Virtual Screening for Bioactive Molecules

Edited by
Hans-Joachim Böhm and Gisbert Schneider

WILEY-VCH

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

This Page Intentionally Left Blank

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

This Page Intentionally Left Blank

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.

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

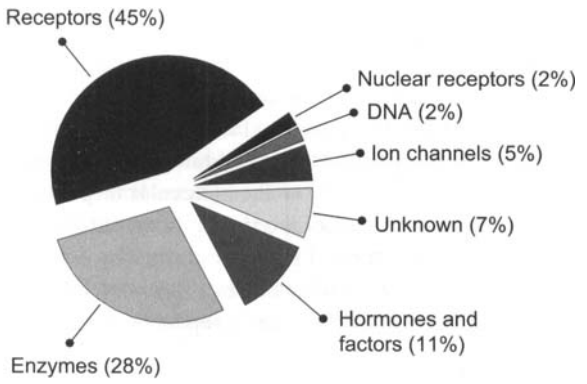

Figure 1.1. Classes of 485 current drug targets. Adapted from reference [20].

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.