarities between the log$P$ and molecular weight distributions of the selected compounds and the reference WDI set, respectively. As with most problems of

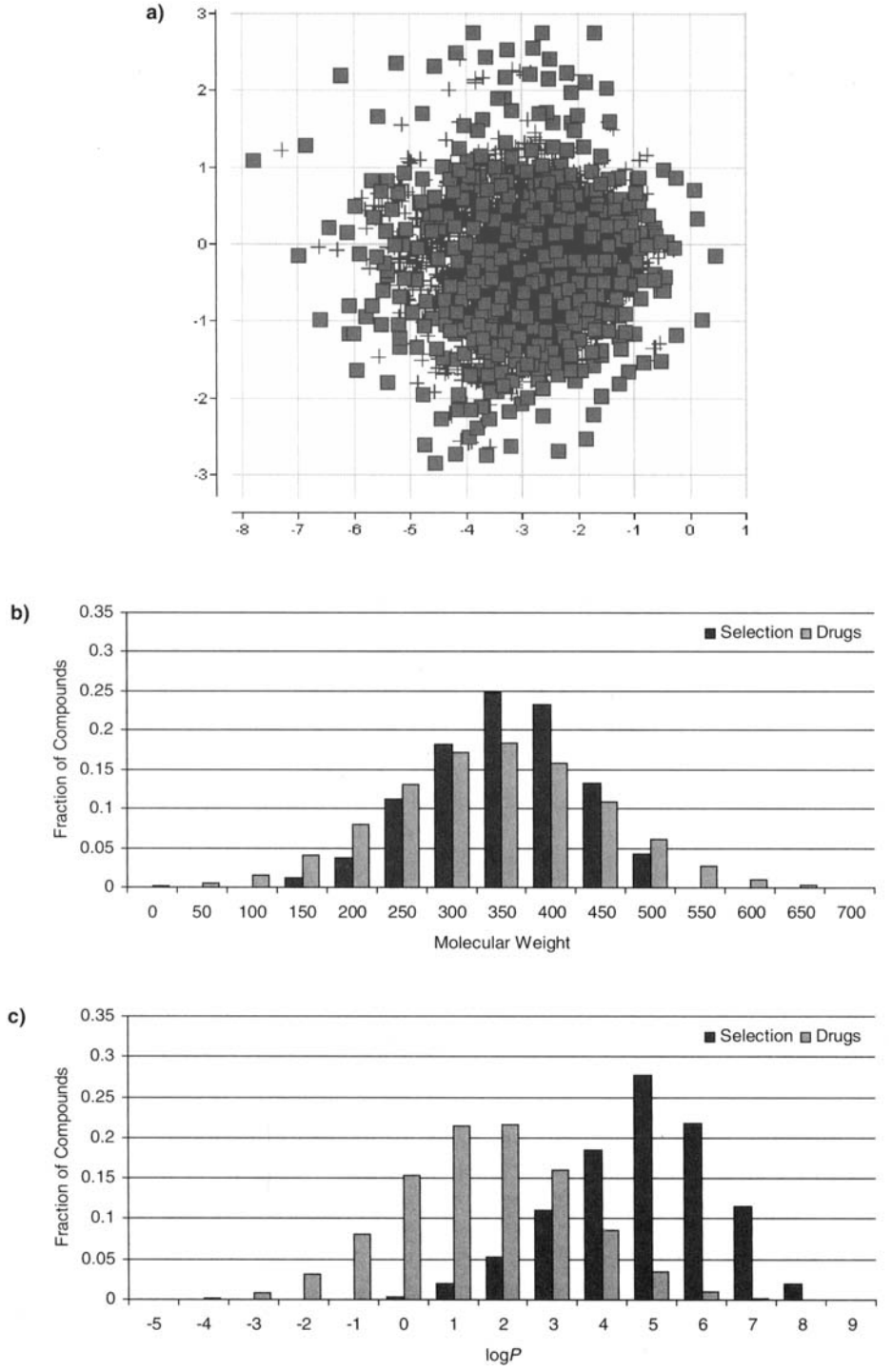


Figure 12.15. Diversity-based selection of a 20 × 20 array from the reductive amination library: **a** non-linear projection, **b** molecular weight distribution, **c** log*P* distribution. Reprinted from [56].

this type, the difficulty lies in assigning a meaningful set of coefficients used to weight the individual objectives. We have found that the easiest and most reliable way to perform this assignment is to identify the maximum values that each criterion can assume independently, and scale them according to the influence that each of them should have in the final selection. In the case at hand, the mean nearest neighbor distance, D, in the diversity-based selection in Figure 12.15 was 0.18, while the value of K for the selections based exclusively on molecular weight or log$P$ was ~0.9. These values suggested that for the three criteria to be placed on an equal footing, the value of K had to be scaled down by a factor of approximately five. In pathological cases where the energy landscapes (i.e. the distributions of scores) of the individual criteria are very different, alternative, more complex objective functions can be devised.

As shown in Figure 12.16, when the selection is carried out using Eq. (12.43), the selected compounds' molecular weight and log$P$ distributions approximate very nicely the respective distributions of known drugs, and this occurs without a significant impact on the diversity of the design (Figure 12.16a). Moreover, because the selection was carried out as an array, the design requires a small number of reagents (20 amines and 20 aldehydes, compared to 154 amines and 162 aldehydes required by an equivalent singles selection [56]) and can be easily executed on robotic hardware. While the discussion in this section has been limited to diversity and property distributions, the system can accommodate a wide variety of selection criteria, and gives the medicinal chemist full control over the precise form of the objective function and the coefficients used to scale the individual objectives.

12.6 Conclusions

In this review, we have attempted to provide the readers with a clear and concise description of the basic elements involved in the measurement and sampling of molecular diversity. These elements include a measure of chemical distance that relates molecules to each other, a measure of molecular diversity that quantifies the diversity of a given set of compounds, and a sampling method for identifying a diverse subset of compounds from a large combinatorial library. Contrary to popular belief, molecular diversity is not a means to increase the number of hits identified in a high-throughput screening experiment, but only the number of different structural classes represented by them. The design of a combinatorial library should take into account a number of additional factors, such as the physicochemical profile and drug-likeness of the resulting compounds, the degree of consistency with known trends regarding the biological targets of interest, the cost of the experiment and the ease by which it can be executed, and many others. This requires a departure from conventional compound selection methodologies, and can be easily accomplished using a modern approach to experimental design based on the principles of multi-objective optimization.

Acknowledgements

The authors are grateful to Drs. Edward P. Jaeger and Renee DesJarlais for many useful discussions, and Dr. Raymond F. Salemme of 3-Dimensional Pharmaceuticals, Inc. for his insightful comments and support of this work.

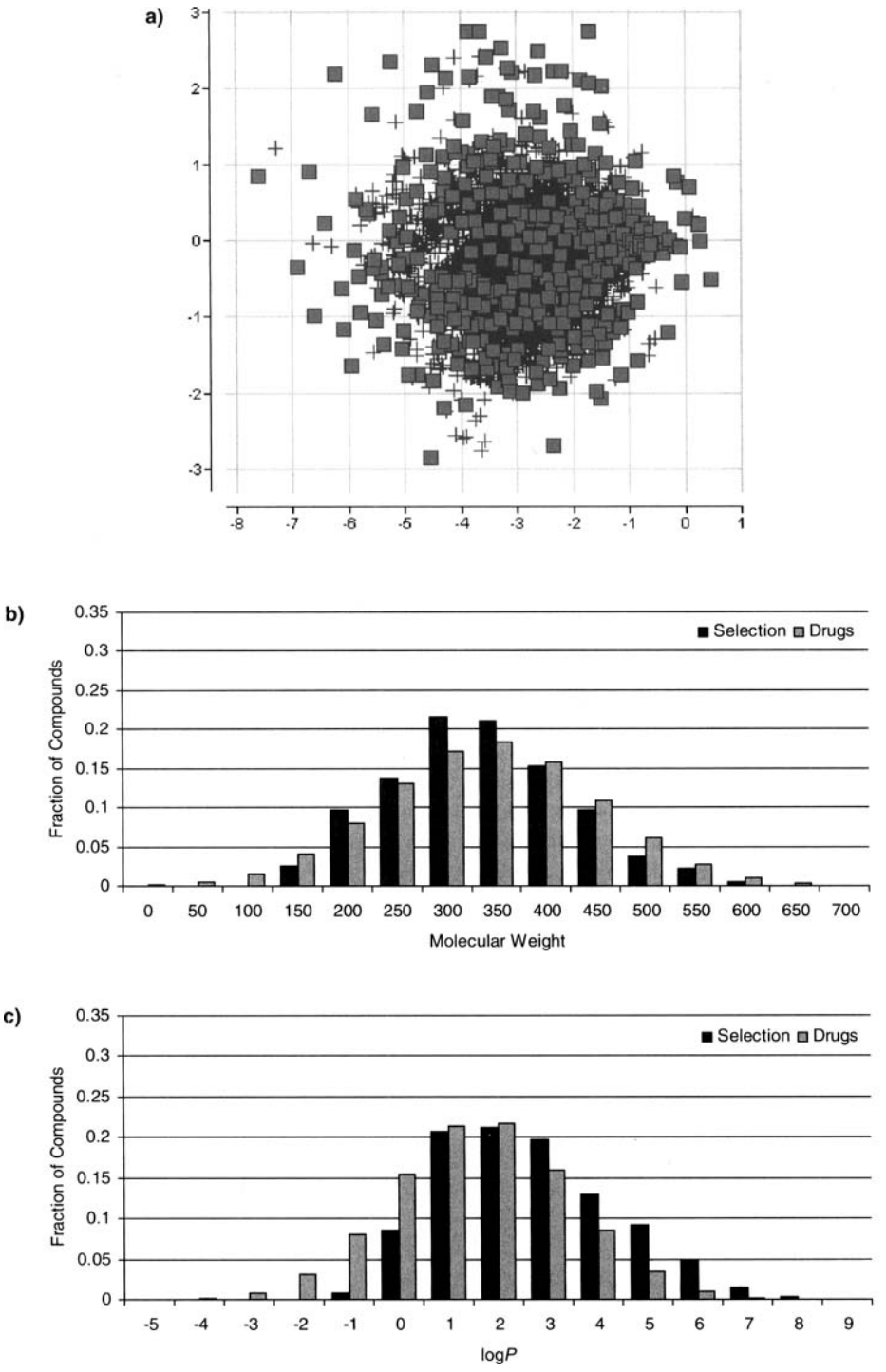


Figure 12.16. Multi-objective selection of a 20×20 array based on diversity, molecular weight and log$P$ (Eq. 12.43). **a** Non-linear projection, **b** molecular weight distribution, **c** log$P$ distribution. Reprinted from [56].

References

[1] D. K. Agrafiotis, in The Encyclopedia of Computational Chemistry, P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, P. R. Schreiner (Eds.) John Wiley & Sons, *Chichester* **1998**, pp. 742–761.
[2] D. K. Agrafiotis, J. C. Myslik, F. R. Salemme, *Mol. Divers.* **1999**, *4*, 1–22.
[3] E. J. Martin, D. C. Spellmeyer, R. E. Critchlow Jr., J. M. Blaney, *Reviews in Computational Chemistry, Vol. 10*, K. B. Lipkowitz, D. B. Boyd (Eds.), VCH, Weinheim **1997**.
[4] D. K. Agrafiotis, *J. Chem. Info. Comput. Sci.* **1997**, *37*, 841–851.
[5] D. K. Agrafiotis, 3rd Electronic Computational Chemistry Conference, http://hackberry.chem.niu. edu/ECCC3/paper48 **1996**.
[6] M. A. Johnson, G. M. Maggiora, G. M. *Concepts and Applications of Molecular Similarity*, Wiley, New York **1990**.
[7] P. Willett, *Similarity and Clustering in Chemical Information Systems*, Research Studies Press, Letchworth **1987**.
[8] S. K. Kearsley, S. Sallmack, E. M. Fluder, J. D. Andose, R. T. Mosley, R. P. Sheridan, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118–127.
[9] E. J. Martin, J. M. Blaney, M. A. Siani, D. C. Spellmeyer, A. K. Wong, W. H. Moos, *J. Med. Chem.* **1995**, *38*, 1431–1436.
[10] M. S. Lajiness, in *QSAR: Rational Aproaches to the Design of Bioactive Compounds*, C. Silipo, A. Vittoria (Eds.), Elsevier, Amsterdam **1991**, pp. 201–204.
[11] M. Hassan, J. P. Bielawski, J. C. Hempel, M. Waldman, *Mol. Divers.* **1996**, *2*, 64–74.
[12] D. K. Agrafiotis, V. S. Lobanov, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 51–58.
[13] D. B. Turner, S. M. Tyrrell, P. Willett, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18–22.
[14] J. Mount, J. Ruppert, W. Welch, A. N. Jain, *J. Med. Chem.* **1999**, *42*, 60–66.
[15] S. K. Lin, *Molecules* **1996**, *1*, 57–67.
[16] D. K. Agrafiotis, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 576–580.
[17] R. S. Pearlman, K. M. Smith, *J. Chem. Inf. Comput. Sci.*, in press.
[18] E. A. Jamois, M. Hassan, M. Waldman, *J. Chem. Inf. Comput. Sci.*, in press.
[19] R. S. Pearlman, *DiverseSolutions User's Manual*, University of Texas, Austin, TX **1995**.
[20] D. J. Cummins, C. W. Andrews, J. A. Bentley, M. Cory, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750–763.
[21] N. E. Shemetulsksis, D. Weininger, C. J. Blankley, J. J. Yang, C. Humblet, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862–871.
[22] S. Pickett, J. S. Mason, I. M. McLay, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214–1223.
[23] E. K. Davies, C. Briant, *Network Science* **1995**, http://www.awod.com/netsci/issues/.
[24] D. C. Montgomery, *Design and Analysis of Experiments*, 4th Edition, John Wiley & Sons **1996**.
[25] *Molconn-X*, Haney Associates, Mercer Island, WA.
[26] L. B. Kier, L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, Wiley, New York **1986**.
[27] R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
[28] G. Moreau, P. Broto, *Nouv. J. Chim.* **1980**, *4*, 359–360.
[29] H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205–1213.
[30] M. Wagener, J. Sadowski, J. Gasteiger, *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.
[31] J. Sadowski, M. Wagener, J. Gasteiger, *Angew. Chem. Int. Ed. Engl.* **1996**, *34*, 23–24.
[32] F. R. Burden, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225–227.
[33] G. M. Downs, P. Willett, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094–1102.
[34] R. D. Brown, Y. C. Martin, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
[35] J. Sadowski, J. Gasteiger, G. Klebe, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.
[36] *Unity Chemical Information Software*, Tripos Associates, St. Louis, MO.
[37] N. W. Murrall, E. K. Davies, *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312–316.
[38] R. P. Sheridan, R. Nilikantan, A. Rusinko, N. Bauman, K. Haraki, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255–260.
[39] R. D. Cramer, R. D. Clark, D. E. Patterson, A. M. Ferguson, *J. Med. Chem.* **1996**, *39*, 3060–3069.
[40] H. Kubinyi, in *Methods and Principles in Medicinal Chemistry*, Vol. 1, R. Manhold, P. Krogsgaard-Larsen, H. Timmermann (Eds.), VCH, Weinheim **1993**, pp. 21–36.
[41] R. Lewis, I. M. McLay, J. S. Mason, *Chem. Des. Autom. News* **1995**, *10*, 37–38.
[42] R. D. Brown, Y. C. Martin, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
[43] W. Cooley, P. Lohnes, *Multivariate Data Analysis*, Wiley, New York **1971**.

[44] S. Gibson, R. McGuire, D. C. Rees, *J. Med. Chem.* **1996**, *39*, 4065–4072.
[45] W. S. Torgerson, *Psychometrika* **1952**, *17*, 401–419.
[46] J. B. Kruskal, *Phychometrika* **1964**, *29*, 115–129.
[47] J. W. Sammon, *IEEE Trans. Comp.* **1969**, *C-18*, 401–409.
[48] D. K. Agrafiotis, V. S. Lobanov, F. R. Salemme, patents pending.
[49] D. K. Agrafiotis, V. S. Lobanov, J. Chem. Inf. Comput. Sci., in press.
[50] A. Polinsky, R. D. Feinstein, S. Shi, A. Kuki, *Molecular Diversity and Combinatorial Chemistry*, I. M. Chaiken, K. D. Janda (Eds.), ACS, Washington **1996**, pp. 219–232.
[51] D. Chapman, *J. Comput. Aided Mol. Design* **1996**, *10*, 501–512.
[52] R. D. Clark, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181–1188.
[53] R. Taylor, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59–67.
[54] V. J. Gillet, P. Willett, J. Bradshaw, *J. Chem. Inf. Comput. Sci.* **1997**, **37**, 731–740.
[55] D. K. Agrafiotis, R. F. Bone, F. R. Salemme, R. M. Soll, United States Patents 5,463,564 **1995**; 5,574,656 **1996**; 5,684,711 **1997**; 5,901,069 **1999**.
[56] D. N. Rassokhin, D. K. Agrafiotis, *J. Mol. Graph. Model.*, in press.
[57] D. S. Dhanoa, V. Gupta, A. Sapienza, R. M. Soll, Poster 26, American Chemical Society National Meeting, Anaheim **1999**.
[58] Available Chemicals Directory is marketed by MDL Information Systems, Inc., 140 Catalina Street, San Leandro, CA 94577.
[59] *The Mt Toolkit: An Object-Oriented C++ Class Library for Molecular Simulations,* Copyright 3-Dimensional Pharmaceuticals, Inc., 1994–2000.
[60] A. K. Ghose, V. N. Viswanadhan, J. J. Wendoloski, *J. Phys. Chem. A* **1998**, *102*, 3762–3772.
[61] R. von Mises, *Mathematical Theory of Probability and Statistics*, Academic Press, New York **1997**.
[62] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeny, *Advanced Drug Delivery Reviews* **1997**, *23*, 3–25.

Virtual Screening for Bioactive Molecules
Edited by Hans-Joachim Böhm and Gisbert Schneider

Index

WILEY-VCH

Virtual Screening for Bioactive Molecules

Edited by H.-J. Böhm and G. Schneider

Methods and Principles in Medicinal Chemistry

Volume 10

Recent progress in high-throughput screening, combinatorial chemistry and molecular biology has radically changed the drug discovery process in the pharmaceutical industry. Currently, typically one hundred thousand to one million molecules have to be tested within a short period of time and, therefore, highly effective screening methods are a must for todayís researchers. New challenges in synthesis result in new analytical methods. The time of preparing and characterizing one compound after another belongs to the past. Intelligent, computer-based search agents are needed, and "Virtual Screening" provides solutions to many problems. Such decision support systems comprise computational techniques designed to turn raw data into valuable chemical information, and assist in extracting the relevant molecular features.

This book aims to bring together the various efforts in the field of Virtual Screening in a unique way and to provide the necessary methodological framework for more effective research. Leading experts give a thorough introduction to the state-of-the-art in Virtual Screening along with a critical assessment of both successful applications and drawbacks. Experienced scientists, as well as novices, working in medicinal chemistry and related disciplines will benefit from this conceptual approach to the topic. The information collated in this volume will be indispensable as a basis for future developments in drug research.

ISBN 3-527-30153-4

9 783527 301539